# Optimization of Photonic Band Structures

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

Dipl.–Math. techn. Markus Richter
aus Frankfurt am Main

Tag der mündlichen Prüfung: 10. November 2010

Referent: Prof. Dr. Willy Dörfler
Koreferent: Prof. Dr. Christian Wieners

# Preface

This dissertation is the result of a research project I conducted at the Institute for Applied Mathematics and Numerical Analysis of the Karlsruhe Institute of Technology (KIT) in the years 2006 to 2010. During that time I worked as a research assistant at the institute under the supervision of Prof. Dr. Willy Dörfler. The project was embedded within the research program of the DFG Research Training Group 1294 "Analysis, Simulation and Design of Nanotechnological Processes" at the Department of Mathematics.

At this point I would like to take the opportunity to thank a number of people, who supported me during my research. First and foremost I would like to thank my advisors, Prof. Dr. Willy Dörfler and Prof. Dr. Christian Wieners, for giving me the opportunity to work on an interesting and challenging problem, and for guiding me along my way towards this dissertation. I am particularly thankful for the many times I was able to discuss my ideas with them, or to seek their advice. I would also like to mention Prof. Wieners' personal efforts in helping me with his numerical software library M++.

Furthermore, I would like to thank my colleagues at the institute and my fellow members of the Research Training Group 1294 for the good collaboration and for the many interesting discussions. I am especially indebted to Dr. Alexander Bulovyatov, who implemented some numerical algorithms, on which I could base my own ones. I also would like to mention Branimir Anić, Markus Bürg, Dr. Thomas Dohnal, Dr. Christian Engström, Markus Feist, Dr. Thomas Gauss, Dr. Vu Hoang, Florian Keller, and Daniel Maurer. A special thanks goes to Benjamin Exner, André Fuetterer, Dr. Desirée Hilbring, and Andreas Kuhn for proofreading parts of this dissertation.

I thank my family for their constant encouragement and support. I would also like to thank my closest friends for their company and for the many hours spent talking and laughing together. Finally, I thank Melanie for her kind, loving support.

Markus Richter
Karlsruhe, October 2010

# Contents


**1 Introduction** **6**
- 1.1 Motivation . . . 6
- 1.2 Aims of This Work . . . 9
- 1.3 Literature Review . . . 10
- 1.4 Outline . . . 12

**2 Preliminaries** **13**
- 2.1 Notations and Conventions . . . 13
- 2.2 The Cross Product . . . 15
- 2.3 Local Lipschitz Continuity . . . 16
- 2.4 Symmetry . . . 17
- 2.5 Periodicity . . . 19

**3 The Mathematical Model** **21**
- 3.1 Modelling Photonic Crystals . . . 21
- 3.2 Crystal Symmetries . . . 23
- 3.3 Wave Propagation in Linear Dielectrics . . . 30
- 3.4 Bloch Modes in Periodic Media . . . 37
- 3.5 The Two-Dimensional Case . . . 40

**4 Spectral Theory** **45**
- 4.1 The Formal Setting . . . 45
- 4.2 Sobolev Spaces of Periodic Functions . . . 46
- 4.3 The Weak Formulation . . . 59
- 4.4 Riesz–Schauder Theory . . . 65
- 4.5 Auchmuty's Principle . . . 72
- 4.6 Photonic Band Structures and Band Diagrams . . . 77
- 4.7 The Two-Dimensional Case . . . 84

**5 Photonic Band Structure Optimization** **90**
- 5.1 The Formal Setting . . . 90

5.2 Regularity of the Goal Functionals . . . 92
5.3 Existence of Optima in the TM Setting . . . 96
5.4 Existence of Optima in Other Settings . . . 103
5.5 Optimization Goals . . . 107

**6 Nonsmooth Analysis 110**
6.1 Generalized Differentials . . . 110
6.2 Generalized Differential Calculus . . . 116
6.3 Differentiability of the Gap Width Functionals . . . 120
6.4 The Two-Dimensional Case . . . 125

**7 A Generalized Gradient Method 127**
7.1 Basic Concepts . . . 127
7.2 Discretization . . . 130
7.3 Incorporating Optimization Constraints . . . 133
7.4 Preserving Symmetries . . . 136
7.5 The Algorithm . . . 138
7.6 Choosing Descent Directions . . . 140
7.7 Remarks . . . 147

**8 A Level-Set Method 148**
8.1 Basic Concepts . . . 148
8.2 Motivation . . . 150
8.3 The Algorithm . . . 154
8.4 Remarks . . . 157

**9 Numerical Results 158**
9.1 Maximizing TM Band Gaps . . . 158
9.1.1 Results of the Generalized Gradient Method . . . 159
9.1.2 Results of the Level Set Method . . . 163
9.2 Maximizing TE Band Gaps . . . 168
9.2.1 Results of the Generalized Gradient Method . . . 168
9.2.2 Results of the Level Set Method . . . 171
9.3 Maximizing Band Gaps of a 3D Photonic Crystal . . . 173

**10 Summary, Conclusions and Outlook 176**
10.1 Summary . . . 176
10.2 Comparison of the Optimization Algorithms . . . 178
10.3 Open Problems . . . 178
10.4 Final Remarks and Outlook . . . 179

**A A FEM Toolbox for MATLAB** **180**
A.1 General Remarks . . . . . 180
A.2 Data Structures and Algorithms . . . . . 182
A.3 Some Customized Features . . . . . 185
A.4 Implementation Examples . . . . . 186

**Frequently Used Symbols** **191**

**About The Author** **192**

**Bibliography** **194**

# Chapter 1

# Introduction

## 1.1 Motivation

*Photonic crystals* are materials, which are composed of two or more different dielectrics or metals, and which exhibit a spatially periodic structure, typically at the length scale of hundred nanometers. Depending on whether the periodicity extends into one, two or three space dimensions, a photonic crystal is called one-, two-, three-dimensional. Photonic crystals can be fabricated using nano-technological processes such as photolithography or vertical deposition methods. They also occur in nature, e.g. in the microscopic structure of certain bird feathers, butterfly wings, or beetle shells (see e.g. [12], [46]).

A characteristic feature of photonic crystals is that they strongly affect the propagation of light waves at certain optical frequencies. This is due to the fact that the optical density inside a photonic crystal varies periodically on the length scale of about 400 to 800 nanometers. One finds that the so-called *optical wavelengths* of light waves lie in precisely the same length scale. Light waves that penetrate a photonic crystal, are therefore subject to periodic, multiple diffraction, which leads to coherent wave interference inside the crystal. Depending on the frequency of the incident light wave this interference can either be constructive or destructive. In the latter case the light wave is not able to propagate inside the photonic crystal at all. Typically, this phenomenon only occurs for a bounded range of optical wave frequencies, if it does occur at all. Such a range of inhibited wave frequencies is called a *photonic band gap*. Light waves with frequencies inside a photonic band gap are totally reflected by the photonic crystal. It is this effect, which causes e.g. to the iridescend colors of peacock feathers (see [54]).

Whether or not photonic band gaps occur, strongly depends on the geometric structure of the photonic crystal, as well as on the contrast in optical density between the different materials the photonic crystal is built of. Apart from photonic

band gaps a photonic crystal may also exhibit other optical phenomena, such as a large refractive index or the ability to slow down the group velocity of light pulses considerably. These phenomena are also caused by periodic diffraction and coherent wave interference inside the crystal.

In the late 1980s Eli Yablonovitch and Sajeev John discovered that the band gap of a photonic crystal can be used to inhibit spontaneous photon emissions (see [43], [80]). It was also then that the term *photonic crystal* was coined. About a century earlier Lord Rayleigh studied the electromagnetic transmissibility of multi-layered dielectrics, and found that for certain frequencies of the incident wave total reflection can occur (see [63]). This phenomenon, which also stems from coherent wave interference inside the multi-layered material, was used to create frequency-selective optical mirrors, the so-called *distributed Bragg reflectors*. In essence these devices consist of periodically alternating layers of different dielectric materials. From a modern perspective, distributed Bragg reflectors can be viewed as one-dimensional photonic crystals. Two- and three-dimensional photonic crystals can be used to guide light waves along a defect in the periodic structure. It is expected that photonic crystals will play a key role in the development of nanometer-scale all-optical communication devices, such as nanometer wave-guides, optical multiplexers, logical gates or frequency filters. Some of these devices could already been realized in the laboratory. However, an industry-scale production is not yet in sight.

Many optical properties of a photonic crystal are determined by its so-called *photonic band structure*. The band structure consists of countably many of so-called *photonic bands*. Each band discloses the dispersion relation of a time-harmonic electromagnetic wave, which is able to propagate in the crystal. Mathematically a photonic band is the graph of a function $\boldsymbol{k} \mapsto \omega(\boldsymbol{k})$, where the vector $\boldsymbol{k}$ varies over a compact subset $\mathbb{B}$ of $\mathbb{R}^3$, the so-called first *Brillouin zone*. For every vector $\boldsymbol{k} \in \mathbb{B}$ the corresponding function value $\omega(\boldsymbol{k})$ is given by a solution of an eigenvalue problem, which is of the form

$$\begin{cases} (\nabla + \mathrm{i}\boldsymbol{k}) \times \big[\rho(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}\big] = \omega^2 \boldsymbol{u} & \text{in } \Omega, \\ (\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{u} = 0 & \text{in } \Omega. \end{cases} \tag{1.1}$$

Here, $\Omega$ denotes a *fundamental cell of periodicity* of the photonic crystal. A fundamental cell of periodicity is a bounded section of space within a photonic crystal. The defining property of a fundamental cell is that the entire photonic crystal structure can be reproduced by repeating the structure within the fundamental cell periodically. The crystal structure itself is represented by the real-valued function $\rho$. Typically $\rho$ is discontinuous and takes only a finite number of different function values. The eigenfunctions $\boldsymbol{u}$ are required to satisfy periodic boundary conditions. It can be shown, that for every vector $\boldsymbol{k} \in \mathbb{B}$ there exists a countable

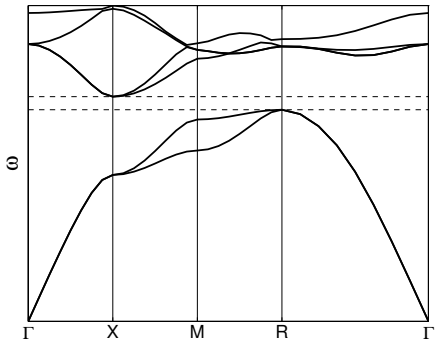



**Figure 1.1:** Photonic band structure exhibiting a band gap (dashed lines) between the second and third photonic band.

number of positive, real numbers $\omega_1(\boldsymbol{k}) \leq \omega_2(\boldsymbol{k}) \leq \cdots$, etc., such that $\omega_j(\boldsymbol{k})^2$ is an eigenvalue of (1.1) for every $j \in \mathbb{N}$. Given an index $j \in \mathbb{N}$, the graph of the function $\boldsymbol{k} \mapsto \omega_j(\boldsymbol{k})$ is commonly referred to as the $j$-th *photonic band*, a photonic band gap exists between the $j$-th and the $(j+1)$-th photonic band, if and only if

$$\min_{\boldsymbol{k}\in\mathbb{B}} \omega_{j+1}(\boldsymbol{k}) - \max_{\boldsymbol{k}\in\mathbb{B}} \omega_j(\boldsymbol{k}) > 0. \tag{1.2}$$

The so-called *gap width* is given by the left-hand side of (1.2). Photonic band structures are usually depicted in so-called *photonic band diagrams*. In such a band diagram each photonic band is represented by a curve. Photonic band gaps are revealed as horizontal regions in the diagram, which are not crossed by any curve (see Figure 1.1). The vertical extent of such a region corresponds to the gap width.

For many applications, it is desirable to have photonic crystals with large band gaps as large as possible. Given two different materials, the question arises how a photonic crystal should be built from these two materials in order to exhibit a maximal band gap. Recall that the structure of a photonic crystal is represented by the function $\rho$ in the eigenvalue problem (1.1). A photonic crystal, which consists of exactly two different materials, is represented by a two-valued function $\rho$ on $\Omega$ with values $\rho_0$ and $\rho_1$. Conversely, every two-valued function on $\Omega$ with values $\rho_0$ and $\rho_1$ represents the structure of a photonic crystal built from the same two materials. We shall call such functions *admissible functions*. For every admissible function $\rho$ and every vector $\boldsymbol{k} \in \mathbb{B}$, one can solve (1.1) and obtain solutions $\omega_1(\rho, \boldsymbol{k}) \leq \omega_2(\rho, \boldsymbol{k}) \leq \cdots$, etc., which determine the band structure of the photonic crystal represented by $\rho$. Given a fixed index $j \in \mathbb{N}$ the width of the band gap between the $j$-th and $(j+1)$-th band of that crystal, in case it exists, is

given by

$$w_j(\rho) := \min_{\boldsymbol{k}\in\mathbb{B}} \omega_{j+1}(\rho, \boldsymbol{k}) - \max_{\boldsymbol{k}\in\mathbb{B}} \omega_j(\rho, \boldsymbol{k}).$$

With this, the problem of finding a photonic crystal structure, which exhibits a maximal band gap between the $j$-th and $(j+1)$-th photonic band can be formulated as follows: Find an admissible function $\rho^\star$, such that $w_j(\rho^\star) \geq w_j(\rho)$ for every admissible function $\rho$. A problem of this type is commonly referred to as a *photonic band gap maximization problem* (hereafter abbreviated as PBGMP).

More generally, let $J$ be a functional, such that for every admissible function $\rho$ the corresponding value $J(\rho)$ of the functional is determined by the functions $\boldsymbol{k} \mapsto \omega_1(\rho, \boldsymbol{k})$, $\boldsymbol{k} \mapsto \omega_2(\rho, \boldsymbol{k}), \ldots$, etc. An optimization problem could then consist in finding an admissible function $\rho^\star$, such that $J(\rho^\star) \leq J(\rho)$ for every admissible function $\rho$. We shall call a problem of this type a *photonic band structure optimization problem* (hereafter abbreviated as PBSOP). The functional $J$ is commonly referred to as the *goal functional* of the problem. Clearly, photonic band gap maximization problems are those PBSOPs, whose goal functionals are given by $J = -w_j$ for some $j \in \mathbb{N}$.

Photonic band structure optimization problems present a number of interesting challenges for a mathematician. In many relevant cases the goal functional $J$ fails to be any more regular than Lipschitz continuous. Moreover, the set of admissible functions is infinite-dimensional and non-compact, which makes it difficult to prove the existence of optimal solutions. Computing the band structure of a photonic crystal is a challenging task from a numerical point of view, especially when the crystal is three-dimensional. One of the reasons for this is that admissible functions, which are by definition two-valued, enter as discontinuous coefficients in the eigenvalue problem (1.1). Therefore, sophisticated numerical schemes need to be employed in order to obtain accurate results. Finally, the development and analysis of numerical optimization algorithms for PBSOPs provides a field of research that mathematicians can contribute to.

In this work we discuss in detail the mathematical theory, which relates the geometric structure of a photonic crystal to its band structure, and develop different optimization algorithms for PBSOPs. A more general introduction to photonic crystals can be found in the books [42], [60], or [66]. A survey on mathematical concepts related to photonic crystals is given in [50].

## 1.2 Aims of This Work

The aim of this thesis is the mathematical study of photonic band structure optimization problems (PBSOPs), as well as the developement of suitable optimization algorithms for these problems. Moreover, we aim at finding solutions of photonic

band gap maximization problems (PBGMPs) for two- and three-dimensional photonic crystals.

## 1.3 Literature Review

Photonic band structure optimization problems are, in essence, eigenfrequency optimization problems. Problems of this type also occur e.g. in structural engineering and acoustics. As an example, we mention the problem of finding the optimal design of a two-density composite membrane, where the optimization goal typically is the minimization or maximization of certain eigenfrequencies. This problem was studied in detail by Cox and McLaughlin (cf. [31], [32]), who proved the existence of optimal designs and characterized them in terms of certain level-sets of eigenfunctions. Some fifty years earlier, Krein was able to identify two-density composite strings with minimal and maximal eigenfrequencies (cf. [48]). The eigenvalue problem in both cases is of the form

$$\begin{cases} -\Delta u = \lambda \rho u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where $\Omega$ is a bounded domain. The function $\rho$ is two-valued on $\Omega$ and models the density distribution of the string or membrane. As will turn out in Section 3.5, eigenvalue problems of similar type also arise for certain PBSOPs involving one- and two-dimensional photonic crystals. In Section 5.4 we will show how some results found by Cox, McLaughlin and Krein carry over to these problems.

Another optimization problem, which is in a way related to PBSOPs, is an optimal design problem concerning the temperature distribution in two-phase conductors. Under certain model assumptions the temperature distribution is determined by an eigenvalue problem of the form

$$\begin{cases} -\nabla \cdot [\rho \nabla u] = \lambda u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where $\rho$, again, is a two-valued function on a bounded domain $\Omega$. The problem was analysed by Cox and Lipman (cf. [27]) and further studied by Alvio, Lion and Trombetti (cf. [5]), as well as by Conca, Mahadevan and Sanz (cf. [26]). Obviously, the eigenvalue problem (1.3) bares some similarity to the the eigenvalue problem (1.1), which determines the band structure of a photonic crystal. In Section 5.4 we will discuss the findings of the authors mentioned above and illustrate how these findings relate to PBSOPs.

Photonic band gap maximization problems have already been considered in the literature, though only for two-dimensional photonic crystals. Cox and Dobson developed an optimization algorithm based on a generalized gradient descent method

(cf. [29]). This algorithm served as a prototype for the optimization algorithms we develop in Chapter 7. It should be noted, however, that the optimization problems considered by Cox and Dobson differ slightly from PBGMPs as defined in the previous Section 1.1. In their paper the authors choose a fixed frequency lying inside a photonic band gap and aim at pushing the nearest photonic bands away from that frequency. In doing so Cox and Dobson are able to find photonic crystals, which exhibit a large band gap with a prescribed frequency at the center. In a second paper Cox and Dobson refine their algorithm allowing them to open up photonic band gaps between specific bands (cf. [30]).

Using a level-set method (see e.g. [58]), Kao, Osher and Yablonovitch were able to find two-dimensional photonic crystals with maximal band gaps (cf. [44]). While providing an extensive list of optimized crystal structures, the authors do not establish a rigorous analysis of their method. A non-rigorous justification can be found, however, in a paper by Osher and Santosa on the optimization of two-density composite membranes via level-set methods (cf. [59]).

Yet another approach was presented by Sigmund and Søndergaard, who treated PBGMPs and PBSOPs as topology optimization problems (see e.g. [11]). In a paper on the systematic design of two-dimensional photonic crystals (cf. [69]) the authors compute optimized crystal structures using the method of moving asymptodes (see [71]). The authors remark, however, that this method could only be applied to rather coarse discretizations of the problem due to computing-time and storage limitations. Preble, M. Lipson, and H. Lipson used evolutionary algorithms in order to design two-dimensional photonic crystals with maximal band gaps (cf. [61]). Each evolutionary algorithm needed to create more than 1000 generations in order to find optimized crystal structures, which is why the primitive cell of each photonic crystal was discretized by a $32 \times 32$ grid only.

As we mentioned before, there are no published results neither on PBGMPs nor on PBSOPs for three-dimensional photonic crystals so far. A possible reason for this may be the fact that these problems are computationally more demanding than their two-dimensional analogues. As will be discussed in Section 3.5, photonic band structures of two-dimensional crystals are determined by Laplace-type or second-order divergence-type eigenvalue problems. Such eigenvalue problems can be discretized using standard finite difference or finite element methods, and band structure computations can be carried out on an average computer. The situation is much different for three-dimensional photonic crystals, whose band structures are determined by second-order curl-curl-type eigenvalue problems of the form (1.1). In order to obtain suitable discretizations of these problems, specialized finite element schemes have to be employed. Moreover, efficient numerical algorithms and high-performance computer architectures are generally needed in order to carry out band structure computations within reasonable time. Thus, the

numerical results presented in Chapter 9 were only made possible by the efforts of Bulovyatov (cf. [17]), who developed a parallel multi-grid algorithm for band structure computations of three-dimensional photonic crystals based on a finite element library developed by Wieners (see [76], [77]).

## 1.4 Outline

This work is organized as follows. In Chapter 2 we present the notation, certain function spaces and fundamental concepts that are used in this work. In Chapter 3 we introduce the mathematical model for wave propagation in photonic crystals, focussing on the underlying physical principles and model assumptions. A family of eigenvalue problems governing the propagation of so-called Bloch modes in photonic crystals is given. The formulation of the eigenvalue problems is made mathematically precise in Chapter 4. There, we also provide the appropriate spectral theory and introduce the band structure of a photonic crystal, as well as photonic band gaps. The chapter is completed by a discussion of the Floquet–Bloch theory, which relates the band structure of a photonic crystal to its transmissibility for electromagnetic radiation. In Chapter 5 we address the problem of optimizing the geometrical structure of a photonic crystal with respect to certain properties of its band structure. We establish a generic minimization problem and analyse its properties. In Chapter 5 we also state some model problems, which motivated this work. In Chapter 6 we review important concepts of nonsmooth analysis. In particular, the concept of generalized differentials is introduced. In Chapter 7 we develop an algorithm for the numerical solution of photonic band structure optimization problems. In essence, this algorithm is based on the idea of gradient descent methods. A slight drawback inherent to the algorithm is that it fails to produce admissible crystal structures in general. In Chapter 8 we therefore develop another optimization algorithm based on a level-set method. In Chapter 9 we present and discuss some numerical results of the algorithms. Conclusions and an outlook on future work are given in Chapter 10.

# Chapter 2

# Preliminaries

In this chapter we introduce some fundamental concepts, notations and conventions that will be used throughout the rest of this work. Basic notations and conventions are introduced in Section 2.1. Important results concerning the vector cross product are listed in Section 2.2. In Section 2.3 we discuss the notion of local Lipschitz continuity. In Section 2.4 we review the concept of symmetry and related aspects of group theory. Periodicity is a special case of symmetry and will be discussed in Section 2.5. In both Sections 2.4 and 2.5 we introduce fundamental mathematical structures, such as symmetry groups or Bravais lattices, which will be used in Chapter 3 to build a mathematical model for the medium structures of photonic crystals.

## 2.1 Notations and Conventions

In this work we denote by $\mathbb{N}$ the set of natural numbers not including zero, and by $\mathbb{N}_0$ the set $\mathbb{N} \cup \{0\}$. Euler's number and the imaginary unit are denoted by the upright letters e and i, respectively. We use boldface letters to denote vector- or matrix-valued quantities. Unless stated otherwise, $\boldsymbol{u}_i$ denotes the $i$-th component of the vector-valued quantity $\boldsymbol{u}$ for a given index $i \in \mathbb{N}$. Similarly, $\boldsymbol{A}_{ij}$ denotes the component in the $i$-th row and $j$-th column of the matrix-valued quantity $\boldsymbol{A}$ for given indices $i \in \mathbb{N}$ and $j \in \mathbb{N}$.

Given a complex-valued quantity $z$, we denote by $\overline{z}$ its complex conjugate. Sesquilinear forms on a complex vector space are understood to be conjugate-linear in the first (left-hand) argument and linear in the second (right-hand) argument. This applies, in particular, to the standard inner product on $\mathbb{C}^n$, $n \in \mathbb{N}$, which we denote by $\langle \cdot \, , \, \cdot \rangle$. In contrast to this, we denote by $\cdot$ the dot product on $\mathbb{C}^n$. The connection between the inner product and the dot product is given by $\langle \boldsymbol{x} \, , \boldsymbol{y} \rangle = \overline{\boldsymbol{x}} \cdot \boldsymbol{y}$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$. If $X$ is a real or complex topological vector space,

then $X^*$ denotes its topological dual, i.e., the space $X^*$ consists of all continuous linear functionals on $X$. Given another topological vector space $Y$ and an operator $T : X \to Y$, we sometimes denote the image of a point $x \in X$ under $T$ by $Tx$, if $T$ is linear or conjugate-linear, or by $T[x]$, if $T$ is linear. If $X$ and $Y$ are normed spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, and if $T$ is a continuous, linear or conjugate-linear operator, we denote by $\|T\|_{X,Y}$ the operator norm of $T$, given by

$$\|T\|_{X,Y} := \sup_{x \in X \setminus \{0\}} \frac{\|Tx\|_Y}{\|x\|_X}.$$

Given a set $S \subset X$, we denote by $\mathrm{conv}(S)$ the open, convex hull of $S$ and by $\overline{\mathrm{conv}}(S)$ its closure.

Given a set $\Omega \subseteq \mathbb{R}^n$, where $n \in \mathbb{N}$, a Banach space $Y$ with norm $\|\cdot\|_Y$, and a real number $p \in \mathbb{R}$, with $1 \leq p < \infty$, we define the norms $\|\cdot\|_{\Omega,p}$ and $\|\cdot\|_{\Omega,\infty}$ by

$$\begin{aligned} \|f\|_{\Omega,p} &:= \left( \int_\Omega \|f\|_Y^p \right)^{1/p} && \text{for all } f \in L^p(\Omega, Y), \\ \|f\|_{\Omega,\infty} &:= \operatorname*{ess\,sup}_\Omega \big( \|f\|_Y \big) && \text{for all } f \in L^\infty(\Omega, Y). \end{aligned}$$

As a convention, we further constitute that $\|\cdot\|_\Omega := \|\cdot\|_{\Omega,2}$. If $Y$ is a Hilbert space with inner product $\langle \cdot\ ,\ \cdot \rangle_Y$, we define the inner product $\langle \cdot\ ,\ \cdot \rangle_\Omega$ by

$$\langle f\ , g \rangle_\Omega := \int_\Omega \langle f\ , g \rangle_Y \qquad \text{for all } f, g \in L^2(\Omega, Y).$$

Whenever $F(\Omega, Y)$ denotes a set of functions mapping $\Omega$ into $Y$, we denote by $F(\Omega)$ the respective set of functions mapping $\Omega$ into $\mathbb{C}$. This convention applies, in particular, to the standard Lebesgue, Sobolev, and Hölder spaces $L^p(\Omega)$, $H^s(\Omega)$, and $C^{m,\alpha}(\Omega)$.

By $\boldsymbol{H}(\mathrm{curl}; \Omega)$ and $\boldsymbol{H}(\mathrm{div}; \Omega)$ we denote the Sobolev spaces, which consist of all functions from $\Omega$ into $\mathbb{C}^3$ that admit weak curl or divergence fields in $L^2(\Omega)^3$ or $L^2(\Omega)$, respectively. A detailed characterization of the Sobolev space $\boldsymbol{H}(\mathrm{curl}; \Omega)$ can be found in Chapter 7, Section 4.1 in [34]. The Sobolev space $\boldsymbol{H}(\mathrm{div}; \Omega)$ is discussed in detail in Section 2.1 in [38]. The spaces $\boldsymbol{H}(\mathrm{curl}; \Omega)$, $\boldsymbol{H}(\mathrm{div}; \Omega)$ and $H^1(\Omega)$ are equipped with the inner products $\langle \cdot\ ,\ \cdot \rangle_{\mathrm{curl},\Omega}$, $\langle \cdot\ ,\ \cdot \rangle_{\mathrm{div},\Omega}$, and $\langle \cdot\ ,\ \cdot \rangle_{1,\Omega}$, respectively, which are given by

$$\begin{aligned} \langle \boldsymbol{u}\ , \boldsymbol{v} \rangle_{\mathrm{curl},\Omega} &:= \langle \nabla \times \boldsymbol{u}\ , \nabla \times \boldsymbol{v} \rangle_\Omega + \langle \boldsymbol{u}\ , \boldsymbol{v} \rangle_\Omega && \text{for all } \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{H}(\mathrm{curl}; \Omega), \\ \langle \boldsymbol{u}\ , \boldsymbol{v} \rangle_{\mathrm{div},\Omega} &:= \langle \nabla \cdot \boldsymbol{u}\ , \nabla \cdot \boldsymbol{v} \rangle_\Omega + \langle \boldsymbol{u}\ , \boldsymbol{v} \rangle_\Omega && \text{for all } \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{H}(\mathrm{div}; \Omega), \\ \langle p\ , q \rangle_{1,\Omega} &:= \langle \nabla p\ , \nabla q \rangle_\Omega + \langle p\ , q \rangle_\Omega && \text{for all } p, q \in H^1(\Omega). \end{aligned}$$

Here as in the following, we denote by $\nabla\times$, $\nabla\cdot$, and $\nabla$ the curl, divergence and gradient operator. The norms, which are induced by the inner products, are denoted by $\|\cdot\|_{\mathrm{curl};\Omega}$, $\|\cdot\|_{\mathrm{div};\Omega}$, and $\|\cdot\|_{1;\Omega}$, respectively.

## 2.2 The Cross Product

In this work we quite perform vector calculations involving the cross product. Therefore, we find it useful to mention some important results related to this vector operation here.

First, let us recall some vector identities. An important identity, which concerns so-called *vector triple products*, reads

$$\boldsymbol{x} \times (\boldsymbol{y} \times \boldsymbol{z}) = (\boldsymbol{x} \cdot \boldsymbol{z})\boldsymbol{y} - (\boldsymbol{x} \cdot \boldsymbol{y})\boldsymbol{z} \qquad \text{for all } \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{C}^3. \tag{2.1}$$

For so-called *scalar triple products* we have the well-known identity

$$\boldsymbol{x} \cdot (\boldsymbol{y} \times \boldsymbol{z}) = \boldsymbol{y} \cdot (\boldsymbol{z} \times \boldsymbol{x}) = \boldsymbol{z} \cdot (\boldsymbol{x} \times \boldsymbol{y}) \qquad \text{for all } \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{C}^3, \tag{2.2}$$

which is often verbalized by saying the the scalar triple product is invariant under cyclic permutations of the vectors. Furthermore, *Lagrange's identity* states that

$$\langle \boldsymbol{x} \times \boldsymbol{y}\,, \widetilde{\boldsymbol{x}} \times \widetilde{\boldsymbol{y}} \rangle = \langle \boldsymbol{x}\,, \widetilde{\boldsymbol{x}} \rangle \langle \boldsymbol{y}\,, \widetilde{\boldsymbol{y}} \rangle + \langle \boldsymbol{y}\,, \widetilde{\boldsymbol{x}} \rangle \langle \boldsymbol{x}\,, \widetilde{\boldsymbol{y}} \rangle \qquad \text{for all } \boldsymbol{x}, \widetilde{\boldsymbol{x}}, \boldsymbol{y}, \widetilde{\boldsymbol{y}} \in \mathbb{C}^3. \tag{2.3}$$

Letting $\widetilde{\boldsymbol{x}} = \boldsymbol{x}$ and $\widetilde{\boldsymbol{y}} = \boldsymbol{y}$ in (2.3), we find that

$$|\boldsymbol{x}|^2 |\boldsymbol{y}|^2 = |\langle \boldsymbol{x}\,, \boldsymbol{y} \rangle|^2 + |\boldsymbol{x} \times \boldsymbol{y}|^2 \qquad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^3, \tag{2.4}$$

where $|\cdot|$ denotes the Euclidean norm on $\mathbb{C}^3$. This norm identity in particular implies the following Cauchy–Schwarz-like inequality, which reads

$$|\boldsymbol{x} \times \boldsymbol{y}| \leq |\boldsymbol{x}||\boldsymbol{y}| \qquad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^3.$$

Next, we introduce a useful concept for representing the action of cross products on $\mathbb{C}^3$. Given a vector $\boldsymbol{z} \in \mathbb{C}^3$, we define the matrix $[\boldsymbol{z}]_\times \in \mathbb{C}^{3\times 3}$ by

$$[\boldsymbol{z}]_\times := \begin{pmatrix} 0 & -\boldsymbol{z}_3 & \boldsymbol{z}_2 \\ \boldsymbol{z}_3 & 0 & -\boldsymbol{z}_1 \\ -\boldsymbol{z}_2 & \boldsymbol{z}_1 & 0 \end{pmatrix}. \tag{2.5}$$

We refer to the matrix $[\boldsymbol{z}]_\times$ as the the *cross product matrix* of $\boldsymbol{z}$. One easily verifies that

$$[\boldsymbol{z}]_\times \boldsymbol{x} = \boldsymbol{z} \times \boldsymbol{x} \qquad \text{for all } \boldsymbol{x} \in \mathbb{C}^3.$$

For $\boldsymbol{z} \in \mathbb{C}^3 \setminus \{\boldsymbol{0}\}$ one finds that the null space and the image space of $[\boldsymbol{z}]_\times$ are given by

$$\begin{aligned} \ker\big([\boldsymbol{z}]_\times\big) &= \operatorname{span}\{\boldsymbol{z}\}, \\ \operatorname{im}\big([\boldsymbol{z}]_\times\big) &= \big(\operatorname{span}\{\boldsymbol{z}\}\big)^\perp, \end{aligned}$$

where $(\operatorname{span}\{\boldsymbol{z}\})^{\perp}$ denotes the orthogonal complement of $\operatorname{span}\{\boldsymbol{z}\}$ in $\mathbb{C}^3$.

Finally, we mention a vector identity concerning the behaviour of cross products under linear coordinate transformations. Given a matrix $\boldsymbol{A} \in \mathbb{C}^{3\times 3}$, one can show by direct computation that

$$(\boldsymbol{A}\boldsymbol{x}) \times (\boldsymbol{A}\boldsymbol{y}) = \operatorname{cof}(\boldsymbol{A})(\boldsymbol{x} \times \boldsymbol{y}) \qquad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^3, \tag{2.6}$$

where $\operatorname{cof}(\boldsymbol{A})$ denotes the *cofactor matrix* of $\boldsymbol{A}$. The cofactor matrix is given by

$$\operatorname{cof}(\boldsymbol{A})_{ij} := (-1)^{i+j} \det\big(\boldsymbol{A}^{[i,j]}\big) \qquad \text{for all } i, j = 1, 2, 3, \tag{2.7}$$

where $\boldsymbol{A}^{[i,j]}$ denotes the matrix, which is obtained by omitting the $i$-th row and the $j$-th column in the matrix $\boldsymbol{A}$ for $i, j = 1, 2, 3$. Recall that the cofactor matrix $\operatorname{cof}(\boldsymbol{A})$ is the transposed of the so-called *adjugate matrix* $\operatorname{adj}(\boldsymbol{A})$ of $\boldsymbol{A}$. We thus have

$$\operatorname{cof}(\boldsymbol{A}) = \det(\boldsymbol{A})\boldsymbol{A}^{-\mathrm{T}} \qquad \text{for all } \boldsymbol{A} \in \mathrm{GL}_3(\mathbb{C}), \tag{2.8}$$

where $\mathrm{GL}_3(\mathbb{C})$ denotes the set consisting of all invertible $3 \times 3$ matrices with complex components.

## 2.3 Local Lipschitz Continuity

In this section we review the concept of locally Lipschitz continuous functions on Banach spaces. It will turn out that such functions play an important role in optimization problems involving band structures of photonic crystals.

Given two real or complex Banach spaces $X$ and $Y$ with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, a function $f : X \to Y$ is called *Lipschitz continuous*, if there exists a non-negative number $c_f \geq 0$, which depends on $f$ only, such that

$$\|f(x) - f(y)\|_Y \leq c_f \|x - y\|_X \quad \text{for all } x, y \in X. \tag{2.9}$$

Given a constant $c > 0$, we shall call the function $f$ above *c-Lipschitz continuous*, if (2.9) holds for $c_f = c$. Every number $c_f \geq 0$, which satisfies (2.9), is called a *Lipschitz constant* of $f$. It is well-known that Lipschitz continuity implies uniform continuity.

Given a point $x_0 \in X$, a function $f : X \to Y$ is called *Lipschitz continuous near* $x_0$, if there exists a neighbourhood $U \subseteq X$ of $x_0$, such that $f$ is Lipschitz continuous on $U$, i.e., if there exists a constant $c_{f,U} > 0$, which only depends on $f$ and the neighbourhood $U$, such that

$$\|f(x) - f(y)\|_Y \leq c_{f,U} \|x - y\|_X \quad \text{for all } x, y \in U. \tag{2.10}$$

Given a constant $c > 0$, we shall call the function $f$ above *c-Lipschitz continuous near* $x_0$, if there exists a neighbourhood $U$ of $x_0$, such that (2.10) holds for $c_{f,U} = c$. A function $f : X \to Y$ is called *locally Lipschitz continuous*, if it is Lipschitz continuous near every point $x \in X$.

Clearly, every Lipschitz continuous function is also locally Lipschitz continuous. It is also easy to see that local Lipschitz continuity implies continuity. It should be noted, however, that local Lipschitz continuity does not imply uniform continuity. To see this, consider e.g. the function $x \mapsto 1/x$ from $\mathbb{R}_{>0}$ into $\mathbb{R}$. Using the Mean-Value Theorem, one can show that this function is locally Lipschitz continuous, but not uniformly continuous. Finally, we remark that a function defined on a compact set is Lipschitz continuous if and only if it is locally Lipschitz continuous.

## 2.4 Symmetry

In this section we briefly review some mathematical concepts related to symmetry. As is well-known, symmetry can be described in terms of groups of isometries acting on metric spaces. Recall that a group $G$ with neutral element $e$ acts on a non-empty set $X$ by virtue of a mapping $G \times X \to X$, which maps each pair $(g, x) \in G \times X$ to an element $x^g \in X$, and which satisfies $x^{gh} = (x^g)^h$, as well as $x^e = x$ for all $g, h \in G$ and all $x \in X$. This mapping $(g, x) \mapsto x^g$ is called the *group action of* $G$ *on* $X$. Given a group element $g \in G$, an element $x \in X$ is said to be *invariant under the action of* $g$ *on* $X$, if $x^g = x$. Furthermore, $x$ is said to be *invariant under the action of* $G$ *on* $X$, if $x^g = x$ for all $g \in G$. For each element $x \in X$ the set $G_x := \{g \in G \mid x^g = x\}$ is called the *stabilizer of* $x$ *in* $G$. The stabilizer of $x$ in $G$ is a subgroup of $G$ (notation: $G_x \leq G$). More precisely, $G_x$ is the maximal subgroup of $G$ under whose action on $X$ the element $x$ is invariant. In general, $G_x$ is not a normal subgroup of of $G$.

Given a metric space $X$, we denote by $\mathrm{Iso}(X)$ the *isometry group of* $X$, which consists of all isometries from the metric space $X$ onto itself with the function composition as group operation. The isometry group $\mathrm{Iso}(X)$ canonically acts on $X$, on the power set $\mathcal{P}(X)$, and on every set of functions from $X$ into another set $Y$. The respective group actions $\mathrm{Iso}(X) \times X \to X$, $\mathrm{Iso}(X) \times \mathcal{P}(X) \to \mathcal{P}(X)$, and $\mathrm{Iso}(X) \times Y^X \to Y^X$ are given by

$$\begin{aligned} x^\varphi &:= \varphi(x) && \text{for all } \varphi \in \mathrm{Iso}(X),\ x \in X, \\ S^\varphi &:= \varphi(S) && \text{for all } \varphi \in \mathrm{Iso}(X),\ S \in \mathcal{P}(X), \\ f^\varphi &:= f \circ \varphi && \text{for all } \varphi \in \mathrm{Iso}(X),\ f \in Y^X. \end{aligned}$$

It is well-known that the isometry group $\mathrm{Iso}(\mathbb{R}^n)$ of the $n$-dimensional Euclidean space, where $n \in \mathbb{N}$, consists exactly of those affine mappings $\boldsymbol{\varphi} : \mathbb{R}^n \to \mathbb{R}^n$, which

are of the form

$$\boldsymbol{\varphi}(\boldsymbol{x}) := \boldsymbol{\Theta}\boldsymbol{x} + \boldsymbol{a},$$

where $\boldsymbol{\Theta} \in \mathrm{O}_n(\mathbb{R})$ is an orthogonal $n \times n$ matrix with components in $\mathbb{R}$, and where $\boldsymbol{a} \in \mathbb{R}^n$ is a vector. Depending on the nature of $\boldsymbol{\Theta}$ and $\boldsymbol{a}$, specific elements $\boldsymbol{\varphi}$ of $\mathrm{Iso}(\mathbb{R}^n)$ can be identified e.g. as *rotations*, *reflexions*, *inversions*, or *translations* in $\mathbb{R}^n$.

Given a subgroup $G$ of the isometry group $\mathrm{Iso}(X)$ of an arbitrary metric space $X$, a subset $S$ of $X$ is called *symmetric with respect to* $G$ or simply $G$*-symmetric*, if $S$ is invariant under the action of $G$ on $\mathcal{P}(X)$, i.e., if

$$\varphi(S) = S \qquad \text{for all } \varphi \in G.$$

Similarly, a function $f \in Y^X$, where $Y$ is a non-empty set, is called $G$-symmetric, if $f$ is invariant under the action of $G$ on $Y^X$, i.e., if

$$f \circ \varphi = f \qquad \text{for all } \varphi \in G.$$

The corresponding stabilizers $\mathrm{Iso}(X)_S$ and $\mathrm{Iso}(X)_f$ are called the *symmetry groups* of the set $S$ and the function $f$, respectively. Elements which belong to the symmetry group of an object are called *symmetry operations* of the object. If $X = \mathbb{R}^n$ for some $n \in \mathbb{N}$, one can think of a symmetry operation as a Euclidean motion, which takes an object into itself. The object's symmetry group is precisely the subgroup of $\mathrm{Iso}(\mathbb{R}^n)$, which consists of all such operations.

Given a subgroup $G$ of $\mathrm{Iso}(X)$, a connected subset $F$ of $X$ is called a *fundamental region* of $G$, if it satisfies

$$\begin{aligned} \varphi(F) \cap \psi(F) &= \emptyset \qquad \text{for all } \varphi, \psi \in G \text{ with } \varphi \neq \psi, \\ \bigcup_{\varphi \in G} \varphi(F) &= X. \end{aligned}$$

Hence, a fundamental region of $G$ is a set whose images under the actions of $G$ cover the entire metric space without overlapping. If follows that every $G$-symmetric function is determined by its restriction to a fundamental region of $G$. The interior of a fundamental region of $G$ is called a *fundamental domain* of $G$.

Given a finite subgroup $G$ of $\mathrm{Iso}(X)$ and a function $f \in Y^X$, where $Y$ is a vector space over $\mathbb{Q}$, we define the function $f^{(G)} \in Y^X$ by

$$f^{(G)} := \frac{1}{|G|} \sum_{\varphi \in G} (f \circ \varphi). \tag{2.11}$$

The function $f^{(G)}$ is called the $G$*-symmetrization of* $f$. Obviously, for every function $f \in Y^X$ the corresponding $G$-symmetrization $f^{(G)}$ is a $G$-symmetric function. Furthermore, one can show that $f^{(G)}$ coincides with $f$, if and only if $f$ is $G$-symmetric.

## 2.5 Periodicity

*Periodicity* is a special case of symmetry, namely that of symmetry with respect to a finitely generated group of translations. Given a vector $\boldsymbol{a} \in \mathbb{R}^n$, $n \in \mathbb{N}$, the mapping $\boldsymbol{\tau}_{\boldsymbol{a}} : \mathbb{R}^n \to \mathbb{R}^n$, which is defined by

$$\boldsymbol{\tau}_{\boldsymbol{a}}(\boldsymbol{x}) := \boldsymbol{x} + \boldsymbol{a} \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^n \tag{2.12}$$

is called the *translation in* $\mathbb{R}^n$ *by* $\boldsymbol{a}$. One easily verifies that the set of all translations in $\mathbb{R}^n$ constitutes a normal subgroup of $\mathrm{Iso}(\mathbb{R}^n)$, the isometry group of $\mathbb{R}^n$. A discrete subset $\Lambda$ of $\mathbb{R}^n$, which is of the form

$$\Lambda = \{z_1\boldsymbol{a}^{(1)} + \cdots + z_r\boldsymbol{a}^{(r)} \mid z_1, \ldots, z_r \in \mathbb{Z}\},$$

where $\boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(r)} \in \mathbb{R}^n$ are linearly independent vectors, is called a *Bravais lattice* or simply a *lattice in* $\mathbb{R}^n$ *of rank* $r$. The vectors $\boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(r)}$, which form a $\mathbb{Z}$-linear basis of $\Lambda$, are often referred to as *primitive vectors of* $\Lambda$. Here as in the following, the term $\mathbb{Z}$-linear refers to the algebraic structure of a vector space over the set of integer numbers $\mathbb{Z}$. The elements of a lattice are called *lattice points* or *lattice vectors*. Note that every lattice contains the origin of the coordinate system as a lattice point.

If $\Lambda_1$ and $\Lambda_2$ are two Bravais lattices in $\mathbb{R}^n$, such that $\Lambda_1 \subseteq \Lambda_2$, we call $\Lambda_1$ a *sublattice of* $\Lambda_2$ and write $\Lambda_1 \leq \Lambda_2$. Clearly, $\Lambda_1$ is a sublattice of $\Lambda_2$, if and only if there exists a set of primitive vectors of $\Lambda_1$, such that every vector is equal to a $\mathbb{Z}$-linear combination of primitive vectors of $\Lambda_2$.

Every Bravais lattice $\Lambda \subset \mathbb{R}^n$ is a finitely generated, free, Abelian additive subgroup of $\mathbb{R}^n$. The corresponding set of *lattice translations*

$$\mathrm{Trn}(\Lambda) := \{\boldsymbol{\tau}_{\boldsymbol{a}} \mid \boldsymbol{a} \in \Lambda\} \tag{2.13}$$

is a finitely generated, free, Abelian subgroup of the isometry group $\mathrm{Iso}(\mathbb{R}^n)$. Clearly, every sublattice of $\Lambda$ is a normal subgroup of $\Lambda$, and the corresponding subgroup of lattice translation is normal in $\mathrm{Trn}(\Lambda)$. Furthermore, we remark that every lattice $\Lambda \subset \mathbb{R}^n$ defines an equivalence relation $\equiv_\Lambda$ on $\mathbb{R}^n$, which is given by

$$\boldsymbol{x} \equiv_\Lambda \boldsymbol{y} \;:\Longleftrightarrow\; \boldsymbol{x} - \boldsymbol{y} \in \Lambda \qquad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \Lambda. \tag{2.14}$$

The corresponding factor group is commonly denoted by $\mathbb{R}^n/\Lambda$. If $\Lambda$ is a lattice of maximal rank $n$, this factor group is called a *torus* of dimension $n$.

With the concept of lattices at hand, the notion of periodicity can be made precise as follows. Given a Bravais lattice $\Lambda \subset \mathbb{R}^n$ and a non-empty set $Y$, a function $f : \mathbb{R}^n \to Y$ is called *periodic with respect to* $\Lambda$ or simply $\Lambda$*-periodic*, if it is invariant under the action of $\mathrm{Trn}(\Lambda)$, i.e., if

$$f \circ \boldsymbol{\tau}_{\boldsymbol{a}} = f \qquad \text{for all } \boldsymbol{a} \in \Lambda.$$

For every Bravais lattice $\Lambda$ there exist fundamental regions of the corresponding group of lattice translations $\mathrm{Trn}(\Lambda)$. The interior of such a fundamental region is usually referred to as a *primitive domain* of $\Lambda$. The closure is referred to as a *primitive cell* of $\Lambda$. The primitive cells of a lattice $\Lambda \subset \mathbb{R}^n$ are bounded sets, if and only if $\Lambda$ is a lattice of maximal rank $n$. All primitive cells of such a lattice $\Lambda$ share the same measure. This measure is commonly referred to as the *discriminant* of $\Lambda$. A primitive cell, which is uniquely defined for every Bravais lattice, is the so-called *Wigner–Seitz cell*. Given a lattice $\Lambda \subset \mathbb{R}^3$, the Wigner–Seitz cell $W_\Lambda$ of $\Lambda$ is given by

$$W_\Lambda := \big\{ \boldsymbol{x} \in \mathbb{R}^n \;\big|\; |\boldsymbol{x}| \leq |\boldsymbol{x} - \boldsymbol{a}| \text{ for all } \boldsymbol{a} \in \Lambda \big\}. \tag{2.15}$$

The set $W_\Lambda$ consists of all points, which are closer to the origin than to any other point in the lattice $\Lambda$. Hence, $W_\Lambda$ is equal to the closure of the origin's Voronoi region with respect to $\Lambda$ (see e.g. Section 1.1 in [47]). Since every Voronoi region can be characterized as the intersection of finitely many open half-spaces, it follows that every Wigner–Seitz cell is a closed, convex polytope. Finally, we remark that every Wigner–Seitz cell is point-symmetric with respect to the origin. This is due to the fact that every Bravais lattice is point-symmetric with respect to each of its lattice points. For every lattice $\Lambda$ we consequently have that

$$|\boldsymbol{x}| \leq \frac{1}{2} \operatorname{diam}\big(W_\Lambda\big) \qquad \text{for all } \boldsymbol{x} \in W_\Lambda.$$

Finally, we introduce the concept of reciprocal lattices. For every Bravais lattice $\Lambda \in \mathbb{R}^n$ of rank $r$ there exists a uniquely defined lattice $\widehat{\Lambda} \subset \mathbb{R}^n$, which is also of rank $r$, such that

$$\boldsymbol{a} \cdot \boldsymbol{b} \in 2\pi\mathbb{Z} \qquad \text{for all } \boldsymbol{a} \in \Lambda,\ \boldsymbol{b} \in \widehat{\Lambda}$$

This lattice $\widehat{\Lambda}$ is called the *reciprocal lattice of* $\Lambda$. If $\boldsymbol{A} \in \mathbb{R}^{n \times r}$ is a matrix, whose columns form a $\mathbb{Z}$-linear basis of $\Lambda$, then the columns of the matrix $2\pi \boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}$ form a $Z$-linear basis of the reciprocal lattice $\widehat{\Lambda}$. Reciprocal lattices appear e.g. in the definition of Fourier series (see Section 4.2). As will be discussed in Section 3.4, they also play an important role in the definition of so-called Bloch modes, which form a special class of solutions of differential equations with periodic coefficients.

# Chapter 3

# The Mathematical Model

In this chapter we develop a mathematical model, which will allow us to describe the medium structure of photonic crystals, as well as the propagation of light waves in photonic crystals. We start off by making certain assumptions on the medium structure of photonic crystals in Section 3.1. In Section 3.2 we investigate in detail the symmetries, which can arise for these medium structures. We also introduce some terminology from crystallography, since this terminology is often used in the literature on photonic crystals. In Section 3.3 we present the mathematical model, which describes the propagation of light waves in certain dielectric media. A specific ansatz for light waves in spatially periodic media is introduced in Section 3.4. This so-called Bloch ansatz leads to a family of eigenvalue problems, which determines whether or not a light wave is able to propagate inside a photonic crystal. In Section 3.5 we consider the special case of two-dimensional photonic crystals and introduce simplified models for electromagnetic wave propagation therein.

## 3.1 Modelling Photonic Crystals

Throughout this work we assume that a photonic crystal is a medium, which occupies the entire three-dimensional Euclidean space $\mathbb{R}^3$. In doing so we focus on the propagation of light waves inside the photonic crystal, neglecting any effects occurring at its boundary. Furthermore we assume that a photonic crystal consists of a finite number of different materials. The crystal's *medium structure*, i.e. the distribution of the different materials within the crystal, can then be represented by a function $\chi : \mathbb{R}^3 \to \mathbb{N}$, in the sense that every function value $\chi(\boldsymbol{x})$ represents the material at the corresponding point $\boldsymbol{x} \in \mathbb{R}^3$. In the following we shall call this function the *crystal function*. We remark that many artificially created photonic crystals simply consist of a dielectric background material, such as silicon, into which some sort of holes are placed periodically. Such crystals can be represented

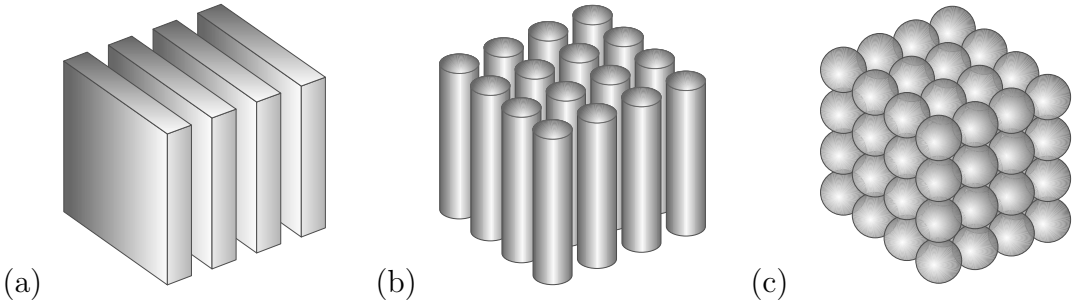



**Figure 3.1:** Examples of a one-dimensional (a), two-dimensional (b), and three-dimensional (c) photonic crystal.

by a two-valued crystal function with values in the set $\{1, 2\}$, where 1 models air and 2 the dielectric.

As characterizing feature, every photonic crystal exhibits some kind of spatial periodicity, which is captured by its crystal function. A photonic crystal is called *three-dimensional*, if its crystal function $\chi$ is periodic with respect to a Bravais lattice $\Lambda \subset \mathbb{R}^3$ of rank 3. Recall that such a lattice is given as the $\mathbb{Z}$-linear hull of three linearly independent primitive lattice vectors (see Section 2.5). The function $\chi$ is $\Lambda$-periodic, if and only if it is invariant under all lattice translations, i.e., if

$$\chi \circ \boldsymbol{\tau}_{\boldsymbol{a}} = \chi \qquad \text{for all } \boldsymbol{a} \in \Lambda,$$

where $\boldsymbol{\tau}_{\boldsymbol{a}}$ denotes the translation operator defined in (2.12). A photonic crystal is called *one-dimensional* or *two-dimensional*, if its crystal function is periodic with respect to a Bravais lattice of rank 1 or 2, respectively, and if the crystal function is furthermore invariant under all translations by vectors that are perpendicular to Bravais lattice. In summary, a photonic crystal is called $r$-dimensional, $r \in \{1, 2, 3\}$, if there exists a Bravais lattice $\Lambda \subset \mathbb{R}^3$ of rank $r$, such that the crystal function $\chi$ satisfies

$$\chi \circ \boldsymbol{\tau}_{\boldsymbol{y}} = \chi \qquad \text{for all } \boldsymbol{y} \in \Lambda \oplus \Lambda^{\perp}.$$

Note that the $\mathbb{R}$-linear hull of a lattice $\Lambda \subset \mathbb{R}^3$ of rank $r$ is a linear subspace of $\mathbb{R}^3$ of dimension $r$. Therefore, the orthogonal complement $\Lambda^{\perp}$ of $\Lambda$ is a linear subspace of $\mathbb{R}^3$ of dimension $3-r$. Simply put, a photonic crystal is called three-dimensional, if the periodicity of its medium structure extends into all three space dimensions. A photonic crystal is called two-dimensional, if its medium structure is *homogeneous* in one spatial direction while being periodic in all directions, which are perpendicular to that direction. Finally, a photonic crystal is called one-dimensional if its medium structure is periodic in one spatial direction and homogeneous on every plane, which is perpendicular to that direction. Examples of one-dimensional, two-dimensional, and three-dimensional photonic crystals are depicted Figure 3.1.

Without loss of generality, we assume in the following that the Bravais lattice $\Lambda$, which describes the spatial periodicity of an $r$-dimensional photonic crystal, can be chosen such that

$$\operatorname{span}(\Lambda) = \{\boldsymbol{x} \in \mathbb{R}^3 \mid \boldsymbol{x}_i = 0 \text{ for all } i > r\},$$

where $\operatorname{span}(\Lambda)$ denotes the $\mathbb{R}$-linear hull of $\Lambda$. With this the lattice Bravais lattice can be identified with a lattice in $\mathbb{R}^r$ of maximal rank $r$. Furthermore, the crystal function can be identified with a function mapping $\mathbb{R}^r$ into the set of material indices. These identification allow us to treat $r$-dimensional photonic crystals as $r$-dimensional media.

## 3.2 Crystal Symmetries

In this section we discuss certain aspects related to the symmetry of photonic crystals. By definition, every photonic crystal exhibits some kind of periodicity, which can be described in terms of a Bravais lattice. Aside from this translational symmetry the medium structures of most photonic crystals exhibit further symmetries. Knowing these non-translational symmetries is essential when analysing the crystal's optical properties.

The scientific discipline, which is primarily concerned with the study and classification of symmetries in spatially periodic media, is *crystallography*. In crystallography, sophisticated classification systems and specialized notations are introduced in order to characterize and designate the symmetry groups of periodic media. In this section we introduce some of the crystallographic terminology, which is often used in the literature on photonic crystals. The main intent of this section is to provide a dictionary, which translates certain notions from crystallography into group theoretic concepts. A more detailed discussion can be found e.g. in [6] and [62]. Specialized notation systems, such as the Schönflies notation or the Hermann–Mauguin notation, which are commonly used in crystallography, are not introduced in this section. They are covered by most standard textbooks on crystallography, such as [15].

In Section 2.5 we introduced *primitive cells* of Bravais lattices as the closures of fundamental regions of the group of lattice translations. A related concept in crystallography is that of unit cells. Let $\Lambda \subset \mathbb{R}^n$ be a Bravais lattice of rank $r \leq n$. Then, a set $P \subset \mathbb{R}^n$ of the form

$$P := \left\{ x_1 \boldsymbol{a}^{(1)} + \ldots + x_r \boldsymbol{a}^{(r)} \;\middle|\; x_1, \ldots, x_r \in [0,1] \right\},$$

where $\boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(r)} \in \Lambda$ are linearly independent lattice vectors, is called a *unit cell* of $\Lambda$. Note that the unit cells of a lattice of rank 3 are parallelepipeds, while

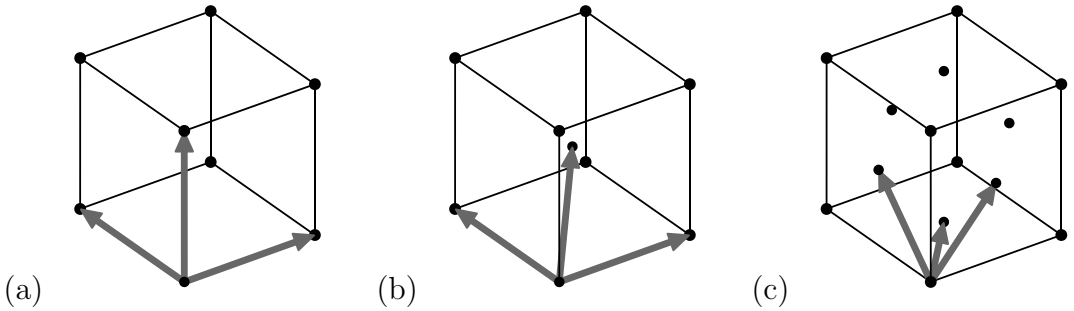



**Figure 3.2:** The standard unit cells of a simple cubic (a), body-centered cubic (b), and face-centered cubic (c) lattice. The arrows indicate primitive lattice vectors.

the unit cells of a lattice of rank 2 are parallelograms. A unit cell of a lattice $\Lambda$ is called *primitive*, if it is spanned by primitive vectors of $\Lambda$ only (see Section 2.5). Clearly, every primitive unit cell of $\Lambda$ is a primitive cell of $\Lambda$. A unit cell, which is not primitive, is called *centered*.

In crystallography, Bravais lattices are categorized according to the shape of their primitive unit cells. For every lattice type, crystallography defines a so-called *standard unit cell*. For many lattice types these standard unit cells are primitive or centered unit cells in the sense that was defined above. For some lattice types, however, other polytopes are considered as standard unit cells. In three space dimensions there exist 14 different *crystallographic lattice types* in total. Each lattice type belongs to exactly one of 7 so-called *crystallographic lattice systems*. In this work, we only consider the three lattice types, which constitute to the so-called *cubic* lattice system. These ones are the so-called simple cubic, body-centered cubic (bcc), and face-centered cubic (fcc) lattices.

A Bravais lattice $\Lambda$ of rank 3 is called *simple cubic*, if there exist primitive vectors $\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}, \boldsymbol{a}^{(3)}$ of $\Lambda$ that satisfy

$$\begin{aligned} \left|\boldsymbol{a}^{(1)}\right| = \left|\boldsymbol{a}^{(2)}\right| = \left|\boldsymbol{a}^{(3)}\right| =: a > 0, \\ \angle\left(\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}\right) = \angle\left(\boldsymbol{a}^{(2)}, \boldsymbol{a}^{(3)}\right) = \angle\left(\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(3)}\right) = \frac{\pi}{2}. \end{aligned}$$

Clearly, a Bravais lattice is a simple cubic lattice, if and only if its primitive unit cells are cubes. The length $a$ of the cubes' edges is called the *lattice constant*. Note that every simple cubic lattice is similar to the *standard simple cubic lattice* $\mathbb{Z}^3$. Choosing $\boldsymbol{e}^{(1)}$, $\boldsymbol{e}^{(2)}$, and $\boldsymbol{e}^{(3)}$ as primitive vectors, where $\boldsymbol{e}^{(i)}$ denotes the $i$-th standard basis vector in $\mathbb{R}^3$ for $i = 1, 2, 3$, one obtains the primitive unit cell $P_3$ of $\mathbb{Z}^3$, which is given by

$$P_3 := [0, 1]^3. \tag{3.1}$$

In crystallography, $P_3$ is the standard unit cell of $\mathbb{Z}^3$.

Next, we introduce body-centered cubic lattices. A Bravais lattice $\Lambda$ of rank 3 is called *body-centered cubic* or *bcc*, if there exist primitive vectors $\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}, \boldsymbol{a}^{(3)}$ of $\Lambda$, such that

$$\begin{aligned} \big|\boldsymbol{a}^{(1)}\big| = \big|\boldsymbol{a}^{(2)}\big| =: a > 0, \quad \big|\boldsymbol{a}^{(3)}\big| = \frac{\sqrt{3}a}{2}, \\ \angle\big(\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}\big) = \frac{\pi}{2}, \quad \angle\big(\boldsymbol{a}^{(2)}, \boldsymbol{a}^{(3)}\big) = \angle\big(\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(3)}\big) = \frac{\pi}{4}. \end{aligned}$$

The primitive cells of a bcc lattice are parallelepipeds with exactly two parallel square faces. The edge lengths of these faces are given by the lattice constant $a$. The distance between the square faces is equal to $a/2$. One quickly discovers that every bcc lattice can be characterized as the union of two staggered simple cubic lattices. More precisely, every bcc lattice is similar to the *standard bcc lattice* $\Lambda_{\mathrm{bcc}}$, which is defined as

$$\Lambda_{\mathrm{bcc}} := \Big\{ z_1\boldsymbol{e}^{(1)} + z_2\boldsymbol{e}^{(2)} + \frac{z_3}{2}\big(\boldsymbol{e}^{(1)} + \boldsymbol{e}^{(2)} + \boldsymbol{e}^{(3)}\big) \;\Big|\; z_1, z_2, z_3 \in \mathbb{Z} \Big\}. \tag{3.2}$$

One easily verifies that $\mathbb{Z}^3$ is a sublattice of $\Lambda_{\mathrm{bcc}}$ (see Section 2.5). Because of this, the standard unit cell $P_3$ of $\mathbb{Z}^3$ is a centered unit cell of $\Lambda_{\mathrm{bcc}}$. In crystallography, $P_3$ is considered to be the standard unit cell of $\Lambda_{\mathrm{bcc}}$.

The third class of Bravais lattices, which belong to the cubic lattice system, consists of the so-called face-centered cubic lattices. A Bravais lattice $\Lambda$ of rank 3 is called *face-centered cubic* or *fcc*, if there exist primitive vectors $\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}, \boldsymbol{a}^{(3)}$ of $\Lambda$, such that

$$\begin{aligned} \big|\boldsymbol{a}^{(1)}\big| = \big|\boldsymbol{a}^{(2)}\big| = \big|\boldsymbol{a}^{(3)}\big| := \frac{\sqrt{2}a}{2}, \quad a > 0, \\ \angle\big(\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}\big) = \angle\big(\boldsymbol{a}^{(2)}, \boldsymbol{a}^{(3)}\big) = \angle\big(\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(3)}\big) = \frac{\pi}{3}. \end{aligned}$$

The primitive unit cells of a fcc lattice are rhombohedra. The edge lengths of these rhombohedra is given by $\sqrt{2}a/2$, where $a$ denotes the lattice constant. Every fcc lattice is therefore similar to the *standard fcc lattice* $\Lambda_{\mathrm{fcc}}$, which is defined by

$$\Lambda_{\mathrm{fcc}} := \Big\{ \frac{z_1}{2}\big(\boldsymbol{e}^{(1)} + \boldsymbol{e}^{(2)}\big) + \frac{z_2}{2}\big(\boldsymbol{e}^{(1)} + \boldsymbol{e}^{(3)}\big) + \frac{z_3}{2}\big(\boldsymbol{e}^{(2)} + \boldsymbol{e}^{(3)}\big) \;\Big|\; z_1, z_2, z_3 \in \mathbb{Z} \Big\}. \tag{3.3}$$

Since $\mathbb{Z}^3$ is also a sublattice of $\Lambda_{\mathrm{fcc}}$, the cube $P_3$ is also considered to be the standard unit cell of $\Lambda_{\mathrm{fcc}}$. In Figure 3.2 we depict the standard unit cells of all three lattice types, which belong to the cubic lattice system.

In two space dimensions, crystallography distinguishes 5 different lattice types. Here, we only consider square and hexagonal lattices. A Bravais lattice $\Lambda$ of rank 2 is called *square*, if there exist primitive vectors $\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}$ of $\Lambda$, such that

$$\big|\boldsymbol{a}^{(1)}\big| = \big|\boldsymbol{a}^{(2)}\big| =: a > 0,$$

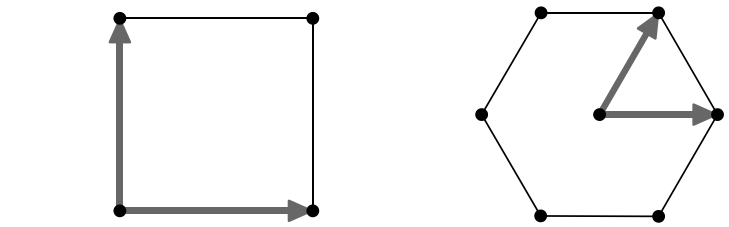



**Figure 3.3:** The standard unit cells of a square (a) and a hexagonal lattice. The arrows indicate primitive lattice vectors.

$$\angle\big(\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}\big) = \frac{\pi}{2}.$$

As the name indicates, square lattices are exactly those Bravais lattices, which feature square primitive unit cells. The side lengths of these squares are given by the lattice constant $a$. Every square lattice is similar to the *standard square lattice* $\mathbb{Z}^2$. The standard unit cell of $\mathbb{Z}^2$ is the square $P_2$, which is defined by

$$P_2 := [0,1]^2. \tag{3.4}$$

Clearly, $P_2$ is a primitive unit cell of $\mathbb{Z}^2$.

Finally, we introduce the so-called hexagonal lattices of rank 2. A Bravais lattice $\Lambda$ of rank 2 is called *hexagonal*, if one can find primitive vectors $\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}$ of $\Lambda$, such that

$$\begin{aligned} \big|\boldsymbol{a}^{(1)}\big| = \big|\boldsymbol{a}^{(2)}\big| =: a > 0, \\ \angle\big(\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}\big) = \frac{\pi}{3}. \end{aligned}$$

The primitive unit cells of a hexagonal lattice are rhombi with side lengths equal to the lattice constant $a$ and angles measuring 60° and 120°. Every hexagonal lattice is hence similar to the *standard hexagonal lattice* $\Lambda_{\text{hex}}$, given by

$$\Lambda_{\text{hex}} := \Big\{ z_1 \boldsymbol{e}^{(1)} + z_2 \Big( \frac{1}{2}\, \boldsymbol{e}^{(1)} + \frac{\sqrt{3}}{2}\, \boldsymbol{e}^{(2)} \Big) \;\Big|\; z_1, z_2 \in \mathbb{Z} \Big\}, \tag{3.5}$$

where $\boldsymbol{e}^{(1)}$ and $\boldsymbol{e}^{(2)}$ denote the standard basis vectors of $\mathbb{R}^2$. The name *hexagonal* alludes to the fact that all lattice points in a hexagonal lattice are equidistant to their nearest neighbours. Hence, by taking the convex hull of the nearest neighbours of a lattice point, one obtains a regular hexagon. The hexagon with the origin at its center is defined as the standard unit cell of $\Lambda_{\text{hex}}$. In Figure 3.3 we

depict the standard unit cells of square and hexagonal lattices. It should be noted that there are also Bravais lattices of rank 3, which are called hexagonal. Such lattices are obtained through periodic translations of a hexagonal lattice of rank 2 in along the direction, which is perpendicular to that lattice.

Recall that the medium structure of an $r$-dimensional photonic crystal can be described by a crystal function $\chi : \mathbb{R}^r \to \mathbb{N}$ (see Section 3.1). The periodicity of the medium structure is represented by a Bravais lattice $\Lambda$ of rank $r$, with respect to which $\chi$ is $\Lambda$-periodic. In the following, we assume that the lattice $\Lambda$ is chosen maximally, i.e., we assume that there are no translations under which $\chi$ is invariant apart from those contained in $\mathrm{Trn}(\Lambda)$. Recall that $\mathrm{Trn}(\Lambda)$ denotes the group of translations, which are generated by $\Lambda$ (see Section 2.5). One can easily verify that $\mathrm{Trn}(\Lambda)$ is a normal subgroup of $\mathrm{Iso}(\mathbb{R}^r)_\chi$, the symmetry group of $\chi$. In contrast to $\mathrm{Trn}(\Lambda)$, which only represents the translational symmetry of the photonic crystal's medium structure, $\mathrm{Iso}(\mathbb{R}^r)_\chi$ represents its complete symmetry. Therefore, $\mathrm{Iso}(\mathbb{R}^r)_\chi$ can contain symmetry operations, which correspond to non-translational symmetries arising from the way the different materials of the photonic crystal are arranged within a primitive cell of $\Lambda$. Note, however, that $\mathrm{Iso}(\mathbb{R}^r)_\chi$ is always a subgroup of $\mathrm{Iso}(\mathbb{R}^r)_\Lambda$, the symmetry group of $\Lambda$, since every symmetry operation of $\chi$ in particular must take $\Lambda$ into itself. In summary, we therefore have

$$\mathrm{Trn}(\Lambda) \trianglelefteq \mathrm{Iso}(\mathbb{R}^r)_\chi \leq \mathrm{Iso}(\mathbb{R}^r)_\Lambda,$$

where $\trianglelefteq$ indicates the normal subgroup relation.

In crystallography, the complete symmetry of a periodic medium is characterized in terms of so-called *crystallographic space groups*. The complete symmetry of an $r$-dimensional photonic crystal's medium structure is said to be given by the space group $G_{\mathrm{space}}$, if there exists a crystal function $\chi$, such that its symmetry group $\mathrm{Iso}(\mathbb{R}^r)_\chi$ is isomorphic to $G_{\mathrm{space}}$. It turns out that there exist exactly 230 different crystallographic space groups in three space dimensions. This means that the complete symmetry of the medium structure of every three-dimensional photonic crystal is given by exactly one of these 230 groups. It is further known that in two space dimensions there exists a total of 17 different space groups, which are sometimes referred to as the 17 *wallpaper groups*.

As was stated before, space groups represent the complete symmetry of a periodic medium. The complete symmetry of such a medium usually consists of some sort of translational symmetry, which is represented by a Bravais lattice, as well as of other, non-translational symmetries. In crystallography, the non-translational symmetries of a periodic medium are characterized in terms of so-called *crystallographic point groups*. The non-translational symmetries of an $r$-dimensional photonic crystal's medium structure, whose complete symmetry is given by the space group $G_{\mathrm{space}}$ and whose translational symmetry is given by a Bravais lattice

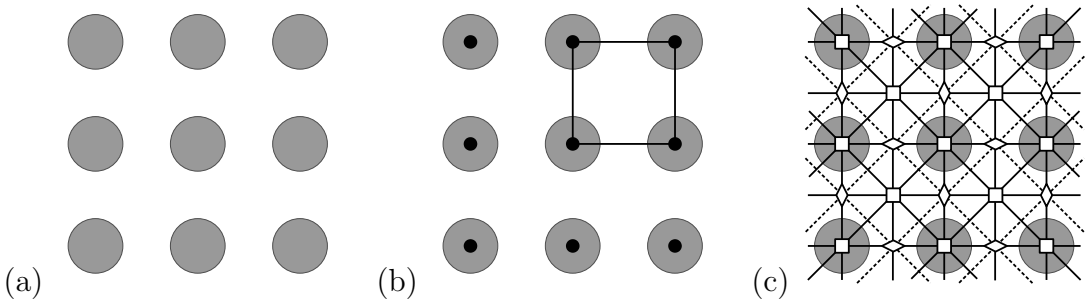



**Figure 3.4:** Example of a two-dimensional photonic crystal's medium structure (a) with its translational (b) and non-translational symmetries (c). The space group is isomorphic to $\mathbb{Z}^2 \rtimes \mathrm{O}_2(\mathbb{Z})$.

$\Lambda$, are said to be given by the point group $G_{\text{point}}$, if there exists a crystal function $\chi$, such that $\mathrm{Iso}(\mathbb{R}^r)_\chi$ is isomorphic to $G_{\text{space}}$ and such that $\mathrm{Iso}(\mathbb{R}^r)_\chi / \mathrm{Trn}(\Lambda)$ is isomorphic to $G_{\text{point}}$. It turns out that there exist exactly 32 different crystallographic point groups in three space dimensions. In two space dimensions there exist exactly 10 different point groups.

In the following, we present two examples, which illustrate the concept of point groups and space groups in two space dimensions. First, we consider a two-dimensional photonic crystal, whose medium structure is depicted in Figure 3.4(a). Without loss of generality, we can assume that the medium structure is represented by a crystal function $\chi$, which is periodic with respect to the standard square lattice $\mathbb{Z}^2$. Figure 3.4(b) depicts the structure's translational symmetry. Lattice points are indicated by black dots (•), the standard unit cell is indicated by a square. Apart from the translational symmetry the medium structure also exhibits a number of *rotational symmetries*, *reflection symmetries* and *glide-reflection symmetries*. In Figure 3.4(c) these symmetries are indicated by graphical symbols. Squares (□) indicate centers of rotations by $\pi/2$, $\pi$, and $3\pi/2$ which take the medium structure into itself. Such rotation centers are called *four-fold centers of rotation*. Centers of rotations by $\pi$ that take the medium structure into itself are called *two-fold centers of rotation* and indicated by lozenges (◊). Solid lines indicate *axes of reflections*, and dashed lines indicate *axis of glide-reflections* that take the medium structure into itself. A glide-reflection is defined as a reflection, which is followed by a translation along the axis of reflection. One finds that the symmetry group of $\chi$ is given by

$$\mathrm{Iso}(\mathbb{R}^2)_\chi = \left\{ \boldsymbol{\varphi} : \mathbb{R}^2 \to \mathbb{R}^2 \;\middle|\; \boldsymbol{\varphi}(\boldsymbol{x}) = \boldsymbol{\Theta}\boldsymbol{x} + \boldsymbol{a},\ \boldsymbol{\Theta} \in \mathrm{O}_2(\mathbb{Z}),\ \boldsymbol{a} \in \mathbb{Z}^2 \right\},$$

where $\mathrm{O}_2(\mathbb{Z})$ denotes the group of orthogonal $2 \times 2$-matrices with integer components. Furthermore, one can easily show that $\mathrm{Iso}(\mathbb{R}^2)_\chi$ is isomorphic to the

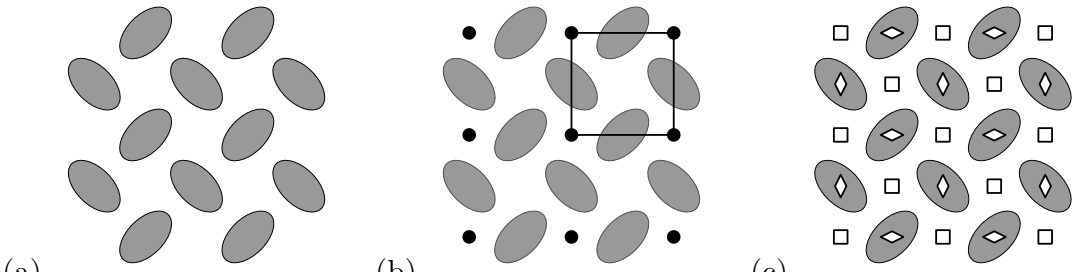



**Figure 3.5:** Example of a two-dimensional photonic crystal's medium structure (a) with its translational (b) and non-translational symmetries (c). The space group is isomorphic to $\mathbb{Z}^2 \rtimes \mathrm{SO}_2(\mathbb{Z})$.

standard semidirect product $\mathbb{Z}^2 \rtimes \mathrm{O}_2(\mathbb{Z})$. Recall that $\mathbb{Z}^2 \rtimes \mathrm{O}_2(\mathbb{Z})$ is defined as the Cartesian product of $\mathbb{Z}^2$ and $\mathrm{O}_2(\mathbb{Z})$ endowed with the group operation, which is given by

$$\big(\boldsymbol{a}^{(1)}, \boldsymbol{\Theta}^{(1)}\big)\big(\boldsymbol{a}^{(2)}, \boldsymbol{\Theta}^{(2)}\big) := \big(\boldsymbol{a}^{(1)} + \boldsymbol{\Theta}^{(1)}\boldsymbol{a}^{(2)}, \boldsymbol{\Theta}^{(1)}\boldsymbol{\Theta}^{(2)}\big)$$

for all $\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)} \in \mathbb{Z}^2$ and all $\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)} \in \mathrm{O}_2(\mathbb{Z})$. The space group, which corresponds to $\mathrm{Iso}(\mathbb{R}^2)_\chi$, is the wallpaper group p4mm (cf. Chapter 26 in [6]). The space group of the medium structure is isomorphic to $\mathrm{O}_2(\mathbb{Z})$. We remark that the three-dimensional analogue of the medium structure depicted in Figure 3.4(a) is a $\mathbb{Z}^3$-periodic arrangement of balls. The corresponding space group, which is denoted by $\mathrm{Pm}\overline{3}\mathrm{m}$ (cf. Chapter 10 in [15]), is isomorphic to $\mathbb{Z}^3 \rtimes \mathrm{O}_3(\mathbb{Z})$, the point group being isomorphic to $\mathrm{O}_3(\mathbb{Z})$.

As a second example, we briefly consider a two-dimensional photonic crystal, whose medium structure is depicted by Figure 3.5(a). As in the previous example, the translational symmetry is given by the standard square lattice $\mathbb{Z}^2$. The lattice and its standard unit cell are depicted in Figure 3.5(b). The medium structure exhibits the same rotational symmetries as the structure depicted in Figure 3.4(a). However, it lacks any reflection symmetries. Thus, we do not find any solid or dashed lines in Figure 3.5(c). The symmetry group of the crystal function $\chi$ is given by

$$\mathrm{Iso}(\mathbb{R}^2)_\chi = \big\{\boldsymbol{\varphi} : \mathbb{R}^2 \to \mathbb{R}^2 \,\big|\, \boldsymbol{\varphi}(\boldsymbol{x}) = \boldsymbol{\Theta}\boldsymbol{x} + \boldsymbol{a},\ \boldsymbol{\Theta} \in \mathrm{SO}_2(\mathbb{Z}),\ \boldsymbol{a} \in \mathbb{Z}^2\big\}.$$

Here as in the following, $\mathrm{SO}_2(\mathbb{Z})$ denotes the group of orthogonal $2 \times 2$-matrices, whose determinant equals 1. The symmetry group is isomorphic to $\mathbb{Z}^2 \rtimes \mathrm{SO}_2(\mathbb{Z})$, and the corresponding space group is denoted by p4 (cf. Chapter 26 in [6]). The point group of the medium structure is isomorphic to $\mathrm{SO}_2(\mathbb{Z})$.

We conclude with the remark that the representation of point groups by subgroups of $\mathrm{O}_r(\mathbb{R})$, $r$ being the dimension of the photonic crystal, is particularly

useful, when symmetrization operations have to be realized. Such operations will play an important role in the Chapters 7 and 8.

## 3.3 Wave Propagation in Linear Dielectrics

In this section we introduce a mathematical model for the propagation of light waves in certain dielectric media. This model will serve as a basis for more specific models, which are developed in Sections 3.4 and 3.5.

The standard theory, which describes the propagation of electromagnetic waves, such as light waves, is *classical electromagnetism*. Central to this theory are Maxwell's equations, a system consisting of two pairs of coupled partial differential equations, which capture fundamental properties of electromagnetic waves. The system of Maxwell's equations alone, however, does not determine the propagation of an electromagnetic wave completely. Further constitutive relations, which describe the wave's interaction with the ambient medium, have to be introduced in order to close the system. For each set of constitutive relations one obtains a different mathematical model for electromagnetic wave propagation. Here, we present the standard model for time-harmonic wave propagation in non-magnetic, non-dispersive, non-absorptive, linear, isotropic dielectrics, which is commonly used for photonic crystals (see e.g. [42], [43], [60], [66]). A detailed derivation of this model can be found in most standard textbooks on classical electromagnetism, such as [41], or in the original work of Maxwell [55]. Here, as in the remaining sections of this chapter, we focus on a formal derivation of the relevant formulas, leaving aside questions concerning the existence and uniqueness of solutions.

In classical electromagnetism electromagnetic waves are described by four vector fields $\boldsymbol{E}, \boldsymbol{H}, \boldsymbol{D}, \boldsymbol{B} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$, which are solutions of *Maxwell's equations*. Each vector field is a function in the variables $(t, \boldsymbol{x})$, where $t \in \mathbb{R}$ represents a point in time and $\boldsymbol{x} \in \mathbb{R}^3$ a position in space. The vector fields $\boldsymbol{E}$ and $\boldsymbol{H}$ represent the wave's *electric* and *magnetic field*, respectively. The vector field $\boldsymbol{D}$ represents the wave's *electric displacement field*, and the vector field $\boldsymbol{B}$ the wave's *magnetic flux density*. The fields $\boldsymbol{D}$ and $\boldsymbol{B}$ depend on the medium, in which the electromagnetic wave propagates, while the fields $\boldsymbol{E}$ and $\boldsymbol{H}$ describe the wave in free space. Under the assumption that all physical quantities are given in SI units (cf. [18]), the so-called macroscopic form of Maxwell's equations reads

$$\begin{cases} \nabla \times \boldsymbol{E} + \dot{\boldsymbol{B}} = \boldsymbol{0} & \text{in } \mathbb{R} \times \mathbb{R}^3 \quad \text{(Faraday's law)} \\ \nabla \times \boldsymbol{H} - \dot{\boldsymbol{D}} = \boldsymbol{J}_{\mathrm{f}} & \text{in } \mathbb{R} \times \mathbb{R}^3 \quad \text{(Ampère's law)} \\ \nabla \cdot \boldsymbol{D} = \rho_{\mathrm{f}} & \text{in } \mathbb{R} \times \mathbb{R}^3 \quad \text{(Gauss's law)} \\ \nabla \cdot \boldsymbol{B} = 0 & \text{in } \mathbb{R} \times \mathbb{R}^3 \quad \text{(Gauss's law for magnetism)} \end{cases} \tag{3.6}$$

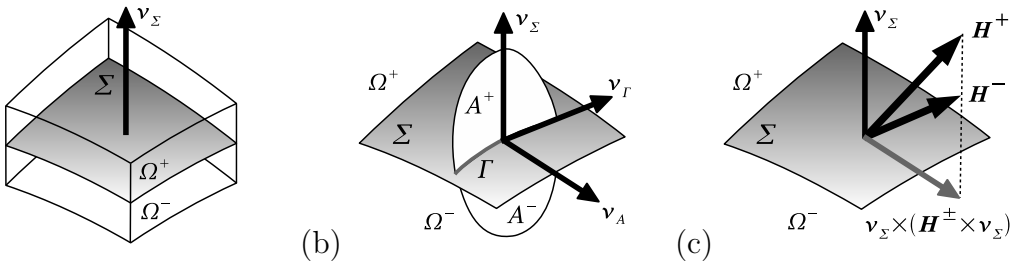



**Figure 3.6:** The layers $\Omega^+$ and $\Omega^-$ are separated by the surface $\Sigma$ (a). The surface $A = A^+ \cup A^-$ intersects $\Sigma$ perpendicularly in $\Gamma$ (b). In general, only the tangential component of the magnetic field $\boldsymbol{H}$ is continuous at $\Sigma$ (c).

Here, $\dot{\boldsymbol{D}}$ and $\dot{\boldsymbol{B}}$ denote the partial derivatives of the vector fields $\boldsymbol{D}$ and $\boldsymbol{B}$ with respect to the time variable $t$. The curl and divergence operators in (3.6) are understood to act on the vector fields with respect to the space variable $\boldsymbol{x}$. The vector field $\boldsymbol{J}_\mathrm{f} : \mathbb{R}\times\mathbb{R}^3 \to \mathbb{R}^3$, as well as the scalar field $\rho_\mathrm{f} : \mathbb{R}\times\mathbb{R}^3 \to \mathbb{R}$ are given functions, which model the *free current density* and the *free charge density* inside the medium. In this work we only consider media, in which neither free currents nor free charges exist. As a consequence, we have

$$\boldsymbol{J}_\mathrm{f} = \boldsymbol{0}, \tag{3.7}$$

$$\rho_\mathrm{f} = 0. \tag{3.8}$$

From top to bottom, the equations in (3.6) are known as *Faraday's law*, *Ampère's law*, *Gauss' law*, and *Gauss' law for magnetism*. The latter one, in particular, states that there are no magnetic monopoles.

Ampère's law implies an important property of the magnetic field $\boldsymbol{H}$ concerning its transition at an interface. To show this, let us assume that $\Sigma$ is smooth surface in $\mathbb{R}^3$, which is oriented by a *unit normal field* $\boldsymbol{\nu}_\Sigma : \Sigma \to \mathbb{S}^2$, where $\mathbb{S}^2$ denotes the three-dimensional unit sphere. Such a surface could be, for example, the interface between two different media. Given some positive number $\eta > 0$, we define the layers $\Omega^+, \Omega^- \subset \mathbb{R}^3$ by

$$\Omega^+ := \big\{\boldsymbol{x} + \delta\boldsymbol{\nu}_\Sigma(\boldsymbol{x}) \,\big|\, \boldsymbol{x}\in\Sigma,\ 0\le\delta\le\eta\big\},$$
$$\Omega^- := \big\{\boldsymbol{x} - \delta\boldsymbol{\nu}_\Sigma(\boldsymbol{x}) \,\big|\, \boldsymbol{x}\in\Sigma,\ 0\le\delta\le\eta\big\},$$

as well as the set $\Omega := \Omega^+\cup\Omega^-$. Note that the intersection of $\Omega^+$ and $\Omega^-$ coincides with $\Sigma$. For illustration, see Figure 3.6(a). Given a vector field $\boldsymbol{F} : \mathbb{R}\times\mathbb{R}^3 \to \mathbb{R}^3$, we define $\boldsymbol{F}^+ : \mathbb{R}\times\Omega^+ \to \mathbb{R}^3$ and $\boldsymbol{F}^- : \mathbb{R}\times\Omega^- \to \mathbb{R}^3$ by

$$\boldsymbol{F}^+(t,\boldsymbol{x}) := \begin{cases} \lim\limits_{\delta\to 0+} \boldsymbol{F}\big(t, \boldsymbol{x}+\delta\boldsymbol{\nu}_\Sigma(\boldsymbol{x})\big) & \text{if } \boldsymbol{x}\in\Sigma, \\ \boldsymbol{F}(t,\boldsymbol{x}) & \text{else} \end{cases} \qquad \text{for all } t\in\mathbb{R}, \boldsymbol{x}\in\Omega^+, \tag{3.9}$$

$$\boldsymbol{F}^-(t,\boldsymbol{x}) := \begin{cases} \lim\limits_{\delta\to 0+} \boldsymbol{F}\big(t, \boldsymbol{x} - \delta\boldsymbol{\nu}_\Sigma(\boldsymbol{x})\big) & \text{if } \boldsymbol{x}\in\Sigma, \\ \boldsymbol{F}(t,\boldsymbol{x}) & \text{else} \end{cases} \qquad \text{for all } t\in\mathbb{R}, \boldsymbol{x}\in\Omega^-. \tag{3.10}$$

Let us now assume that $\Gamma \subset \Sigma$ is a smooth curve, which is oriented by a *unit velocity field* $\boldsymbol{\nu}_\Gamma : \Gamma \to \mathbb{S}^2$. Assuming that $\Gamma$ is neither closed nor self-intersecting, one can find a simply connected, bounded, smooth surface $A \subset \mathbb{R}^3$, which is contained in the set $\Omega$, and which intersects the surface $\Sigma$ exactly in $\Gamma$. We further assume that the sets $A^+ := \Omega^+ \cap A$ and $A^- := \Omega^- \cap A$ are non-empty, and that the surface $A$ is oriented by a unit normal field $\boldsymbol{\nu}_A : A \to \mathbb{S}^2$ (see Figure 3.6(b)). Now, let $t\in\mathbb{R}$ be an arbitrary point in time. Then, we obtain by Stokes' theorem and Ampère's law that

$$\begin{aligned} &\int_\Gamma \boldsymbol{\nu}_\Gamma \cdot \big(\boldsymbol{H}^+(t,\,\cdot\,) - \boldsymbol{H}^-(t,\,\cdot\,)\big) \\ &\quad= \int_{\partial A^+} \boldsymbol{\nu}_{\partial A^+}\cdot \boldsymbol{H}^+(t,\,\cdot\,) + \int_{\partial A^-} \boldsymbol{\nu}_{\partial A^-}\cdot \boldsymbol{H}^-(t,\,\cdot\,) - \int_{\partial A} \boldsymbol{\nu}_{\partial A}\cdot \boldsymbol{H}(t,\,\cdot\,) \\ &\quad= \int_{A^+} \boldsymbol{\nu}_A \cdot \nabla\times \boldsymbol{H}^+(t,\,\cdot\,) + \int_{A^-} \boldsymbol{\nu}_A \cdot \nabla\times \boldsymbol{H}^-(t,\,\cdot\,) - \int_{A} \boldsymbol{\nu}_A \cdot \nabla\times \boldsymbol{H}(t,\,\cdot\,) \\ &\quad= \int_{A^+} \boldsymbol{\nu}_A \cdot \dot{\boldsymbol{D}}^+(t,\,\cdot\,) + \int_{A^-} \boldsymbol{\nu}_A \cdot \dot{\boldsymbol{D}}^-(t,\,\cdot\,) - \int_{A} \boldsymbol{\nu}_A \cdot \dot{\boldsymbol{D}}(t,\,\cdot\,) \\ &\quad= 0. \end{aligned} \tag{3.11}$$

Here $\boldsymbol{\nu}_{\partial A^+}$, $\boldsymbol{\nu}_{\partial A^-}$, and $\boldsymbol{\nu}_{\partial A}$ denote the unit velocity fields on the boundary curves $\partial A^+$, $\partial A^-$, and $\partial A$ of $A^+$, $A^-$, and $A$, respectively. Note that the identity (3.11) only holds under the assumption stated in (3.7), namely that there are no free currents, and under the assumption that the vector field $\dot{\boldsymbol{B}}$ is smooth up to the surface $\Sigma$. Since (3.11) can be established for every oriented, smooth curve $\Gamma$ on $\Sigma$, we conclude that

$$\boldsymbol{\tau}\cdot(\boldsymbol{H}^+ - \boldsymbol{H}^-)|_\Sigma = 0 \tag{3.12}$$

holds in the sense of integrals along smooth curves on $\Sigma$ for every unit vector field $\boldsymbol{\tau} : \Sigma \to \mathbb{S}^2$ that is *tangential* to $\Sigma$, i.e., that satisfies

$$\boldsymbol{\tau}\cdot\boldsymbol{\nu}_\Sigma = 0. \tag{3.13}$$

It follows from (3.12) that the vector field $(\boldsymbol{H}^+ - \boldsymbol{H}^-)|_\Sigma$ is *normal* on $\Sigma$, which means that

$$(\boldsymbol{H}^+ - \boldsymbol{H}^-)|_\Sigma = \big(\boldsymbol{\nu}_\Sigma\cdot(\boldsymbol{H}^+ - \boldsymbol{H}^-)|_\Sigma\big)\boldsymbol{\nu}_\Sigma.$$

Hence, one can easiyl verify that

$$\boldsymbol{\nu}_\Sigma \times (\boldsymbol{H}^+ - \boldsymbol{H}^-)|_\Sigma = \boldsymbol{0},$$

and that

$$\boldsymbol{\nu}_\Sigma \times \big((\boldsymbol{H}^+ - \boldsymbol{H}^-)|_\Sigma \times \boldsymbol{\nu}_\Sigma\big) = \mathbf{0}.$$

The latter identity implies, that the transition of the so-called *tangential component* $\boldsymbol{\nu} \times (\boldsymbol{H} \times \boldsymbol{\nu})$ of the magnetic field $\boldsymbol{H}$ is continuous at the interface $\Sigma$. Here, $\boldsymbol{\nu}$ denotes a continuous extension of the unit normal field $\boldsymbol{\nu}_\Sigma$ to $\Omega$. As we will see below, the so-called *normal component* $\boldsymbol{\nu} \cdot \boldsymbol{H}$ of $\boldsymbol{H}$ can be discontinuous at $\Sigma$. Figure 3.6(c) illustrates the generic transitional behavior of the magnetic field at an interface.

We now return to the system of Maxwell's equations (3.6). As one can see, this system is not closed. In order to close the system, one needs to introduce constitutive relations, which relate the fields $\boldsymbol{D}$ and $\boldsymbol{B}$ to the fields $\boldsymbol{E}$ and $\boldsymbol{H}$. Such constitutive relations model the response of the ambient medium to the electromagnetic wave. In this work, we only consider non-dispersive, non-absorptive, linear, and isotropic dielectrics. For these media the constitutive relations are of the form

$$\boldsymbol{D} = \varepsilon_0 \varepsilon_r \boldsymbol{E}, \tag{3.14}$$

$$\boldsymbol{B} = \mu_0 \mu_r \boldsymbol{H}, \tag{3.15}$$

where $\varepsilon_r, \mu_r : \mathbb{R}^3 \to \mathbb{R}_{>0}$ are scalar fields that only depend on the space variable $\boldsymbol{x}$. Typically, $\varepsilon_r$ and $\mu_r$ are positive, bounded, and uniformly bounded away from zero. The real numbers $\varepsilon_0, \mu_0 \in \mathbb{R}_{>0}$ are physical constants, which are called the *vacuum electric permittivity* and the *vacuum magnetic permeability*. For these constants the relation

$$\varepsilon_0 \mu_0 = \frac{1}{c_0^2} \tag{3.16}$$

holds, where $c_0$ denotes the *vacuum speed of light*. The scalar fields $\varepsilon_r$ and $\mu_r$ represent the medium's *relative electric permittivity* and *relative magnetic permeability*. Both fields $\varepsilon_r$ and $\mu_r$ determine the medium's optical density, which is measured by the space-dependent *refractive index* $n : \mathbb{R}^3 \to \mathbb{R}$. The refractive index is given by

$$n := \sqrt{\varepsilon_r \mu_r}. \tag{3.17}$$

We remark that for *non-isotropic* media the relative electric permittivity, as well as the magnetic permeability are represented by matrix-valued functions. For *dispersive*, linear media the constitutive relations between $\boldsymbol{D}$ and $\boldsymbol{E}$, as well as between $\boldsymbol{B}$ and $\boldsymbol{H}$ are given in terms of convolutions (see e.g. Section 7.10 in [41]). Here, "dispersive" means that the properties of the medium depend on the phase frequency of the wave.

Let us, once again, consider the transitional behaviour of the magnetic field $\boldsymbol{H}$ at an interface $\Sigma$. Under the above assumptions, let $\Delta \subset \Sigma$ be an oriented surface

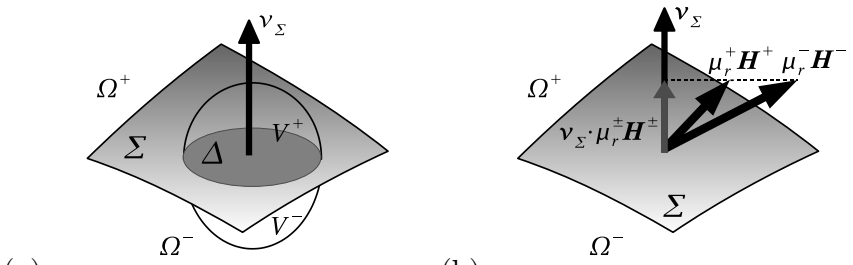



**Figure 3.7:** The volume $V = V^+ \cup V^-$ intersects the surface $\Sigma$ in $\Delta$ (a). In general, only the normal component of $\mu_r \boldsymbol{H}$ is continuous at $\Sigma$ (b).

on $\Sigma$, whose unit normal field is identical to that of $\Sigma$. Furthermore, let $V \subset \mathbb{R}^3$ be a smoothly bounded volume, which is contained in $\Omega$ and which intersects $\Sigma$ exactly in $\Delta$, such that $V^+ := V \cap \Omega^+$ and $V^- := V \cap \Omega^-$ are nonempty sets. For illustration, see Figure 3.7. Letting $\mu_r^+ : \Omega^+ \to \mathbb{R}$ and $\mu_r^- : \Omega^- \to \mathbb{R}$ be defined in analogy to (3.9) and (3.10), the constitutive relation (3.15), Gauss' theorem, and Gauss' law for magnetism imply that

$$\begin{aligned}
&\int_\Delta \boldsymbol{\nu}_\Sigma \cdot \big(\mu_r^+ \boldsymbol{H}^+(t,\,\cdot\,) - \mu_r^- \boldsymbol{H}^-(t,\,\cdot\,)\big) \\
&\quad = \int_{\partial V^+} \boldsymbol{\nu}_{\partial V^+} \cdot \boldsymbol{B}^+(t,\,\cdot\,) + \int_{\partial V^-} \boldsymbol{\nu}_{\partial V^-} \cdot \boldsymbol{B}^-(t,\,\cdot\,) - \int_{\partial V} \boldsymbol{\nu}_{\partial V} \cdot \mu_r \boldsymbol{H}(t,\,\cdot\,) \\
&\quad = \int_{V^+} \nabla \cdot \big[\boldsymbol{B}^+(t,\cdot)\big] + \int_{V^-} \nabla \cdot \big[\boldsymbol{B}^-(t,\cdot)\big] - \int_V \nabla \cdot \big[\boldsymbol{B}(t,\cdot)\big] \\
&\quad = 0.
\end{aligned} \tag{3.18}$$

Here, $\boldsymbol{\nu}_{\partial V^+}$, $\boldsymbol{\nu}_{\partial V^-}$, and $\boldsymbol{\nu}_{\partial V}$ denote the outer unit normal field on the boundaries $\partial V^+$, $\partial V^-$, and $\partial V$ of the volumes $V^+$, $V^-$, and $V$, respectively. Since (3.18) can be established for every surface $\Delta$, we conclude that

$$\boldsymbol{\nu}_\Sigma \cdot \big(\mu_r^+ \boldsymbol{H}^+ - \mu_r^- \boldsymbol{H}^-\big)|_\Sigma = 0 \tag{3.19}$$

in the sense of integrals over surfaces on $\Sigma$. It hence follows that the so-called *normal component* $\boldsymbol{\nu} \cdot \mu_r \boldsymbol{H}$ of the vector field $\mu_r \boldsymbol{H}$ is continuous across interfaces. As before, we denote by $\boldsymbol{\nu}$ a continuous extension of the unit normal field $\boldsymbol{\nu}_\Sigma$ to $\Omega$. This implies, of course, that the normal component $\boldsymbol{\nu} \cdot \boldsymbol{H}$ of the magnetic field $\boldsymbol{H}$ must be discontinuous at interfaces, where the relative magnetic permeability $\mu_r$ changes discontinuously. Figure 3.7(b) depicts the generic behaviour of the vector field $\mu_r \boldsymbol{H}$.

Under the assumptions (3.7), (3.8), (3.14), and (3.15), Maxwell's equations become a fully coupled system of first-order partial differential equations, which reads

$$\begin{cases} \nabla \times \boldsymbol{E} + \mu_0 \mu_r \dot{\boldsymbol{H}} = 0 & \text{in } \mathbb{R} \times \mathbb{R}^3, \\ \nabla \times \boldsymbol{H} - \varepsilon_0 \varepsilon_r \dot{\boldsymbol{E}} = 0 & \text{in } \mathbb{R} \times \mathbb{R}^3, \\ \nabla \cdot \big[\varepsilon_r \boldsymbol{E}\big] = 0 & \text{in } \mathbb{R} \times \mathbb{R}^3, \\ \nabla \cdot \big[\mu_r \boldsymbol{H}\big] = 0 & \text{in } \mathbb{R} \times \mathbb{R}^3. \end{cases} \tag{3.20}$$

One can extract a second-order system for the magnetic field $\boldsymbol{H}$ from (3.20). This can be done by differentiating Faraday's law with respect to the time variable $t$, and by then substituting the identity

$$\dot{\boldsymbol{E}} = \frac{1}{\varepsilon_0 \varepsilon_r} \nabla \times \boldsymbol{H}, \tag{3.21}$$

which follows from Ampère's law, into the resulting equation. Together with (3.16) and Gauss' law for magnetism this yields

$$\begin{cases} \nabla \times \left[\frac{1}{\varepsilon_r} \nabla \times \boldsymbol{H}\right] + \frac{1}{c_0^2} \mu_r \ddot{\boldsymbol{H}} = 0 & \text{in } \mathbb{R} \times \mathbb{R}^3, \\ \nabla \cdot \big[\mu_r \boldsymbol{H}\big] = 0 & \text{in } \mathbb{R} \times \mathbb{R}^3. \end{cases} \tag{3.22}$$

Intent on studying the propagation of electromagnetic waves with a single *phase frequency* $\omega \in \mathbb{R}$, we make a so-called *time-harmonic ansatz* for the magnetic field $\boldsymbol{H}$. This time-harmonic ansatz is given by

$$\boldsymbol{H}(t, \boldsymbol{x}) = \operatorname{Re}\big[\boldsymbol{h}_\omega(\boldsymbol{x}) \mathrm{e}^{\mathrm{i}\omega t}\big] \qquad \text{for all } (t, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^3, \tag{3.23}$$

where $\boldsymbol{h}_\omega : \mathbb{R}^3 \to \mathbb{C}^3$ is a complex vector field. Plugging the time-harmonic ansatz into (3.22), one obtains

$$\begin{cases} \nabla \times \left[\frac{1}{\varepsilon_r} \nabla \times \boldsymbol{h}_\omega\right] = \frac{\omega^2}{c_0^2} \mu_r \boldsymbol{h}_\omega & \text{in } \mathbb{R}^3, \\ \nabla \cdot \big[\mu_r \boldsymbol{h}_\omega\big] = 0 & \text{in } \mathbb{R}^3. \end{cases} \tag{3.24}$$

In contrast to the system of equations (3.22), the time-harmonic system of equations (3.24) is posed in $\mathbb{R}^3$ rather than in $\mathbb{R} \times \mathbb{R}^3$. Notice that (3.24) can be viewed as a constrained eigenvalue problem for the eigenvalue $\omega^2/c_0^2$ and corresponding eigenfunctions $\boldsymbol{h}_\omega$. The divergence constraint in the second line of (3.24) is natural for the problem in the following sense: Suppose that $\boldsymbol{h}_\omega$ is a complex vector field that solves the eigenvalue equation in the first line of (3.24) for some non-vanishing eigenvalue $\omega^2/c_0^2$. Then, $\boldsymbol{h}_\omega$ is a curl field, which is why it also satisfies the divergence constrained. For this reason the divergence equation is often omitted in the

literature. We shall keep this equation, nevertheless, as it will play an important role in the analysis and the numerical discretization of the problem.

The constrained eigenvalue problem (3.24) can be further simplified by assuming that the medium, in which the wave propagates, is *non-magnetic*. For such media the relative magnetic permeability is given by

$$\mu_r = 1. \tag{3.25}$$

The system (3.24) hence becomes

$$\begin{cases} \nabla \times \left[ \dfrac{1}{\varepsilon_r} \nabla \times \boldsymbol{h}_\omega \right] = \dfrac{\omega^2}{c_0^2} \boldsymbol{h}_\omega & \text{in } \mathbb{R}^3, \\ \nabla \cdot \boldsymbol{h}_\omega = 0 & \text{in } \mathbb{R}^3. \end{cases} \tag{3.26}$$

So far, we only considered the magnetic field of an electromagnetic wave propagating in a non-magnetic, non-dispersive, non-absorptive, linear, isotropic dielectric. For completeness, we also introduce the mathematical model for the associated electric field in the following.

When investigating the transition properties of the electric field $\boldsymbol{E}$ at a smooth surface $\Sigma$, which is oriented by the unit normal field $\boldsymbol{\nu}_\Sigma : \Sigma \to \mathbb{S}^2$, one finds that

$$\boldsymbol{\tau} \cdot \big( \boldsymbol{E}^+ - \boldsymbol{E}^- \big)|_\Sigma = 0 \tag{3.27}$$

holds in the sense of integrals along smooth curves on $\Sigma$ for every unit vector field $\boldsymbol{\tau} : \Sigma \to \mathbb{S}^2$, which is *tangential* to $\Sigma$. Recall that the latter notion is made precise by (3.13). As is the case of the magnetic field, we derive from (3.27) that

$$\boldsymbol{\nu}_\Sigma \times (\boldsymbol{E}^+ - \boldsymbol{E}^-)|_\Sigma = \boldsymbol{0},$$

and that

$$\boldsymbol{\nu}_\Sigma \times \big( (\boldsymbol{E}^+ - \boldsymbol{E}^-)|_\Sigma \times \boldsymbol{\nu}_\Sigma \big) = \boldsymbol{0}.$$

One can also show that the electric field satisfies

$$\boldsymbol{\nu}_\Sigma \cdot \big( \varepsilon_r^+ \boldsymbol{E}^+ - \varepsilon_r^- \boldsymbol{E}^- \big)|_\Sigma = 0 \tag{3.28}$$

in the sense of integrals over surfaces on $\Sigma$. Clearly, (3.27) and (3.28) are analogues of (3.12) and (3.19), which state that only the tangential component of the electric field is continuous at medium interfaces. The normal component, however, is discontinuous at interfaces, where the relative electric permittivity $\varepsilon_r$ changes discontinuously.

One can also derive a second-order, time-harmonic system for the electric field. This is easily done by differentiating Ampère's law in (3.20) with respect to the time variable $t$, and by then plugging in the identity

$$\dot{\boldsymbol{H}} = -\frac{1}{\mu_0} \nabla \times \boldsymbol{E}, \tag{3.29}$$

which follows from Faraday's law. The resulting system reads

$$\begin{cases} \nabla \times \nabla \times \boldsymbol{E} + \dfrac{1}{c_0^2}\, \varepsilon_r \ddot{\boldsymbol{E}} = 0 & \text{in } \mathbb{R} \times \mathbb{R}^3, \\ \nabla \cdot \big[\varepsilon_r \boldsymbol{E}\big] = 0 & \text{in } \mathbb{R} \times \mathbb{R}^3. \end{cases}, \tag{3.30}$$

With the time-harmonic ansatz

$$\boldsymbol{E}(t, \boldsymbol{x}) = \mathrm{Re}\big[\boldsymbol{e}_\omega(\boldsymbol{x}) \mathrm{e}^{\mathrm{i}\omega t}\big] \qquad \text{for all } (t, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^3, \tag{3.31}$$

where $\boldsymbol{e}_\omega : \mathbb{R}^3 \to \mathbb{C}^3$ is a complex vector field, one obtains the system of equations

$$\begin{cases} \nabla \times \nabla \times \boldsymbol{e}_\omega = \dfrac{\omega^2}{c_0^2}\, \varepsilon_r \boldsymbol{e}_\omega & \text{in } \mathbb{R}^3, \\ \nabla \cdot \big[\varepsilon_r \boldsymbol{e}_\omega\big] = 0 & \text{in } \mathbb{R}^3. \end{cases} \tag{3.32}$$

Clearly, the system of equations (3.32) poses a constrained eigenvalue problem for the eigenvalue $\omega^2/c_0^2$ and corresponding eigenfunctions $\boldsymbol{e}_\omega$, which is similar to that posed by the system of equations for the magnetic field (3.24).

Finally, we remark that the theory of classical electromagnetism can also be expressed in terms of differential forms on smooth manifolds. In this framework the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{H}$ are described by 1-forms which, in essence, represent curve integrals. This alludes to the fact that in practice electric and magnetic fields are always measured with respect to oriented curves in physical space. Recall also that the identities (3.12) and (3.27) should be understood in the sense of integrals along smooth curves at an interface. The electric displacement field $\boldsymbol{D}$ and the magnetic flux density $\boldsymbol{B}$ are described by 2-forms, which represent surface integrals. This embodies the fact that the corresponding physical quantities are measured with respect to oriented surfaces in physical space. In view of this, it is not surprising that the identities (3.19) and (3.28) should be understood in the sense of surface integrals, since according to (3.14) and (3.15) the vector fields $\varepsilon_r \boldsymbol{E}$ and $\mu_r \boldsymbol{H}$ are proportional to $\boldsymbol{D}$ and $\boldsymbol{B}$, respectively. Expressing electromagnetic waves in terms of differential forms can also be useful, when studying finite element discretizations of Maxwell's equations. By viewing the electric field and the magnetic field as 1-forms, for example, one is able to identify the so-called *Nédélec elements* (cf. [57]) as the class of finite elements, which is best suited for discretizing these fields. It is beyond the scope of this work to make this notion precise. For details, we refer to the works of Arnold [7] and Hiptmair [39].

## 3.4 Bloch Modes in Periodic Media

The mathematical model which was introduced in the previous section, is valid for time-harmonic wave propagation in any non-magnetic, non-dispersive, non-

absorptive, linear, isotropic, dielectric medium. In this section we consider the special case of wave propagation inside a photonic crystal, whose medium structure, in addition, exhibits some spatial periodicity.

In the following, we assume that the photonic crystal's spatial periodicity is represented by a Bravais lattice $\Lambda \in \mathbb{R}^3$ of rank 3. In Section 3.1 we discussed that the crystal's medium structure can be represented by a $\Lambda$-periodic function $\chi : \mathbb{R}^3 \to \mathbb{N}$, such that every function value $\chi(\boldsymbol{x})$ represents the material, which is present at the position $\boldsymbol{x} \in \mathbb{R}$. Since each material has its own specific electric permittivity, it follows that the relative electric permittivity field $\varepsilon_r : \mathbb{R}^3 \to \mathbb{R}$ is also $\Lambda$-periodic, i.e.,

$$\varepsilon_r \circ \boldsymbol{\tau}_{\boldsymbol{a}} = \varepsilon_r \qquad \text{for all } \boldsymbol{a} \in \Lambda.$$

Consequently, (3.26) and (3.32) are systems of partial differential equations with $\Lambda$-periodic coefficients for photonic crystals. Such partial differential equations are the subject of the *Floquet–Bloch theory*. This theory is a generalization of the *Floquet theory*. One of the main results of the Floquet theory is Floquet's theorem, which characterizes the fundamental matrices of ordinary, linear differential equations as products of periodic functions and matrix exponentials (cf. [37]). In his work on the motion of electrons in a periodic potential, Bloch showed that the eigenfunctions of Schrödinger-type operators with periodic coefficients are products of periodic functions and plane waves (see [13]). Such functions are now called *Bloch functions*. The findings of Bloch are, in essence, an analog of Floquet's theorem for Schrödinger-type operators. It could be shown that variants of Floquet's theorem apply to a large class of linear partial differential operators, and especially to elliptic ones (see e.g. [49]).

Reconsidering the time-harmonic system of equations for the magnetic field (3.26), the Floquet–Bloch theory suggests that the vector field $\boldsymbol{h}_\omega$ is the product of a $\Lambda$-periodic vector field and a plane wave function. More precisely, we make the ansatz

$$\boldsymbol{h}_\omega = \boldsymbol{h}_{\omega,\boldsymbol{k}}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{k}, \cdot \rangle}, \tag{3.33}$$

where $\boldsymbol{h}_{\omega,\boldsymbol{k}} : \mathbb{R}^3 \to \mathbb{C}^3$ is a $\Lambda$-periodic vector field, and where $\boldsymbol{k}$ is a vector belonging to the so-called *first Brillouin zone* $\mathbb{B}$ of $\Lambda$. As was discussed in Section 2.5, there exists a so-called reciprocal latttice $\widehat{\Lambda} \subset \mathbb{R}^3$ of $\Lambda$. The reciprocal lattice $\widehat{\Lambda}$ is a lattice of rank 3, such that

$$\boldsymbol{a} \cdot \boldsymbol{b} \in 2\pi\mathbb{Z} \qquad \text{for all } \boldsymbol{a} \in \Lambda,\ \boldsymbol{b} \in \widehat{\Lambda}.$$

The first Brillouin zone $\mathbb{B}$ of $\Lambda$ is defined as the Wigner–Seitz cell of $\widehat{\Lambda}$, i.e.,

$$\mathbb{B} := W_{\widehat{\Lambda}}. \tag{3.34}$$

Since the rank of the reciprocal lattice $\widehat{\Lambda}$ is maximal, the first Brillouin zone is a closed, bounded, convex polyhedron, which is point-symmetric with respect to the origin. The ansatz (3.33) is called a *Bloch ansatz*. The vector $\boldsymbol{k}$ is called the *quasimomentum vector* of the Bloch function $\boldsymbol{h}_\omega$. One easily verifies that the Bloch function $\boldsymbol{h}_\omega$ solves the time-harmonic system (3.26), if and only if the function $\boldsymbol{h}_{\omega,\boldsymbol{k}}$ satisfies

$$\begin{cases} (\nabla + \mathrm{i}\boldsymbol{k}) \times \left[ \dfrac{1}{\varepsilon_r} (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{h}_{\omega,\boldsymbol{k}} \right] = \dfrac{\omega^2}{c_0^2} \boldsymbol{h}_{\omega,\boldsymbol{k}} & \text{in } \mathbb{R}^3, \\ (\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{h}_{\omega,\boldsymbol{k}} = 0 & \text{in } \mathbb{R}^3. \end{cases} \tag{3.35}$$

The operator $(\nabla + \mathrm{i}\boldsymbol{k})\times$ is referred to as a *modified curl operator*. The operator $(\nabla + \mathrm{i}\boldsymbol{k})\cdot$ is called a *modified divergence operator*. The system of equations (3.35) constitutes a family of constrained eigenvalue problems, which is parametrized by the quasimomentum vector $\boldsymbol{k}$. Since the vector field $\boldsymbol{h}_{\omega,\boldsymbol{k}}$ and the relative electric permittivity function $\varepsilon_r$ in (3.35) are $\Lambda$-periodic by assumption, it suffices to solve this family of eigenvalue problems in a primitive cell of $\Lambda$ only. Given such a primitive cell $\Omega \subset \mathbb{R}^3$, we hence consider the family of constrained eigenvalue problems

$$\begin{cases} (\nabla + \mathrm{i}\boldsymbol{k}) \times \left[ \dfrac{1}{\varepsilon_r} (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{h}_{\omega,\boldsymbol{k}} \right] = \dfrac{\omega^2}{c_0^2} \boldsymbol{h}_{\omega,\boldsymbol{k}} & \text{in } \Omega, \\ (\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{h}_{\omega,\boldsymbol{k}} = 0 & \text{in } \Omega, \end{cases} \tag{3.36}$$

for the eigenvalue $\omega^2/c_0^2$ and corresponding eigenfunctions $\boldsymbol{h}_{\omega,\boldsymbol{k}}$. This family of constrained eigenvalue problems is indexed by the quasimomentum vector $\boldsymbol{k}$, which varies over the first Brillouin zone $\mathbb{B}$ of $\Lambda$.

Every eigensolution $(\omega^2/c_0^2, \boldsymbol{h}_{\omega,\boldsymbol{k}})$ of (3.36) determines via (3.33) a so-called *Bloch mode* of the photonic crystal with frequency $\omega$ and quasimomentum vector $\boldsymbol{k}$. In Chapter 4 we will discuss the existence of such Bloch modes. Bloch modes play an essential role in the investigation of the optical properties of photonic crystals. Roughly speaking, an electromagnetic wave with a given frequency $\omega$ is able to propagate inside a photonic crystal, if there exists a Bloch mode with the same frequency. In Chapter 4 we will show that for a given quasimomentum vector $\boldsymbol{k} \in \mathbb{B}$ there exists a countable set of non-negative, real eigenvalues of (3.36) and hence also a countable set of Bloch mode frequencies. The variation of these frequencies with respect to the quasimomentum vector $\boldsymbol{k}$ is continuous. Hence, the Bloch mode frequencies can be interpreted as continuous functions of the quasimomentum vector. The graphs of these functions are referred to as *photonic bands*. In Chapter 5 we shall study optimization problems related to these photonic bands.

So far, we only considered the magnetic field of an electromagnetic wave propagating in a periodic medium. In the following, we briefly study the associated

electric field. By virtue of the Bloch ansatz

$$\boldsymbol{e}_{\omega} = \boldsymbol{e}_{\omega,\boldsymbol{k}}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{k}, \cdot \rangle}, \tag{3.37}$$

where $\boldsymbol{e}_{\omega,\boldsymbol{k}} : \mathbb{R}^3 \to \mathbb{C}^3$ is a $\Lambda$-periodic vector field, and where $\boldsymbol{k} \in \mathbb{B}$ is a quasimomentum vector, the time-harmonic system of equations for the electric field (3.32) can be rewritten as

$$\begin{cases} (\nabla + \mathrm{i}\boldsymbol{k}) \times (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{e}_{\omega,\boldsymbol{k}} = \dfrac{\omega^2}{c_0^2}\, \varepsilon_r \boldsymbol{e}_{\omega,\boldsymbol{k}} & \text{in } \Omega, \\ (\nabla + \mathrm{i}\boldsymbol{k}) \cdot \left[\varepsilon_r \boldsymbol{e}_{\omega,\boldsymbol{k}}\right] = 0 & \text{in } \Omega. \end{cases} \tag{3.38}$$

Clearly, the system of equations (3.38) constitutes a family of constrained eigenvalue problems indexed by the quasimomentum vector $\boldsymbol{k}$ for the eigenvalue $\omega^2/c_0^2$ and corresponding eigenfunctions $\boldsymbol{e}_{\omega,\boldsymbol{k}}$. In this work we shall not consider this family of constrained eigenvalue problems for the electric field any further, since for the eigensolutions of (3.38) are determined by those of (3.36), and vice versa. More precisely, one can show that if $(\omega^2/c_0^2, \boldsymbol{h}_{\omega,\boldsymbol{k}})$ is an eigensolution of (3.36) with non-vanishing frequency $\omega$ for a given quasimomentum vector $\boldsymbol{k} \in \mathbb{B}$, then $(\omega^2/c_0^2, \boldsymbol{e}_{\omega,\boldsymbol{k}})$, where the eigenfunction $\boldsymbol{e}_{\omega,\boldsymbol{k}}$ is given by

$$\boldsymbol{e}_{\omega,\boldsymbol{k}} = \frac{1}{\mathrm{i}\omega\varepsilon_0\varepsilon_r}(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{h}_{\omega,\boldsymbol{k}},$$

is an eigensolution of (3.38). Conversely, if $(\omega^2/c_0^2, \boldsymbol{e}_{\omega,\boldsymbol{k}})$ is an eigensolution of (3.38) with non-vanishing frequency $\omega$ for a given quasimomentum vector $\boldsymbol{k} \in \mathbb{B}$, then $(\omega^2/c_0^2, \boldsymbol{h}_{\omega,\boldsymbol{k}})$ with

$$\boldsymbol{h}_{\omega,\boldsymbol{k}} = \frac{1}{\mathrm{i}\omega\mu_0}(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{e}_{\omega,\boldsymbol{k}}$$

is an eigensolution of (3.36). It hence suffices to only consider the family of constrained eigenvalue problems for the magnetic field (3.36) in order to determine Bloch mode frequencies for three-dimensional photonic crystals.

## 3.5 The Two-Dimensional Case

In this section we consider time-harmonic wave propagation in two-dimensional photonic crystals and deduce simplified models for waves with specific polarizations.

In Section 3.1, two-dimensional photonic crystals were characterized as media, which are periodic in two spatial dimensions while being homogeneous in the third

one. By choosing a suitable coordinate system, the crystal function of a two-dimensional photonic crystal can be defined as a function from $\mathbb{R}^2$ into $\mathbb{N}$, which is periodic with respect to a Bravais lattice $\Lambda \subset \mathbb{R}^2$ of rank 2. Consequently, the relative electric permittivity $\varepsilon_r$ can be identified with a $\Lambda$-periodic function mapping from $\mathbb{R}^2$ into $\mathbb{R}$. Furthermore, since the electric permittivity only depends on two spatial coordinates, it is reasonable to assume the same for the electric and magnetic field of a wave, which propagates inside a two-dimensional photonic crystal. In this section, we hence assume that the electric and magnetic field are represented by vector fields $\boldsymbol{E}, \boldsymbol{H} : \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}^3$, which depend on a time variable $t \in \mathbb{R}$ and a spatial variable $\boldsymbol{x} \in \mathbb{R}^2$. A time-harmonic ansatz for the two vector fields is given by

$$\boldsymbol{E}(t, \boldsymbol{x}) = \mathrm{Re}\big[\boldsymbol{e}_\omega(\boldsymbol{x})\mathrm{e}^{\mathrm{i}\omega t}\big] \qquad \text{for all } (t, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^2, \tag{3.39}$$

$$\boldsymbol{H}(t, \boldsymbol{x}) = \mathrm{Re}\big[\boldsymbol{h}_\omega(\boldsymbol{x})\mathrm{e}^{\mathrm{i}\omega t}\big] \qquad \text{for all } (t, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^2 \tag{3.40}$$

where $\boldsymbol{e}_\omega, \boldsymbol{h}_\omega : \mathbb{R}^2 \to \mathbb{C}^3$ are complex vector field, and where $\omega \in \mathbb{R}$ is a phase frequency. With this ansatz, two constrained eigenvalue problems in $\mathbb{R}^2$ can be established for the vector fields $\boldsymbol{e}_\omega$ and $\boldsymbol{h}_\omega$ in analogy to the time-harmonic systems (3.32) and (3.26).

It turns out the the constrained eigenvalue problems can be simplified significantly, when time-harmonic waves with specific *polarizations* are considered. Here, the term polarization refers to a property of electromagnetic waves, which characterizes the directions of oscillation of the wave's electric and magnetic field. For instance, an electromagnetic wave is called *TM-polarized* or simply a *TM-wave*, if the first two components of its electric field vanish identically. With regard to the time-harmonic ansatz (3.39)–(3.40), an electromagnetic wave is TM-polarized, if and only if the complex vector field $\boldsymbol{e}_\omega$ is of the form

$$\boldsymbol{e}_\omega = \begin{pmatrix} 0 \\ 0 \\ e_\omega^{\mathrm{TM}} \end{pmatrix}, \tag{3.41}$$

where $e_\omega^{\mathrm{TM}} : \mathbb{R}^2 \to \mathbb{C}$ is a complex scalar field. According to (3.21) the magnetic field of a time-harmonic TM-wave is determined by the complex vector field

$$\boldsymbol{h}_\omega^{\mathrm{TM}} := -\frac{1}{\mu_0 \omega} \begin{pmatrix} \partial_2 e_\omega^{\mathrm{TM}} \\ -\partial_1 e_\omega^{\mathrm{TM}} \\ 0 \end{pmatrix}, \tag{3.42}$$

where $\partial_i e_\omega^{\mathrm{TM}}$ denotes the partial derivative of the function $e_\omega^{\mathrm{TM}}$ with respect to the $i$-th spatial coordinate for $i = 1, 2$. One easily verifies that the time-harmonic

system (3.32) posed in $\mathbb{R}^2$ reduces for TM-polarized waves to a scalar Laplace-type eigenvalue problem, which reads

$$-\Delta e_\omega^{\mathrm{TM}} = \frac{\omega^2}{c_0^2}\,\varepsilon_r e_\omega^{\mathrm{TM}} \quad \text{in } \mathbb{R}^2. \tag{3.43}$$

Notice, in particular, that the divergence constraint is automatically fulfilled, since neither $\varepsilon_r$ nor $e_\omega^{\mathrm{TM}}$ depend on the third spatial coordinate.

Another polarization, which affords a simplification of the time-harmonic system (3.26), is the so-called TE-polarization. An electromagnetic wave is called *TE-polarized* or simply a *TE-wave*, if the first two components of its magnetic field vanish identically. Hence, time-harmonic, TE-polarized waves are given by complex vector fields $\boldsymbol{h}_\omega$, which are of the form

$$\boldsymbol{h}_\omega = \begin{pmatrix} 0 \\ 0 \\ h_\omega^{\mathrm{TE}} \end{pmatrix}, \tag{3.44}$$

where $h_\omega^{\mathrm{TE}} : \mathbb{R}^2 \to \mathbb{C}$ is a complex scalar field. According to (3.29), the electric field of a time-harmonic TE-wave is determined by the complex vector field

$$\boldsymbol{e}_\omega^{\mathrm{TE}} := \frac{1}{\varepsilon_0 \varepsilon_r \omega} \begin{pmatrix} \partial_2 h_\omega^{\mathrm{TE}} \\ -\partial_1 h_\omega^{\mathrm{TE}} \\ 0 \end{pmatrix}. \tag{3.45}$$

For TE-polarized waves the constrained eigenvalue problem (3.26) can be simplified to a second-order, divergence-type eigenvalue problem, which is given by

$$-\nabla \cdot \left[\frac{1}{\varepsilon_r} \nabla h_\omega^{\mathrm{TE}}\right] = \frac{\omega^2}{c_0^2}\, h_\omega^{\mathrm{TE}} \quad \text{in } \mathbb{R}^2. \tag{3.46}$$

Again, the divergence constraint is fulfilled due to the fact that the scalar field $h_\omega^{\mathrm{TE}}$ does not depend on the third spatial coordinate. Looking at (3.41), (3.42), (3.44), and (3.45), one realizes that every electromagnetic wave, which propagates inside a two-dimensional photonic crystal, can be represented as the sum of a TM-wave and a TE-wave. Therefore, it is common to consider the simplified eigenvalue problems (3.43) and (3.46) for two-dimensional photonic crystals instead of the more general eigenvalue problems (3.26) and (3.32) posed in $\mathbb{R}^2$.

Notice that the eigenvalue problems for the TM- and TE-waves (3.43) and (3.46) are given in terms of partial differential equations with the $\Lambda$-periodic coefficients $\varepsilon_r$ and $1/\varepsilon_r$, respectively. Therefore, we make a Bloch ansatz

$$e_\omega^{\mathrm{TM}} = e_{\omega,\boldsymbol{k}}^{\mathrm{TM}}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{k}, \cdot \rangle}, \tag{3.47}$$

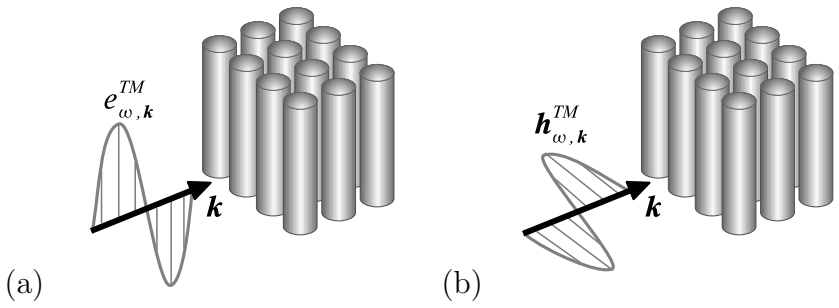



**Figure 3.8:** Directions of oscillation for the electric field (a) and the magnetic field (b) of a TM-polarized wave in the sense of two-dimensional photonic crystals.

$$h_\omega^{\mathrm{TE}} = h_{\omega,\boldsymbol{k}}^{\mathrm{TE}}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{k}, \cdot \rangle}, \tag{3.48}$$

where $e_{\omega,\boldsymbol{k}}, h_{\omega,\boldsymbol{k}} : \mathbb{R}^2 \to \mathbb{C}$ are $\Lambda$-periodic functions, and where $\boldsymbol{k}$ is a quasimomentum vector belonging to the first Brillouin-zone $\mathbb{B}$ of $\widehat{\Lambda}$. Note that for two-dimensional photonic crystals the first Brillouin $\mathbb{B}$ is a closed, bounded, convex polygon, since the Bravais lattice $\Lambda$ has maximal rank. Plugging the Bloch ansatz (3.47) into (3.43) yields

$$-(\nabla + \mathrm{i}\boldsymbol{k}) \cdot (\nabla + \mathrm{i}\boldsymbol{k}) e_{\omega,\boldsymbol{k}}^{\mathrm{TM}} = \frac{\omega^2}{c_0^2}\, \varepsilon_r\, e_{\omega,\boldsymbol{k}}^{\mathrm{TM}} \quad \text{in } \Omega, \tag{3.49}$$

where $\Omega$ is an arbitrary primitive cell of $\Lambda$. It suffices to solve the eigenvalue problem (3.49) in a primitive cell of $\Lambda$, since the functions $e_{\omega,\boldsymbol{k}}^{\mathrm{TM}}$ and $\varepsilon_r$ are $\Lambda$-periodic by assumption. Note that (3.49) constitutes a family of eigenvalue problems for TM-waves with the quasimomentum vector $\boldsymbol{k}$ varying in the first Brillouin zone $\mathbb{B}$. Similarly, by plugging the Bloch ansatz (3.48) into (3.46), one obtains a family of eigenvalue problems for TE-waves, which reads

$$-(\nabla + \mathrm{i}\boldsymbol{k}) \cdot \left[ \frac{1}{\varepsilon_r} (\nabla + \mathrm{i}\boldsymbol{k}) h_{\omega,\boldsymbol{k}}^{\mathrm{TE}} \right] = \frac{\omega^2}{c_0^2}\, h_{\omega,\boldsymbol{k}}^{\mathrm{TE}} \quad \text{in } \Omega. \tag{3.50}$$

As in the case of three-dimensional photonic crystals, a solution $(\omega, e_{\omega,\boldsymbol{k}}^{\mathrm{TM}})$ of (3.49) determines the electric field of a TM-polarized Bloch mode of the photonic crystal with frequency $\omega$. The corresponding magnetic field $\boldsymbol{h}_{\omega,\boldsymbol{k}}^{\mathrm{TM}}$ can be obtained from (3.42) and (3.47). Similarly, a solution $(\omega, e_{\omega,\boldsymbol{k}}^{\mathrm{TE}})$ determines the magnetic field of a TE-polarized Bloch mode, and the corresponding electric field $\boldsymbol{e}_{\omega,\boldsymbol{k}}^{\mathrm{TE}}$ can be obtained from (3.45) and (3.48). An electromagnetic wave with a specific frequency is able to propagate inside a two-dimensional photonic crystal if there exists a TM-polarized Bloch mode or a TE-polarized Bloch mode with the same frequency.

Finally, we mention that TM and TE are abbreviations for *transverse magnetic* and *transverse electric*. The names for the corresponding polarizations originally came from the theory of electromagnetic wave guides, were the propagation of an electromagnetic wave is always assumed in the direction of the third spatial coordinate. The magnetic fields of TM-waves hence oscillate in a plane, which is perpendicular (or transverse) to the direction of propagation. For two-dimensional photonic crystals, however, the term transverse magnetic is rather misleading, since the magnetic fields of TM-waves actually oscillate in the plane of wave propagation as we illustrated in Figure 3.8. The same misconception can arise for TE-waves. Nevertheless, the abbreviations TM and TE are commonly used in the literature on photonic crystals in order to indicate that the polarization of an electromagnetic wave is given by (3.41) or (3.44), respectively.

# Chapter 4

# Spectral Theory

In Sections 3.3 and 3.4 we developed a mathematical model for time-harmonic wave propagation in three-dimensional photonic crystals. This model leads to a family of constrained eigenvalue problems for the wave's magnetic field, which was given by (3.36). As is turns out the crystal's relative electric permittivity function $\varepsilon_r$ enters as a coefficient in each eigenvalue problem. In this chapter, we shall consider the constrained eigenvalue problems from a more general point of view. In Section 4.1 we therefore define a generic family of constrained eigenvalue problems of similar type. In Section 4.2 we introduce suitable function spaces for the corresponding eigenfunctions. Section 4.3 is concerned with establishing a weak formulation of the eigenvalue problems. This weak formulation will be the basis for the spectral theory we present in Section 4.4. In Section 4.5 we introduce a less well-known characterization principle for eigenvalues of an elliptic partial differential operator. This principle will play an important role in the discussion on the existence of solutions of photonic band structure optimization problems in Chapter 5. In Section 4.6 we relate the generic family of eigenvalue problems introduced in Section 4.1 to the band structures of photonic crystals. In particular, we discuss the effect of the non-translational symmetries of a photonic crystal's medium structure on its photonic band structure. In Section 4.7 we briefly comment on how the results of this chapter can be transferred to two-dimensional photonic crystals.

## 4.1 The Formal Setting

Throughout this chapter, we assume that $\Lambda \subset \mathbb{R}^3$ is a Bravais lattice of rank 3, which represents the translational symmetry of a three-dimensional photonic crystal. By $\Omega$ we denote the interior of a bounded, convex, primitive cell of $\Lambda$ (see Section 2.5). Since the lattice $\Lambda$ has maximal rank, $\Omega$ is an bounded, convex

domain in $\mathbb{R}^3$. Furthermore, we denote by $\mathbb{B}$ the first Brillouin zone of $\Lambda$. The first Brillouin zone is the Wigner–Seitz cell of the reciprocal lattice $\widehat{\Lambda}$ of $\Lambda$ and as such a closed, bounded, convex subset of $\mathbb{R}^3$. We then consider a family of eigenvalue problems, which is given as follows.

**Problem 4.1.** *Given a coefficient $\rho : \Omega \to \mathbb{R}$, which is positive, bounded, and uniformly bounded away from zero almost everywhere on $\Omega$, and a vector $\boldsymbol{k} \in \mathbb{B}$, find eigenvalues $\lambda \in \mathbb{C}$ and corresponding, $\Lambda$-periodic eigenfunctions $\boldsymbol{u} : \mathbb{R}^3 \to \mathbb{C}^3$, with $\boldsymbol{u} \neq \boldsymbol{0}$, such that*

$$\begin{cases} (\nabla + \mathrm{i}\boldsymbol{k}) \times \big[\rho(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}\big] = \lambda \boldsymbol{u} & \text{in } \Omega, \\ (\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{u} = 0 & \text{in } \Omega. \end{cases} \tag{4.1}$$

Clearly, the constrained eigenvalue problem (4.1) is of the same form as (3.36). In (4.1), the coefficient $\rho$ takes the place of the reciprocal relative electric permittivity $1/\varepsilon_r$. More precisely $\rho$ can be identified with $1/\varepsilon_r|_\Omega$. The unknown eigenvalue $\lambda$ takes the place of the quantity $\omega^2/c_0^2$ and the unknown eigenfunction $\boldsymbol{u}$ that of the complex vector field $\boldsymbol{h}_{\omega,\boldsymbol{k}}$, which determines the corresponding Bloch mode with frequency $\omega = c_0\sqrt{\lambda}$ and quasimomentum vector $\boldsymbol{k}$.

In Section 3.1 we assumed that a photonic crystal consists of finitely many different materials. Since every dielectric material possess its own specific electric permittivity, the function $\varepsilon_r$ is discontinuous in most cases. Therefore, we only assume $\rho$ as a function, which is bounded almost everywhere on $\Omega$. Because of this, the eigenvalue equation in (4.1) can only be understood in the classical sense, where all functions are assumed to be sufficiently smooth. Therefore, we develop a mathematical framework in the following sections, which allows us to treat (4.1) in a weak sense.

## 4.2 Sobolev Spaces of Periodic Functions

In this section we introduce suitable function spaces for the analysis of the family of constrained eigenvalue problems stated in Problem 4.1. Recall that the unknown eigenfunctions in (4.1) were required to be $\Lambda$-periodic, where $\Lambda$ is the Bravais lattice describing the translational symmetry of the medium. Since (4.1) is posed in a primitive domain $\Omega$ of $\Lambda$ only, one has to ensure that a non-vanishing vector field, which solves (4.1) for some $\lambda \in \mathbb{C}$, also has a suitable $\Lambda$-periodic extension to $\mathbb{R}^3$. Furthermore, we require the eigenfunctions to satisfy some weak regularity assumptions, in order to establish a weak formulation of (4.1). Having this in mind, we introduce certain Sobolev spaces of periodic functions in this section. Our aim is to give a consistent characterization of these function spaces in terms of Fourier

series. Moreover, we list a number of important results, which are related to the family of constrained eigenvalue problems stated in Problem 4.1.

In the following, we assume that $\Omega$ is a primitive domain of a Bravais lattice $\Lambda \subset \mathbb{R}^3$ of rank 3. We define the Sobolev spaces

$$\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega) := \big\{ \boldsymbol{w}|_\Omega \;\big|\; \boldsymbol{w} \in \boldsymbol{H}_{\mathrm{loc}}(\mathrm{curl};\mathbb{R}^3),\ \boldsymbol{w} \text{ is } \Lambda\text{-periodic} \big\}, \tag{4.2}$$

$$\boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega) := \big\{ \boldsymbol{f}|_\Omega \;\big|\; \boldsymbol{f} \in \boldsymbol{H}_{\mathrm{loc}}(\mathrm{div};\mathbb{R}^3),\ \boldsymbol{f} \text{ is } \Lambda\text{-periodic} \big\}, \tag{4.3}$$

$$H^1_{\mathrm{per}}(\Omega) := \big\{ q|_\Omega \;\big|\; q \in H^1_{\mathrm{loc}}(\mathbb{R}^3),\ q \text{ is } \Lambda\text{-periodic} \big\}. \tag{4.4}$$

The spaces $\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)$, $\boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega)$, and $H^1_{\mathrm{per}}(\Omega)$ are Hilbert spaces, which are equipped with the inner products $\langle \cdot\,, \cdot \rangle_{\mathrm{curl},\Omega}$, $\langle \cdot\,, \cdot \rangle_{\mathrm{div},\Omega}$, and $\langle \cdot\,, \cdot \rangle_{1,\Omega}$, respectively. Clearly, they are proper linear subspaces of $\boldsymbol{H}(\mathrm{curl};\Omega)$, $\boldsymbol{H}(\mathrm{div};\Omega)$, and $H^1(\Omega)$. Next, our aim is to derive a suitable characterization of these spaces.

Recall that every $\Lambda$-periodic function in $L^2_{\mathrm{loc}}(\mathbb{R}^3)$ is represented by its *Fourier expansion* (see e.g. Section 4.19 in [10]). More precisely,

$$f = \sum_{\boldsymbol{b} \in \widehat{\Lambda}} \widehat{f}_{\boldsymbol{b}}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{b}, \cdot \rangle} \qquad \text{for all } f \in L^2_{\mathrm{loc}}(\mathbb{R}^3)$$

in the sense of $L^2_{\mathrm{loc}}(\Omega)$. The so-called *Fourier coefficients* $\widehat{f}_{\boldsymbol{b}}$ of $f$ are given by

$$\widehat{f}_{\boldsymbol{b}} := \frac{1}{\mathrm{meas}(\Omega)} \int_\Omega f\, \mathrm{e}^{-\mathrm{i}\langle \boldsymbol{b}, \cdot \rangle} \qquad \text{for all } \boldsymbol{b} \in \widehat{\Lambda}. \tag{4.5}$$

According to *Parseval's identity*, we also have that

$$\|f\|^2_\Omega = \mathrm{meas}(\Omega) \sum_{\boldsymbol{b} \in \widehat{\Lambda}} |\widehat{f}_{\boldsymbol{b}}|^2 \qquad \text{for all } f \in L^2_{\mathrm{loc}}(\mathbb{R}^3).$$

It should be noted that Fourier coefficients, as defined by (4.5), can be computed for every $\Lambda$-periodic function in $L^1_{\mathrm{loc}}(\mathbb{R}^3)$. This is because the restriction of a function in $L^1_{\mathrm{loc}}(\mathbb{R}^3)$ to the primitive domain $\Omega$ belongs to $L^1(\Omega)$. Parseval's identity implies that the restriction of a $\Lambda$-periodic function $f \in L^1_{\mathrm{loc}}(\mathbb{R}^3)$ to $\Omega$ belongs to $L^2(\Omega)$, if and only if its Fourier coefficients satisfy

$$\sum_{\boldsymbol{b} \in \widehat{\Lambda}} |\widehat{f}_{\boldsymbol{b}}|^2 < \infty. \tag{4.6}$$

If that is the case, the function $f$ obviously belongs to $L^2_{\mathrm{loc}}(\mathbb{R}^3)$ due to the $\Lambda$-periodicity. It follows, that a $\Lambda$-periodic function $f : \mathbb{R}^3 \to \mathbb{C}$ belongs to $L^2_{\mathrm{loc}}(\mathbb{R}^3)$, if and only if its Fourier coefficients $\widehat{f}_{\boldsymbol{b}}$ are well-defined, and if they satisfy (4.6).

Now suppose that $g : \Omega \to \mathbb{C}$ is a some given function. The question is, whether or not $g$ can be understood as the restriction to $\Omega$ of a $\Lambda$-periodic function $L^2_{\mathrm{loc}}(\Omega)$. If $g$ belongs to $L^1(\Omega)$, one can compute Fourier coefficients $\widehat{g}_{\boldsymbol{b}}$ of $g$ according to (4.5). By Parseval's identity, the function $g$ even belongs to $L^2(\Omega)$, if and only if

$$\sum_{\boldsymbol{b}\in\widehat{\Lambda}} |\widehat{g}_{\boldsymbol{b}}|^2 < \infty. \tag{4.7}$$

Provided that the above estimate holds, one can define a function $\widetilde{g} \in L^2_{\mathrm{loc}}(\mathbb{R}^3)$ by

$$\widetilde{g} := \sum_{\boldsymbol{b}\in\widehat{\Lambda}} \widehat{g}_{\boldsymbol{b}}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{b}, \cdot\, \rangle}.$$

Obviously, the function $\widetilde{g}$ is $\Lambda$-periodic, and we have that the restriction of $\widetilde{g}$ to the primitive domain $\Omega$ coincides with $g$. Hence, a function $g : \Omega \to \mathbb{C}$ can be understood as the restriction to $\Omega$ of a $\Lambda$-periodic function in $L^2_{\mathrm{loc}}(\Omega)$, if and only if its Fourier coefficients $\widehat{g}_{\boldsymbol{b}}$ are well-defined, and if they satisfy (4.7).

The above discussion shows that the restrictions of $\Lambda$-periodic functions in $L^2_{\mathrm{loc}}(\mathbb{R}^3)$ to a primitive domain $\Omega$ can be characterized by their Fourier coefficients. In the following, this principle is used to characterize the Sobolev spaces Sobolev spaces $\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)$, $\boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega)$, and $H^1_{\mathrm{per}}(\Omega)$, which were defined by (4.2)–(4.4).

Given some $\Lambda$-periodic functions $\boldsymbol{w} \in \boldsymbol{H}_{\mathrm{loc}}(\mathrm{curl};\mathbb{R}^3)$, $\boldsymbol{f} \in \boldsymbol{H}_{\mathrm{loc}}(\mathrm{div};\mathbb{R}^3)$, and $q \in H^1_{\mathrm{loc}}(\mathbb{R}^3)$, we have that $\nabla\times\boldsymbol{w}$, $\nabla\cdot\boldsymbol{f}$, and $\nabla q$ belong to $L^2_{\mathrm{loc}}(\mathbb{R}^3)^3$, $L^2_{\mathrm{loc}}(\mathbb{R}^3)$, and $L^2_{\mathrm{loc}}(\mathbb{R}^3)^3$, respectively. Furthermore, we have that these functions are $\Lambda$-periodic. Hence, they are represented by their Fourier expansions. More precisely, we have that the identities

$$\nabla\times\boldsymbol{w} = \sum_{\boldsymbol{b}\in\widehat{\Lambda}} \mathrm{i}\boldsymbol{b}\times\widehat{\boldsymbol{w}}_{\boldsymbol{b}}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{b}, \cdot\, \rangle}, \tag{4.8}$$

$$\nabla\cdot\boldsymbol{f} = \sum_{\boldsymbol{b}\in\widehat{\Lambda}} \mathrm{i}\boldsymbol{b}\cdot\widehat{\boldsymbol{f}}_{\boldsymbol{b}}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{b}, \cdot\, \rangle}, \tag{4.9}$$

$$\nabla q = \sum_{\boldsymbol{b}\in\widehat{\Lambda}} \mathrm{i}\boldsymbol{b}\,\widehat{q}_{\boldsymbol{b}}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{b}, \cdot\, \rangle} \tag{4.10}$$

hold in the sense of $L^2_{\mathrm{loc}}(\mathbb{R}^3)^3$, $L^2_{\mathrm{loc}}(\mathbb{R}^3)$, and $L^2_{\mathrm{loc}}(\mathbb{R}^3)^3$, respectively. By Parseval's identity we also have that

$$\|\nabla\times\boldsymbol{w}\|^2_\Omega = \mathrm{meas}(\Omega) \sum_{\boldsymbol{b}\in\widehat{\Lambda}} |\boldsymbol{b}\times\widehat{\boldsymbol{w}}_{\boldsymbol{b}}|^2, \tag{4.11}$$

$$\|\nabla\cdot\boldsymbol{f}\|^2_\Omega = \mathrm{meas}(\Omega) \sum_{\boldsymbol{b}\in\widehat{\Lambda}} |\boldsymbol{b}\cdot\widehat{\boldsymbol{f}}_{\boldsymbol{b}}|^2, \tag{4.12}$$

$$\|\nabla q\|_\Omega^2 = \mathrm{meas}(\Omega) \sum_{\boldsymbol{b}\in\widehat{\Lambda}} |\boldsymbol{b}|^2 |\widehat{q}_{\boldsymbol{b}}|^2. \tag{4.13}$$

Since the restrictions of $\nabla \times \boldsymbol{w}$, $\nabla \cdot \boldsymbol{f}$, and $\nabla q$ to the primitive domain $\Omega$ belong to $L^2(\Omega)^3$, $L^2(\Omega)$, and $L^2(\Omega)^3$, it follows that

$$\sum_{\boldsymbol{b}\in\widehat{\Lambda}} |\boldsymbol{b} \times \widehat{\boldsymbol{w}}_{\boldsymbol{b}}|^2 < \infty, \tag{4.14}$$

$$\sum_{\boldsymbol{b}\in\widehat{\Lambda}} |\boldsymbol{b} \cdot \widehat{\boldsymbol{f}}_{\boldsymbol{b}}|^2 < \infty, \tag{4.15}$$

$$\sum_{\boldsymbol{b}\in\widehat{\Lambda}} |\boldsymbol{b}|^2 |\widehat{q}_{\boldsymbol{b}}|^2 < \infty. \tag{4.16}$$

Conversely, suppose that $\boldsymbol{w}$, $\boldsymbol{f}$ and $q$ are functions in $L^2(\Omega)^3$, in $L^2(\Omega)^3$, and in $L^2(\Omega)$, whose Fourier coefficients satisfy (4.14)–(4.16). Then, one can extend these functions $\Lambda$-periodically to $\mathbb{R}^3$ using their Fourier series, and thus obtain functions that belong to $\boldsymbol{H}_{\mathrm{loc}}(\mathrm{curl};\mathbb{R}^3)$, $\boldsymbol{H}_{\mathrm{loc}}(\mathrm{div};\mathbb{R}^3)$, and $H^1_{\mathrm{loc}}(\mathbb{R}^3)$. Hence, the Sobolev spaces $\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)$, $\boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega)$, and $H^1_{\mathrm{per}}(\Omega)$ an be characterizes as

$$\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega) = \left\{ \boldsymbol{w} \in L^2(\Omega)^3 \;\middle|\; \sum_{\boldsymbol{b}\in\widehat{\Lambda}} \big(|\widehat{\boldsymbol{w}}_{\boldsymbol{b}}|^2 + |\boldsymbol{b} \times \widehat{\boldsymbol{w}}_{\boldsymbol{b}}|^2\big) < \infty \right\},$$

$$\boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega) = \left\{ \boldsymbol{f} \in L^2(\Omega)^3 \;\middle|\; \sum_{\boldsymbol{b}\in\widehat{\Lambda}} \big(|\widehat{\boldsymbol{f}}_{\boldsymbol{b}}|^2 + |\boldsymbol{b} \cdot \widehat{\boldsymbol{f}}_{\boldsymbol{b}}|^2\big) < \infty \right\},$$

$$H^1_{\mathrm{per}}(\Omega) = \left\{ q \in L^2(\Omega) \;\middle|\; \sum_{\boldsymbol{b}\in\widehat{\Lambda}} \big(1 + |\boldsymbol{b}|^2\big) |\widehat{q}_{\boldsymbol{b}}|^2 < \infty \right\}.$$

The fact that all functions, which belong to the above Sobolev spaces, are represented by their Fourier expansions greatly simplifies the proofs of the results, which are given below. As we shall see, some of these proofs can be accomplished by relatively simple computations involving Fourier coefficients. Consider, for example, the proof of the following lemma.

**Lemma 4.2.** *For every vector field $\boldsymbol{\psi} \in H^1_{\mathrm{per}}(\Omega)^3$ the following identity holds,*

$$\|\boldsymbol{\psi}\|_{1,\Omega}^2 = \|\boldsymbol{\psi}\|_\Omega^2 + \|\nabla \cdot \boldsymbol{\psi}\|_\Omega^2 + \|\nabla \times \boldsymbol{\psi}\|_\Omega^2.$$

*Proof.* Let $\boldsymbol{\psi} = (\psi^{(1)}, \psi^{(2)}, \psi^{(3)}) \in H^1_{\mathrm{per}}(\Omega)^3$ be an arbitrarily chosen vector field. Then, we obtain by (4.13) that

$$\|\boldsymbol{\psi}\|_{1,\Omega}^2 = \|\boldsymbol{\psi}\|_\Omega^2 + \|\nabla\psi^{(1)}\|_\Omega^2 + \|\nabla\psi^{(2)}\|_\Omega^2 + \|\nabla\psi^{(3)}\|_\Omega^2$$

$$
\begin{aligned}
&= \|\boldsymbol{\psi}\|_\Omega^2 + \operatorname{meas}(\Omega) \sum_{\boldsymbol{b}\in\widehat{\Lambda}} |\boldsymbol{b}|^2 \Big( |\widehat{\psi^{(1)}}_{\boldsymbol{b}}|^2 + |\widehat{\psi^{(2)}}_{\boldsymbol{b}}|^2 + |\widehat{\psi^{(3)}}_{\boldsymbol{b}}|^2 \Big) \\
&= \|\boldsymbol{\psi}\|_\Omega^2 + \operatorname{meas}(\Omega) \sum_{\boldsymbol{b}\in\widehat{\Lambda}} |\boldsymbol{b}|^2 \Big( \big|\widehat{\psi^{(1)}}_{\boldsymbol{b}}\big|^2 + \big|\widehat{\psi^{(2)}}_{\boldsymbol{b}}\big|^2 + \big|\widehat{\psi^{(3)}}_{\boldsymbol{b}}\big|^2 \Big) \\
&= \|\boldsymbol{\psi}\|_\Omega^2 + \operatorname{meas}(\Omega) \sum_{\boldsymbol{b}\in\widehat{\Lambda}} |\boldsymbol{b}|^2 |\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2 .
\end{aligned}
$$

Using the vector norm identity (2.4) on page 15, one obtains that $|\boldsymbol{b}|^2|\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2 = |\boldsymbol{b}\cdot\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2 + |\boldsymbol{b}\times\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2$ holds for every $\boldsymbol{b}\in\widehat{\Lambda}$. Hence, by (4.11) and (4.12), we have that

$$
\begin{aligned}
\|\boldsymbol{\psi}\|_{1,\Omega}^2 &= \|\boldsymbol{\psi}\|_\Omega^2 + \operatorname{meas}(\Omega) \sum_{\boldsymbol{b}\in\widehat{\Lambda}} \Big( |\boldsymbol{b}\cdot\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2 + |\boldsymbol{b}\times\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2 \Big) \\
&= \|\boldsymbol{\psi}\|_\Omega^2 + \|\nabla\cdot\boldsymbol{\psi}\|_\Omega^2 + \|\nabla\times\boldsymbol{\psi}\|_\Omega^2 ,
\end{aligned}
$$

which completes the proof. □

It should be noted that Lemma 4.2 implies that

$$
H^1_{\mathrm{per}}(\Omega)^3 = \boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega) \cap \boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega).
$$

Next, we investigate some properties of the modified curl, divergence, and gradient operators $(\nabla+\mathrm{i}\boldsymbol{k})\times$, $(\nabla+\mathrm{i}\boldsymbol{k})\cdot$ and $(\nabla+\mathrm{i}\boldsymbol{k})$, where $\boldsymbol{k}$ is a vector belonging to the first Brillouin zone $\mathbb{B}$ of $\Lambda$. As a first result, we state the following lemma, which establishes important norm estimates.

**Lemma 4.3.** *For every vector $\boldsymbol{k}\in\mathbb{B}$ we have the norm estimates*

$$
\begin{aligned}
\|(\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{w}\|_\Omega &\le \sqrt{\beta_0(\boldsymbol{k})}\,\|\boldsymbol{w}\|_{\mathrm{curl},\Omega} && \text{for all } \boldsymbol{w}\in\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega), \\
\|(\nabla+\mathrm{i}\boldsymbol{k})\cdot\boldsymbol{f}\|_\Omega &\le \sqrt{\beta_0(\boldsymbol{k})}\,\|\boldsymbol{f}\|_{\mathrm{div},\Omega} && \text{for all } \boldsymbol{f}\in\boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega), \\
\|(\nabla+\mathrm{i}\boldsymbol{k})q\|_\Omega &\le \sqrt{\beta_0(\boldsymbol{k})}\,\|q\|_{1,\Omega} && \text{for all } q\in H^1_{\mathrm{per}}(\Omega),
\end{aligned}
$$

*where the function $\beta_0:\mathbb{B}\to\mathbb{R}_{>0}$ is given by*

$$
\beta_0(\boldsymbol{k}) := 2\max\{1,|\boldsymbol{k}|^2\} \qquad \text{for all } \boldsymbol{k}\in\mathbb{B}.
$$

*Proof.* Given an arbitrary function $\boldsymbol{w}\in\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl},\Omega)$, one obtains

$$
\|(\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{w}\|_\Omega^2 = \int_\Omega |(\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{w}|^2 \le \int_\Omega \big(|\nabla\times\boldsymbol{w}| + |\boldsymbol{k}\times\boldsymbol{w}|\big)^2
$$

$$\begin{aligned}
&\leq 2\biggl(\int_\Omega |\nabla \times \boldsymbol{w}|^2 + \int_\Omega |\boldsymbol{k} \times \boldsymbol{w}|^2\biggr) \\
&= 2\biggl(\int_\Omega |\nabla \times \boldsymbol{w}|^2 + |\boldsymbol{k}|^2 \int_\Omega |\boldsymbol{w}|^2\biggr) \\
&\leq 2\, \max\bigl\{1, |\boldsymbol{k}|^2\bigr\}\, \|\boldsymbol{w}\|_{\mathrm{curl},\Omega}^2,
\end{aligned}$$

which proves the first norm estimate. The remaining norm estimates can be verified analogously. □

Lemma 4.3 in particular implies that the modified curl, divergence, and gradient operators are continuous, linear operators from $\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl}, \Omega)$, $\boldsymbol{H}_{\mathrm{per}}(\mathrm{div}, \Omega)$, and $H^1(\Omega)$ into $L^2(\Omega)^3$, $L^2(\Omega)^3$, and $L^2(\Omega)$, respectively. Hence, given some arbitrary functions $\boldsymbol{w} \in \boldsymbol{H}_{\mathrm{per}}(\mathrm{curl}; \Omega)$, $\boldsymbol{f} \in \boldsymbol{H}_{\mathrm{per}}(\mathrm{div}; \Omega)$, and $q \in H^1_{\mathrm{per}}(\Omega)$, we have that the functions $(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{w}$, $(\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{f}$, and $(\nabla + \mathrm{i}\boldsymbol{k})q$ are represented by their Fourier expansions, i.e., we have that

$$(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{w} = \sum_{\boldsymbol{b} \in \widehat{\Lambda}} \mathrm{i}(\boldsymbol{b} + \boldsymbol{k}) \times \widehat{\boldsymbol{w}}_{\boldsymbol{b}}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{b}, \cdot \rangle}, \tag{4.17}$$

$$(\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{f} = \sum_{\boldsymbol{b} \in \widehat{\Lambda}} \mathrm{i}(\boldsymbol{b} + \boldsymbol{k}) \cdot \widehat{\boldsymbol{f}}_{\boldsymbol{b}}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{b}, \cdot \rangle}, \tag{4.18}$$

$$(\nabla + \mathrm{i}\boldsymbol{k})q = \sum_{\boldsymbol{b} \in \widehat{\Lambda}} \mathrm{i}(\boldsymbol{b} + \mathrm{i}\boldsymbol{k})\, \widehat{q}_{\boldsymbol{b}}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{b}, \cdot \rangle} \tag{4.19}$$

in the sense of $L^2(\Omega)^3$, $L^2(\Omega)$, and $L^2(\Omega)^3$. Using Parseval's identity again, we also find that

$$\|(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{w}\|_\Omega^2 = \sum_{\boldsymbol{b} \in \widehat{\Lambda}} \bigl|(\boldsymbol{b} + \boldsymbol{k}) \times \widehat{\boldsymbol{w}}_{\boldsymbol{b}}\bigr|^2, \tag{4.20}$$

$$\|(\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{f}\|_\Omega^2 = \sum_{\boldsymbol{b} \in \widehat{\Lambda}} \bigl|(\boldsymbol{b} + \boldsymbol{k}) \cdot \widehat{\boldsymbol{f}}_{\boldsymbol{b}}\bigr|^2, \tag{4.21}$$

$$\|(\nabla + \mathrm{i}\boldsymbol{k})q\|_\Omega^2 = \sum_{\boldsymbol{b} \in \widehat{\Lambda}} \bigl|(\boldsymbol{b} + \boldsymbol{k})\, \widehat{q}_{\boldsymbol{b}}\bigr|^2. \tag{4.22}$$

The following lemma provides *Gårding*-type inequalities for the modified curl, divergence and gradient operators. We remark that the proof of this lemma relies on the fact that the first Brillouin zone $\mathbb{B}$ is defined as the Wigner–Seitz cell of the reciprocal lattice $\widehat{\Lambda}$. This implies that $\mathbb{B}$ is point symmetric with respect to the origin (see Section 2.5). Therefore, we have that

$$|\boldsymbol{k}| \leq \frac{1}{2}\, \mathrm{diam}(\mathbb{B}) \qquad \text{for all } \boldsymbol{k} \in \mathbb{B}, \tag{4.23}$$

where $\operatorname{diam}(\mathbb{B})$ denotes the diameter of $\mathbb{B}$.

**Lemma 4.4.** *For every vector $\boldsymbol{k} \in \mathbb{B}$ we have the inequalities*

$$\begin{aligned}
\|(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{w}\|_\Omega^2 &\geq \alpha_0(\boldsymbol{k}) \, \|\boldsymbol{w}\|_{\mathrm{curl},\Omega}^2 - \kappa_0(\boldsymbol{k}) \, \|\boldsymbol{w}\|_\Omega^2 && \text{for all } \boldsymbol{w} \in \boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega), \\
\|(\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{f}\|_\Omega^2 &\geq \alpha_0(\boldsymbol{k}) \, \|\boldsymbol{f}\|_{\mathrm{div},\Omega}^2 - \kappa_0(\boldsymbol{k}) \, \|\boldsymbol{f}\|_\Omega^2 && \text{for all } \boldsymbol{f} \in \boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega), \\
\|(\nabla + \mathrm{i}\boldsymbol{k}) q\|_\Omega^2 &\geq \alpha_0(\boldsymbol{k}) \, \|q\|_{1,\Omega}^2 - \kappa_0(\boldsymbol{k}) \, \|q\|_\Omega^2 && \text{for all } q \in H^1_{\mathrm{per}}(\Omega),
\end{aligned}$$

*where the functions $\alpha_0 : \mathbb{B} \to \mathbb{R}_{>0}$ and $\kappa_0 : \mathbb{B} \to \mathbb{R}_{>0}$ are given by*

$$\begin{aligned}
\alpha_0(\boldsymbol{k}) &:= \left(1 - \frac{|\boldsymbol{k}|}{\operatorname{diam}(\mathbb{B})}\right) && \text{for all } \boldsymbol{k} \in \mathbb{B}. \\
\kappa_0(\boldsymbol{k}) &:= \alpha_0(\boldsymbol{k}) + \operatorname{diam}(\mathbb{B})|\boldsymbol{k}| - |\boldsymbol{k}|^2 && \text{for all } \boldsymbol{k} \in \mathbb{B}.
\end{aligned}$$

*Proof.* Let $\boldsymbol{w} \in \boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)$ be chosen arbitrarily. Then, we have

$$\begin{aligned}
\|(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{w}\|_\Omega^2 &= \int_\Omega |(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}|^2 \\
&= \int_\Omega \Big( |\nabla \times \boldsymbol{u}|^2 + 2 \operatorname{Re}\big[\overline{\nabla \times \boldsymbol{u}} \cdot (\mathrm{i}\boldsymbol{k} \times \boldsymbol{u})\big] + |\boldsymbol{k} \times \boldsymbol{u}|^2 \Big) \\
&\geq \int_\Omega \Big( |\nabla \times \boldsymbol{u}|^2 - 2|\nabla \times \boldsymbol{u}||\boldsymbol{k} \times \boldsymbol{u}| + |\boldsymbol{k} \times \boldsymbol{u}|^2 \Big)
\end{aligned}$$

By Young's inequality, we have that

$$|\nabla \times \boldsymbol{u}|_\Omega \, |\boldsymbol{k} \times \boldsymbol{u}|_\Omega \leq \frac{1}{2}\left( \frac{1}{\theta} \, |\nabla \times \boldsymbol{u}|_\Omega^2 + \theta \, |\boldsymbol{k} \times \boldsymbol{u}|_\Omega^2 \right) \qquad \text{for all } \theta > 0.$$

Letting $\theta := \operatorname{diam}(\mathbb{B})/|\boldsymbol{k}|$, we have that $\theta \geq 2$ according to (4.23). Using the vector norm estimate (2.4) on page 15, we hence get

$$\begin{aligned}
\|(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{w}\|_\Omega^2 &\geq \left(1 - \frac{1}{\theta}\right) \int_\Omega |\nabla \times \boldsymbol{u}|^2 - (\theta - 1) \int_\Omega |\boldsymbol{k} \times \boldsymbol{u}|^2 \\
&\geq \left(1 - \frac{1}{\theta}\right) \int_\Omega |\nabla \times \boldsymbol{u}|^2 - (\theta - 1) \int_\Omega |\boldsymbol{k}|^2 |\boldsymbol{u}|^2 \\
&= \left(1 - \frac{|\boldsymbol{k}|}{\operatorname{diam}(\mathbb{B})}\right) \|\nabla \times \boldsymbol{u}\|_\Omega^2 - \big(\operatorname{diam}(\mathbb{B})|\boldsymbol{k}| - |\boldsymbol{k}|^2\big) \|\boldsymbol{u}\|_\Omega^2
\end{aligned}$$

which establishes the inequality stated in (a). The remaining inequalities can be established through a similar line of argument. □

Note that the functions $\alpha_0$ and $\kappa_0$ in Lemma 4.4 are indeed positive everywhere on $\mathbb{B}$. This is due to the estimate given in (4.23), which implies that

$$\alpha_0(\boldsymbol{k}) \geq \frac{1}{2} \qquad \text{for all } \boldsymbol{k} \in \mathbb{B},$$
$$\kappa_0(\boldsymbol{k}) \geq \frac{1}{2} \qquad \text{for all } \boldsymbol{k} \in \mathbb{B}.$$

We remark that the parameter $\theta$ in the proof of Lemma 4.4 should be chosen such that $\alpha_0(\mathbf{0}) = \kappa_0(\mathbf{0}) = 1$. With this, the asserted inequalities become equalities for the case $\boldsymbol{k} = \mathbf{0}$ as one would expect.

In order to establish a weak formulation of the eigenvalue problem stated in Problem 4.1, it is desirable to have *Green*-type identities for the modified curl, divergence and gradient operators. Such identities are established in the following theorem.

**Theorem 4.5.** *The following assertions hold for every vector $\boldsymbol{k} \in \mathbb{B}$.*

(a) *For every $\boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{H}_{\mathrm{per}}(\mathrm{curl}; \Omega)$ we have*

$$\int_\Omega \overline{(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{w}} \cdot \boldsymbol{v} = \int_\Omega \overline{\boldsymbol{w}} \cdot (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{v}.$$

(b) *For every $\boldsymbol{f} \in \boldsymbol{H}_{\mathrm{per}}(\mathrm{div}; \Omega)$ and $q \in H^1_{\mathrm{per}}(\Omega)$ we have*

$$\int_\Omega \overline{(\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{f}}\, q = -\int_\Omega \overline{\boldsymbol{f}} \cdot (\nabla + \mathrm{i}\boldsymbol{k}) q.$$

*Proof.* To prove the identities, we expand all functions according (4.17)–(4.19). Because of (4.20)–(4.22) integration and summation can be interchanged. Since

$$\int_\Omega \mathrm{e}^{\mathrm{i}\langle \boldsymbol{b} - \widetilde{\boldsymbol{b}}, \cdot \rangle} = \begin{cases} \mathrm{meas}(\Omega) & \text{if } \boldsymbol{b} = \widetilde{\boldsymbol{b}}, \\ 0 & \text{else} \end{cases} \qquad \text{for all } \boldsymbol{b}, \widetilde{\boldsymbol{b}} \in \widehat{\Lambda},$$

we obtain

$$\int_\Omega \overline{(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{w}} \cdot \boldsymbol{v} = -\,\mathrm{meas}(\Omega)\,\mathrm{i} \sum_{\boldsymbol{b} \in \widehat{\Lambda}} \big((\boldsymbol{b} + \boldsymbol{k}) \times \overline{\widehat{\boldsymbol{w}}_{\boldsymbol{b}}}\big) \cdot \widehat{\boldsymbol{v}}_{\boldsymbol{b}},$$
$$\int_\Omega \overline{\boldsymbol{w}} \cdot (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{v} = \mathrm{meas}(\Omega)\,\mathrm{i} \sum_{\boldsymbol{b} \in \widehat{\Lambda}} \overline{\widehat{\boldsymbol{w}}_{\boldsymbol{b}}} \cdot \big((\boldsymbol{b} + \boldsymbol{k}) \times \widehat{\boldsymbol{v}}_{\boldsymbol{b}}\big),$$
$$\int_\Omega \overline{(\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{f}}\, q = -\,\mathrm{meas}(\Omega)\,\mathrm{i} \sum_{\boldsymbol{b} \in \widehat{\Lambda}} (\boldsymbol{b} + \boldsymbol{k}) \cdot \overline{\widehat{\boldsymbol{f}}_{\boldsymbol{b}}}\, \widehat{q}_{\boldsymbol{b}},$$

$$-\int_\Omega \overline{\boldsymbol{f}} \cdot (\nabla + \mathrm{i}\boldsymbol{k})q = \mathrm{meas}(\Omega)\,\mathrm{i}\sum_{\boldsymbol{b}\in\widehat{\Lambda}} \overline{\widehat{\boldsymbol{f}}_{\boldsymbol{b}}} \cdot (\boldsymbol{b}+\boldsymbol{k})\widehat{q}_{\boldsymbol{b}}.$$

The assertions then follow from the vector identity (2.2) and the fact that $\boldsymbol{x}\times\boldsymbol{y} = -\boldsymbol{y}\times\boldsymbol{x}$ for all $\boldsymbol{x},\boldsymbol{y}\in\mathbb{C}^3$. □

It should be noted that in Theorem 4.5 integrals over the boundary $\partial\Omega$ of the fundamental cell $\Omega$ do not appear, neither in the integral identities nor in the proof. Again, this is due to the fact that all functions are represented by corresponding Fourier expansions, so that the integral identities reduce to simple vector identities for the Fourier coefficients. Using the same principle, it is easy to verify the following lemma, which extends some well-known results about certain compositions of curl, divergence and gradient operators to their modified counterparts.

**Lemma 4.6.** *The following assertions hold for every vector* $\boldsymbol{k}\in\mathbb{B}$.

(a) *For every function* $q\in H^1_{\mathrm{per}}(\Omega)$ *we have* $(\nabla+\mathrm{i}\boldsymbol{k})q\in\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)$ *and*

$$(\nabla+\mathrm{i}\boldsymbol{k})\times(\nabla+\mathrm{i}\boldsymbol{k})q = \boldsymbol{0}.$$

(b) *For every function* $\boldsymbol{w}\in\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)$ *we have* $(\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{w}\in\boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega)$ *and*

$$(\nabla+\mathrm{i}\boldsymbol{k})\cdot(\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{w} = 0.$$

*Proof.* Let $q\in H^1_{\mathrm{per}}(\Omega)$ and $\boldsymbol{w}\in\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)$ be chosen arbitrarily. Then, by (4.17) and (4.19) the Fourier coefficients of the functions $\boldsymbol{u} := (\nabla+\mathrm{i}\boldsymbol{k})q$ and $\boldsymbol{f} := (\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{w}$ are given by $\widehat{\boldsymbol{u}}_{\boldsymbol{b}} = \mathrm{i}(\boldsymbol{b}+\boldsymbol{k})\widehat{q}_{\boldsymbol{b}}$ and $\widehat{\boldsymbol{f}}_{\boldsymbol{b}} = \mathrm{i}(\boldsymbol{b}+\boldsymbol{k})\times\widehat{\boldsymbol{w}}_{\boldsymbol{b}}$ for all $\boldsymbol{b}\in\widehat{\Lambda}$, where $\widehat{q}_{\boldsymbol{b}}$ and $\widehat{\boldsymbol{w}}_{\boldsymbol{b}}$ denote the respective Fourier coefficients of $q$ and $\boldsymbol{w}$. The functions $\boldsymbol{u}$ and $\boldsymbol{f}$ belong to $\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)$ and $\boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega)$, respectively, since

$$\begin{aligned}\sum_{\boldsymbol{b}\in\widehat{\Lambda}}\bigl(|\widehat{\boldsymbol{u}}_{\boldsymbol{b}}|^2+|\boldsymbol{b}\times\widehat{\boldsymbol{u}}_{\boldsymbol{b}}|^2\bigr) &= \sum_{\boldsymbol{b}\in\widehat{\Lambda}}\bigl(|(\boldsymbol{b}+\boldsymbol{k})\widehat{q}_{\boldsymbol{b}}|^2+|\boldsymbol{b}\times\boldsymbol{k}\widehat{q}_{\boldsymbol{b}}|^2\bigr)\\ &\le \sum_{\boldsymbol{b}\in\widehat{\Lambda}}|(\boldsymbol{b}+\boldsymbol{k})\widehat{q}_{\boldsymbol{b}}|^2+|\boldsymbol{k}|^2\sum_{\boldsymbol{b}\in\widehat{\Lambda}}|\boldsymbol{b}|^2|\widehat{q}_{\boldsymbol{b}}|^2<\infty,\end{aligned}$$

and

$$\begin{aligned}\sum_{\boldsymbol{b}\in\widehat{\Lambda}}\bigl(|\widehat{\boldsymbol{f}}_{\boldsymbol{b}}|^2+|\boldsymbol{b}\cdot\widehat{\boldsymbol{f}}_{\boldsymbol{b}}|^2\bigr) &= \sum_{\boldsymbol{b}\in\widehat{\Lambda}}\Bigl(\bigl|(\boldsymbol{b}+\boldsymbol{k})\times\widehat{\boldsymbol{w}}_{\boldsymbol{b}}\bigr|^2+\bigl|\boldsymbol{b}\cdot\bigl((\boldsymbol{b}+\boldsymbol{k})\times\widehat{\boldsymbol{w}}_{\boldsymbol{b}}\bigr)\bigr|^2\Bigr)\\ &= \sum_{\boldsymbol{b}\in\widehat{\Lambda}}\bigl(|(\boldsymbol{b}+\boldsymbol{k})\times\widehat{\boldsymbol{w}}_{\boldsymbol{b}}|^2+|\boldsymbol{k}\cdot(\boldsymbol{b}\times\widehat{\boldsymbol{w}}_{\boldsymbol{b}})|^2\bigr)\end{aligned}$$

$$\leq \sum_{\boldsymbol{b}\in\widehat{\Lambda}} |(\boldsymbol{b}+\boldsymbol{k})\times\widehat{\boldsymbol{w}}_{\boldsymbol{b}}|^2 + |\boldsymbol{k}|^2 \sum_{\boldsymbol{b}\in\widehat{\Lambda}} |\boldsymbol{b}\times\widehat{\boldsymbol{w}}_{\boldsymbol{b}}|^2 < \infty.$$

To obtain the above inequality we used the same vector identities as in the proof of Theorem 4.8. By (4.17) and (4.18), the Fourier coefficients of the function $\boldsymbol{v} := (\nabla + \mathrm{i}\boldsymbol{k})\times\boldsymbol{u}$ and $g := (\nabla+\mathrm{i}\boldsymbol{k})\cdot\boldsymbol{f}$ are given by $\widehat{\boldsymbol{v}}_{\boldsymbol{b}} = \mathrm{i}(\boldsymbol{b}+\boldsymbol{k})\times\widehat{\boldsymbol{u}}_{\boldsymbol{b}} = -(\boldsymbol{b}+\boldsymbol{k})\times(\boldsymbol{b}+\boldsymbol{k})\widehat{q}_{\boldsymbol{b}} = \boldsymbol{0}$, and $\widehat{g}_{\boldsymbol{b}} = \mathrm{i}(\boldsymbol{b}+\boldsymbol{k})\cdot\widehat{\boldsymbol{f}}_{\boldsymbol{b}} = -(\boldsymbol{b}+\boldsymbol{k})\cdot\big((\boldsymbol{b}+\boldsymbol{k})\times\widehat{\boldsymbol{w}}_{\boldsymbol{b}}\big) = -\widehat{\boldsymbol{w}}_{\boldsymbol{b}}\cdot\big((\boldsymbol{b}+\boldsymbol{k})\times(\boldsymbol{b}+\boldsymbol{k})\big) = 0$ for all $\boldsymbol{b}\in\widehat{\Lambda}$. Hence, $\boldsymbol{v}=\boldsymbol{0}$ and $g=0$. □

**Corollary 4.7.** *For every vector $\boldsymbol{k}\in\mathbb{B}$ we have*

$$(\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega) \perp (\nabla+\mathrm{i}\boldsymbol{k})H^1_{\mathrm{per}}(\Omega),$$

*where $\perp$ denotes the orthogonality relation in $L^2(\Omega)^3$.*

*Proof.* By Lemma 4.6 and Theorem 4.8 we have

$$\begin{aligned}\langle(\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{w}\,,(\nabla+\mathrm{i}\boldsymbol{k})q\rangle_\Omega &= \int_\Omega \overline{(\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{w}}\cdot(\nabla+\mathrm{i}\boldsymbol{k})q\\ &= \int_\Omega \overline{\boldsymbol{w}}\cdot(\nabla+\mathrm{i}\boldsymbol{k})\times(\nabla+\mathrm{i}\boldsymbol{k})q = 0\end{aligned}$$

for every $\boldsymbol{w}\in\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)$ and every $q\in H^1_{\mathrm{per}}(\Omega)$. □

Before we state our next theorem, we find it convenient to introduce the function space

$$L^2_\diamond(\Omega) := \left\{ f\in L^2(\Omega) \;\middle|\; \int_\Omega f = 0\right\}. \tag{4.24}$$

Clearly, $L^2_\diamond(\Omega)$ is a closed linear subspace of $L^2(\Omega)$ and thus a Hilbert space equipped with the inner product $\langle\,\cdot\,,\,\cdot\,\rangle_\Omega$. According to (4.5), we have

$$\widehat{f}_{\boldsymbol{0}} = \frac{1}{\mathrm{meas}(\Omega)}\int_\Omega f \qquad \text{for all } f\in L^2(\Omega).$$

Therefore, the function space $L^2_\diamond(\Omega)$ can be characterized as

$$L^2_\diamond(\Omega) = \big\{ f\in L^2(\Omega) \;\big|\; \widehat{f}_{\boldsymbol{0}} = 0\big\}.$$

It is furthermore well-known that the Hilbert space $L^2_\diamond(\Omega)$ is isometrically linearly isomorphic to the linear factor space $L^2(\Omega)/\mathbb{C}$. From this follows the orthogonal space decomposition $L^2(\Omega) = L^2_\diamond(\Omega)\oplus\mathbb{C}$. The function spaces

$$\boldsymbol{H}_{\mathrm{per},\diamond}(\mathrm{curl};\Omega) := \boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)\cap L^2_\diamond(\Omega)^3,$$

$$\begin{aligned}\boldsymbol{H}_{\mathrm{per},\diamond}(\mathrm{div};\Omega) &:= \boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega) \cap L^2_\diamond(\Omega)^3,\\ H^1_{\mathrm{per},\diamond}(\Omega) &:= H^1_{\mathrm{per}}(\Omega) \cap L^2_\diamond(\Omega),\end{aligned}$$

are closed linear subspaces of $\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)$, $\boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega)$, and $H^1_{\mathrm{per}}(\Omega)$, respectively.

**Theorem 4.8.** *The following assertions hold for every vector field $\boldsymbol{f} \in L^2(\Omega)^3$.*

(a) *For every vector $\boldsymbol{k} \in \mathbb{B} \setminus \{\boldsymbol{0}\}$ there exist uniquely determined functions $\boldsymbol{\psi} \in H^1_{\mathrm{per}}(\Omega)^3$ and $p \in H^1_{\mathrm{per}}(\Omega)$, as well as a constant $c > 0$, such that*

$$\begin{aligned}(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{\psi} + (\nabla + \mathrm{i}\boldsymbol{k})p &= \boldsymbol{f},\\ (\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{\psi} &= 0,\\ \|\boldsymbol{\psi}\|_{1,\Omega} + \|p\|_{1,\Omega} &\leq c\|\boldsymbol{f}\|_\Omega.\end{aligned}$$

(b) *There exist uniquely determined functions $\boldsymbol{\psi} \in H^1_{\mathrm{per},\diamond}(\Omega)^3$ and $p \in H^1_{\mathrm{per},\diamond}(\Omega)$, as well as a constant $c > 0$, such that*

$$\begin{aligned}\nabla \times \boldsymbol{\psi} + \nabla p + \widehat{\boldsymbol{f}}_{\boldsymbol{0}} &= \boldsymbol{f},\\ \nabla \cdot \boldsymbol{\psi} &= 0,\\ \|\boldsymbol{\psi}\|_{1,\Omega} + \|p\|_{1,\Omega} + \sqrt{\mathrm{meas}(\Omega)}\,|\widehat{\boldsymbol{f}}_{\boldsymbol{0}}| &\leq c\|\boldsymbol{f}\|_\Omega.\end{aligned}$$

*Proof.* A proof for (a) is given in Section 3 in [33], but can also be derived with the techniques that are presented in the following. In order to proof (b), we consider an arbitrary function $\boldsymbol{f} \in L^2(\Omega)^3$ with

$$\boldsymbol{f} = \sum_{\boldsymbol{b} \in \widehat{\Lambda}} \widehat{\boldsymbol{f}}_{\boldsymbol{b}}\, \mathrm{e}^{\mathrm{i}\langle \boldsymbol{b}, \cdot \rangle}.$$

Notice that for every vector $\boldsymbol{b} \in \widehat{\Lambda} \setminus \{\boldsymbol{0}\}$ there exists a uniquely defined complex number $\widehat{q}_{\boldsymbol{b}} \in \mathbb{C}$ and a uniquely defined vector $\widehat{\boldsymbol{\varphi}}_{\boldsymbol{b}} \in (\mathrm{span}\{\boldsymbol{b}\})^\perp$, such that $\widehat{\boldsymbol{f}}_{\boldsymbol{b}} = \boldsymbol{b}\widehat{q}_{\boldsymbol{b}} + \widehat{\boldsymbol{\varphi}}_{\boldsymbol{b}}$.

In Section 2.2 we introduced the cross product matrix $[\boldsymbol{z}]_\times \in \mathbb{C}^{3\times 3}$ of a complex vector $\boldsymbol{z} \in \mathbb{C}^3$. For non-vanishing $\boldsymbol{z}$ we showed that the image space of $[\boldsymbol{z}]_\times$ coincides with $(\mathrm{span}\{\boldsymbol{z}\})^\perp$. It follows, that the pseudo-inverse $[\boldsymbol{z}]^{\mathrm{P}}_\times$ of $[\boldsymbol{z}]_\times$ constitutes a linear isomorphism from $(\mathrm{span}\{\boldsymbol{z}\})^\perp$ onto itself.

With the above definitions, we set $\widehat{p}_{\boldsymbol{b}} := -\mathrm{i}\widehat{q}_{\boldsymbol{b}}$ and $\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}} := -\mathrm{i}[\boldsymbol{b}]^{\mathrm{P}}_\times \widehat{\boldsymbol{\varphi}}_{\boldsymbol{b}}$. With this we obtain $\widehat{\boldsymbol{f}}_{\boldsymbol{b}} = \mathrm{i}\boldsymbol{b}\widehat{p}_{\boldsymbol{b}} + \mathrm{i}\boldsymbol{b} \times \widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}$ for all $\boldsymbol{b} \in \widehat{\Lambda} \setminus \{\boldsymbol{0}\}$. Furthermore, since

$\mathrm{im}([\boldsymbol{b}]_\times^{\mathrm{P}}) = \mathrm{im}([\boldsymbol{b}]_\times^{\mathrm{T}}) = \mathrm{im}([\boldsymbol{b}]_\times)$, we have $\boldsymbol{b}\cdot\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}} = \langle \boldsymbol{b}\,,\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}\rangle = \langle \boldsymbol{b}\,,\widehat{\boldsymbol{\varphi}}_{\boldsymbol{b}}\rangle = 0$ for all $\boldsymbol{b}\in\widehat{\Lambda}$. Next, we define the functions $p:\Omega\to\mathbb{C}$ and $\boldsymbol{\psi}:\Omega\to\mathbb{C}^3$ by

$$p := \sum_{\boldsymbol{b}\in\widehat{\Lambda}\setminus\{\boldsymbol{0}\}} \widehat{p}_{\boldsymbol{b}}\,\mathrm{e}^{\mathrm{i}\langle \boldsymbol{b},\,\cdot\,\rangle}, \quad \boldsymbol{\psi} := \sum_{\boldsymbol{b}\in\widehat{\Lambda}\setminus\{\boldsymbol{0}\}} \widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}\,\mathrm{e}^{\mathrm{i}\langle \boldsymbol{b},\,\cdot\,\rangle}.$$

By Parseval's identity we have

$$\begin{aligned}
\|p\|_{1,\Omega}^2 + \|\boldsymbol{\psi}\|_{1,\Omega}^2 &= \mathrm{meas}(\Omega)\sum_{\boldsymbol{b}\in\widehat{\Lambda}\setminus\{\boldsymbol{0}\}}(1+|\boldsymbol{b}|^2)\big(|\widehat{p}_{\boldsymbol{b}}|^2+|\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2\big)\\
&\le 2\,\mathrm{meas}(\Omega)\sum_{\boldsymbol{b}\in\widehat{\Lambda}\setminus\{\boldsymbol{0}\}}\big(|\widehat{p}_{\boldsymbol{b}}|^2|\boldsymbol{b}|^2+|\boldsymbol{b}|^2|\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2\big)\\
&= 2\,\mathrm{meas}(\Omega)\sum_{\boldsymbol{b}\in\widehat{\Lambda}\setminus\{\boldsymbol{0}\}}\big(|\widehat{p}_{\boldsymbol{b}}\boldsymbol{b}|^2+|\boldsymbol{b}\times\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2\big)\\
&= 2\,\mathrm{meas}(\Omega)\sum_{\boldsymbol{b}\in\widehat{\Lambda}\setminus\{\boldsymbol{0}\}}\big(|\widehat{p}_{\boldsymbol{b}}\boldsymbol{b}|^2+|\widehat{\boldsymbol{\varphi}}_{\boldsymbol{b}}|^2\big)\\
&= 2\,\mathrm{meas}(\Omega)\sum_{\boldsymbol{b}\in\widehat{\Lambda}\setminus\{\boldsymbol{0}\}}|\widehat{\boldsymbol{f}}_{\boldsymbol{b}}|^2\\
&= 2\|\boldsymbol{f}\|_\Omega^2 - 2\,\mathrm{meas}(\Omega)|\widehat{\boldsymbol{f}}_{\boldsymbol{0}}|^2.
\end{aligned}$$

Since $\boldsymbol{f}$ belongs to $L^2(\Omega)^3$, it follows that $p$ and $\boldsymbol{\psi}$ belong to $H^1_{\mathrm{per}}(\Omega)$ and $H^1_{\mathrm{per}}(\Omega)^3$, respectively. According to (4.8)–(4.10) we also have that $\nabla\times\boldsymbol{\psi}+\nabla p+\widehat{\boldsymbol{f}}_{\boldsymbol{0}} = \boldsymbol{f}$, as well as $\nabla\cdot\boldsymbol{\psi}=0$. Finally, the above inequality implies that $\|\boldsymbol{\psi}\|_{1,\Omega}+\|p\|_{1,\Omega}+\sqrt{\mathrm{meas}(\Omega)}|\widehat{\boldsymbol{f}}_{\boldsymbol{0}}| \le \sqrt{2}\|\boldsymbol{f}\|_\Omega$. □

The functions $\boldsymbol{\psi}$ and $p$ in Theorem 4.8 are commonly referred to as *vector potentials* and *scalar potentials* of the function $\boldsymbol{f}$. An additive decomposition of a vector field into a curl field of a vector potential and a gradient field of a scalar potential is called a *Helmholtz decomposition*. In analogy to this, Part (a) in Theorem 4.8 provides a Helmholtz decomposition with modified vector and scalar potentials $(\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{\psi}$ and $(\nabla+\mathrm{i}\boldsymbol{k})p$, where $\boldsymbol{k}\in\mathbb{B}\setminus\{\boldsymbol{0}\}$. It is also well-known, that for every vector field $\boldsymbol{f}\in L^2(\Omega)^3$ there exists a Helmholtz decomposition $\boldsymbol{f}=\nabla\times\boldsymbol{\psi}+\nabla p$, such that $\nabla\cdot\boldsymbol{\psi}=0$, with uniquely determined functions $\boldsymbol{\psi}\in H^1_\diamond(\Omega)^3$ and $p\in H^1_\diamond(\Omega)$ (see e.g. Theorem and Remarks 3.3 in [38]). Part (b) of Theorem 4.8, however, states that such a decomposition with $\Lambda$-periodic vector and scalar potentials $\boldsymbol{\psi}\in H^1_{\mathrm{per},\diamond}(\Omega)^3$ and $p\in H^1_{\mathrm{per},\diamond}(\Omega)$ exists, if and only if $\boldsymbol{f}\in L^2_\diamond(\Omega)^3$.

Notice that Part (b) of Theorem 4.8 implies a fortiori that for every $\boldsymbol{f}\in L^2(\Omega)$ there exist functions $\boldsymbol{\psi}\in H^1_{\mathrm{per}}(\Omega)^3$ and $p\in H^1_{\mathrm{per}}(\Omega)$ with $\nabla\cdot\boldsymbol{\psi}=0$, as well as a uniquely defined complex vector $\boldsymbol{z}\in\mathbb{C}^3$, such that $\boldsymbol{f}=\nabla\times\boldsymbol{\psi}+\nabla p+\boldsymbol{z}$. The

functions $\boldsymbol{\psi}$ and $p$ are only uniquely determined modulo additive constants in this case. From Theorem 4.5 and Theorem 4.8 we hence deduce the following corollary.

**Corollary 4.9.** *For every vector $\boldsymbol{k} \in \mathbb{B} \setminus \{\mathbf{0}\}$ we have the following orthogonal space decompositions.*

(a) $L^2(\Omega)^3 = (\nabla + \mathrm{i}\boldsymbol{k}) \times H^1_{\mathrm{per}}(\Omega)^3 \oplus (\nabla + \mathrm{i}\boldsymbol{k}) H^1_{\mathrm{per}}(\Omega)$.

(b) $L^2(\Omega)^3 = \nabla \times H^1_{\mathrm{per},\diamond}(\Omega)^3 \oplus \nabla H^1_{\mathrm{per},\diamond}(\Omega) \oplus \mathbb{C}^3$.

(c) $L^2_\diamond(\Omega)^3 = \nabla \times H^1_{\mathrm{per},\diamond}(\Omega)^3 \oplus \nabla H^1_{\mathrm{per},\diamond}(\Omega)$.

Note that Part (b) of Corollary 4.9 in particular states that the curl of any vector field in $H^1_{\mathrm{per}}(\Omega)^3$ can only be constant, if it vanishes identically, and that the same holds true for the gradient of any function in $H^1_{\mathrm{per}}(\Omega)$. Another important corollary is obtained from Lemma 4.6 and Theorem 4.8.

**Corollary 4.10.** *The following identities hold for every vector $\boldsymbol{k} \in \mathbb{B} \setminus \{\mathbf{0}\}$.*

(a) $\ker\big((\nabla + \mathrm{i}\boldsymbol{k}) \times |_{\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)}\big) = (\nabla + \mathrm{i}\boldsymbol{k}) H^1_{\mathrm{per}}(\Omega)$.

(b) $\ker\big((\nabla + \mathrm{i}\boldsymbol{k}) \cdot |_{\boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega)}\big) = (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)$.

(c) $\ker\big(\nabla \times |_{\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)}\big) = \nabla H^1_{\mathrm{per}}(\Omega) \oplus \mathbb{C}^3$.

(d) $\ker\big(\nabla \cdot |_{\boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega)}\big) = \nabla \times \boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega) \oplus \mathbb{C}^3$.

(e) $\ker\big(\nabla \times |_{\boldsymbol{H}_{\mathrm{per},\diamond}(\mathrm{curl};\Omega)}\big) = \nabla H^1_{\mathrm{per},\diamond}(\Omega)$.

(f) $\ker\big(\nabla \cdot |_{\boldsymbol{H}_{\mathrm{per},\diamond}(\mathrm{div};\Omega)}\big) = \nabla \times \boldsymbol{H}_{\mathrm{per},\diamond}(\mathrm{curl};\Omega)$.

In the literature the identities in Part (a) and (b) of Corollary 4.10 are often summarized in stating that for every vector $\boldsymbol{k} \in \mathbb{B} \setminus \{\mathbf{0}\}$ the sequence

$$H^1_{\mathrm{per}}(\Omega) \overset{(\nabla+\mathrm{i}\boldsymbol{k})}{\longrightarrow} \boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega) \overset{(\nabla+\mathrm{i}\boldsymbol{k})\times}{\longrightarrow} \boldsymbol{H}_{\mathrm{per}}(\mathrm{div};\Omega) \overset{(\nabla+\mathrm{i}\boldsymbol{k})\cdot}{\longrightarrow} L^2(\Omega) \tag{4.25}$$

is exact (see e.g. [14], [33]). Here, "exact" simply means that the image of every operator in the sequence coincides with the kernel of the next operator. From Part (e) and (f) of Corollary 4.10 we further deduce that the sequence

$$H^1_{\mathrm{per},\diamond}(\Omega) \overset{\nabla}{\longrightarrow} \boldsymbol{H}_{\mathrm{per},\diamond}(\mathrm{curl};\Omega) \overset{\nabla\times}{\longrightarrow} \boldsymbol{H}_{\mathrm{per},\diamond}(\mathrm{div};\Omega) \overset{\nabla\cdot}{\longrightarrow} L^2_\diamond(\Omega) \tag{4.26}$$

is also exact. Both sequences are analogues of the *de Rham sequence*

$$H^1(\Omega) \overset{\nabla}{\longrightarrow} \boldsymbol{H}(\mathrm{curl};\Omega) \overset{\nabla\times}{\longrightarrow} \boldsymbol{H}(\mathrm{div};\Omega) \overset{\nabla\cdot}{\longrightarrow} L^2(\Omega), \tag{4.27}$$

which is also known to be exact (see e.g. Proposition 8 and Theorem 8 in [20]). We remark that the sequences in (4.27) can be interpreted as a variant of de Rham's exact cochain complex of differential forms on $\mathbb{R}^3$ (see e.g. Section 1 in [16]). When interpreted this way, the function spaces $H^1(\Omega)$, $\boldsymbol{H}(\mathrm{curl};\Omega)$, $\boldsymbol{H}(\mathrm{div};\Omega)$, and $L^2(\Omega)$ represent certain classes of differential forms of order 0, 1, 2, and 3, respectively. The gradient, curl and divergence operators represent exterior derivative operators acting on the respective differential forms. Clearly, the sequences in (4.25) and (4.27) can be interpreted in the same way. Thus, all three sequences reveal certain geometric relationships between the respective function spaces.

## 4.3 The Weak Formulation

In this section we establish a weak formulation of the family of eigenvalue problems stated in Problem 4.1. It turns out that the weak formulation of each individual eigenvalue problem bears some similarity to the weak formulation of standard Laplace-type or second-order divergence type eigenvalue problems. In particular, one is able to prove coercivity and, for non-vanishing vectors $\boldsymbol{k} \in \mathbb{B}$, even ellipticity of the corresponding sesquilinear forms.

In order to establish the weak formulation, we define the complex Hilbert spaces

$$\boldsymbol{W} := \boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega), \tag{4.28}$$

$$Q := H^1_{\mathrm{per}}(\Omega), \tag{4.29}$$

$$\boldsymbol{Z} := L^2(\Omega)^3, \tag{4.30}$$

as well as the real Banach space

$$\mathcal{E} := L^\infty(\Omega,\mathbb{R}), \tag{4.31}$$

and the function set

$$\mathcal{D} := \Big\{ \rho \in \mathcal{E} \;\Big|\; \exists\, \delta > 0 \text{ such that } \operatorname*{ess\,inf}_{\Omega}(\rho) \geq \delta \Big\}. \tag{4.32}$$

The set $\mathcal{D}$ consists of all essentially bounded functions, which are positive almost everywhere on $\Omega$. Clearly, the function set $\mathcal{D}$ is an open subset of $\mathcal{E}$. Furthermore, every function in $\mathcal{D}$ satisfies the required properties of the coefficient $\rho$ in Problem 4.1. Therefore, we define for every function $\rho \in \mathcal{D}$ and every vector $\boldsymbol{k} \in \mathbb{B}$ the sesquilinear forms $a_{\boldsymbol{k}}(\rho) : \boldsymbol{W} \times \boldsymbol{W} \to \mathbb{C}$, $b_{\boldsymbol{k}} : \boldsymbol{W} \times Q \to \mathbb{C}$, and $m : \boldsymbol{Z} \times \boldsymbol{Z} \to \mathbb{C}$ by

$$a_{\boldsymbol{k}}(\rho)(\boldsymbol{w},\boldsymbol{v}) := \int_\Omega \rho \overline{(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{w}} \cdot (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{v} \qquad \text{for all } \boldsymbol{w},\boldsymbol{v} \in \boldsymbol{W}, \tag{4.33}$$

$$b_{\boldsymbol{k}}(\boldsymbol{f},q) := \int_\Omega \overline{\boldsymbol{f}} \cdot (\nabla + \mathrm{i}\boldsymbol{k})q \qquad \text{for all } \boldsymbol{f} \in \boldsymbol{Z}, q \in Q, \quad (4.34)$$

$$m(\boldsymbol{f},\boldsymbol{g}) := \int_\Omega \overline{\boldsymbol{f}} \cdot \boldsymbol{g} \qquad \text{for all } \boldsymbol{f},\boldsymbol{g} \in \boldsymbol{Z}. \quad (4.35)$$

Furthermore, we define for every vector $\boldsymbol{k} \in \mathbb{B}$ the linear function space

$$\boldsymbol{V}_{\boldsymbol{k}} := \{\boldsymbol{w} \in \boldsymbol{W} \mid b_{\boldsymbol{k}}(\boldsymbol{w},q) = 0 \text{ for all } q \in Q\}. \quad (4.36)$$

Clearly, for every vector $\boldsymbol{k} \in \mathbb{B}$, the function space $\boldsymbol{V}_{\boldsymbol{k}}$ is a closed subspace of $\boldsymbol{W}$ and thus a complex Hilbert space when equipped with the inner product $\langle \cdot\,, \cdot \rangle_{\mathrm{curl},\Omega}$. Each such Hilbert space $\boldsymbol{V}_{\boldsymbol{k}}$ can be characterized as follows.

**Proposition 4.11.** *For every vector $\boldsymbol{k} \in \mathbb{B}$ we have the identity*

$$\boldsymbol{V}_{\boldsymbol{k}} = \begin{cases} \boldsymbol{W} \cap \big((\nabla + \mathrm{i}\boldsymbol{k}) \times H^1_{\mathrm{per}}(\Omega)^3\big) & \text{if } \boldsymbol{k} \neq \boldsymbol{0}, \\ \boldsymbol{W} \cap \big(\nabla \times H^1_{\mathrm{per}}(\Omega)^3 \oplus \mathbb{C}^3\big) & \text{if } \boldsymbol{k} = \boldsymbol{0}. \end{cases}$$

*Proof.* First, we consider the case $\boldsymbol{k} \neq 0$. Let $\boldsymbol{w} \in \boldsymbol{W}$, with $\boldsymbol{w} = (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{\psi}$ for some $\boldsymbol{\psi} \in H^1_{\mathrm{per}}(\Omega)^3$. By Corollary 4.7 we then have

$$b_{\boldsymbol{k}}(\boldsymbol{w},q) = \langle (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{\psi}\,, (\nabla + \mathrm{i}\boldsymbol{k})q \rangle_\Omega = 0 \qquad \text{for all } q \in Q,$$

and hence $\boldsymbol{w} \in \boldsymbol{V}_{\boldsymbol{k}}$. Now, we choose a function $\boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}}$ arbitrarily. Then, by Theorem 4.8, there exist functions $\boldsymbol{\psi} \in \boldsymbol{H}^1_{\mathrm{per}}(\Omega)^3$ and $p \in Q$, such that $\boldsymbol{u} = (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{\psi} + (\nabla + \mathrm{i}\boldsymbol{k})p$. Furthermore, we have

$$\begin{aligned} \|(\nabla + \mathrm{i}\boldsymbol{k})p\|^2_\Omega &= \langle (\nabla + \mathrm{i}\boldsymbol{k})p\,, (\nabla + \mathrm{i}\boldsymbol{k})p \rangle_\Omega \\ &= \langle (\nabla + \mathrm{i}\boldsymbol{k})p\,, (\nabla + \mathrm{i}\boldsymbol{k})p \rangle_\Omega + \langle (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{\psi}\,, (\nabla + \mathrm{i}\boldsymbol{k})p \rangle_\Omega \\ &= b_{\boldsymbol{k}}((\nabla + \mathrm{i}\boldsymbol{k})p, p) + b_{\boldsymbol{k}}((\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{\psi}, p) \\ &= b_{\boldsymbol{k}}(\boldsymbol{u}, p) = 0, \end{aligned}$$

which implies $(\nabla + \mathrm{i}\boldsymbol{k})p = \boldsymbol{0}$. This completes the proof for the case $\boldsymbol{k} \neq \boldsymbol{0}$. The proof for the case $\boldsymbol{k} = \boldsymbol{0}$ is established analogously. Let $\boldsymbol{w_0} \in \boldsymbol{W}$, with $\boldsymbol{w_0} = \nabla \times \boldsymbol{\psi_0} + \boldsymbol{z_0}$ for some function $\boldsymbol{\psi_0} \in H^1_{\mathrm{per}}(\Omega)^3$ and some vector $\boldsymbol{z_0} \in \mathbb{C}^3$. By Theorem 4.8 and Corollary 4.7 we then obtain

$$b_{\boldsymbol{0}}(\boldsymbol{w_0},q) = \langle \nabla \times \boldsymbol{\psi_0}\,, \nabla q \rangle_\Omega + \langle \nabla \cdot \boldsymbol{z_0}\,, q \rangle_\Omega = 0 \qquad \text{for all } q \in Q,$$

which implies that $\boldsymbol{w_0} \in \boldsymbol{V_0}$. Choosing $\boldsymbol{u_0} \in \boldsymbol{V_0}$ arbitrarily, we have by Theorem 4.8 that there exist functions $\boldsymbol{\psi_0} \in H^1_{\mathrm{per}}(\Omega)^3$ and $p_{\boldsymbol{0}} \in Q$, as well as a vector $\boldsymbol{z_0} \in \mathbb{C}^3$, such that $\boldsymbol{u_0} = \nabla \times \boldsymbol{\psi_0} + \nabla p_{\boldsymbol{0}} + \boldsymbol{z}_0$. Furthermore, we have

$$\|\nabla p_{\boldsymbol{0}}\|^2_\Omega = \langle \nabla p_{\boldsymbol{0}}\,, \nabla p_{\boldsymbol{0}} \rangle_\Omega$$

$$\begin{aligned}
&= \langle \nabla p_{\mathbf{0}} \, , \nabla p_{\mathbf{0}} \rangle_\Omega + \langle \nabla \times \boldsymbol{\psi}_{\mathbf{0}} \, , \nabla p_{\mathbf{0}} \rangle_\Omega + \langle \boldsymbol{z}_{\mathbf{0}} \, , \nabla p_{\mathbf{0}} \rangle \\
&= b_{\boldsymbol{k}}(\nabla p_{\mathbf{0}}, p_{\mathbf{0}}) + b_{\boldsymbol{k}}(\nabla \times \boldsymbol{\psi}_{\mathbf{0}}, p_{\mathbf{0}}) + b_{\boldsymbol{k}}(\boldsymbol{z}_{\mathbf{0}}, p_{\mathbf{0}}) \\
&= b_{\boldsymbol{k}}(\boldsymbol{u}_{\mathbf{0}}, p_{\mathbf{0}}) = 0,
\end{aligned}$$

which implies $\nabla p_{\mathbf{0}} = \mathbf{0}$. □

In view of the eigenvalue problem stated in Problem 4.1 we have by Theorem 4.5 that each function space $\boldsymbol{V}_{\boldsymbol{k}}$ consists of those functions $\boldsymbol{u} \in \boldsymbol{W}$, which satisfy the divergence constraint $(\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{u} = 0$ in the weak sense. Another relevant aspect of the function spaces $\boldsymbol{V}_{\boldsymbol{k}}$ is the existence of the following norm estimate for $\boldsymbol{k} \neq \mathbf{0}$.

**Lemma 4.12.** *For every vector $\boldsymbol{k} \in \mathbb{B} \setminus \{\mathbf{0}\}$ we have the norm estimate*

$$\|(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}\|_\Omega \geq |\boldsymbol{k}| \, \|\boldsymbol{u}\|_\Omega \qquad \textit{for all } \boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}}.$$

*Proof.* Let $\boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}}$ be chosen arbitrarily. Then, according to Theorem 4.8 there exists a uniquely determined function $\boldsymbol{\psi} \in H^1_{\mathrm{per}}(\Omega)^3$, such that $\boldsymbol{u} = (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{\psi}$ and $(\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{\psi} = 0$. Note that the latter identity implies $(\boldsymbol{b} + \boldsymbol{k}) \cdot \widehat{\boldsymbol{\psi}}_{\boldsymbol{b}} = 0$ for all $\boldsymbol{b} \in \widehat{\Lambda}$, where $\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}$ denotes the respective Fourier coefficient of $\boldsymbol{\psi}$ as defined in (4.5). Using (2.4) and Parseval's identity, we obtain

$$\begin{aligned}
\frac{1}{\mathrm{meas}(\Omega)} \, \|\boldsymbol{u}\|_\Omega^2 &= \frac{1}{\mathrm{meas}(\Omega)} \, \|(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{\psi}\|_\Omega^2 \\
&= \sum_{\boldsymbol{b} \in \widehat{\Lambda}} |(\boldsymbol{b} + \boldsymbol{k}) \times \widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2 \\
&= \sum_{\boldsymbol{b} \in \widehat{\Lambda}} \big( |(\boldsymbol{b} + \boldsymbol{k})|^2 |\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2 - |(\boldsymbol{b} + \boldsymbol{k}) \cdot \widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2 \big) \\
&= \sum_{\boldsymbol{b} \in \widehat{\Lambda}} |(\boldsymbol{b} + \boldsymbol{k})|^2 |\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2 .
\end{aligned}$$

Note that $|\boldsymbol{b} + \boldsymbol{k}| \geq |\boldsymbol{k}|$ for all $\boldsymbol{b} \in \widehat{\Lambda}$. For $\boldsymbol{k} = \mathbf{0}$ this assertion is trivial. For $\boldsymbol{k} \neq \mathbf{0}$ it follows from the fact that $\mathbb{B}$ is the Wigner–Seitz cell of the reciprocal lattice (see Section 3.4). As was discussed in Section 2.5 the Wigner–Seitz cell of a lattice is point symmetric with respect to the origin and consists of all points, which are closer to the origin than to any other lattice point. Using the vector identity stated in (2.1), we find

$$\begin{aligned}
\frac{1}{\mathrm{meas}(\Omega)} \, \|(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}\|_\Omega^2 &= \frac{1}{\mathrm{meas}(\Omega)} \, \|(\nabla + \mathrm{i}\boldsymbol{k}) \times (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{\psi}\|_\Omega^2 \\
&= \sum_{\boldsymbol{b} \in \widehat{\Lambda}} \big| (\boldsymbol{b} + \boldsymbol{k}) \times \big( (\boldsymbol{b} + \boldsymbol{k}) \times \widehat{\boldsymbol{\psi}}_{\boldsymbol{b}} \big) \big|^2
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\boldsymbol{b}\in\widehat{\Lambda}} \left| \big((\boldsymbol{b}+\boldsymbol{k})\cdot\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}\big)(\boldsymbol{b}+\boldsymbol{k}) - |\boldsymbol{b}+\boldsymbol{k}|^2\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}} \right|^2 \\
&= \sum_{\boldsymbol{b}\in\widehat{\Lambda}} |\boldsymbol{b}+\boldsymbol{k}|^4 |\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2 \\
&\geq |\boldsymbol{k}|^2 \sum_{\boldsymbol{b}\in\widehat{\Lambda}} |\boldsymbol{b}+\boldsymbol{k}|^2 |\widehat{\boldsymbol{\psi}}_{\boldsymbol{b}}|^2,
\end{aligned}
$$

which completes the proof. □

Lemma 4.12 provides a *Poincaré–Friedrichs*-type inequality for the restriction of the modified curl operator $(\nabla + \mathrm{i}\boldsymbol{k})\times$ to the Hilbert space $\boldsymbol{V}_{\boldsymbol{k}}$ for every vector $\boldsymbol{k}\in\mathbb{B}\backslash\{\boldsymbol{0}\}$. Since the modified curl operators also satisfy a Gårding-type inequality (see Lemma 4.4), we can state the following result.

**Lemma 4.13.** *For every vector $\boldsymbol{k}\in\mathbb{B}$ we have the norm estimate*

$$
\|(\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{u}\|_\Omega \geq \sqrt{\gamma_0(\boldsymbol{k})}\|\boldsymbol{u}\|_{\mathrm{curl},\Omega} \qquad \text{for all } \boldsymbol{u}\in\boldsymbol{V}_{\boldsymbol{k}}.
$$

*With the definitions in Lemma 4.4, the function $\gamma_0:\mathbb{B}\to\mathbb{R}_{>0}$ is given by*

$$
\gamma_0(\boldsymbol{k}) := \frac{\alpha_0(\boldsymbol{k})|\boldsymbol{k}|^2}{|\boldsymbol{k}|^2+\kappa_0(\boldsymbol{k})} \qquad \text{for all } \boldsymbol{k}\in\mathbb{B}. \tag{4.37}
$$

*Proof.* For $\boldsymbol{k}=\boldsymbol{0}$ the assertion is trivial. We hence turn our attention to the case, where $\boldsymbol{k}\in\mathbb{B}\setminus\{\boldsymbol{0}\}$. Choosing a function $\boldsymbol{u}\in\boldsymbol{V}_{\boldsymbol{k}}$ arbitrarily, we have

$$
\begin{aligned}
\alpha_0(\boldsymbol{k})\|\boldsymbol{u}\|_{\mathrm{curl},\Omega}^2 &\leq \|(\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{u}\|_\Omega^2 + \kappa_0(\boldsymbol{k})\|\boldsymbol{u}\|_\Omega^2 \\
&= \|(\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{u}\|_\Omega^2 + \frac{\kappa_0(\boldsymbol{k})}{|\boldsymbol{k}|^2}\,|\boldsymbol{k}|^2\|\boldsymbol{u}\|_\Omega^2 \\
&\leq \left(1+\frac{\kappa_0(\boldsymbol{k})}{|\boldsymbol{k}|^2}\right)\|(\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{u}\|_\Omega^2
\end{aligned}
$$

by Lemma 4.4 and Lemma 4.12. The norm estimate is obtained by dividing the above inequality by the parenthesized term and taking the square root. □

Lemma 4.13 and Lemma 4.3 imply, that $\|(\nabla+\mathrm{i}\boldsymbol{k})\times(\,\cdot\,)\|_\Omega$ constitutes a norm on $\boldsymbol{V}_{\boldsymbol{k}}$ for all $\boldsymbol{k}\in\mathbb{B}\setminus\{\boldsymbol{0}\}$, which is equivalent to the norm $\|\cdot\|_{\mathrm{curl},\Omega}$. This is an analogon to the well-known fact that $\|\nabla(\,\cdot\,)\|_\Omega$, the seminorm on $H^1(\Omega)$, constitutes a norm on $H_0^1(\Omega)$, which is equivalent to the norm $\|\cdot\|_{1,\Omega}$. Note, however, that this analogon only holds for non-vanishing $\boldsymbol{k}$. In fact, one easily verifies that $\|\nabla\times\cdot\|_\Omega$ only constitutes a seminorm on $\boldsymbol{V}_{\boldsymbol{0}}$.

The following propositions state important properties of the sesquilinear forms $a_{\boldsymbol{k}}(\rho)$, which were defined in (4.33).

**Proposition 4.14.** *For every coefficient $\rho \in \mathcal{D}$, and every vector $\boldsymbol{k} \in \mathbb{B}$ the sesquilinear form $a_{\boldsymbol{k}}(\rho)$ is conjugate-symmetric, bounded, $\boldsymbol{W}$–$\boldsymbol{Z}$-coercive, and positive semidefinite. More precisely, we have*

$$\begin{aligned} a_{\boldsymbol{k}}(\rho)(\boldsymbol{w},\boldsymbol{v}) &= \overline{a_{\boldsymbol{k}}(\rho)(\boldsymbol{v},\boldsymbol{w})} && \text{for all } \boldsymbol{w},\boldsymbol{v} \in \boldsymbol{W},\\ \big|a_{\boldsymbol{k}}(\rho)(\boldsymbol{w},\boldsymbol{v})\big| &\le \beta(\rho,\boldsymbol{k})\|\boldsymbol{w}\|_{\mathrm{curl},\Omega}\|\boldsymbol{v}\|_{\mathrm{curl},\Omega} && \text{for all } \boldsymbol{w},\boldsymbol{v} \in \boldsymbol{W},\\ a_{\boldsymbol{k}}(\rho)(\boldsymbol{w},\boldsymbol{w}) &\ge \alpha(\rho,\boldsymbol{k})\|\boldsymbol{w}\|^2_{\mathrm{curl},\Omega} - \kappa(\rho,\boldsymbol{k})\|\boldsymbol{w}\|^2_{\Omega} && \text{for all } \boldsymbol{w} \in \boldsymbol{W},\\ a_{\boldsymbol{k}}(\rho)(\boldsymbol{w},\boldsymbol{w}) &\ge 0 && \text{for all } \boldsymbol{w} \in \boldsymbol{W}. \end{aligned}$$

*With the definitions in Lemma 4.3 and Lemma 4.4 the functions $\alpha : \mathcal{D}\times\mathbb{B} \to \mathbb{R}_{>0}$, $\beta : \mathcal{D}\times\mathbb{B} \to \mathbb{R}_{>0}$, and $\kappa : \mathcal{D}\times\mathbb{B} \to \mathbb{R}_{>0}$ are given by*

$$\alpha(\rho,\boldsymbol{k}) := \operatorname*{ess\,inf}_{\Omega}(\rho)\,\alpha_0(\boldsymbol{k}) \qquad \text{for all } \rho \in \mathcal{D},\ \boldsymbol{k} \in \mathbb{B}, \tag{4.38}$$

$$\beta(\rho,\boldsymbol{k}) := \|\rho\|_{\Omega,\infty}\,\beta_0(\boldsymbol{k}) \qquad \text{for all } \rho \in \mathcal{D},\ \boldsymbol{k} \in \mathbb{B}, \tag{4.39}$$

$$\kappa(\rho,\boldsymbol{k}) := \operatorname*{ess\,inf}_{\Omega}(\rho)\,\kappa_0(\boldsymbol{k}) \qquad \text{for all } \rho \in \mathcal{D},\ \boldsymbol{k} \in \mathbb{B}. \tag{4.40}$$

*Proof.* The conjugate-symmetry directly follows from the definition of the sesquilinear form $a_{\boldsymbol{k}}(\rho)$. Furthermore, choosing $\rho \in \mathcal{D}$, $\boldsymbol{k} \in \mathbb{B}$, and $\boldsymbol{w},\boldsymbol{v} \in \boldsymbol{W}$ arbitrarily, we have

$$a_{\boldsymbol{k}}(\boldsymbol{w},\boldsymbol{v}) \le \|\rho\|_{\Omega,\infty}\,\|(\nabla + \mathrm{i}\boldsymbol{k})\times\boldsymbol{w}\|_{\Omega}\,\|(\nabla + \mathrm{i}\boldsymbol{k})\times\boldsymbol{v}\|_{\Omega}.$$

Boundedness thus follows from Lemma 4.3. Moreover, we have

$$a_{\boldsymbol{k}}(\boldsymbol{w},\boldsymbol{w}) \ge \operatorname*{ess\,inf}_{\Omega}(\rho)\|(\nabla + \mathrm{i}\boldsymbol{k})\times\boldsymbol{w}\|^2_{\Omega}.$$

Hence, $\boldsymbol{W}$–$\boldsymbol{Z}$-coercivity follows from Lemma 4.4. The above inequality also implies the positive semidefiniteness of the sesquilinear form $a_{\boldsymbol{k}}$. □

**Proposition 4.15.** *For every coefficient $\rho \in \mathcal{D}$ and every vector $\boldsymbol{k} \in \mathbb{B}\setminus\{\boldsymbol{0}\}$ the sesquilinear form $a_{\boldsymbol{k}}(\rho)$ is $\boldsymbol{V}_{\boldsymbol{k}}$-elliptic. More precisely, we have*

$$a_{\boldsymbol{k}}(\rho)(\boldsymbol{u},\boldsymbol{u}) \ge \gamma(\rho,\boldsymbol{k})\|\boldsymbol{u}\|^2_{\mathrm{curl},\Omega} \qquad \text{for all } \boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}}.$$

*With the definitions in Lemma 4.13 the function $\gamma : \mathcal{D}\times\mathbb{B} \to \mathbb{R}_{>0}$ is given by*

$$\gamma(\rho,\boldsymbol{k}) := \operatorname*{ess\,inf}_{\Omega}(\rho)\,\gamma_0(\boldsymbol{k}) \qquad \text{for all } \rho \in \mathcal{D},\ \boldsymbol{k} \in \mathbb{B}. \tag{4.41}$$

*Proof.* Given a coefficient $\rho \in \mathcal{D}$, a vector $\boldsymbol{k} \in \mathbb{B}\setminus\{\boldsymbol{0}\}$, and a function $\boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}}$, we have

$$a_{\boldsymbol{k}}(\boldsymbol{u},\boldsymbol{u}) \ge \operatorname*{ess\,inf}_{\Omega}(\rho)\|(\nabla + \mathrm{i}\boldsymbol{k})\times\boldsymbol{u}\|^2_{\Omega}.$$

The assertion hence follows from Lemma 4.13. □

In the above Propositions 4.14 and 4.15 the terms *coercive* and *elliptic* were used as defined by Wloka (cf. Definitions 17.3 and 17.4 in [78]). Note that Proposition 4.15 implies that, for every $\boldsymbol{k} \in \mathbb{B} \setminus \{\boldsymbol{0}\}$, the corresponding sesquilinear form $a_{\boldsymbol{k}}$ is positive definite on $\boldsymbol{V}_{\boldsymbol{k}}$.

Having stated some important properties of the function spaces $\boldsymbol{V}_{\boldsymbol{k}}$ and the sesquilinear forms $a_{\boldsymbol{k}}(\rho)$, we are finally in a position to state the weak formulation of Problem 4.1.

**Problem 4.16.** *Given a coefficient $\rho \in \mathcal{D}$, and a vector $\boldsymbol{k} \in \mathbb{B}$, find eigenvalues $\lambda \in \mathbb{C}$ and corresponding eigenfunctions $\boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}} \setminus \{\boldsymbol{0}\}$, such that*

$$a_{\boldsymbol{k}}(\rho)(\boldsymbol{u}, \boldsymbol{v}) = \lambda\, m(\boldsymbol{u}, \boldsymbol{v}) \qquad \textit{for all } \boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}. \tag{4.42}$$

Note that the function spaces $\boldsymbol{V}_{\boldsymbol{k}}$ are sufficient as test spaces for the weak eigenvalue problem stated in Problem 4.16. Suppose that (4.42) holds for some coefficient $\rho \in \mathcal{D}$, some vector $\boldsymbol{k} \in \mathbb{B}$, some vector field $\boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}}$, and some complex number $\lambda \in \mathbb{C}$. Choosing an arbitrary test function $\boldsymbol{w} \in \boldsymbol{W}$, we have by Theorem 4.8 and Proposition 4.11 that there exist a vector field $\boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}$ and a function $q \in Q$, such that

$$\boldsymbol{w} = \boldsymbol{v} + (\nabla + \mathrm{i}\boldsymbol{k})q.$$

By Lemma 4.6 we than have that

$$a_{\boldsymbol{k}}(\rho)\big(\boldsymbol{u}, (\nabla + \mathrm{i}\boldsymbol{k})q\big) = 0.$$

and hence

$$a_{\boldsymbol{k}}(\rho)\big(\boldsymbol{u}, \boldsymbol{w}\big) = a_{\boldsymbol{k}}(\rho)\big(\boldsymbol{u}, \boldsymbol{v}\big).$$

By definition of the sequilinear forms $m$ and $b_{\boldsymbol{k}}$, we also have that

$$m\big(\boldsymbol{u}, (\nabla + \mathrm{i}\boldsymbol{k})q\big) = b_{\boldsymbol{k}}(\boldsymbol{u}, q) = 0,$$

since $\boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}}$. The latter identity implies that

$$m\big(\boldsymbol{u}, \boldsymbol{w}\big) = m\big(\boldsymbol{u}, \boldsymbol{v}\big).$$

It thus follows that (4.42) holds for some coefficient $\rho \in \mathcal{D}$, some vector $\boldsymbol{k} \in \mathbb{B}$, some vector field $\boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}}$, and some complex number $\lambda \in \mathbb{C}$, if and only if

$$a_{\boldsymbol{k}}(\rho)(\boldsymbol{u}, \boldsymbol{w}) = \lambda\, m(\boldsymbol{u}, \boldsymbol{w}) \qquad \text{for all } \boldsymbol{w} \in \boldsymbol{W}.$$

It is therefore sufficient to choose $\boldsymbol{V}_{\boldsymbol{k}}$ as the test space in Problem 4.16.

## 4.4 Riesz–Schauder Theory

In this section we review the spectral theory for the weakly formulated family of eigenvalue problems stated in Problem 4.16. The theory we present here is standard and can be found in many textbooks (see e.g. Chapter 17 in [78]). We still discuss this theory here in detail, since some aspects are non-trivial for our particular family of eigenvalue problems.

In essence, we aim at applying the Spectral Theorem of Riesz–Schauder for compact, self-adjoint operators on infinite-dimensional Hilbert spaces. Our first step, therefore, is to devise such an operator. To this end, we define for every coefficient $\rho \in D$, and every vector $\boldsymbol{k} \in \mathbb{B}$ the conjugate-linear operators $A_{\boldsymbol{k}}(\rho) : \boldsymbol{V}_{\boldsymbol{k}} \to \boldsymbol{V}_{\boldsymbol{k}}^*$, $M_{\boldsymbol{k}} : \boldsymbol{Z} \to \boldsymbol{V}_{\boldsymbol{k}}^*$, and $I_{\boldsymbol{k}} : \boldsymbol{V}_{\boldsymbol{k}} \to \boldsymbol{Z}$ by

$$A_{\boldsymbol{k}}(\rho)\boldsymbol{u}[\boldsymbol{v}] := a_{\boldsymbol{k}}(\rho)(\boldsymbol{u}, \boldsymbol{v}) \qquad \text{for all } \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}, \tag{4.43}$$

$$M_{\boldsymbol{k}}\boldsymbol{f}[\boldsymbol{v}] := m(\boldsymbol{f}, \boldsymbol{v}) \qquad \text{for all } \boldsymbol{f} \in \boldsymbol{Z},\ \boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}, \tag{4.44}$$

$$I_{\boldsymbol{k}}\boldsymbol{v} := \boldsymbol{v} \qquad \text{for all } \boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}. \tag{4.45}$$

Clearly, $I_{\boldsymbol{k}}$ is the identical embedding from $\boldsymbol{V}_{\boldsymbol{k}}$ into $\boldsymbol{Z}$ for every $\boldsymbol{k} \in \mathbb{B}$. Note that the operator $M_{\boldsymbol{k}}$ is continuous for each $\boldsymbol{k} \in \mathbb{B}$ with

$$\|M_{\boldsymbol{k}}\|_{\boldsymbol{Z}, \boldsymbol{V}_{\boldsymbol{k}}^*} \leq 1. \tag{4.46}$$

Moreover, one easily verifies that

$$\ker(M_{\boldsymbol{k}}) = (\nabla + \mathrm{i}\boldsymbol{k})Q \qquad \text{for all } \boldsymbol{k} \in \mathbb{B}. \tag{4.47}$$

This implies, in particular, that the operators $M_{\boldsymbol{k}}$ are not injective. Restricting these operators to $\boldsymbol{V}_{\boldsymbol{k}}$, however, yields injective operators for all $\boldsymbol{k} \in \mathbb{B}$, since $(\nabla + \mathrm{i}\boldsymbol{k})Q$ is $\boldsymbol{Z}$-orthogonal to $\boldsymbol{V}_{\boldsymbol{k}}$ according to Corollary 4.9 and Proposition 4.11.

With the above definitions, the weakly formulated eigenvalue problem stated in Problem 4.16 is equivalent to

**Problem 4.17.** *Given a coefficient $\rho \in D$, and a vector $\boldsymbol{k} \in \mathbb{B}$, find eigenvalues $\lambda \in \mathbb{C}$ and corresponding eigenfunctions $\boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}} \setminus \{\boldsymbol{0}\}$, such that*

$$A_{\boldsymbol{k}}(\rho)\boldsymbol{u} = \lambda\, M_{\boldsymbol{k}}\boldsymbol{u}.$$

Next, we define for every coefficient $\rho \in D$, and every vector $\boldsymbol{k} \in \mathbb{B}$ the conjugate-linear operator $L_{\boldsymbol{k}}(\rho) : \boldsymbol{V}_{\boldsymbol{k}} \to \boldsymbol{V}_{\boldsymbol{k}}^*$ by

$$L_{\boldsymbol{k}}(\rho) := A_{\boldsymbol{k}}(\rho) + \kappa(\rho, \boldsymbol{k})\, M_{\boldsymbol{k}}|_{\boldsymbol{V}_{\boldsymbol{k}}}, \tag{4.48}$$

where $\kappa(\rho, \boldsymbol{k}) \in \mathbb{R}_{>0}$ is the constant defined in Proposition 4.14. It follows that every operator $L_{\boldsymbol{k}}(\rho)$ induces a conjugate-symmetric, bounded, and $\boldsymbol{V}_{\boldsymbol{k}}$-elliptic sesquilinear form $l_{\boldsymbol{k}}(\rho) : \boldsymbol{V}_{\boldsymbol{k}} \times \boldsymbol{V}_{\boldsymbol{k}} \to \mathbb{C}$ by

$$l_{\boldsymbol{k}}(\rho)(\boldsymbol{u}, \boldsymbol{v}) := L_{\boldsymbol{k}}(\rho)\boldsymbol{u}[\boldsymbol{v}] \qquad \text{for all } \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}. \tag{4.49}$$

More precisely, each sesquilinear form $l_{\boldsymbol{k}}(\rho)$ satisfies

$$\begin{aligned} l_{\boldsymbol{k}}(\rho)(\boldsymbol{u}, \boldsymbol{v}) &= \overline{l_{\boldsymbol{k}}(\rho)(\boldsymbol{v}, \boldsymbol{u})} && \text{for all } \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}, \\ \left| l_{\boldsymbol{k}}(\rho)(\boldsymbol{u}, \boldsymbol{v}) \right| &\leq \beta(\rho, \boldsymbol{k}) \|\boldsymbol{u}\|_{\mathrm{curl},\Omega} \, \|\boldsymbol{v}\|_{\mathrm{curl},\Omega} && \text{for all } \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}, \\ l_{\boldsymbol{k}}(\rho)(\boldsymbol{u}, \boldsymbol{u}) &\geq \alpha(\rho, \boldsymbol{k}) \|\boldsymbol{u}\|^2_{\mathrm{curl},\Omega} && \text{for all } \boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}}, \end{aligned}$$

where $\alpha(\rho, \boldsymbol{k})$ and $\beta(\rho, \boldsymbol{k})$ were defined in Proposition 4.14. The Lemma of Lax–Milgram then implies that every operator $L_{\boldsymbol{k}}(\rho)$ has a conjugate-linear, continuous inverse $L_{\boldsymbol{k}}(\rho)^{-1} : \boldsymbol{V}_{\boldsymbol{k}}^* \to \boldsymbol{V}_{\boldsymbol{k}}$ with

$$\|L_{\boldsymbol{k}}(\rho)^{-1}\|_{\boldsymbol{V}_{\boldsymbol{k}}^*, \boldsymbol{V}_{\boldsymbol{k}}} \leq \frac{1}{\alpha(\rho, \boldsymbol{k})} \qquad \text{for all } \rho \in \mathcal{D}, \ \boldsymbol{k} \in \mathbb{B}. \tag{4.50}$$

Therefore, we are able to define for every coefficient $\rho \in \mathcal{D}$, and every vector $\boldsymbol{k} \in \mathbb{B}$ the linear operator $G_{\boldsymbol{k}}(\rho) : \boldsymbol{Z} \to \boldsymbol{Z}$ by

$$G_{\boldsymbol{k}}(\rho) := I_{\boldsymbol{k}} L_{\boldsymbol{k}}(\rho)^{-1} M_{\boldsymbol{k}} \qquad \text{for all } \boldsymbol{k} \in \mathbb{B}, \tag{4.51}$$

In the literature, the operator $G_{\boldsymbol{k}}(\rho)$ is sometimes referred to as the *Green solution operator* for $L_{\boldsymbol{k}}(\rho)$ (see e.g. Definition 17.5 in [78]). The following theorem states that the identical embeddings $I_{\boldsymbol{k}}$ from $\boldsymbol{V}_{\boldsymbol{k}}$ into $\boldsymbol{Z}$ are compact operators.

**Theorem 4.18.** *For every vector $\boldsymbol{k} \in \mathbb{B}$ the complex Hilbert space $\boldsymbol{V}_{\boldsymbol{k}}$ is compactly embedded in $\boldsymbol{Z}$.*

*Proof.* The following proof is an adaptation of the proof of Theorem 2.1 in [74]. Let $\{\boldsymbol{u}^{(n)}\}_{n \in \mathbb{N}}$ be a sequence in $\boldsymbol{V}_{\boldsymbol{k}}$, which is bounded in $\boldsymbol{V}_{\boldsymbol{k}}$ by a constant $c_1 > 0$. By Theorem 4.8 and Proposition 4.11 there exists a sequence $\{\boldsymbol{\psi}^{(n)}\}_{n \in \mathbb{N}}$ in $H^1_{\mathrm{per}}(\Omega)^3$, a sequence $\{\boldsymbol{z}_n\}_{n \in \mathbb{N}}$ in $\mathbb{C}^3$, and a constant $c > 0$, such that

$$\begin{aligned} (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{\psi}^{(n)} + \boldsymbol{z}_n &= \boldsymbol{u}^{(n)} && \text{for all } n \in \mathbb{N}, \\ \|\boldsymbol{\psi}^{(n)}\|_{1,\Omega} + \sqrt{\mathrm{meas}(\Omega)} |\boldsymbol{z}_n| &\leq c \|\boldsymbol{u}^{(n)}\|_{\Omega} && \text{for all } n \in \mathbb{N}. \end{aligned}$$

Since $\|\boldsymbol{u}^{(n)}\|_{\Omega} \leq \|\boldsymbol{u}^{(n)}\|_{\mathrm{curl},\Omega} \leq c_1$ for all $n \in \mathbb{N}$, the sequences $\{\boldsymbol{\psi}^{(n)}\}_{n \in \mathbb{N}}$ and $\{\boldsymbol{z}_n\}_{n \in \mathbb{N}}$ are bounded in $H^1(\Omega)^3$ and $\mathbb{C}^3$, respectively. According to the Theorem of Rellich–Kondrachov (see e.g. Theorem 6.3 in [1]) the identical embedding from $H^1(\Omega)^3$ into $\boldsymbol{Z}$ is compact. Therefore, there exists a subsequence $\{\boldsymbol{\psi}^{(1,n)}\}_{n \in \mathbb{N}}$

of $\{\boldsymbol{\psi}^{(n)}\}_{n\in\mathbb{N}}$, which converges in $\boldsymbol{Z}$. According to the Theorem of Heine–Borel there also exists a subsequence $\{\boldsymbol{z}_{(2,n)}\}_{n\in\mathbb{N}}$ of $\{\boldsymbol{z}_{(1,n)}\}_{n\in\mathbb{N}}$, which converges in $\mathbb{C}^3$. We now consider the corresponding subsequence $\{\boldsymbol{u}^{(2,n)}\}_{n\in\mathbb{N}}$ of $\{\boldsymbol{u}^{(n)}\}_{n\in\mathbb{N}}$. To simplify notations, we define $\boldsymbol{u}^{[m,n]} := \boldsymbol{u}^{(2,m)} - \boldsymbol{u}^{(2,n)}$, $\boldsymbol{\psi}^{[m,n]} := \boldsymbol{\psi}^{(2,m)} - \boldsymbol{\psi}^{(2,n)}$, and $\boldsymbol{z}^{[m,n]} := \boldsymbol{z}_{(2,m)} - \boldsymbol{z}_{(2,n)}$ for all $m, n \in \mathbb{N}$. By Proposition 4.3 and Theorem 4.5, we obtain

$$
\begin{aligned}
&\|\boldsymbol{u}^{(2,m)} - \boldsymbol{u}^{(2,n)}\|_\Omega^2 \\
&\quad= \int_\Omega \overline{\boldsymbol{u}^{[m,n]}} \cdot \boldsymbol{u}^{[m,n]} \\
&\quad= \int_\Omega \overline{\boldsymbol{u}^{[m,n]}} \cdot \big((\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{\psi}^{[m,n]} + \boldsymbol{z}^{[m,n]}\big) \\
&\quad= \int_\Omega \overline{(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}^{[m,n]}} \cdot \boldsymbol{\psi}^{[m,n]} + \int_\Omega \overline{\boldsymbol{u}^{[m,n]}} \cdot \boldsymbol{z}^{[m,n]} \\
&\quad\le \|(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}^{[m,n]}\|_\Omega \|\boldsymbol{\psi}^{[m,n]}\|_\Omega + \sqrt{\mathrm{meas}(\Omega)}\|\boldsymbol{u}^{[m,n]}\|_\Omega |\boldsymbol{z}^{[m,n]}| \\
&\quad\le \|\boldsymbol{u}^{[m,n]}\|_{\mathrm{curl},\Omega}\big(\beta_0(\boldsymbol{k})\|\boldsymbol{\psi}^{[m,n]}\|_\Omega + \sqrt{\mathrm{meas}(\Omega)}|\boldsymbol{z}^{[m,n]}|\big) \\
&\quad\le 2c_1\big(\beta_0(\boldsymbol{k})\|\boldsymbol{\psi}^{[m,n]}\|_\Omega + \sqrt{\mathrm{meas}(\Omega)}|\boldsymbol{z}^{[m,n]}|\big) \\
&\quad\le c_2\big(\|\boldsymbol{\psi}^{(2,m)} - \boldsymbol{\psi}^{(2,n)}\|_\Omega + |\boldsymbol{z}_{(2,m)} - \boldsymbol{z}_{(2,n)}|\big) \to 0 \quad \text{as } m, n \to \infty,
\end{aligned}
$$

where $c_2 := 2c_1 \max\big\{\beta_0(\boldsymbol{k}), \sqrt{\mathrm{meas}(\Omega)}\big\}$. Hence, $\{\boldsymbol{u}^{(2,n)}\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $\boldsymbol{Z}$ and therefore convergent in $\boldsymbol{Z}$. □

The following proposition establishes important properties of the operators

**Proposition 4.19.** *The following assertions hold for every coefficient $\rho \in D$ and every vector $\boldsymbol{k} \in \mathbb{B}$.*

(a) *The operator $G_{\boldsymbol{k}}(\rho)$ is compact, self-adjoint and positive semidefinite.*

(b) $\ker(G_{\boldsymbol{k}}(\rho)) = \boldsymbol{V}_{\boldsymbol{k}}^\perp$, *where $\boldsymbol{V}_{\boldsymbol{k}}^\perp$ denotes the $\boldsymbol{Z}$-orthogonal complement of $\boldsymbol{V}_{\boldsymbol{k}}$.*

*Proof.* In the following we use the fact that $\langle \boldsymbol{f}\,, \boldsymbol{v}\rangle_\Omega = M_{\boldsymbol{k}}\boldsymbol{f}[\boldsymbol{v}]$ for all $\boldsymbol{f} \in \boldsymbol{Z}$ and all $\boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}$. Furthermore, we use that fact that $I_{\boldsymbol{k}}$ is the identity on $\boldsymbol{V}_{\boldsymbol{k}}$.
(a) Compactness directly follows from the definition of $G_{\boldsymbol{k}}(\rho)$. In order to prove self-adjoinedness and positive semi-definiteness, we choose two functions $\boldsymbol{f}, \boldsymbol{g} \in \boldsymbol{Z}$ arbitrarily. Then, there exist uniquely determined functions $\boldsymbol{u}, \boldsymbol{w} \in \boldsymbol{V}_{\boldsymbol{k}}$, such that $L_{\boldsymbol{k}}(\rho)\boldsymbol{u} = M_{\boldsymbol{k}}\boldsymbol{f}$ and $L_{\boldsymbol{k}}(\rho)\boldsymbol{w} = M_{\boldsymbol{k}}\boldsymbol{g}$. We hence obtain

$$\langle G_{\boldsymbol{k}}(\rho)\boldsymbol{f}\,, \boldsymbol{g}\rangle_\Omega = \langle I_{\boldsymbol{k}} L_{\boldsymbol{k}}(\rho)^{-1} M_{\boldsymbol{k}}\boldsymbol{f}\,, \boldsymbol{g}\rangle_\Omega = \langle L_{\boldsymbol{k}}(\rho)^{-1} M_{\boldsymbol{k}}\boldsymbol{f}\,, \boldsymbol{g}\rangle_\Omega = \langle \boldsymbol{u}\,, \boldsymbol{g}\rangle_\Omega$$

$$
\begin{aligned}
&= \overline{\langle \boldsymbol{g}\,,\,\boldsymbol{u}\rangle_\Omega} = \overline{M_{\boldsymbol{k}}\boldsymbol{g}[\boldsymbol{u}]} = \overline{L_{\boldsymbol{k}}(\rho)\boldsymbol{w}[\boldsymbol{u}]} = L_{\boldsymbol{k}}(\rho)\boldsymbol{u}[\boldsymbol{w}] = M_{\boldsymbol{k}}\boldsymbol{f}[\boldsymbol{w}] \\
&= \langle \boldsymbol{f}\,,\,\boldsymbol{w}\rangle_\Omega = \langle \boldsymbol{f}\,,\,L_{\boldsymbol{k}}(\rho)^{-1}M_{\boldsymbol{k}}\boldsymbol{g}\rangle_\Omega = \langle \boldsymbol{f}\,,\,I_{\boldsymbol{k}}L_{\boldsymbol{k}}(\rho)^{-1}M_{\boldsymbol{k}}\boldsymbol{g}\rangle_\Omega \\
&= \langle \boldsymbol{f}\,,\,G_{\boldsymbol{k}}(\rho)\boldsymbol{g}\rangle_\Omega,
\end{aligned}
$$

which proves the self-adjoinedness of $G_{\boldsymbol{k}}(\rho)$. Similarly, we deduce the inequality

$$\langle G_{\boldsymbol{k}}(\rho)\boldsymbol{f}\,,\,\boldsymbol{f}\rangle_\Omega = L_{\boldsymbol{k}}(\rho)\boldsymbol{u}[\boldsymbol{u}] \geq \alpha(\rho,\boldsymbol{k})\|\boldsymbol{u}\|_{\mathrm{curl},\Omega} \geq 0,$$

which proves the positive semidefiniteness of $G_{\boldsymbol{k}}(\rho)$.
(b) For every $\boldsymbol{f} \in \boldsymbol{Z}$, and every $\boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}$ we have the identity

$$
\begin{aligned}
\langle \boldsymbol{f}\,,\,\boldsymbol{v}\rangle_\Omega &= M_{\boldsymbol{k}}\boldsymbol{f}[\boldsymbol{v}] = (L_{\boldsymbol{k}}(\rho)L_{\boldsymbol{k}}(\rho)^{-1}M_{\boldsymbol{k}}\boldsymbol{f})[\boldsymbol{v}] = L_{\boldsymbol{k}}(\rho)(L_{\boldsymbol{k}}(\rho)^{-1}M_{\boldsymbol{k}}\boldsymbol{f})[\boldsymbol{v}] \\
&= L_{\boldsymbol{k}}(\rho)(I_{\boldsymbol{k}}L_{\boldsymbol{k}}(\rho)^{-1}M_{\boldsymbol{k}}\boldsymbol{f})[\boldsymbol{v}] = L_{\boldsymbol{k}}(\rho)(G_{\boldsymbol{k}}(\rho)\boldsymbol{f})[\boldsymbol{v}].
\end{aligned}
$$

Since $L_{\boldsymbol{k}}(\rho)$ is a bijective operator, $\langle \boldsymbol{f}\,,\,\boldsymbol{v}\rangle_\Omega = 0$ holds for all $\boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}$, if and only if $G_{\boldsymbol{k}}(\rho)\boldsymbol{f} = 0$. Hence, $\ker(G_{\boldsymbol{k}}(\rho)) = \boldsymbol{V}_{\boldsymbol{k}}^{\perp}$. □

Since the operators $G_{\boldsymbol{k}}(\rho)$ are compact, they are subject to the Spectral Theorem of Riesz–Schauder (see e.g. Section 12.1 in [78]). Hence, the following statements hold for every coefficient $\rho \in \mathcal{D}$, and every vector $\boldsymbol{k} \in \mathbb{B}$. First, we have the orthogonal decomposition

$$\boldsymbol{Z} = \ker(G_{\boldsymbol{k}}(\rho)) \oplus \overline{\mathrm{im}(G_{\boldsymbol{k}}(\rho))}.$$

By Part (b) of Proposition 4.19 we hence have

$$\boldsymbol{V}_{\boldsymbol{k}} = \overline{\mathrm{im}(G_{\boldsymbol{k}}(\rho))}. \tag{4.52}$$

Furthermore, since the operator $G_{\boldsymbol{k}}(\rho)$ is self-adjoined and positive semidefinite, we have that there exists a sequence $\{\mu_j(\rho,\boldsymbol{k})\}_{j\in\mathbb{N}}$, which consists of positive, real eigenvalues of $G_{\boldsymbol{k}}(\rho)$, and which converges to zero. In the following we shall assume that the eigenvalues of $G_{\boldsymbol{k}}(\rho)$ are enumerated, such that

$$\mu_1(\rho,\boldsymbol{k}) \geq \mu_2(\rho,\boldsymbol{k}) \geq \mu_3(\rho,\boldsymbol{k}) \geq \cdots$$

The sequence of corresponding eigenfunctions $\{\boldsymbol{u}_j(\rho,\boldsymbol{k})\}_{j\in\mathbb{N}}$ forms a complete $\boldsymbol{Z}$-orthogonal system of $\boldsymbol{V}_{\boldsymbol{k}}$. Here, as in the following, we denote by $\boldsymbol{u}_j(\rho,\boldsymbol{k})$ the eigenfunction that corresponds to the $j$-th largest eigenvalue of $G_{\boldsymbol{k}}(\rho)$, i.e.,

$$G_{\boldsymbol{k}}(\rho)\boldsymbol{u}_j(\rho,\boldsymbol{k}) = \mu_j(\rho,\boldsymbol{k})\,\boldsymbol{u}_j(\rho,\boldsymbol{k}). \tag{4.53}$$

The eigenspace of each eigenvalue is finite-dimensional. The spectrum of the operator $G_{\boldsymbol{k}}(\rho)$ consists precisely of the elements of the sequence $\{\mu_j(\rho,\boldsymbol{k})\}_{j\in\mathbb{N}}$ and

zero, which is also an eigenvalue of $G_{\boldsymbol{k}}(\rho)$. Finally, we have that the operator $G_{\boldsymbol{k}}(\rho)$ acts on $\boldsymbol{Z}$ as

$$G_{\boldsymbol{k}}(\rho)\boldsymbol{f} := \sum_{j=1}^{\infty} \mu_j(\rho,\boldsymbol{k}) \, \frac{\langle \boldsymbol{u}_j(\rho,\boldsymbol{k}) \, , \boldsymbol{f} \rangle_\Omega}{\|\boldsymbol{u}_j(\rho,\boldsymbol{k})\|_\Omega^2} \, \boldsymbol{u}_j(\rho,\boldsymbol{k}) \qquad \text{for all } \boldsymbol{f} \in \boldsymbol{Z}.$$

The $j$-th largest eigenvalue can be characterized according to the *Min-Max Principle* as

$$\mu_j(\rho,\boldsymbol{k}) = \max_{\substack{\boldsymbol{U} \sqsubseteq \boldsymbol{Z} \\ \dim \boldsymbol{U} = j}} \min_{\boldsymbol{f} \in \boldsymbol{U} \setminus \{\boldsymbol{0}\}} \frac{\langle G_{\boldsymbol{k}}(\rho)\boldsymbol{f} \, , \boldsymbol{f} \rangle_\Omega}{\|\boldsymbol{f}\|_\Omega^2} \qquad \text{for all } j \in \mathbb{N}, \tag{4.54}$$

where $\sqsubseteq$ denotes the linear subspace relation (see Theorem XIII.1 in [64]). For convenience, we define for every coefficient $\rho \in \mathcal{D}$, every vector $\boldsymbol{k} \in \mathbb{B}$, and every index $j \in \mathbb{N}$ the linear function space

$$\boldsymbol{V}_{\boldsymbol{k},j}(\rho) := \operatorname{span}\big\{ \boldsymbol{u}_1(\rho,\boldsymbol{k}), \ldots, \boldsymbol{u}_j(\rho,\boldsymbol{k}) \big\}. \tag{4.55}$$

Then, one can show that the maxima and minima in (4.54) are attained precisely for $\boldsymbol{U} = \boldsymbol{V}_{\boldsymbol{k},j}(\rho)$. An equivalent characterization of the $j$-th largest eigenvalue is given by *Rayleigh's principle*, which reads

$$\mu_j(\rho,\boldsymbol{k}) = \max_{\substack{\boldsymbol{f} \in \boldsymbol{Z} \setminus \{\boldsymbol{0}\} \\ P_{\boldsymbol{k},j-1}(\rho)\boldsymbol{f} = \boldsymbol{0}}} \frac{\langle G_{\boldsymbol{k}}(\rho)\boldsymbol{f} \, , \boldsymbol{f} \rangle_\Omega}{\|\boldsymbol{f}\|_\Omega^2}, \tag{4.56}$$

Here, $P_{\boldsymbol{k},j}(\rho)$ is the linear, $\boldsymbol{Z}$-orthogonal projection operator from $\boldsymbol{Z}$ onto the linear function space $\boldsymbol{V}_{\boldsymbol{k},j}(\rho)$. for every coefficient $\rho \in \mathcal{D}$, every vector $\boldsymbol{k} \in \mathbb{B}$, and every index $j \in \mathbb{N}$ the operator $P_{\boldsymbol{k},j}(\rho) : \boldsymbol{Z} \to \boldsymbol{V}_{\boldsymbol{k}}$ is given by

$$P_{\boldsymbol{k},j}(\rho)\boldsymbol{f} := \sum_{i=1}^{j} \frac{\langle \boldsymbol{u}_i(\rho,\boldsymbol{k}) \, , \boldsymbol{f} \rangle_\Omega}{\|\boldsymbol{u}_i(\rho,\boldsymbol{k})\|_\Omega^2} \, \boldsymbol{u}_i(\rho,\boldsymbol{k}) \qquad \text{for all } \boldsymbol{f} \in \boldsymbol{Z}. \tag{4.57}$$

Now, let $\mu_j(\rho,\boldsymbol{k}) > 0$ be a positive eigenvalue of $G_{\boldsymbol{k}}(\rho)$, and let $\boldsymbol{u}_j(\rho,\boldsymbol{k})$ be a corresponding eigenfunction, i.e.,

$$G_{\boldsymbol{k}}(\rho) \, \boldsymbol{u}_j(\rho,\boldsymbol{k}) = \mu_j(\rho,\boldsymbol{k}) \, \boldsymbol{u}_j(\rho,\boldsymbol{k}). \tag{4.58}$$

According to (4.48) and (4.51) we can rewrite (4.58) equivalently as

$$\left( \frac{1}{\mu_j(\rho,\boldsymbol{k})} - \kappa(\rho,\boldsymbol{k}) \right) M_{\boldsymbol{k}} \boldsymbol{u}_j(\rho,\boldsymbol{k}) = A_{\boldsymbol{k}}(\rho) \boldsymbol{u}_j(\rho,\boldsymbol{k}). \tag{4.59}$$

Since the sequence of eigenfunctions $\{\boldsymbol{u}_j(\rho,\boldsymbol{k})\}_{j\in\mathbb{N}}$ of the operator $G_{\boldsymbol{k}}(\rho)$ form a complete orthogonal system of $\boldsymbol{V}_{\boldsymbol{k}}$, it follows that each eigenpair of $G_{\boldsymbol{k}}(\rho)$ corresponds to a solution of the eigenvalue problem stated in Problem 4.17 and thus to a solution of the weak eigenvalue problem stated in Problem 4.16. In particular, the elements of the sequence $\{\lambda_j(\rho,\boldsymbol{k})\}_{j\in\mathbb{N}}$, defined by

$$\lambda_j(\rho,\boldsymbol{k}) := \frac{1}{\mu_j(\rho,\boldsymbol{k})} - \kappa(\rho,\boldsymbol{k}) \qquad \text{for all } j\in\mathbb{N}, \tag{4.60}$$

are the eigenvalues. In summary, we have the following proposition.

**Proposition 4.20.** *For every coefficient $\rho\in\mathcal{D}$ and every vector $\boldsymbol{k}\in\mathbb{B}$ there exists an increasing sequence $\{\lambda_j(\rho,\boldsymbol{k})\}_{j\in\mathbb{N}}$ of non-negative, real numbers tending to $\infty$, which consists of the eigenvalues of the weakly formulated eigenvalue problem (4.42). The sequence of corresponding eigenfunctions $\{\boldsymbol{u}_j(\rho,\boldsymbol{k})\}_{j\in\mathbb{N}}$ forms a complete, $\boldsymbol{Z}$-orthogonal system of $\boldsymbol{V}_{\boldsymbol{k}}$. The eigenspace of each eigenvalue is finite-dimensional. The $j$-th smallest eigenvalue can be characterized as*

$$\lambda_j(\rho,\boldsymbol{k}) = \min_{\substack{\boldsymbol{U}\sqsubseteq\boldsymbol{V}_{\boldsymbol{k}}\\ \dim\boldsymbol{U}=j}} \max_{\boldsymbol{u}\in\boldsymbol{U}\setminus\{\boldsymbol{0}\}} \frac{a_{\boldsymbol{k}}(\rho)(\boldsymbol{u},\boldsymbol{u})}{m(\boldsymbol{u},\boldsymbol{u})} \qquad \textit{for all } j\in\mathbb{N}. \tag{4.61}$$

*or equivalently as*

$$\lambda_j(\rho,\boldsymbol{k}) = \min_{\substack{\boldsymbol{u}\in\boldsymbol{V}_{\boldsymbol{k}}\\ P_{\boldsymbol{k},j-1}(\rho)\boldsymbol{u}=\boldsymbol{0}}} \frac{a_{\boldsymbol{k}}(\rho)(\boldsymbol{u},\boldsymbol{u})}{m(\boldsymbol{u},\boldsymbol{u})} \qquad \textit{for all } j\in\mathbb{N}. \tag{4.62}$$

*Zero is an eigenvalue, if and only if $\boldsymbol{k}=0$.*

*Proof.* The existence of the sequences $\{\lambda_j(\rho,\boldsymbol{k})\}_{j\in\mathbb{N}}$ and $\{\boldsymbol{u}_j(\rho,\boldsymbol{k})\}_{j\in\mathbb{N}}$ follows from the equivalence of (4.58) and (4.59). Notice also, that every eigenvalue $\lambda_j(\rho,\boldsymbol{k})$ is an eigenvalue of the weakly formulated eigenvalue problem (4.42).

In order to show (4.61), we define for every $\rho\in\mathcal{D}$ and every $\boldsymbol{k}\in\mathbb{B}$ the operators $F_{\boldsymbol{k}}(\rho):\boldsymbol{V}_{\boldsymbol{k}}\to\boldsymbol{V}_{\boldsymbol{k}}$ by $F_{\boldsymbol{k}} := G_{\boldsymbol{k}}|_{\boldsymbol{V}_{\boldsymbol{k}}}$, as well as the operator $F_{\boldsymbol{k}}^{1/2}(\rho):\boldsymbol{V}_{\boldsymbol{k}}\to\boldsymbol{V}_{\boldsymbol{k}}$ by

$$F_{\boldsymbol{k}}^{1/2}(\rho)\,\boldsymbol{v} := \sum_{j=1}^{\infty} \sqrt{\mu_j(\rho,\boldsymbol{k})}\,\frac{\langle \boldsymbol{u}_i(\rho,\boldsymbol{k})\,,\,\boldsymbol{v}\rangle_\Omega}{\|\boldsymbol{u}_i(\rho,\boldsymbol{k})\|_\Omega^2}\,\boldsymbol{u}_i(\rho,\boldsymbol{k}) \qquad \text{for all } \boldsymbol{u}\in\boldsymbol{V}_{\boldsymbol{k}}.$$

One easily verifies that the continuous, linear operators $F_{\boldsymbol{k}}(\rho)$ and $F_{\boldsymbol{k}}^{1/2}(\rho)$ are symmetric with respect to the inner product $\langle\,\cdot\,,\,\cdot\,\rangle_\Omega$ and continuously invertible for all $\rho\in\mathcal{D}$ and all $\boldsymbol{k}\in\mathbb{B}$. The operators also satisfy $F_{\boldsymbol{k}}(\rho) = F_{\boldsymbol{k}}^{1/2}(\rho)F_{\boldsymbol{k}}^{1/2}(\rho)$ and $G_{\boldsymbol{k}}(\rho)F_{\boldsymbol{k}}(\rho)^{-1} = \mathrm{id}_{\boldsymbol{V}_{\boldsymbol{k}}}$, where $\mathrm{id}_{\boldsymbol{V}_{\boldsymbol{k}}}$ denotes the identity on $\boldsymbol{V}_{\boldsymbol{k}}$, for all $\rho\in\mathcal{D}$ and

all $\boldsymbol{k} \in \mathbb{B}$. Choosing an arbitrary function $\boldsymbol{u} \in \boldsymbol{V_k}$, we define $\boldsymbol{v} := F_{\boldsymbol{k}}(\rho)^{-1}\boldsymbol{u}$ and $\boldsymbol{w} := L_{\boldsymbol{k}}(\rho)^{-1}M_{\boldsymbol{k}}\boldsymbol{v}$ for some $\rho \in \mathcal{D}$ and some $\boldsymbol{k} \in \mathbb{B}$. We then find that

$$\boldsymbol{w} = L_{\boldsymbol{k}}(\rho)^{-1}M_{\boldsymbol{k}}F_{\boldsymbol{k}}(\rho)^{-1}\boldsymbol{u} = I_{\boldsymbol{k}}L_{\boldsymbol{k}}(\rho)^{-1}M_{\boldsymbol{k}}F_{\boldsymbol{k}}(\rho)^{-1}\boldsymbol{u} = G_{\boldsymbol{k}}(\rho)F_{\boldsymbol{k}}(\rho)^{-1}\boldsymbol{u} = \boldsymbol{u},$$

and hence,

$$\begin{aligned}\langle F_{\boldsymbol{k}}(\rho)^{-1}\boldsymbol{u}\,,\,\boldsymbol{u}\rangle_\Omega &= \langle \boldsymbol{v}\,,\,F_{\boldsymbol{k}}(\rho)\boldsymbol{v}\rangle_\Omega = \langle \boldsymbol{v}\,,\,G_{\boldsymbol{k}}(\rho)\boldsymbol{v}\rangle_\Omega = \langle \boldsymbol{v}\,,\,I_{\boldsymbol{k}}L_{\boldsymbol{k}}(\rho)^{-1}M_{\boldsymbol{k}}\boldsymbol{v}\rangle_\Omega \\ &= \langle \boldsymbol{v}\,,\,L_{\boldsymbol{k}}(\rho)^{-1}M_{\boldsymbol{k}}\boldsymbol{v}\rangle_\Omega = \langle \boldsymbol{v}\,,\,\boldsymbol{w}\rangle_\Omega = M_{\boldsymbol{k}}\boldsymbol{v}[\boldsymbol{w}] = L_{\boldsymbol{k}}(\rho)\boldsymbol{w}[\boldsymbol{w}] \\ &= L_{\boldsymbol{k}}(\rho)\boldsymbol{u}[\boldsymbol{u}].\end{aligned}$$

Notice that this identity can be established for all functions $\boldsymbol{u} \in \boldsymbol{V_k}$. According to the Min-Max Principle (4.54), we have

$$\begin{aligned}\frac{1}{\mu_j(\rho,\boldsymbol{k})} &= \left(\max_{\substack{\boldsymbol{U}\sqsubseteq\boldsymbol{Z}\\ \dim\boldsymbol{U}=j}}\min_{\boldsymbol{f}\in\boldsymbol{U}\setminus\{\boldsymbol{0}\}}\frac{\langle G_{\boldsymbol{k}}(\rho)\boldsymbol{f}\,,\,\boldsymbol{f}\rangle_\Omega}{\|\boldsymbol{f}\|_\Omega^2}\right)^{-1} \\ &= \min_{\substack{\boldsymbol{U}\sqsubseteq\boldsymbol{Z}\\ \dim\boldsymbol{U}=j}}\max_{\boldsymbol{f}\in\boldsymbol{U}\setminus\{\boldsymbol{0}\}}\frac{\|\boldsymbol{f}\|_\Omega^2}{\langle G_{\boldsymbol{k}}(\rho)\boldsymbol{f}\,,\,\boldsymbol{f}\rangle_\Omega} \\ &= \min_{\substack{\boldsymbol{U}\sqsubseteq\boldsymbol{V_k}\\ \dim\boldsymbol{U}=j}}\max_{\boldsymbol{u}\in\boldsymbol{U}\setminus\{\boldsymbol{0}\}}\frac{\|\boldsymbol{u}\|_\Omega^2}{\langle F_{\boldsymbol{k}}(\rho)\boldsymbol{u}\,,\,\boldsymbol{u}\rangle_\Omega} \\ &= \min_{\substack{\boldsymbol{U}\sqsubseteq\boldsymbol{V_k}\\ \dim\boldsymbol{U}=j}}\max_{\boldsymbol{u}\in\boldsymbol{U}\setminus\{\boldsymbol{0}\}}\frac{\langle \boldsymbol{u}\,,\,\boldsymbol{u}\rangle_\Omega}{\langle F_{\boldsymbol{k}}^{1/2}(\rho)\boldsymbol{u}\,,\,F_{\boldsymbol{k}}^{1/2}(\rho)\boldsymbol{u}\rangle_\Omega} \\ &= \min_{\substack{\boldsymbol{U}\sqsubseteq\boldsymbol{V_k}\\ \dim\boldsymbol{U}=j}}\max_{\boldsymbol{u}\in\boldsymbol{U}\setminus\{\boldsymbol{0}\}}\frac{\langle F_{\boldsymbol{k}}^{1/2}(\rho)^{-1}\boldsymbol{u}\,,\,F_{\boldsymbol{k}}^{1/2}(\rho)^{-1}\boldsymbol{u}\rangle_\Omega}{\langle \boldsymbol{u}\,,\,\boldsymbol{u}\rangle_\Omega} \\ &= \min_{\substack{\boldsymbol{U}\sqsubseteq\boldsymbol{V_k}\\ \dim\boldsymbol{U}=j}}\max_{\boldsymbol{u}\in\boldsymbol{U}\setminus\{\boldsymbol{0}\}}\frac{\langle F_{\boldsymbol{k}}(\rho)^{-1}\boldsymbol{u}\,,\,\boldsymbol{u}\rangle_\Omega}{\langle \boldsymbol{u}\,,\,\boldsymbol{u}\rangle_\Omega} \\ &= \min_{\substack{\boldsymbol{U}\sqsubseteq\boldsymbol{V_k}\\ \dim\boldsymbol{U}=j}}\max_{\boldsymbol{u}\in\boldsymbol{U}\setminus\{\boldsymbol{0}\}}\frac{L_{\boldsymbol{k}}(\rho)\boldsymbol{u}[\boldsymbol{u}]}{M_{\boldsymbol{k}}\boldsymbol{u}[\boldsymbol{u}]} \\ &= \min_{\substack{\boldsymbol{U}\sqsubseteq\boldsymbol{V_k}\\ \dim\boldsymbol{U}=j}}\max_{\boldsymbol{u}\in\boldsymbol{U}\setminus\{\boldsymbol{0}\}}\frac{a_{\boldsymbol{k}}(\rho)(\boldsymbol{u},\boldsymbol{u})}{m(\boldsymbol{u},\boldsymbol{u})} + \kappa(\rho,\boldsymbol{k}) \qquad \text{for all } j \in \mathbb{N}.\end{aligned}$$

This establishes the characterization given in (4.61). The characterization given in (4.62) can be shown analogously using Rayleigh's principle (4.56).

Finally, we show that zero is an eigenvalue if and only if $\boldsymbol{k} \in \mathbb{B}$ vanishes. According to Proposition 4.15 the sesquilinear forms $a_{\boldsymbol{k}}(\rho)$ are $\boldsymbol{V_k}$-elliptic for all

$\rho \in \mathcal{D}$ and all $\boldsymbol{k} \in \mathbb{B} \setminus \{\mathbf{0}\}$. Therefore, zero cannot be an eigenvalue, if $\boldsymbol{k} \neq \mathbf{0}$. By Proposition 4.11, however, the three-dimensional space of constant functions from $\Omega$ into $\mathbb{C}^3$ is a linear subspace of $\boldsymbol{V_0}$. One easily verifies that a non-vanishing constant function $\boldsymbol{u} : \Omega \to \mathbb{C}^3$ satisfies $a_{\mathbf{0}}(\boldsymbol{u}, \boldsymbol{u}) = 0$. Hence, zero is an eigenvalue for $\boldsymbol{k} = \mathbf{0}$. □

Proposition 4.20 makes a conclusive statement about the existence of Bloch modes in photonic crystals. Recall that Problem 4.16 was motivated by the family of constrained eigenvalue problems (3.36), which arose from the Bloch ansatz for the magnetic field of a time-harmonic wave propagating inside a photonic crystal (see Section 4.1). The coefficient $\rho$ represents the restriction to $\Omega$ of the crystal's reciprocal relative electric permittivity function $1/\varepsilon_r$. The vector $\boldsymbol{k} \in \mathbb{B}$ coincides with the quasimomentum vector of a Bloch mode. Proposition 4.20 now states that in every three-dimensional photonic crystal there exists a countable set of Bloch modes for each quasimomentum vector $\boldsymbol{k}$. The frequency of each Bloch mode is determined by the corresponding eigenvalue $\lambda$ of (4.42) via $\omega = c_0\sqrt{\lambda}$. The Bloch modes for a given quasimomentum vector $\boldsymbol{k}$ form a complete $\boldsymbol{Z}$-orthogonal system.

## 4.5 Auchmuty's Principle

The identities (4.61) and (4.62) in Proposition 4.20 allow us to characterize the eigenvalues of Problem 4.16 in terms of the sesquilinear forms $a_{\boldsymbol{k}}(\rho)$, as defined by (4.33). The first identity is a variant of the Min-Max Principle, the second one is a variant of Rayleigh's principle. In this section we introduce yet another, however less well-known, characterization principle which goes back to works of Auchmuty (cf. [8], [9]). We present this principle here, because it will play an important role in proving the existence of optimal solutions for photonic band gap optimization problems.

In essence, Auchmuty's principle stems from a duality principle for certain types of optimization problems. Before we can make this notion precise, we need to introduce some fundamental concepts of convex analysis. The concepts we present here are discussed in detail in standard textbooks on convex analysis such as [35] or [65]. In the following we assume that $X$ and $Y$ are real Banach spaces, and that $X^*$ and $Y^*$ denote their normed duals. A convex functional from $X$ into the set $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ is called *proper* if it is not identical to $-\infty$ or $\infty$. Given a proper, lower semicontinuous, convex functional $f : X \to \overline{\mathbb{R}}$, the so-called *convex conjugate* functional $f^* : X^* \to \overline{\mathbb{R}}$ of $f$ is defined by

$$f^*(p) := \sup_{x \in X} \big(p(x) - f(x)\big) \qquad \text{for all } p \in X^*, \tag{4.63}$$

where $X^*$ denotes the normed dual of $X$. One can show that $f^*$ is also a proper, lower semicontinuous, convex functional.

Given two proper, lower semicontinuous, convex functionals $f : X \to \overline{\mathbb{R}}$ and $g : Y \to \overline{\mathbb{R}}$, we define the functional $F : X \to \overline{\mathbb{R}}$ by

$$F(x) := f(x) - g(Tx) \qquad \text{for all } x \in X,$$

where $T : X \to Y$ is a bijective, continuous linear operator. Notice that the functional $F$ is not convex in general. Our aim is to find the infimum

$$\inf_{x \in X} F(x). \tag{4.64}$$

One observes that $F$ can be characterized as

$$F(x) = \inf_{q \in Y^*} \mathcal{L}(x, q) \qquad \text{for all } x \in X, \tag{4.65}$$

where the functional $\mathcal{L} : X \times Y^* \to \overline{\mathbb{R}}$ is given by

$$\mathcal{L}(x, q) := f(x) + g^*(-q) + q(Tx) \qquad \text{for all } x \in X,\ q \in Y^*.$$

The functional $\mathcal{L}$ is called a *Lagrangian of type II* associated with $F$. Here, "of type II" means that the functional $F$ is determined by the infima of $\mathcal{L}$ with respect to the Lagrange parameter $q$ according to (4.65). In his work, Auchmuty shows that the following *duality principle* holds (see Theorem 3.3 in [8]),

$$\inf_{x \in X} \inf_{q \in Y^*} \mathcal{L}(x, q) = \inf_{q \in Y^*} \inf_{x \in X} \mathcal{L}(x, q). \tag{4.66}$$

One can show, that the functional $G : Y^* \to \overline{\mathbb{R}}$, given by

$$G(q) := \inf_{x \in X} \mathcal{L}(x, q) \qquad \text{for all } q \in Y^*,$$

is characterized as

$$G(q) = g^*(-q) - f^*(-T^*q) \qquad \text{for all } q \in Y^*.$$

Here, $T^*$ denotes the *dual operator* of $T$ in the sense of functional analysis. We hence find, that the infimum (4.64) coincides with

$$\inf_{q \in Y^*} G(q). \tag{4.67}$$

Moreover, the duality principle (4.66) can be rewritten as

$$\inf_{x \in X} \big(f(x) - g(Tx)\big) = \inf_{q \in Y^*} \big(g^*(-q) - f^*(-T^*q)\big). \tag{4.68}$$

We remark that similar duality principles are well-known for convex functionals $F : X \to \overline{\mathbb{R}}$, which are of the form $F = f + g \circ T$ (see e.g. Chapter 3, Section 4 in [35]).

An interesting application of the duality principle stated in (4.68) is the following characterization principle for the eigenvalues of the weakly formulated eigenvalue problem (4.42).

**Proposition 4.21.** *Under the assumptions of Proposition 4.20, we have*

$$\frac{-1}{2\big(\lambda_j(\rho,\boldsymbol{k})+\kappa(\rho,\boldsymbol{k})\big)} = \min_{\boldsymbol{u}\in\boldsymbol{V}_{\boldsymbol{k}}}\Bigg(\frac{1}{2}\left(a_{\boldsymbol{k}}(\rho)(\boldsymbol{u},\boldsymbol{u})+\kappa(\rho,\boldsymbol{k})m(\boldsymbol{u},\boldsymbol{u})\right) - \left\|\boldsymbol{u}-P_{\boldsymbol{k},j-1}(\rho)\boldsymbol{u}\right\|_{\Omega}\Bigg) \qquad \text{for all } j\in\mathbb{N}.$$

*Proof.* This proof is an adaptation of the line of arguments laid out in [9]. In essence, we aim at taking advantage of the duality principle stated in (4.68). To this end, we define the real Hilbert space $\boldsymbol{\mathcal{Z}} := L^2(\Omega,\mathbb{R})^3$. Furthermore, we define the isometric, linear isomorphism $\Phi : \boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{Z}} \to \boldsymbol{Z}$ by

$$\Phi(\boldsymbol{f}_1,\boldsymbol{f}_2) := \boldsymbol{f}_1 + \mathrm{i}\boldsymbol{f}_2 \qquad \text{for all } \boldsymbol{f}_1,\boldsymbol{f}_2\in\boldsymbol{\mathcal{Z}}.$$

Choosing an arbitrary coefficient $\rho \in \mathcal{D}$ and a vector $\boldsymbol{k} \in \mathbb{B}$, we define for every index $j \in \mathbb{N}$ the set

$$\mathcal{K}_j := \big\{(\boldsymbol{f}_1,\boldsymbol{f}_2)\in\boldsymbol{\mathcal{Z}}\times\boldsymbol{\mathcal{Z}} \;\big|\; P_{\boldsymbol{k},j-1}(\rho)\Phi(\boldsymbol{f}_1,\boldsymbol{f}_2)=\boldsymbol{0}, \|\Phi(\boldsymbol{f}_1,\boldsymbol{f}_2)\|_{\Omega}\le 1\big\}.$$

One observes that for every $j \in \mathbb{N}$ the set $\mathcal{K}_j$ is a closed, convex subset of $\boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{Z}}$. Next, we define for every index $j \in \mathbb{N}$ the functional $\psi_j : \boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{Z}} \to \overline{\mathbb{R}}$ as well as the functional $\sigma : \boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{Z}} \to \mathbb{R}$ by

$$\psi_j(\boldsymbol{f}_1,\boldsymbol{f}_2) := \begin{cases} 0 & \text{if } (\boldsymbol{f}_1,\boldsymbol{f}_2)\in\mathcal{K}_j,\\ \infty & \text{else}\end{cases} \qquad \text{for all } \boldsymbol{f}_1,\boldsymbol{f}_2\in\boldsymbol{\mathcal{Z}},$$

$$\sigma(\boldsymbol{f}_1,\boldsymbol{f}_2) := \frac{1}{2}\left\langle G_{\boldsymbol{k}}(\rho)\Phi(\boldsymbol{f}_1,\boldsymbol{f}_2)\,,\,\Phi(\boldsymbol{f}_1,\boldsymbol{f}_2)\right\rangle_{\Omega} \qquad \text{for all } \boldsymbol{f}_1,\boldsymbol{f}_2\in\boldsymbol{\mathcal{Z}}.$$

Clearly, the functionals $\psi_j$ and $\sigma$ are proper, lower semicontinuous, and convex on $\boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{Z}}$. In particular, the functional $\psi_j$ is the so-called *indicator function* of the set $\mathcal{K}_j$ in the sense of convex analysis (see e.g. page 28 in [65]).

Next, we determine the convex conjugate functionals of $\psi_j$ for $j \in \mathbb{N}$. Notice that the Rieszian isomorphism $R : \boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{Z}} \to (\boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{Z}})^*$ is given by

$$R(\boldsymbol{f}_1,\boldsymbol{f}_2)[\boldsymbol{g}_1,\boldsymbol{g}_2] := \langle \boldsymbol{f}_1\,,\,\boldsymbol{g}_1\rangle_{\Omega} + \langle \boldsymbol{f}_2\,,\,\boldsymbol{g}_2\rangle_{\Omega} \qquad \text{for all } \boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{g}_1,\boldsymbol{g}_2\in\boldsymbol{\mathcal{Z}}.$$

Choosing two functions $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \boldsymbol{\mathcal{Z}}$ arbitrarily, we set $\boldsymbol{f} := \Phi(\boldsymbol{f}_1, \boldsymbol{f}_2)$. Given an arbitrary index $j \in \mathbb{N}$, we then obtain

$$\begin{aligned}
(\psi_j^* \circ R)(\boldsymbol{f}_1, \boldsymbol{f}_2) &= \sup_{(\boldsymbol{g}_1, \boldsymbol{g}_2) \in \boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{Z}}} \Big( \langle \boldsymbol{f}_1 \, , \boldsymbol{g}_1 \rangle_\Omega + \langle \boldsymbol{f}_2 \, , \boldsymbol{g}_2 \rangle_\Omega - \psi_j(\boldsymbol{g}_1, \boldsymbol{g}_2) \Big) \\
&= \sup_{\boldsymbol{g} \in \boldsymbol{Z}} \Big( \mathrm{Re}\big( \langle \boldsymbol{f} \, , \boldsymbol{g} \rangle_\Omega \big) - \psi_j\big( \mathrm{Re}(\boldsymbol{g}), \mathrm{Im}(\boldsymbol{g}) \big) \Big) \\
&= \sup \Big\{ \mathrm{Re}\big( \langle \boldsymbol{f} \, , \boldsymbol{g} \rangle_\Omega \big) \;\Big|\; \boldsymbol{g} \in \boldsymbol{Z},\ \|\boldsymbol{g}\|_\Omega \leq 1,\ P_{\boldsymbol{k},j-1}(\rho)\boldsymbol{g} = \boldsymbol{0} \Big\} \\
&= \sup \Big\{ \mathrm{Re}\big( \langle \boldsymbol{f} \, , \boldsymbol{g} \rangle_\Omega \big) \;\Big|\; \boldsymbol{g} \in \boldsymbol{Z},\ \|\boldsymbol{g}\|_\Omega = 1,\ P_{\boldsymbol{k},j-1}(\rho)\boldsymbol{g} = \boldsymbol{0} \Big\} \\
&= \sup_{\boldsymbol{h} \in \boldsymbol{Z}} \frac{\mathrm{Re}\big( \langle \boldsymbol{f} \, , \boldsymbol{h} - P_{\boldsymbol{k},j-1}(\rho)\boldsymbol{h} \rangle_\Omega \big)}{\|\boldsymbol{h} - P_{\boldsymbol{k},j-1}(\rho)\boldsymbol{h}\|_\Omega} \\
&= \sup_{\boldsymbol{h} \in \boldsymbol{Z}} \frac{\mathrm{Re}\big( \langle \boldsymbol{f} - P_{\boldsymbol{k},j-1}(\rho)\boldsymbol{f} \, , \boldsymbol{h} - P_{\boldsymbol{k},j-1}(\rho)\boldsymbol{h} \rangle_\Omega \big)}{\|\boldsymbol{h} - P_{\boldsymbol{k},j-1}(\rho)\boldsymbol{h}\|_\Omega}.
\end{aligned}$$

According to the Cauchy–Schwarz inequality, the maximum is attained for $\boldsymbol{h} = \boldsymbol{f}$. Hence, for every index $j \in \mathbb{N}$ the convex conjugate functional of $\psi_j$ is given by

$$(\psi_j^* \circ R)(\boldsymbol{f}_1, \boldsymbol{f}_2) = \|\Phi(\boldsymbol{f}_1, \boldsymbol{f}_2) - P_{\boldsymbol{k},j-1}(\rho)\Phi(\boldsymbol{f}_1, \boldsymbol{f}_2)\|_\Omega \qquad \text{for all } \boldsymbol{f}_1, \boldsymbol{f}_2 \in \boldsymbol{\mathcal{Z}}.$$

We remark that $\psi_j^*$ coincides with the so-called *support function* of $\mathcal{K}_j$ in the sense of convex anlysis (see e.g. page 28 in [65]), i.e.,

$$\psi_j^*(q) = \sup_{(\boldsymbol{f}_1, \boldsymbol{f}_2) \in \mathcal{K}_j} q(\boldsymbol{f}_1, \boldsymbol{f}_2) \qquad \text{for all } q \in (\boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{Z}})^*.$$

Now, we determine the convex conjugate functional of $\sigma$. Again, we choose two functions $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \boldsymbol{\mathcal{Z}}$ arbitrarily, and set $\boldsymbol{f} := \Phi(\boldsymbol{f}_1, \boldsymbol{f}_2)$. If $\boldsymbol{f} \in \ker(G_{\boldsymbol{k}}(\rho))$ and $\boldsymbol{f} \neq \boldsymbol{0}$, we find that

$$\begin{aligned}
(\sigma^* \circ R)(\boldsymbol{f}_1, \boldsymbol{f}_2) &= \sup_{(\boldsymbol{g}_1, \boldsymbol{g}_2) \in \boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{Z}}} \Big( \langle \boldsymbol{f}_1 \, , \boldsymbol{g}_1 \rangle_\Omega + \langle \boldsymbol{f}_2 \, , \boldsymbol{g}_2 \rangle_\Omega - \sigma(\boldsymbol{g}_1, \boldsymbol{g}_2) \Big) \\
&= \sup_{\boldsymbol{g} \in \boldsymbol{Z}} \Big( \mathrm{Re}\big( \langle \boldsymbol{f} \, , \boldsymbol{g} \rangle_\Omega \big) - \frac{1}{2} \langle G_{\boldsymbol{k}}(\rho)\boldsymbol{g} \, , \boldsymbol{g} \rangle_\Omega \Big) \\
&\geq \mathrm{Re}\big( \langle \boldsymbol{f} \, , s\boldsymbol{f} \rangle_\Omega \big) - \frac{1}{2} \langle s G_{\boldsymbol{k}}(\rho)\boldsymbol{f} \, , s\boldsymbol{f} \rangle_\Omega \\
&= s\|\boldsymbol{f}\|_\Omega^2,
\end{aligned}$$

where $s > 0$ is an arbitrary positive number. Since $s$ can be chosen arbitrarily, we conclude that $(\sigma^* \circ R)(\boldsymbol{f}_1, \boldsymbol{f}_2) = \infty$ if $\boldsymbol{f} \in \ker(G_{\boldsymbol{k}}(\rho))$. Next, we assume that

$\boldsymbol{f} \in \mathrm{im}(G_{\boldsymbol{k}}(\rho))$. Then, there exists a function $\boldsymbol{g}_{\boldsymbol{f}} \in \boldsymbol{Z}$, such that $G_{\boldsymbol{k}}(\rho)\boldsymbol{g}_{\boldsymbol{f}} = \boldsymbol{f}$. Moreover, we have that the functional $\varphi_{\boldsymbol{f}} : \boldsymbol{Z} \to \mathbb{R}$, defined by

$$\varphi_{\boldsymbol{f}}(\boldsymbol{g}) := \mathrm{Re}\big(\langle \boldsymbol{f}\,, \boldsymbol{g}\rangle_\Omega\big) - \frac{1}{2}\langle G_{\boldsymbol{k}}(\rho)\boldsymbol{g}\,, \boldsymbol{g}\rangle \qquad \text{for all } \boldsymbol{g} \in \boldsymbol{Z},$$

is maximized on the affine subspace $\boldsymbol{g}_{\boldsymbol{f}} + \ker(G_{\boldsymbol{k}}(\rho))$. To see this, one can simply compute the first and second Fréchet derivatives $\mathrm{D}\varphi_{\boldsymbol{f}}$ and $\mathrm{D}^2\varphi_{\boldsymbol{f}}$ of the functional $\varphi_{\boldsymbol{f}}$. Then, it is easy to verify that $\mathrm{D}\varphi_{\boldsymbol{f}}(\boldsymbol{g}) = 0$, if and only if $\boldsymbol{g} \in \boldsymbol{g}_{\boldsymbol{f}} + \ker(G_{\boldsymbol{k}}(\rho))$, and that $\mathrm{D}^2\varphi_{\boldsymbol{f}}(\boldsymbol{v}, \boldsymbol{v}) < 0$ for all $\boldsymbol{v} \in \ker(G_{\boldsymbol{k}}(\rho))^\perp$. Furthermore, since

$$\boldsymbol{f} = G_{\boldsymbol{k}}(\rho)\boldsymbol{g}_{\boldsymbol{f}} = I_{\boldsymbol{k}}L_{\boldsymbol{k}}(\rho)^{-1}M_{\boldsymbol{k}}\boldsymbol{g}_{\boldsymbol{f}} = L_{\boldsymbol{k}}(\rho)^{-1}M_{\boldsymbol{k}}\boldsymbol{g}_{\boldsymbol{f}},$$

we have that

$$\langle \boldsymbol{f}\,, \boldsymbol{g}_{\boldsymbol{f}}\rangle_\Omega = M_{\boldsymbol{k}}\boldsymbol{g}_{\boldsymbol{f}}[\boldsymbol{f}] = L_{\boldsymbol{k}}(\rho)\boldsymbol{f}[\boldsymbol{f}],$$

and hence

$$(\sigma^* \circ R)(\boldsymbol{f}_1, \boldsymbol{f}_2) = \sup_{\boldsymbol{g}\in\boldsymbol{Z}} \varphi_{\boldsymbol{f}}(\boldsymbol{g}) = \varphi_{\boldsymbol{f}}(\boldsymbol{g}_{\boldsymbol{f}}) = \frac{1}{2}\,\langle \boldsymbol{f}\,, \boldsymbol{g}_{\boldsymbol{f}}\rangle_\Omega = \frac{1}{2}\,L_{\boldsymbol{k}}(\rho)\boldsymbol{f}[\boldsymbol{f}].$$

Since the image space of $G_{\boldsymbol{k}}(\rho)$ is dense in $\boldsymbol{V}_{\boldsymbol{k}}$ according to (4.52), we conclude that the convex conjugate functional of $\sigma$ is given by

$$(\sigma^* \circ R)(\boldsymbol{f}_1, \boldsymbol{f}_2) = \begin{cases} \dfrac{1}{2}L_{\boldsymbol{k}}(\rho)\Phi(\boldsymbol{f}_1, \boldsymbol{f}_2)[\Phi(\boldsymbol{f}_1, \boldsymbol{f}_2)] & \text{if } \Phi(\boldsymbol{f}_1, \boldsymbol{f}_2) \in \boldsymbol{V}_{\boldsymbol{k}}, \\ \infty & \text{else} \end{cases}$$

for all $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \boldsymbol{\mathcal{Z}}$. According to Rayleigh's principle (4.56), we have

$$-\frac{1}{2}\,\mu_j(\rho, \boldsymbol{k}) = \min_{(\boldsymbol{f}_1,\boldsymbol{f}_2)\in\boldsymbol{\mathcal{Z}}\times\boldsymbol{\mathcal{Z}}} \Big(\psi_j(\boldsymbol{f}_1, \boldsymbol{f}_2) - \sigma(\boldsymbol{f}_1, \boldsymbol{f}_2)\Big) \qquad \text{for all } j \in \mathbb{N}.$$

The assertion thus follows from (4.60) and the duality principle stated in (4.68). □

By reviewing the line of arguments in the previous Section 4.4, one realizes that the operator $\kappa(\rho, \boldsymbol{k})M_{\boldsymbol{k}}|_{\boldsymbol{V}_{\boldsymbol{k}}}$ was added to $A_{\boldsymbol{k}}(\rho)$ in order to obtain a continuously invertible operator $L_{\boldsymbol{k}}(\rho) = A_{\boldsymbol{k}}(\rho) + \kappa(\rho, \boldsymbol{k})M_{\boldsymbol{k}}|_{\boldsymbol{V}_{\boldsymbol{k}}}$ for all $\rho \in \mathcal{D}$ and all $\boldsymbol{k} \in \mathbb{B}$. The continuous invertibility of $L_{\boldsymbol{k}}(\rho)$ followed from the ellipticity of the associated sesquilinear forms $l_{\boldsymbol{k}}(\rho) = a_{\boldsymbol{k}}(\rho) + \kappa(\rho, \boldsymbol{k})m$. However, Proposition 4.15 states that the sesquilinear forms $a_{\boldsymbol{k}}(\rho)$ are also elliptic for non-vanishing $\boldsymbol{k}$. Hence, the spectral theory in Section 4.4 can also be established with $\kappa(\rho, \boldsymbol{k})$ being replaced by zero, provided that $\boldsymbol{k}$ does not vanish. From this we deduce the following variant of Auchmuty's principle.

**Corollary 4.22.** *Under the assumptions of Proposition 4.20, the following identity holds for all* $\boldsymbol{k} \in \mathbb{B} \setminus \{\mathbf{0}\}$,

$$\frac{-1}{2\lambda_j(\rho,\boldsymbol{k})} = \min_{\boldsymbol{u}\in\boldsymbol{V}_{\boldsymbol{k}}} \left( \frac{1}{2}\, a_{\boldsymbol{k}}(\rho)(\boldsymbol{u},\boldsymbol{u}) - \left\| \boldsymbol{u} - P_{\boldsymbol{k},j-1}(\rho)\boldsymbol{u} \right\|_\Omega \right) \qquad \textit{for all } j \in \mathbb{N}.$$

Note that for $j = 1$ the assertion of Corollary 4.22 can be obtained without using the duality principle (4.68). Given a coefficient $\rho \in \mathcal{D}$ and a vector $\boldsymbol{k} \in \mathbb{B} \setminus \{\mathbf{0}\}$, one simply defines the functional $f : \boldsymbol{V}_{\boldsymbol{k}} \to \mathbb{R}$ by

$$f(\boldsymbol{u}) := \frac{1}{2}\, a_{\boldsymbol{k}}(\rho)(\boldsymbol{u},\boldsymbol{u}) - \|\boldsymbol{u}\|_\Omega \qquad \text{for all } \boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}}.$$

Clearly, the functional $f$ is continuous on $\boldsymbol{V}_{\boldsymbol{k}}$. Since $f(\mathbf{0}) = 0$ and $f(\theta\boldsymbol{v}) \to \infty$ as $\theta \to \infty$ for all $\boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}$, we deduce that $f$ admits a global minimum on $\boldsymbol{V}_{\boldsymbol{k}}$. The Fréchet-derivative of $f$ is given by

$$\mathrm{D}f(\boldsymbol{u})[\boldsymbol{v}] = a_{\boldsymbol{k}}(\rho)(\boldsymbol{u},\boldsymbol{v}) - \frac{m(\boldsymbol{u},\boldsymbol{v})}{\|\boldsymbol{u}\|_\Omega} \qquad \text{for all } \boldsymbol{u},\boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}.$$

From this we see that any stationary point $\boldsymbol{u}_* \in \boldsymbol{V}_{\boldsymbol{k}}$ of $f$ solves the weak eigenvalue equation (4.42) for the eigenvalue $\lambda_* = 1/\|\boldsymbol{u}_*\|_\Omega$. Therefore, we have

$$f(\boldsymbol{u}_*) = -\frac{1}{2\lambda_*}.$$

It follows that the function values of $f$ at its stationary points are given by $-1/(2\lambda)$, where $\lambda$ is a eigenvalue of the weak eigenvalue equation (4.42). The minimal function value at a critical point is therefore given by $-1/(2\lambda_1(\rho,\boldsymbol{k}))$, i.e.,

$$\frac{-1}{2\lambda_1(\rho,\boldsymbol{k})} = \min_{\boldsymbol{u}\in\boldsymbol{V}_{\boldsymbol{k}}} \Big( \frac{1}{2}\, a_{\boldsymbol{k}}(\rho)(\boldsymbol{u},\boldsymbol{u}) - \|\boldsymbol{u}\|_\Omega \Big).$$

We conclude this section with the remark that Auchmuty's principle, as stated in Proposition 4.21 or Corollary 4.22, can be established for every coercive or elliptic sesquilinear form on a complex Hilbert space, respectively.

## 4.6 Photonic Band Structures and Band Diagrams

In Section 4.4, we studied the existence of eigensolutions of the weakly formulated family of constrained eigenvalue problems given by Problem 4.16. We found that for every coefficient $\rho \in \mathcal{D}$ and every vector $\boldsymbol{k} \in \mathbb{B}$ there exists a countable number

of eigensolutions. Provided that the coefficient $\rho$ represents the medium structure of a three-dimensional photonic crystal, each eigensolution corresponds to a Bloch mode of that crystal with quasimomentum vector $\boldsymbol{k}$. In this section we relate the eigensolutions of Problem 4.16 to the so-called band structures of photonic crystals.

According to Proposition 4.20, there exists an increasing sequence $\{\lambda_j(\rho,\boldsymbol{k})\}_{j\in\mathbb{N}}$ of non-negative eigenvalues of (4.42) for every $\rho\in\mathcal{D}$ and every $\boldsymbol{k}\in\mathbb{B}$. Fixing a coefficient $\rho\in\mathcal{D}$, we now consider for every index $j\in\mathbb{N}$ the $j$-th smallest eigenvalue $\lambda_j(\rho,\boldsymbol{k})$ as a function of $\boldsymbol{k}$ from $\mathbb{B}$ into $\mathbb{R}$.

Before we can prove our next result, we need to define some operators. Given a coefficient $\rho\in\mathcal{D}$ and a vector $\boldsymbol{k}\in\mathbb{B}$, we define the operators $\widetilde{A}_{\boldsymbol{k}}(\rho):\boldsymbol{W}\to\boldsymbol{W}^*$, $B_{\boldsymbol{k}}:\boldsymbol{Z}\to Q^*$, and $-\Delta_{\boldsymbol{k}}:Q\to Q^*$ by

$$\widetilde{A}_{\boldsymbol{k}}(\rho)\boldsymbol{w}[\boldsymbol{v}] := a_{\boldsymbol{k}}(\rho)(\boldsymbol{w},\boldsymbol{v}) \qquad \text{for all } \boldsymbol{w},\boldsymbol{v}\in\boldsymbol{W} \tag{4.69}$$

$$B_{\boldsymbol{k}}\boldsymbol{f}[q] := b_{\boldsymbol{k}}(\boldsymbol{f},q) \qquad \text{for all } \boldsymbol{f}\in\boldsymbol{Z}, q\in Q \tag{4.70}$$

$$-\Delta_{\boldsymbol{k}}p[q] := \int_\Omega \overline{(\nabla+\mathrm{i}\boldsymbol{k})p}\cdot(\nabla+\mathrm{i}\boldsymbol{k})q \qquad \text{for all } p,q\in Q. \tag{4.71}$$

Recall that $\boldsymbol{W}$ denotes the complex Hilbert space $\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl};\Omega)$, and that $Q$ denotes the complex Hilbert space $H^1_{\mathrm{per}}(\Omega)$. Notice that the operators $A_{\boldsymbol{k}}(\rho)$, defined by (4.43), differ from the operators $\widetilde{A}_{\boldsymbol{k}}(\rho)$ only in the domain of definition and in the range.

Given a vector $\boldsymbol{k}\in\mathbb{B}\setminus\{\boldsymbol{0}\}$, it is not difficult to show that for every continuous linear functional $f\in Q^*$ there exists a unique function $p\in Q$, such that

$$-\Delta_{\boldsymbol{k}}p = f. \tag{4.72}$$

One can also show that the solution $p$ is bounded in $Q$ and that it depends continuously on $f$. For $\boldsymbol{k}=\boldsymbol{0}$ the solution of (4.72) is only unique up to an additive constant. However, it can be chosen uniquely in the Hilbert space $Q_\diamond := Q\cap L^2_\diamond(\Omega)$. Recall that the Hilbert space $L^2_\diamond(\Omega)$ was defined in (4.24). It follows that for every $\boldsymbol{k}\in\mathbb{B}$ there exists a continuous linear operator $(-\Delta_{\boldsymbol{k}})^{-1}:Q^*\to Q$, which maps every functional $f\in Q^*$ either to the unique solution $p\in Q$ of (4.72), if $\boldsymbol{k}\in\mathbb{B}\setminus\{\boldsymbol{0}\}$, or to the unique solution $p\in Q_\diamond$ of (4.72), if $\boldsymbol{k}=\boldsymbol{0}$. It follows that this operator $(-\Delta_{\boldsymbol{k}})^{-1}$ is the continuous, conjugate-linear inverse of $-\Delta_{\boldsymbol{k}}$.

Finally, we define for every vector $\boldsymbol{k}\in\mathbb{B}$ the operator $\pi_{\boldsymbol{k}}:\boldsymbol{W}\to\boldsymbol{W}$ by

$$\pi_{\boldsymbol{k}} := \mathrm{id}_{\boldsymbol{W}} - (\nabla+\mathrm{i}\boldsymbol{k})(-\Delta_{\boldsymbol{k}})^{-1}B_{\boldsymbol{k}}|_{\boldsymbol{W}}, \tag{4.73}$$

where $\mathrm{id}_{\boldsymbol{W}}$ denotes the identity on $\boldsymbol{W}$. The utility of the operator $\pi_{\boldsymbol{k}}$ is made obvious by the following proposition.

**Proposition 4.23.** *For every $\boldsymbol{k} \in \mathbb{B}$, the operator $\pi_{\boldsymbol{k}}$ is the $\boldsymbol{Z}$-orthogonal projection from $\boldsymbol{W}$ onto $\boldsymbol{V}_{\boldsymbol{k}}$.*

*Proof.* Let $\boldsymbol{w} \in \boldsymbol{W}$ be an arbitrary function. By Theorem 4.8(a) there exist for every $\boldsymbol{k} \in \mathbb{B} \setminus \{\boldsymbol{0}\}$ uniquely determined functions $\boldsymbol{\psi} \in H^1_{\mathrm{per}}(\Omega)^3$ and $p \in Q$, such that $\boldsymbol{w} = (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{\psi} + (\nabla + \mathrm{i}\boldsymbol{k})p$. According to the definition of the operators $B_{\boldsymbol{k}}$ and $(-\Delta_{\boldsymbol{k}})^{-1}$, we have $(-\Delta_{\boldsymbol{k}})^{-1}B_{\boldsymbol{k}}\boldsymbol{w} = p$ and hence $\pi_{\boldsymbol{k}}\boldsymbol{w} = (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{\psi}$. By Proposition 4.11 we have that $\pi_{\boldsymbol{k}}\boldsymbol{w} \in \boldsymbol{V}_{\boldsymbol{k}}$. Furthermore, since $\boldsymbol{w} - \pi_{\boldsymbol{k}}\boldsymbol{w} = (\nabla + \mathrm{i}\boldsymbol{k})p$, we have by Corollary 4.9(a) that $\pi_{\boldsymbol{k}}$ is indeed an $\boldsymbol{Z}$-orthogonal projection and onto.

Next, we consider the case $\boldsymbol{k} = \boldsymbol{0}$. By Theorem 4.8(b) there exist uniquely determined function $\boldsymbol{\psi} \in H^1_{\mathrm{per},\diamond}(\Omega)^3$ and $p \in Q_\diamond$, as well as a uniquely determined vector $\boldsymbol{z} \in \mathbb{C}^3$, such that $\boldsymbol{w} = \nabla \times \boldsymbol{\psi} + \nabla p + \boldsymbol{z}$. We find $(-\Delta_{\boldsymbol{0}})^{-1}B_{\boldsymbol{0}}\boldsymbol{w} = \nabla p$, which implies that $\pi_{\boldsymbol{0}}\boldsymbol{w} = \nabla \times \boldsymbol{\psi} + \boldsymbol{z}$. Again, we have by Proposition 4.11 that $\pi_{\boldsymbol{0}}\boldsymbol{w} \in \boldsymbol{V}_{\boldsymbol{0}}$, and by Corollary 4.9(b) that $\pi_{\boldsymbol{0}}$ is an $\boldsymbol{Z}$-orthogonal projection and onto. □

With the operators $\widetilde{A}_{\boldsymbol{k}}(\rho)$ and $\pi_{\boldsymbol{k}}$ at hand, we are now able to prove the following regularity result for the functions $\boldsymbol{k} \mapsto \lambda_j(\rho, \boldsymbol{k})$.

**Proposition 4.24.** *For every index $j \in \mathbb{N}$ and every coefficient $\rho \in \mathcal{D}$ the functional $\lambda_j(\rho, \cdot\,) : \mathbb{B} \to \mathbb{R}$ is continuous.*

*Proof.* First, we notice that $\widetilde{A}_{\boldsymbol{k}}(\rho) = \pi^*_{\boldsymbol{k}} A_{\boldsymbol{k}}(\rho)\pi_{\boldsymbol{k}}$ for all $\rho \in \mathcal{D}$ and all $\boldsymbol{k} \in \mathbb{B}$, where $\pi^*_{\boldsymbol{k}}$ denotes the dual operator of $\pi_{\boldsymbol{k}}$. From this we deduce that the positive eigenvalues of the operators $\widetilde{A}_{\boldsymbol{k}}(\rho)$ and $A_{\boldsymbol{k}}(\rho)$ coincide for every $\rho \in \mathcal{D}$ and every $\boldsymbol{k} \in \mathbb{B}$. Choosing a fixed coefficient $\rho \in \mathcal{D}$ and an arbitrary vector $\boldsymbol{k} \in \mathbb{B}$, we have

$$\begin{aligned}
\widetilde{A}_{\boldsymbol{k}}(\rho)\boldsymbol{w}[\boldsymbol{v}] &= \int_\Omega \rho\Big(\overline{\nabla \times \boldsymbol{w}} \cdot \nabla \times \boldsymbol{v} + \overline{\nabla \times \boldsymbol{w}} \cdot [\mathrm{i}\boldsymbol{k}]_\times \boldsymbol{v} \\
&\qquad - [\mathrm{i}\boldsymbol{k}]_\times \overline{\boldsymbol{w}} \cdot \nabla \times \boldsymbol{v} - [\mathrm{i}\boldsymbol{k}]_\times \overline{\boldsymbol{w}} \cdot [\mathrm{i}\boldsymbol{k}]_\times \boldsymbol{v}\Big) \\
&= \int_\Omega \rho\Big(\overline{\nabla \times \boldsymbol{w}} \cdot \nabla \times \boldsymbol{v} + [\mathrm{i}\boldsymbol{k}]^*_\times \overline{\nabla \times \boldsymbol{w}} \cdot \boldsymbol{v} \\
&\qquad - [\mathrm{i}\boldsymbol{k}]_\times \overline{\boldsymbol{w}} \cdot \nabla \times \boldsymbol{v} - [\mathrm{i}\boldsymbol{k}]^*_\times [\mathrm{i}\boldsymbol{k}]_\times \overline{\boldsymbol{w}} \cdot \boldsymbol{v}\Big) \qquad \text{for all } \boldsymbol{w}, \boldsymbol{v} \in \boldsymbol{W},
\end{aligned}$$

where $[\mathrm{i}\boldsymbol{k}]_\times$ denotes the cross product matrix corresponding to $\mathrm{i}\boldsymbol{k}$ (see Section 2.2). Now, let $\{\boldsymbol{k}^{(n)}\}_{n\in\mathbb{N}}$ be a sequence in $\mathbb{B}$, which converges to $\boldsymbol{k}$. Choosing an arbitrary number $n \in \mathbb{N}$ we obtain the estimate

$$\big|\big(\widetilde{A}_{\boldsymbol{k}^{(n)}}(\rho) - \widetilde{A}_{\boldsymbol{k}}(\rho)\big)\boldsymbol{w}[\boldsymbol{v}]\big| \le \|\rho\|_\infty (2\, f_n + g_n) \|\boldsymbol{w}\|_{\mathrm{curl},\Omega} \|\boldsymbol{v}\|_{\mathrm{curl},\Omega} \quad \text{for all } \boldsymbol{w}, \boldsymbol{v} \in \boldsymbol{W},$$

where, the numbers $f_n$ and $g_n$ are given by

$$\begin{aligned} f_n &:= \big\| [\boldsymbol{k}^{(n)} - \boldsymbol{k}]_\times \big\|_2 && \text{for all } n \in \mathbb{N}, \\ g_n &:= \big\| [\boldsymbol{k}^{(n)}]_\times^* [\boldsymbol{k}^{(n)}]_\times^* - [\boldsymbol{k}]_\times^* [\boldsymbol{k}]_\times^* \big\|_2 && \text{for all } n \in \mathbb{N}. \end{aligned}$$

Here, we denote by $\| \cdot \|_2$ the spectral norm on $\mathbb{C}^{3\times 3}$. Clearly, we have $f_n \to 0$ and $g_n \to 0$ as $n \to \infty$. The above estimate hence implies

$$\| \widetilde{A}_{\boldsymbol{k}^{(n)}}(\rho) - \widetilde{A}_{\boldsymbol{k}}(\rho) \|_{\boldsymbol{W},\boldsymbol{W}^*} \to 0 \quad \text{as } n \to \infty.$$

We have thus shown that the operator $\widetilde{A}_{\boldsymbol{k}^{(n)}}(\rho)$ converges to $\widetilde{A}_{\boldsymbol{k}}(\rho)$ as $n \to \infty$. This convergence in particular implies the generalized convergence of $\widetilde{A}_{\boldsymbol{k}^{(n)}}(\rho)$ to $\widetilde{A}_{\boldsymbol{k}}(\rho)$ in the sense of Kato, since the assumption of Theorem 2.24 in Chapter IV, Section 2.6 in [45] hold a fortiori. Kato establishes in Chapter IV, Section 3.5 in [45] that the generalized convergence implies the convergence of the operators' spectra. Therefore, we have $\lambda_j(\rho, \boldsymbol{k}^{(n)}) \to \lambda_j(\rho, \boldsymbol{k})$ as $n \to \infty$, which completes this proof. □

Recall that the unknown eigenvalues $\lambda$ in Problem 4.16 represent frequencies $\omega$ of the unknown Bloch modes of a photonic crystal via $\lambda = \omega^2/c_0^2$ (see Section 4.1). Assuming $c_0 = 1$ in the following, we defined for every coefficient $\rho \in \mathcal{D}$ and every index $j \in \mathbb{N}$ the function $\omega_j : \mathcal{D} \times \mathbb{B} \to \mathbb{R}$ by

$$\omega_j(\rho, \boldsymbol{k}) := \sqrt{\lambda_j(\rho, \boldsymbol{k})} \qquad \text{for all } \rho \in \mathcal{D},\ \boldsymbol{k} \in \mathbb{B}. \tag{4.74}$$

By Proposition 4.24 the functions $\omega_j(\rho, \cdot\,) : \mathbb{B} \to \mathbb{R}$ are continuous for every every index $j \in \mathbb{N}$ and every coefficient $\rho \in \mathcal{D}$. Provided that $\rho$ represents the medium structure of a photonic crystal via $\rho = 1/\varepsilon_r$, where $\varepsilon_r$ denotes the crystal's relative electric permittivity field, the graphs of the functions $\omega_1(\rho, \cdot\,), \omega_2(\rho, \cdot\,), \ldots$, etc. are referred to as the *photonic bands* of that crystal. In particular, the graph of the function $\omega_j(\rho, \cdot\,)$ is called the *$j$-th photonic band* for $i \in \mathbb{N}$. The union of all photonic bands is called the *photonic band structure*.

Given an index $j \in \mathbb{N}$ we shall say that the $j$-th and the $(j+1)$-th photonic band are *strictly separated*, if

$$\omega_j(\rho, \boldsymbol{k}) < \omega_{j+1}(\rho, \boldsymbol{k}) \qquad \text{for all } \boldsymbol{k} \in \mathbb{B}. \tag{4.75}$$

We shall say that there exists a *band gap* between the $j$-th and the $(j+1)$-th photonic band, if

$$\max_{\boldsymbol{k}\in\mathbb{B}} \omega_j(\rho, \boldsymbol{k}) < \min_{\boldsymbol{k}\in\mathbb{B}} \omega_{j+1}(\rho, \boldsymbol{k}). \tag{4.76}$$

The corresponding *gap width* $w_j(\rho)$ is given by

$$w_j(\rho) := \min_{\boldsymbol{k}\in\mathbb{B}} \omega_{j+1}(\rho,\boldsymbol{k}) - \max_{\boldsymbol{k}\in\mathbb{B}} \omega_j(\rho,\boldsymbol{k}). \tag{4.77}$$

In Section 3.2 we discussed that the complete symmetry of a photonic crystal's medium structure consists of translational symmetries as well as of non-translational symmetries. The translational symmetries are captured by the crystal's Bravais lattice $\Lambda$. The non-translational symmetries are characterized in terms of so-called crystallographic point groups. Every element of a photonic crystal's point group corresponds to an orthogonal transformation that constitutes a symmetry operation for the photonic crystal. The following proposition reveals, how the non-translational symmetries of a three-dimensional photonic crystal's medium structure influence the corresponding band structure.

**Proposition 4.25.** *Let $\boldsymbol{\theta} : \mathbb{R}^3 \to \mathbb{R}^3$ be an orthogonal transformation, which constitutes a symmetry operation for a given coefficient $\rho \in \mathcal{D}$, as well as for the first Brillouin zone $\mathbb{B}$. Then, we have*

$$\begin{aligned} &\lambda_j(\rho,\boldsymbol{\theta}^{-1}(\boldsymbol{k})) = \lambda_j(\rho,\boldsymbol{k}) && \text{for all } j\in\mathbb{N}, \boldsymbol{k}\in\mathbb{B},\\ &\boldsymbol{u}_j(\rho,\boldsymbol{\theta}^{-1}(\boldsymbol{k})) = \boldsymbol{\theta}^{-1}\big(\boldsymbol{u}_j(\rho,\boldsymbol{k})\circ\boldsymbol{\theta}\big) && \text{for all } j\in\mathbb{N}, \boldsymbol{k}\in\mathbb{B}. \end{aligned}$$

*Proof.* To simplify notations we define for given $j \in \mathbb{N}$, $\rho \in \mathcal{D}$, and $\boldsymbol{k} \in \mathbb{B}$ the real number $\lambda := \lambda_j(\rho,\boldsymbol{k})$, and the vector field $\boldsymbol{u} := \boldsymbol{u}_j(\rho,\boldsymbol{k})$. Furthermore, let $\boldsymbol{\theta}$ be given by $\boldsymbol{\theta}(\boldsymbol{x}) := \boldsymbol{\Theta}\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^3$, where $\boldsymbol{\Theta} \in \mathrm{O}_3(\mathbb{R})$ is an orthogonal matrix. For every diffeomorphism $\boldsymbol{\varphi} : \mathbb{R}^3 \to \mathbb{R}^3$, which satisfies $\boldsymbol{\varphi}(\Omega) = \Omega$, we define the mapping $(\,\cdot\,)_{\boldsymbol{\varphi}} : \boldsymbol{Z} \to \boldsymbol{Z}$ by

$$\boldsymbol{f}_{\boldsymbol{\varphi}} := (\boldsymbol{\varphi}')^{\mathrm{T}}(\boldsymbol{f}\circ\boldsymbol{\varphi}) \qquad \text{for all } \boldsymbol{f}\in\boldsymbol{Z}.$$

Clearly, each such mapping $(\,\cdot\,)_{\boldsymbol{\varphi}}$ is a linear isomorphism, where the inverse mapping of $(\,\cdot\,)_{\boldsymbol{\varphi}}$ is given by $(\,\cdot\,)_{\boldsymbol{\varphi}^{-1}}$. Moreover, it can be shown (see e.g. Lemma 8 in Section A.1 in [25]) that

$$\nabla\times\boldsymbol{w}_{\boldsymbol{\varphi}} = \det(\boldsymbol{\varphi}')\,(\boldsymbol{\varphi}')^{-1}(\nabla\times\boldsymbol{w}\circ\boldsymbol{\varphi}) \qquad \text{for all } \boldsymbol{w}\in\boldsymbol{W}.$$

One observes that $\nabla\times\boldsymbol{w}_{\boldsymbol{\theta}}$ is given by the so-called *Piola transform* of $\nabla\times\boldsymbol{w}$ with respect to $\boldsymbol{\varphi}$ (see Definition 7.18 in [53]). Accordingly, we obtain $\boldsymbol{f}_{\boldsymbol{\theta}} = \boldsymbol{\Theta}^{\mathrm{T}}(\boldsymbol{f}\circ\boldsymbol{\theta})$ for all $\boldsymbol{f}\in\boldsymbol{Z}$, as well as $\nabla\times\boldsymbol{w}_{\boldsymbol{\theta}} = \det(\boldsymbol{\Theta})\boldsymbol{\Theta}^{\mathrm{T}}(\nabla\times\boldsymbol{w}\circ\boldsymbol{\theta})$ for all $\boldsymbol{w}\in\boldsymbol{W}$. It follows that the mapping $(\,\cdot\,)_{\boldsymbol{\theta}}$ constitutes an isometric, linear isomorphism from $\boldsymbol{Z}$ onto $\boldsymbol{Z}$, as well as from $\boldsymbol{W}$ onto $\boldsymbol{W}$. We now show that the mapping $(\,\cdot\,)_{\boldsymbol{\theta}}$ also constitutes an isometric linear isomorphism from $\boldsymbol{V}_{\boldsymbol{k}}$ onto $\boldsymbol{V}_{\boldsymbol{\theta}^{-1}(\boldsymbol{k})}$. To this end, we define for every diffeomorphism $\boldsymbol{\varphi} : \mathbb{R}^3 \to \mathbb{R}^3$, which satisfies $\boldsymbol{\varphi}(\Omega) = \Omega$, the

mapping $(\,\cdot\,)^{\boldsymbol{\varphi}} : Q \to Q$ by $q^{\boldsymbol{\varphi}} := q \circ \boldsymbol{\varphi}$ for all $q \in Q$. Since $\nabla q^{\boldsymbol{\theta}} = \boldsymbol{\Theta}^{\mathrm{T}}((\nabla q) \circ \boldsymbol{\theta})$ for all $q \in Q$, one easily shows that the mapping $(\,\cdot\,)^{\boldsymbol{\theta}}$ is an isometrical, linear isomorphism. Clearly, the inverse mapping of $(\,\cdot\,)^{\boldsymbol{\theta}}$ is given by $(\,\cdot\,)^{\boldsymbol{\theta}^{-1}}$. Choosing $\boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}$ and $q \in Q$ arbitrarily, and letting $r := q^{\boldsymbol{\theta}^{-1}}$, we find that

$$\begin{aligned}
b_{\boldsymbol{\theta}^{-1}(\boldsymbol{k})}(\boldsymbol{v}_{\boldsymbol{\theta}}, q) = b_{\boldsymbol{\theta}^{-1}(\boldsymbol{k})}(\boldsymbol{v}_{\boldsymbol{\theta}}, r^{\boldsymbol{\theta}}) &= \int_{\Omega} \boldsymbol{\Theta}^{\mathrm{T}}(\overline{\boldsymbol{v}} \circ \boldsymbol{\theta}) \cdot (\nabla + \mathrm{i}\boldsymbol{\theta}^{-1}(\boldsymbol{k}))(r \circ \boldsymbol{\theta}) \\
&= \int_{\Omega} \boldsymbol{\Theta}^{\mathrm{T}}(\overline{\boldsymbol{v}} \circ \boldsymbol{\theta}) \cdot \boldsymbol{\Theta}^{\mathrm{T}}\big((\nabla r \circ \boldsymbol{\theta}) + \mathrm{i}\boldsymbol{k}(r \circ \boldsymbol{\theta})\big) \\
&= \int_{\Omega} (\overline{\boldsymbol{v}} \circ \boldsymbol{\theta}) \cdot \big((\nabla + \mathrm{i}\boldsymbol{k})r \circ \boldsymbol{\theta}\big) = \int_{\boldsymbol{\theta}(\Omega)} \overline{\boldsymbol{v}} \cdot (\nabla + \mathrm{i}\boldsymbol{k})r \\
&= b_{\boldsymbol{k}}(\boldsymbol{v}, r) \\
&= 0.
\end{aligned}$$

Since this identity can be established for all $q \in Q$, we have that $\boldsymbol{v}_{\boldsymbol{\theta}} \in \boldsymbol{V}_{\boldsymbol{\theta}^{-1}(\boldsymbol{k})}$. It follows that the image of $\boldsymbol{V}_{\boldsymbol{k}}$ under the mapping $(\,\cdot\,)_{\boldsymbol{\theta}}$ is a subset of $\boldsymbol{V}_{\boldsymbol{\theta}^{-1}(\boldsymbol{k})}$. By a similar computation one can show, that for each function $\boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{\theta}^{-1}(\boldsymbol{k})}$ we have $\boldsymbol{v}_{\boldsymbol{\theta}^{-1}} \in \boldsymbol{V}_{\boldsymbol{k}}$, implying that the restriction of $(\,\cdot\,)_{\boldsymbol{\theta}}$ to $\boldsymbol{V}_{\boldsymbol{k}}$ is surjective. Injectivity, and hence bijectivity, of this restriction follow from the fact that the mapping $(\,\cdot\,)_{\boldsymbol{\theta}}$ is an isometry from $\boldsymbol{W}$ onto $\boldsymbol{W}$.

Since $\boldsymbol{\Theta}$ is an orthogonal matrix, we have $(\boldsymbol{\Theta}\boldsymbol{x}) \times (\boldsymbol{\Theta}\boldsymbol{y}) = \det(\boldsymbol{\Theta})\boldsymbol{\Theta}(\boldsymbol{x} \times \boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^3$ according to (2.6) and (2.8). Choosing $\boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}$ arbitrarily, we obtain

$$\begin{aligned}
a_{\boldsymbol{\theta}^{-1}(\boldsymbol{k})}(\rho)(\boldsymbol{u}_{\boldsymbol{\theta}}, \boldsymbol{v}_{\boldsymbol{\theta}}) &= \int_{\Omega} \rho\overline{(\nabla + \mathrm{i}\boldsymbol{\theta}^{-1}(\boldsymbol{k})) \times \boldsymbol{u}_{\boldsymbol{\theta}}} \cdot (\nabla + \mathrm{i}\boldsymbol{\theta}^{-1}(\boldsymbol{k})) \times \boldsymbol{v}_{\boldsymbol{\theta}} \\
&= \int_{\Omega} \rho\overline{\big(\nabla \times \boldsymbol{u}_{\boldsymbol{\theta}} + \mathrm{i}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{k} \times \boldsymbol{u}_{\boldsymbol{\theta}}\big)} \cdot \big(\nabla \times \boldsymbol{v}_{\boldsymbol{\theta}} + \mathrm{i}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{k} \times \boldsymbol{v}_{\boldsymbol{\theta}}\big) \\
&= \int_{\Omega} \rho\overline{\big((\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u} \circ \boldsymbol{\theta}\big)} \cdot \big((\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{v} \circ \boldsymbol{\theta}\big) \\
&= \int_{\boldsymbol{\theta}(\Omega)} (\rho \circ \boldsymbol{\theta}^{-1})\overline{(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}} \cdot (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{v} \\
&= a_{\boldsymbol{k}}(\rho)(\boldsymbol{u}, \boldsymbol{v}).
\end{aligned}$$

Similarly, we find that $m(\boldsymbol{u}_{\boldsymbol{\theta}}, \boldsymbol{v}_{\boldsymbol{\theta}}) = m(\boldsymbol{u}, \boldsymbol{v})$. Since both identities can be established for all $\boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}$, we have that $a_{\boldsymbol{\theta}^{-1}(\boldsymbol{k})}(\rho)(\boldsymbol{u}_{\boldsymbol{\theta}}, \boldsymbol{v}_{\boldsymbol{\theta}}) - \lambda m(\boldsymbol{u}_{\boldsymbol{\theta}}, \boldsymbol{v}_{\boldsymbol{\theta}}) = 0$ for all $\boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{k}}$. Since $(\,\cdot\,)_{\boldsymbol{\theta}}$ is a linear isomorphism from $\boldsymbol{V}_{\boldsymbol{k}}$ onto $\boldsymbol{V}_{\boldsymbol{\theta}^{-1}(\boldsymbol{k})}$, we also have

$$a_{\boldsymbol{\theta}^{-1}(\boldsymbol{k})}(\rho)(\boldsymbol{u}_{\boldsymbol{\theta}}, \boldsymbol{v}) = \lambda m(\boldsymbol{u}_{\boldsymbol{\theta}}, \boldsymbol{v}) \qquad \text{for all } \boldsymbol{v} \in \boldsymbol{V}_{\boldsymbol{\theta}^{-1}(\boldsymbol{k})},$$

implying that $\lambda_j(\rho, \boldsymbol{\theta}^{-1}(\boldsymbol{k})) = \lambda$, and $\boldsymbol{u}_j(\rho, \boldsymbol{\theta}^{-1}(\boldsymbol{k})) = \boldsymbol{u}_{\boldsymbol{\theta}} = \boldsymbol{\Theta}^{\mathrm{T}}(\boldsymbol{u}_j(\rho, \boldsymbol{k}) \circ \boldsymbol{\theta})$. $\square$

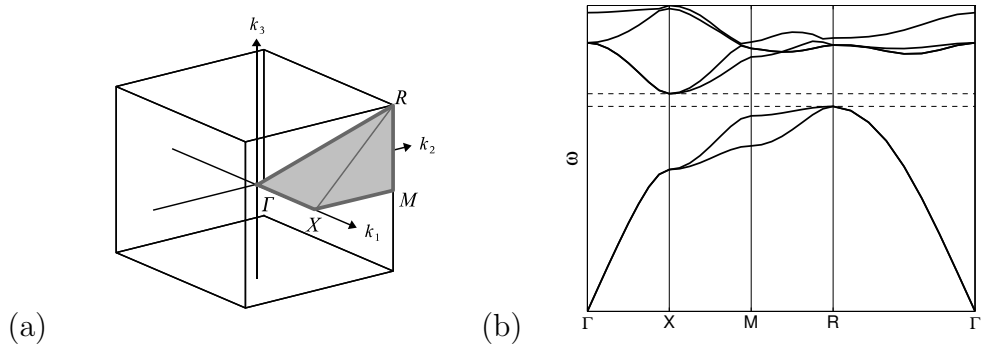



**Figure 4.1:** The first Brillouin zone with the irreducible zone (dark) and the critical points (a). The band diagram (b) shows the band structure along the critical path $\Gamma$–$X$–$M$–$R$–$\Gamma$. The dashed lines indicate a suspected band gap. Here, the space group is isomorphic to the semidirect product $\mathbb{Z}^3 \rtimes \mathrm{O}_3(\mathbb{Z})$.

In the following we illustrate a practical consequence of Proposition 4.25. Suppose that $\varepsilon_r$ is the relative electric permittivity function of a three-dimensional photonic crystal, whose medium structure is periodic with respect to a simple cubic lattice $\Lambda = a\mathbb{Z}^3$. Here $a > 0$ denotes the lattice constant. Furthermore, suppose that the crystal's point group is isomorphic to $\mathrm{O}_3(\mathbb{Z})$. Letting

$$G := \big\{ \boldsymbol{\theta} : \mathbb{R}^3 \to \mathbb{R}^3 \;\big|\; \boldsymbol{\theta}(\boldsymbol{x}) = \boldsymbol{\Theta}\boldsymbol{x},\ \boldsymbol{\Theta} \in \mathrm{O}_3(\mathbb{Z}) \big\},$$

we then have that

$$\varepsilon_r \circ \boldsymbol{\theta} = \varepsilon_r \qquad \text{for all } \boldsymbol{\theta} \in G.$$

The crystal's Wigner–Seitz cell is given by the cube $[-a/2, a/2]^3$. Letting $\Omega := (-a/2, a/2)^3$ and $\rho := 1/\varepsilon_r|_\Omega$, we also have that

$$\rho \circ \boldsymbol{\theta} = \rho \qquad \text{for all } \boldsymbol{\theta} \in G.$$

One easily verifies that the reciprocal lattice of $\Lambda$ is also a simple cubic lattice, which is given by $\widehat{\Lambda} = (2\pi/a)\mathbb{Z}^3$. It thus follows that the first Brillouin zone of $\Lambda$ is given by $\mathbb{B} = [-\pi/a, \pi/a]^3$. Clearly, this first Brillouin zone is symmetric with respect to $G$. By Proposition 4.25 we thus have that

$$\lambda_j(\rho, \boldsymbol{\theta}^{-1}(\boldsymbol{k})) = \lambda_j(\rho, \boldsymbol{k}) \qquad \text{for all } j \in \mathbb{N},\ \boldsymbol{k} \in \mathbb{B},\ \boldsymbol{\theta} \in G. \tag{4.78}$$

From (4.78) we deduce that for every index $j \in \mathbb{N}$ the function $\lambda_j(\rho, \cdot\,) : \mathbb{B} \to \mathbb{R}$ is $G$-symmetric, and hence uniquely determined by its restriction to any fundamental region of $G$. The closure of such a fundamental region is called an *irreducible zone*.

One can show that the tetrahedron

$$K := \overline{\mathrm{conv}}\big\{\boldsymbol{k}_\Gamma, \boldsymbol{k}_X, \boldsymbol{k}_M, \boldsymbol{k}_R\big\},$$

where

$$\boldsymbol{k}_\Gamma := \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{k}_X := \begin{pmatrix} \pi/a \\ 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{k}_M := \begin{pmatrix} \pi/a \\ \pi/a \\ 0 \end{pmatrix}, \quad \boldsymbol{k}_R := \begin{pmatrix} \pi/a \\ \pi/a \\ \pi/a \end{pmatrix},$$

is an irreducible zone for every simple cubic photonic crystal with lattice constant $a > 0$, whose crystallographic point group is isomorphic to $\mathrm{O}_3(\mathbb{Z})$. The vectors $\boldsymbol{k}_\Gamma$, $\boldsymbol{k}_X$, $\boldsymbol{k}_M$ and $\boldsymbol{k}_R$ are referred to as *critical points*. The path line segments, which connect the critical points $\boldsymbol{k}_\Gamma$, $\boldsymbol{k}_X$, $\boldsymbol{k}_M$, $\boldsymbol{k}_R$ and $\boldsymbol{k}_X$ (in this order), is called the *critical path*. We depicted the tetrahedron $K$, the critical point, and the critical path in Figure 4.1(a). By plotting the graphs of the functions $\omega_1(\rho, \cdot), \ldots, \omega_n(\rho, \cdot)$ along the critical path for some $n \in \mathbb{N}$, one obtains a so-called *photonic band diagram*. An example of such a band diagram is depicted in Figure 4.1(b).

In the literature on photonic crystals it is common practice to assume that the functions $\omega_j(\rho, \cdot)$ attain their extrema exactly on the critical path. Provided that this assumption indeed holds true, a band gap is revealed by a region in the band diagram, which is bounded by two horizontal lines such that neither of the two lines is crossed by any of the curves in the band diagram. The distance between the two horizontal lines then corresponds to the gap width.

## 4.7 The Two-Dimensional Case

So far, we only considered the spectral theory for the weakly formulated eigenvalue problems that determine the band structures of three-dimensional photonic crystals. In this section we turn to the eigenvalue problems that arise for TM- and TE-polarized Bloch modes in two-dimensional photonic crystals (see Section 3.5). The spectral theory related to these eigenvalue problems has already been discussed thoroughly in the literature (see e.g. [29], [50]), which is why we only give a brief account of the most important results.

Throughout this section we assume that $\Lambda \subset \mathbb{R}^2$ is a Bravais lattice of rank 2, and that $\Omega \subset \mathbb{R}^2$ is a primitive domain of $\Lambda$. As before, we denote by $\mathbb{B}$ the first Brillouin zone of $\Lambda$. For convenience we define the complex Hilbert spaces

$$V := H^1_{\mathrm{per}}(\Omega), \tag{4.79}$$

$$Z := L^2(\Omega). \tag{4.80}$$

Furthermore, we assume that the real Banach space $\mathcal{E}$ and the open function set $\mathcal{D}$ are defined according to (4.31) and (4.32).

Suppose that the Bravais lattice $\Lambda$ represents the periodicity a two-dimensional photonic crystal's medium structure. In Section 3.5 we derived an eigenvalue problem, which determines the existence of so-called TM-polarized Bloch modes with a given quasimomentum vector $\boldsymbol{k} \in \mathbb{B}$ (see (3.49) on page 43). In order to establish a weak formulation of this eigenvalue problem, we define for every coefficient $\rho \in \mathcal{D}$ and every vector $\boldsymbol{k} \in \mathbb{B}$ the sesquilinear forms $a_{\boldsymbol{k}}^{\mathrm{TM}} : V \times V \to \mathbb{C}$ and $m^{\mathrm{TM}}(\rho) : V \times V \to \mathbb{C}$ by

$$a_{\boldsymbol{k}}^{\mathrm{TM}}(u,v) := \int_\Omega \overline{(\nabla + \mathrm{i}\boldsymbol{k})u} \cdot (\nabla + \mathrm{i}\boldsymbol{k})v \qquad \text{for all } u, v \in V, \tag{4.81}$$

$$m^{\mathrm{TM}}(\rho)(f,g) := \int_\Omega \rho \overline{f} g \qquad \text{for all } f, g \in Z. \tag{4.82}$$

With this, we consider the following problem.

**Problem 4.26.** *Given a coefficient $\rho \in \mathcal{D}$ and a vector $\boldsymbol{k} \in \mathbb{B}$, find eigenvalues $\lambda \in \mathbb{C}$ and corresponding eigenfunctions $u \in V$, such that*

$$a_{\boldsymbol{k}}^{\mathrm{TM}}(u,v) = \lambda\, m^{\mathrm{TM}}(\rho)(u,v) \qquad \textit{for all } v \in V. \tag{4.83}$$

If $\rho = \varepsilon_r|_\Omega$, where $\varepsilon_r$ is the relative electric permittivity function of a two-dimensional photonic crystal, the (4.83) is precisely the weak formulation of the eigenvalue equation (3.49) on page 43, and thus determines the existence of TM-polarized Bloch modes in the photonic crystal. One can show that for every vector $\boldsymbol{k} \in \mathbb{B}$ the sesquilinear form $a_{\boldsymbol{k}}(\rho)$ satisfies

$$\begin{aligned}
a_{\boldsymbol{k}}^{\mathrm{TM}}(v,u) &= \overline{a_{\boldsymbol{k}}^{\mathrm{TM}}(u,v)} && \text{for all } v, u \in V,\\
\left|a_{\boldsymbol{k}}^{\mathrm{TM}}(v,u)\right| &\le \beta^{\mathrm{TM}}(\boldsymbol{k}) \|v\|_{1,\Omega} \|u\|_{1,\Omega} && \text{for all } u, v \in V,\\
a_{\boldsymbol{k}}^{\mathrm{TM}}(v,v) &\ge \alpha^{\mathrm{TM}}(\boldsymbol{k}) \|v\|_{1,\Omega}^2 - \kappa^{\mathrm{TM}}(\boldsymbol{k}) \|v\|_{\Omega}^2 && \text{for all } v \in V,\\
a_{\boldsymbol{k}}^{\mathrm{TM}}(v,v) &\ge 0 && \text{for all } v \in V.
\end{aligned}$$

The functions $\alpha^{\mathrm{TM}}, \beta^{\mathrm{TM}}, \kappa^{\mathrm{TM}} : \mathbb{B} \to \mathbb{R}$ are given by

$$\alpha^{\mathrm{TM}}(\boldsymbol{k}) := \left(1 - \frac{|\boldsymbol{k}|}{\mathrm{diam}(\mathbb{B})}\right) \qquad \text{for all } \boldsymbol{k} \in \mathbb{B}, \tag{4.84}$$

$$\beta^{\mathrm{TM}}(\boldsymbol{k}) := 2\max\{1, |\boldsymbol{k}|^2\} \qquad \text{for all } \boldsymbol{k} \in \mathbb{B}, \tag{4.85}$$

$$\kappa^{\mathrm{TM}}(\boldsymbol{k}) := \alpha^{\mathrm{TM}}(\boldsymbol{k}) + \left(\mathrm{diam}(\mathbb{B})|\boldsymbol{k}| + |\boldsymbol{k}|^2\right) \qquad \text{for all } \boldsymbol{k} \in \mathbb{B}. \tag{4.86}$$

Thus, it follows that the sesquilinear form $a_{\boldsymbol{k}}^{\mathrm{TM}}$ is conjugate-symmetric, bounded, $V$–$Z$-coercive and positive semidefinite for every $\boldsymbol{k} \in \mathbb{B}$.

By a line of argument similar to the one presented in Section 4.4, one can show that for every coefficient $\rho \in \mathcal{D}$ and every vector $\boldsymbol{k} \in \mathbb{B}$ there exists an increasing sequence $\{\lambda_j^{\mathrm{TM}}(\rho, \boldsymbol{k})\}_{j\in\mathbb{N}}$ of non-negative, real eigenvalues of the weak eigenvalue equation (4.83). Moreover, one can show that the sequence of the corresponding eigenfunctions $\{u_j^{\mathrm{TM}}(\rho, \boldsymbol{k})\}_{j\in\mathbb{N}}$ forms a complete system of $V$, which is orthogonal with respect to the inner product $\langle \cdot\,,\,\cdot \rangle_\rho$ on $Z$. For every $\rho \in \mathcal{D}$ the inner product $\langle \cdot\,,\,\cdot \rangle_\rho : Z \times Z \to \mathbb{C}$ is given by

$$\langle f\,, g\rangle_\rho := m^{\mathrm{TM}}(\rho)(f, g) \qquad \text{for all } f, g \in Z. \tag{4.87}$$

By $\|\cdot\|_\rho$, we denote the corresponding norm on $V$. Clearly, for every $\rho \in \mathcal{D}$, one has that

$$\operatorname*{ess\,inf}_\Omega(\rho)\|f\|_\Omega \leq \|f\|_\rho \leq \operatorname*{ess\,sup}_\Omega(\rho)\|f\|_\Omega \qquad \text{for all } f \in Z,$$

i.e., the norms $\|\cdot\|_\Omega$ and $\|\cdot\|_\rho$ are equivalent.

As in the three-dimensional setting, the eigenspace of every eigenvalue of (4.83) is finite-dimensional, and the $j$-th smallest eigenvalues $\lambda_j^{\mathrm{TM}}(\rho, \boldsymbol{k})$ can be characterized by the Min-Max Principle as

$$\lambda_j^{\mathrm{TM}}(\rho, \boldsymbol{k}) = \min_{\substack{U \sqsubseteq V \\ \dim U = j}} \max_{u \in U\setminus\{0\}} \frac{a_{\boldsymbol{k}}^{\mathrm{TM}}(u, u)}{m^{\mathrm{TM}}(\rho)(u, u)} \qquad \text{for all } j \in \mathbb{N} \tag{4.88}$$

or, alternatively, by Auchmuty's principle. In order to formulate the latter one, we define for every coefficient $\rho \in \mathcal{D}$ and every index $j \in \mathbb{N}$ the linear function space

$$V_{\boldsymbol{k},j}^{\mathrm{TM}}(\rho) := \operatorname{span}\big\{u_1^{\mathrm{TM}}(\rho, \boldsymbol{k}), \ldots, u_j^{\mathrm{TM}}(\rho, \boldsymbol{k})\big\}, \tag{4.89}$$

as well as the projection operator $P_{\boldsymbol{k},j}^{\mathrm{TM}}(\rho) : Z \to V_{\boldsymbol{k},j}(\rho)$ by

$$P_{\boldsymbol{k},j}^{\mathrm{TM}}(\rho) f := \sum_{i=1}^{j} \frac{\langle u_i^{\mathrm{TM}}(\rho, \boldsymbol{k})\,,\, f\rangle_\rho}{\|u_i^{\mathrm{TM}}(\rho, \boldsymbol{k})\|_\rho^2}\, u_i^{\mathrm{TM}}(\rho, \boldsymbol{k}) \qquad \text{for all } f \in Z. \tag{4.90}$$

Then, Auchmuty's principle reads

$$\begin{aligned}\frac{-1}{2\big(\lambda_j^{\mathrm{TM}}(\rho, \boldsymbol{k}) + \kappa^{\mathrm{TM}}(\boldsymbol{k})\big)} = \min_{u\in V}\Bigg(&\frac{1}{2}\left(a_{\boldsymbol{k}}^{\mathrm{TM}}(u, u) + \kappa^{\mathrm{TM}}(\boldsymbol{k})\, m^{\mathrm{TM}}(\rho)(u, u)\right) \\ &- \left\|u - P_{\boldsymbol{k},j-1}^{\mathrm{TM}}(\rho)u\right\|_\rho\Bigg) \qquad \text{for all } j \in \mathbb{N}.\end{aligned} \tag{4.91}$$

In Section 3.5 we also derived an eigenvalue problem that determines the existence of so-called TE-polarized Bloch modes (see (3.50) on page 43) in two-dimensional photonic crystals. In order to establish also a weak formulation of this problem, we defined for every coefficient $\rho \in \mathcal{D}$ and every vector $\boldsymbol{k} \in \mathbb{B}$, we define the sesquilinear forms $a_{\boldsymbol{k}}^{\mathrm{TE}}(\rho) : V \times V \to \mathbb{C}$ and $m^{\mathrm{TE}} : Z \times Z \to \mathbb{C}$ by

$$a_{\boldsymbol{k}}^{\mathrm{TE}}(\rho)(u,v) := \int_\Omega \rho \overline{(\nabla + \mathrm{i}\boldsymbol{k})u} \cdot (\nabla + \mathrm{i}\boldsymbol{k})v \qquad \text{for all } u, v \in V, \tag{4.92}$$

$$m^{\mathrm{TE}}(f,g) := \int_\Omega \overline{f} g \qquad \text{for all } f, g \in Z. \tag{4.93}$$

We then consider the following problem.

**Problem 4.27.** *Given a coefficient $\rho \in \mathcal{D}$, and a vector $\boldsymbol{k} \in \mathbb{B}$, find eigenvalues $\lambda \in \mathbb{C}$ and corresponding eigenfunctions $u \in V$, such that*

$$a_{\boldsymbol{k}}^{\mathrm{TE}}(\rho)(u,v) = \lambda\, m^{\mathrm{TE}}(u,v) \qquad \textit{for all } v \in V. \tag{4.94}$$

One notices that (4.94) is precisely the weak formulation of the eigenvalue equation (3.49) on page 43, provided that the coefficient $\rho$ is given by the relative electric permittivity function $\varepsilon_r$ of a two-dimensional photonic crystal according to $\rho = 1/\varepsilon_r|_\Omega$. One can show that the sesquilinear form $a_{\boldsymbol{k}}^{\mathrm{TE}}(\rho)$ is conjugate-symmetric, bounded, $V$–$Z$-coercive and positive semidefinite for every coefficient $\rho \in \mathcal{D}$ and every vector $\boldsymbol{k} \in \mathbb{B}$. It thus follows that for every $\rho \in \mathcal{D}$ and every $\boldsymbol{k} \in \mathbb{B}$ there exists an increasing sequence $\{\lambda_j^{\mathrm{TE}}(\rho, \boldsymbol{k})\}_{j\in\mathbb{N}}$ of non-negative, real eigenvalues of (4.94). The sequence of the corresponding eigenfunctions $\{u_j^{\mathrm{TE}}(\rho, \boldsymbol{k})\}_{j\in\mathbb{N}}$ forms a complete $Z$-orthogonal system of $V$.

In analogy to the definitions in Section 4.6, we define for every index $j \in \mathbb{N}$ the functions $\omega_j^{\mathrm{TM}} : \mathcal{D} \times \mathbb{B} \to \mathbb{R}$ and $\omega_j^{\mathrm{TE}} : \mathcal{D} \times \mathbb{B} \to \mathbb{R}$ by

$$\omega_j^{\mathrm{TM}}(\rho, \boldsymbol{k}) := \sqrt{\lambda_j^{\mathrm{TM}}(\rho, \boldsymbol{k})} \qquad \text{for all } \rho \in \mathcal{D},\ \boldsymbol{k} \in \mathbb{B}, \tag{4.95}$$

$$\omega_j^{\mathrm{TE}}(\rho, \boldsymbol{k}) := \sqrt{\lambda_j^{\mathrm{TE}}(\rho, \boldsymbol{k})} \qquad \text{for all } \rho \in \mathcal{D},\ \boldsymbol{k} \in \mathbb{B}. \tag{4.96}$$

Now suppose that $\varepsilon_r$ is the relative electric permittivity function of a given, two-dimensional photonic crystal. Then, the graphs of the functions $\omega_1^{\mathrm{TM}}(\varepsilon_r|_\Omega, \cdot\,)$, $\omega_2^{\mathrm{TM}}(\varepsilon_r|_\Omega, \cdot\,), \ldots$, etc. are referred to as the *TM bands* of the photonic crystal. Similarly, the graphs of the functions $\omega_1^{\mathrm{TE}}(1/\varepsilon_r|_\Omega, \cdot\,), \omega_2^{\mathrm{TE}}(1/\varepsilon_r|_\Omega, \cdot\,), \ldots$, etc. are referred to as the *TE bands* of the photonic crystal. The union of all TM bands is called the *TM band structure*, and the union of all TE bands is called the *TE band structure* of the photonic crystal.

A two-dimensional photonic crystal with relative electric permittivity function $\varepsilon_r$ is said to exhibit a *TM band gap*, if there exists an index $j \in \mathbb{N}$, such that

$$\max_{\boldsymbol{k}\in\mathbb{B}} \omega_j^{\mathrm{TM}}(\varepsilon_r|_\Omega, \boldsymbol{k}) < \min_{\boldsymbol{k}\in\mathbb{B}} \omega_{j+1}^{\mathrm{TM}}(\varepsilon_r|_\Omega, \boldsymbol{k})$$

The crystal is said to exhibit a *TE band gap*, if there exists an index $j \in \mathbb{N}$, such that

$$\max_{\boldsymbol{k}\in\mathbb{B}} \omega_j^{\mathrm{TE}}(1/\varepsilon_r|_\Omega, \boldsymbol{k}) < \min_{\boldsymbol{k}\in\mathbb{B}} \omega_{j+1}^{\mathrm{TE}}(1/\varepsilon_r|_\Omega, \boldsymbol{k}).$$

It should be noted that a TM band gap only corresponds to a range of inhibited TM wave frequencies, and that a TE band gap only corresponds to a range of inhibited TE wave frequencies. This means that a TM polarized wave with a frequency lying in a TE band gap of a two-dimensional photonic crystal might still be able to propagate inside the crystal. Therefore, two-dimensional photonic crystals only inhibit light propagation for frequencies, which belong to a so-called *complete band gap*. A crystal with a relative electric permittivity function $\varepsilon_r$ is said to exhibit such a complete band gap, if there exist indices $i, j \in \mathbb{N}$, such that

$$\begin{aligned}\max_{\boldsymbol{k}\in\mathbb{B}} \max\Big\{ &\omega_i^{\mathrm{TM}}(\varepsilon_r|_\Omega, \boldsymbol{k}), \omega_j^{\mathrm{TE}}(1/\varepsilon_r|_\Omega, \boldsymbol{k}) \Big\} \\ &< \min_{\boldsymbol{k}\in\mathbb{B}} \min\Big\{ \omega_{i+1}^{\mathrm{TM}}(\varepsilon_r|_\Omega, \boldsymbol{k}), \omega_{j+1}^{\mathrm{TE}}(1/\varepsilon_r|_\Omega, \boldsymbol{k}) \Big\}.\end{aligned}$$

Clearly, a frequency $\omega_0$ belongs to a complete band gap of a two-dimensional photonic crystal, if and only if $\omega_0$ belongs to a TM band gap as well as to a TE band gap of that crystal.

Now, suppose that the relative electric permittivity function $\varepsilon_r$ is periodic with respect to a square lattice $\Lambda = a\mathbb{Z}^2$ with lattice constant $a > 0$, and that the primitive domain $\Omega$ is given by the square $(-a/2, a/2)^2$. We further assume that the point group of $\varepsilon_r$ is isomorphic to $\mathrm{O}_2(\mathbb{Z})$. Then, we have by Lemma 4.1 in [29] that the functions $\omega_j^{\mathrm{TM}}(\varepsilon_r|_\Omega, \cdot\,)$ and $\omega_j^{\mathrm{TE}}(1/\varepsilon_r|_\Omega, \cdot\,)$ are uniquely determined by their restrictions to a fundamental region of the point group. As in the three-dimensional setting, the closure of such a fundamental region is called an irreducible zone.

For every two-dimensional photonic crystals, whose medium structure is periodic with respect to a square lattice with lattice constant $a > 0$, and whose point group is isomorphic to $\mathrm{O}_2(\mathbb{Z})$, an irreducible zone is given by the triangle

$$K := \overline{\mathrm{conv}}\big\{\boldsymbol{k}_\Gamma, \boldsymbol{k}_X, \boldsymbol{k}_M\big\},$$

where the critical points $\boldsymbol{k}_\Gamma$, $\boldsymbol{k}_X$, and $\boldsymbol{k}_M$ are given by

$$\boldsymbol{k}_\Gamma := \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{k}_X := \begin{pmatrix} \pi/a \\ 0 \end{pmatrix}, \quad \boldsymbol{k}_M := \begin{pmatrix} \pi/a \\ \pi/a \end{pmatrix}.$$

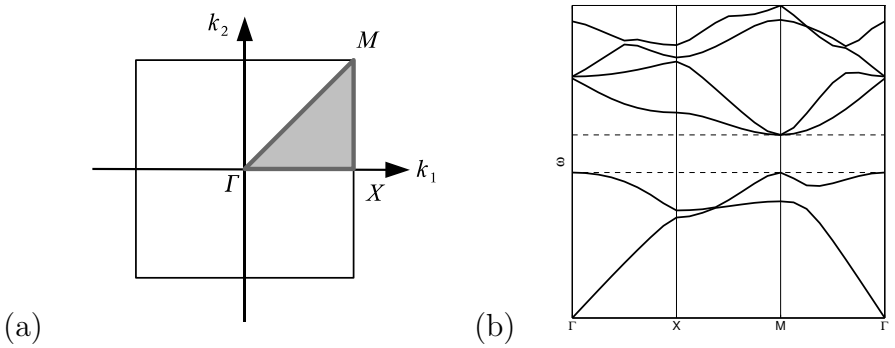



**Figure 4.2:** The first Brillouin zone with the irreducible zone (dark) and the critical points (a). The band diagram (b) shows the band structure along the critical path $\Gamma$–$X$–$M$–$\Gamma$. The dashed lines indicate a suspected band gap. Here, the space group is isomorphic to the semidirect product $\mathbb{Z}^2 \rtimes \mathrm{O}_2(\mathbb{Z})$.

The corresponding critical path is defined as the boundary of $K$. In Figure 4.2(a) we depicted the first Brillouin zone $\mathbb{B}$ and the irreducible zone $K$ with its critical path and points. An example of a corresponding band diagram is depicted in Figure 4.2(b).

# Chapter 5

# Photonic Band Structure Optimization

In this chapter we study some mathematical aspects of photonic band structure optimization problems (PBSOPs). Recall that a PBSOP consists in finding medium structures of photonic crystals, which are optimal with respect to certain features of the crystals' photonic band structures. In Section 5.1 we establish a general framework, which allows us to treat a large class of PBSOPs as minimization problems over an admissible set of essentially bounded functions. In Section 5.2 we prove important results concerning the regularity of the goal functionals, which typically appear in the minimization problems. In particular, we show that these goal functionals are locally Lipschitz continuous and, under certain assumptions, even Lipschitz continuous. The existence of optimal solutions is discussed in the following two sections. In Section 5.3 we show that PBSOPs, which involve TM band structures of a two-dimensional photonic crystals, admit optimal solutions in an extended admissible set. In Section 5.4 we comment on why this result could not be shown so far for PBSOPs that involve TE band structures of two-dimensional photonic crystals or band structures of three-dimensional photonic crystals. Finally, we present some concrete optimization problems in Section 5.5.

## 5.1 The Formal Setting

With the definitions of Section 4.1, let $\mathcal{M}(\Omega)$ be the system of Lebesgue measurable subsets of $\Omega$. Recall that $\Omega$ denotes a primitive cell of a Bravais lattice $\Lambda$ of rank 3. Furthermore, let $\rho_{\min}, \rho_{\max} \in \mathbb{R}$ be two given constants, which satisfy

$$0 < \rho_{\min} < \rho_{\max}. \tag{5.1}$$

We define the set

$$\mathcal{C} := \big\{\rho_{\min}(1-\chi_S) + \rho_{\max}\chi_S \;\big|\; S \in \mathcal{M}(\Omega)\big\}, \tag{5.2}$$

where $\chi_S$ denotes the *characteristic function* of a set $S \subseteq \Omega$. Clearly, the set $\mathcal{C}$ consists of all Lebesgue measurable functions on $\Omega$ with image in the set $\{\rho_0, \rho_1\}$. In particular, $\mathcal{C}$ is a proper subset of the function set $\mathcal{D} \subset \mathcal{E}$, which we introduced in Section 4.3. In the following, we shall refer to the set $C$ as the *admissible set*.

In view of three-dimensional photonic crystals, $\mathcal{C}$ represents the set of all possible $\Lambda$-periodic crystal structures (see Section 3.1). The constants $\rho_{\min}$ and $\rho_{\max}$ correspond to the reciprocal relative electric permittivities of two materials, the photonic crystals consists of. Given a function $\rho \in \mathcal{C}$, with $\rho = \rho_0(1-\chi_S) + \rho_1\chi_S$ and $S \in \mathcal{M}(\Omega)$, the set $S$ represents the region within the primitive domain $\Omega$, which is occupied by the material whose reciprocal relative electric permittivity is given by $\rho_{\max}$. The remaining region $\Omega \setminus S$ is hence occupied by the material with reciprocal relative electric permittivity $\rho_{\min}$. The elements of the set $\mathcal{C}$ will be referred to as *density functions*, henceforth.

As was discussed in Section 4.4, for every density $\rho \in \mathcal{C}$ and every vector $\boldsymbol{k} \in \mathbb{B}$ there exists an increasing sequence $\{\lambda_j(\rho, \boldsymbol{k})\}_{j\in\mathbb{N}}$ of non-negative, real eigenvalues of the eigenvalue problem

$$(\nabla + \mathrm{i}\boldsymbol{k}) \times \big[\rho(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}\big] = \lambda \boldsymbol{u} \quad \text{in } \Omega$$

for $\lambda \in \mathbb{C}$ and $\boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}} \setminus \{\boldsymbol{0}\}$. Given an index $j \in \mathbb{N}$, the $j$-th smallest eigenvalue can be viewed as a function $\lambda_j : \mathcal{D} \times \mathbb{B} \to \mathbb{R}$. According to Proposition 4.24 the function $\lambda_j(\rho, \cdot) : \mathbb{B} \to \mathbb{R}$ is continuous for every density function $\rho \in \mathcal{C}$. The band structure of a photonic crystal, whose medium structure is represented by the density function $\rho \in \mathcal{C}$, is given by the graphs of all functions $\omega_1(\rho, \cdot)$, $\omega_2(\rho, \cdot), \ldots$, etc., where $\omega_j = \sqrt{\lambda_j}$ for every $j \in \mathbb{N}$.

Many optical properties of a photonic crystal, such as its effective refractive index or its transmission spectrum are determined by its band structure. The band structure itself is a function of the photonic crystal structure, which in our setting is given by a density function $\rho$ belonging to the admissible set $\mathcal{C}$. Suppose that somebody wants to fabricate a three-dimensional photonic crystal with a specific optical property. Choosing two different materials, the question arises, how these two materials should be spatially arranged in order to obtain a photonic crystal with the desired optical property. In our setting the choice of the two materials determines the numbers $\rho_{\min}$ and $\rho_{\max}$, and thus the admissible set $\mathcal{C}$. Devising a spatial arrangement of the two materials corresponds to choosing a specific element of $\mathcal{C}$. Thus, the problem is to find an element in $\mathcal{C}$, i.e. a density function, which is optimal with respect to the desired optical property.

In the following, we assume that the desired optical property of the photonic crystal can be expressed in terms of the lowest $n$ bands of the crystal's band structure, where $n \in \mathbb{N}$ is some fixed number. We further assume that there exists a functional $J : \mathcal{D} \to \mathbb{R}$, such that the desired optical property is characterized by a global minimum of $J$. Such a functional is referred to as a *goal functional*, henceforth. In the following, we assume that every goal function $J$ is of the form

$$J(\rho) = \Upsilon\big(\omega_1(\rho, \cdot), \ldots, \omega_n(\rho, \cdot)\big) \quad \text{for all } \rho \in \mathcal{D}, \tag{5.3}$$

where

$$\Upsilon : \underbrace{C^0(\mathbb{B}, \mathbb{R}) \times \cdots \times C^0(\mathbb{B}, \mathbb{R})}_{n\text{-times}} \to \mathbb{R}$$

is a locally Lipschitz continuous functional, which valuates the first $n$ bands of a band structure with respect to the optimization goal. Recall that the notion of local Lipschitz continuity was introduced in Section 2.3. Notice that by Proposition 4.24 on page 79, we have indeed that the functions $\omega_1(\rho, \cdot), \ldots, \omega_n(\rho, \cdot)$ are continuous on $\mathbb{B}$ for every $\rho \in \mathcal{D}$.

With the above definitions, we arrive at the following optimization problem.

**Problem 5.1.** *Given a goal functional $J : D \to \mathbb{R}$ of the form* (5.3)*, find a density function $\rho_* \in \mathcal{C}$, such that*

$$J(\rho_*) \leq J(\rho) \quad \textit{for all } \rho \in \mathcal{C}.$$

Clearly, Problem 5.1 constitutes a minimization problem of the form

$$\operatorname*{minimize}_{\rho \in \mathcal{C}} \; J(\rho).$$

It should be noted that Problem 5.1 is, in fact, only a problem scheme, since the functional $\Upsilon$ was not defined so far. In the following, it is assumed that the term *photonic band structure optimization problem (PBSOP)* refers to an optimization problem, which fits into the scheme of Problem 5.1.

## 5.2 Regularity of the Goal Functionals

In this section we investigate the regularity of the goal functionals $J$ in Problem 5.1. Recall that these goal functionals are assumed to be of the form given in (5.3). Due to this, we are able to show that the goal functionals are locally Lipschitz continuous on $\mathcal{D}$ and, under further assumptions, even Lipschitz continuous.

The following proposition establishes the local Lipschitz continuity of the eigenvalues $\lambda_j(\rho, \boldsymbol{k})$ of the weakly formulated eigenvalue equation (4.42) on page 64 with respect to $\rho$.

**Theorem 5.2.** *For every index $j \in \mathbb{N}$ and every vector $\boldsymbol{k} \in \mathbb{B}$ the functional $\lambda_j(\,\cdot\,, \boldsymbol{k}) : \mathcal{D} \to \mathbb{R}$ is locally Lipschitz continuous.*

*Proof.* In the following we assume that $j \in \mathbb{N}$ and $\boldsymbol{k} \in \mathbb{B}$ are fixed. One easily verifies that the mapping

$$\rho \mapsto \operatorname*{ess\,inf}_{\Omega}(\rho)$$

is Lipschitz continuous as a functional on $\mathcal{E}$ with Lipschitz constant 1. This implies that the functions $\alpha(\,\cdot\,, \boldsymbol{k})$ and $\kappa(\,\cdot\,, \boldsymbol{k})$ are Lipschitz continuous on $\mathcal{D}$. Recall that the functions $\alpha$ and $\kappa$ were defined in (4.38) and (4.40) on page 63. Next, we chose an arbitrary coefficient $\rho_0 \in \mathcal{D}$. Letting

$$\delta_1 := \frac{1}{2} \operatorname*{ess\,inf}_{\Omega}(\rho_0),$$

we denote by $\mathcal{B}_{\delta_1}(\rho_0)$ the open ball in $\mathcal{E}$ with radius $\delta_1$ and center $\rho_0$. By definition of the function $\alpha$ and the estimate given in (4.23) on page 51, we have that

$$\begin{aligned} \alpha(\rho, \boldsymbol{k}) &\geq \alpha(\rho_0, \boldsymbol{k}) - \|\rho - \rho_0\|_{\Omega,\infty} \left(1 - \frac{|\boldsymbol{k}|}{\operatorname{diam}(\mathbb{B})}\right) \\ &\geq \frac{\delta_1}{4} > 0 \qquad \text{for all } \rho \in \mathcal{B}_{\delta_1}(\rho_0). \end{aligned}$$

Given two arbitrary coefficients $\rho_1, \rho_2 \in \mathcal{B}_{\delta_1}(\rho_0)$, we now consider the operators $G_{\boldsymbol{k}}(\rho_1)$ and $G_{\boldsymbol{k}}(\rho_2)$ defined by (4.51) on page 66. Let $\boldsymbol{f} \in \boldsymbol{Z}$ be an arbitrarily chosen function, and let $\boldsymbol{u} := L_{\boldsymbol{k}}(\rho_1)^{-1} M_{\boldsymbol{k}} \boldsymbol{f}$ and $\boldsymbol{v} := L_{\boldsymbol{k}}(\rho_2)^{-1} M_{\boldsymbol{k}} \boldsymbol{f}$. Recall that the operators $M_{\boldsymbol{k}}$ and $L_{\boldsymbol{k}}(\rho_1)$ were defined in (4.44) and (4.48) on page 65. It follows that $\boldsymbol{u}$ and $\boldsymbol{v}$ are vector fields in $\boldsymbol{V}_{\boldsymbol{k}}$. By the norm estimates (4.46) and (4.50) on page 65 and the above estimate for the function $\alpha$, we have that $\boldsymbol{u}$ and $\boldsymbol{v}$ satisfy $\|\boldsymbol{u}\|_{\operatorname{curl},\Omega} \|\boldsymbol{v}\|_{\operatorname{curl},\Omega} \leq (16/\delta_1^2) \|\boldsymbol{f}\|_{\Omega}^2$. Since $G_{\boldsymbol{k}}(\rho_1)$ is self-adjoint, we also have that $\langle G_{\boldsymbol{k}}(\rho_1) \boldsymbol{f}\,,\, \boldsymbol{f} \rangle_{\Omega}$ is a real number. Therefore, we obtain the identity

$$\begin{aligned} \langle G_{\boldsymbol{k}}(\rho_1) \boldsymbol{f}\,,\, \boldsymbol{f} \rangle_{\Omega} &= \langle I_{\boldsymbol{k}} L_{\boldsymbol{k}}(\rho_1)^{-1} M_{\boldsymbol{k}} \boldsymbol{f}\,,\, \boldsymbol{f} \rangle_{\Omega} = \langle L_{\boldsymbol{k}}(\rho_1)^{-1} M_{\boldsymbol{k}} \boldsymbol{f}\,,\, \boldsymbol{f} \rangle_{\Omega} \\ &= \langle \boldsymbol{u}\,,\, \boldsymbol{f} \rangle_{\Omega} = M_{\boldsymbol{k}} \boldsymbol{u}[\boldsymbol{f}] = M_{\boldsymbol{k}} \boldsymbol{f}[\boldsymbol{u}] \\ &= L_{\boldsymbol{k}}(\rho_2) \boldsymbol{v}[\boldsymbol{u}] = L_{\boldsymbol{k}}(\rho_2) \boldsymbol{u}[\boldsymbol{v}] \\ &= a_{\boldsymbol{k}}(\rho_2)(\boldsymbol{u}, \boldsymbol{v}) + \kappa(\rho_2, \boldsymbol{k})\, m(\boldsymbol{u}, \boldsymbol{v}). \end{aligned}$$

Similarly, we obtain that $\langle G_{\boldsymbol{k}}(\rho_2) \boldsymbol{f}\,,\, \boldsymbol{f} \rangle_{\Omega} = a_{\boldsymbol{k}}(\rho_1)(\boldsymbol{u}, \boldsymbol{v}) + \kappa(\rho_1, \boldsymbol{k}) m(\boldsymbol{u}, \boldsymbol{v})$, and from this we deduce the estimate

$$\begin{aligned} \big| \big\langle \big(G_{\boldsymbol{k}}(\rho_2) - G_{\boldsymbol{k}}(\rho_1)\big) \boldsymbol{f}\,,\, \boldsymbol{f} \big\rangle_{\Omega} \big| &\leq \big| a_{\boldsymbol{k}}(\rho_1)(\boldsymbol{u}, \boldsymbol{v}) - a_{\boldsymbol{k}}(\rho_2)(\boldsymbol{u}, \boldsymbol{v}) \big| \\ &\quad + \big| \kappa(\rho_1, \boldsymbol{k}) - \kappa(\rho_2, \boldsymbol{k}) \big|\, \big| m(\boldsymbol{u}, \boldsymbol{v}) \big|. \end{aligned}$$

Furthermore, by Lemma 4.3 on page 50, we have that

$$\begin{aligned}\big|a_{\boldsymbol{k}}(\rho_1)(\boldsymbol{u},\boldsymbol{v}) - a_{\boldsymbol{k}}(\rho_2)(\boldsymbol{u},\boldsymbol{v})\big| &= \big|(a_{\boldsymbol{k}}(\rho_1-\rho_2)(\boldsymbol{u},\boldsymbol{v})\big| \\ &\leq \beta_0(\boldsymbol{k})\|\rho_1-\rho_2\|_{\Omega,\infty}\|\boldsymbol{u}\|_{\mathrm{curl},\Omega}\|\boldsymbol{v}\|_{\mathrm{curl},\Omega}.\end{aligned}$$

Since $\kappa(\,\cdot\,,\boldsymbol{k})$ is Lipschitz continuous on $\mathcal{D}$, there exists a constant $c_1 > 0$, such that $|\kappa(\rho_1,\boldsymbol{k}) - \kappa(\rho_2,\boldsymbol{k})| \leq c_1\|\rho_1-\rho_2\|_{\Omega,\infty}$. Moreover, we have that $|m(\boldsymbol{u},\boldsymbol{v})| \leq \|\boldsymbol{u}\|_{\Omega}\|\boldsymbol{v}\|_{\Omega} \leq \|\boldsymbol{u}\|_{\mathrm{curl},\Omega}\|\boldsymbol{v}\|_{\mathrm{curl},\Omega}$, so that we finally obtain

$$\big|\big\langle \big(G_{\boldsymbol{k}}(\rho_2) - G_{\boldsymbol{k}}(\rho_1)\big)\boldsymbol{f}\,,\boldsymbol{f}\big\rangle_{\Omega}\big| \leq c_2\,\|\rho_1-\rho_2\|_{\Omega,\infty}\|\boldsymbol{f}\|_{\Omega}^2,$$

where $c_2 := 16(\beta_0(\boldsymbol{k}) + c_1)/\delta_1^2$. From this inequality we deduce that

$$\begin{aligned}\frac{\langle G_{\boldsymbol{k}}(\rho_1)\boldsymbol{f}\,,\boldsymbol{f}\rangle_{\Omega}}{\|\boldsymbol{f}\|_{\Omega}^2} &\geq \frac{\langle G_{\boldsymbol{k}}(\rho_2)\boldsymbol{f}\,,\boldsymbol{f}\rangle_{\Omega}}{\|\boldsymbol{f}\|_{\Omega}^2} - c_2\|\rho_1-\rho_2\|_{\Omega,\infty} \qquad \text{for all } \boldsymbol{f}\in\boldsymbol{Z}\setminus\{\boldsymbol{0}\},\\ \frac{\langle G_{\boldsymbol{k}}(\rho_1)\boldsymbol{f}\,,\boldsymbol{f}\rangle_{\Omega}}{\|\boldsymbol{f}\|_{\Omega}^2} &\leq \frac{\langle G_{\boldsymbol{k}}(\rho_2)\boldsymbol{f}\,,\boldsymbol{f}\rangle_{\Omega}}{\|\boldsymbol{f}\|_{\Omega}^2} + c_2\|\rho_1-\rho_2\|_{\Omega,\infty} \qquad \text{for all } \boldsymbol{f}\in\boldsymbol{Z}\setminus\{\boldsymbol{0}\}.\end{aligned}$$

By applying the Min-Max principle, stated in (4.54) on page 69, we find that the $j$-th largest positive eigenvalues of $G_{\boldsymbol{k}}(\rho_1)$ and $G_{\boldsymbol{k}}(\rho_2)$ satisfy $\mu_j(\rho_1,\boldsymbol{k}) \geq \mu_j(\rho_2,\boldsymbol{k}) - c_2\|\rho_1-\rho_2\|_{\Omega,\infty}$ and $\mu_j(\rho_1,\boldsymbol{k}) \leq \mu_j(\rho_2,\boldsymbol{k}) + c_2\|\rho_1-\rho_2\|_{\Omega,\infty}$ for every $j\in\mathbb{N}$. Combining both estimates we obtain the inequality $|\mu_j(\rho_1,\boldsymbol{k}) - \mu_j(\rho_2,\boldsymbol{k})| \leq c_2\|\rho_1-\rho_2\|_{\Omega,\infty}$. Since this inequality can be established for all coefficients $\rho_1,\rho_2\in\mathcal{B}_{\delta_1}(\rho_0)$, we have that the function $\mu_j(\,\cdot\,,\boldsymbol{k})$ is Lipschitz continuous on $\mathcal{B}_{\delta_1}(\rho_0)$. Setting

$$\delta_2 := \min\left\{\delta_1, \frac{\mu_j(\rho_0,\boldsymbol{k})}{2c_2}\right\}$$

we obtain that

$$\mu_j(\rho_1,\boldsymbol{k}) \geq \mu_j(\rho_0,\boldsymbol{k}) - c_2\|\rho_1-\rho_0\|_{\Omega,\infty} \geq \frac{\mu_j(\rho_0,\boldsymbol{k})}{2} > 0 \qquad \text{for all } \rho_1\in\mathcal{B}_{\delta_2}(\rho_0).$$

Hence, the function $\mu_j(\,\cdot\,,\boldsymbol{k})$ is uniformly bounded away from zero by $\mu_j(\rho_0,\boldsymbol{k})/2$ on $\mathcal{B}_{\delta_2}(\rho_0)$. Using the Mean Value Theorem, one easily verifies that the function $1/\mu_j(\,\cdot\,,\boldsymbol{k})$ is Lipschitz continuous on $\mathcal{B}_{\delta_2}(\rho_0)$. By (4.60) on page 70 the function $\lambda_j(\,\cdot\,,\boldsymbol{k})$ is given as the sum of the functions $1/\mu_j(\,\cdot\,,\boldsymbol{k})$ and $-\kappa(\,\cdot\,,\boldsymbol{k})$. Hence, $\lambda_j(\,\cdot\,,\boldsymbol{k})$ is Lipschitz continuous on $\mathcal{B}_{\delta_2}(\rho_0)$, which establishes the assertion of this theorem. □

We can prove an even stronger regularity result for restrictions of the functions $\lambda_j(\,\cdot\,,\boldsymbol{k})$ to a specific subset of $\mathcal{D}$. Before, however, we state the following, auxiliary result.

**Proposition 5.3.** *For every index $j \in \mathbb{N}$ and every vector $\boldsymbol{k} \in \mathbb{B}$ the functional $\lambda_j(\,\cdot\,, \boldsymbol{k}) : \mathcal{D} \to \mathbb{R}$ is monotonically increasing, i.e., for every two coefficients $\rho_1, \rho_2 \in \mathcal{D}$ that satisfy $\rho_1 \leq \rho_2$ almost everywhere on $\Omega$, we have*

$$\lambda_j(\rho_1, \boldsymbol{k}) \leq \lambda_j(\rho_2, \boldsymbol{k}).$$

*Proof.* Let $\rho_1, \rho_2 \in \mathcal{D}$, such that $\rho_1 \leq \rho_2$ almost everywhere on $\Omega$, and let $\boldsymbol{u} \in \boldsymbol{V_k}$. Then, it follows immediately form the definition of the sesquilinear forms $a_{\boldsymbol{k}}(\rho)$, which was given in (4.33) on page 59, that $a_{\boldsymbol{k}}(\rho_1)(\boldsymbol{u}, \boldsymbol{u}) \leq a_{\boldsymbol{k}}(\rho_2)(\boldsymbol{u}, \boldsymbol{u})$. Since this inequality can be established for arbitrary functions $\boldsymbol{u} \in \boldsymbol{V_k}$, we have that

$$\frac{a_{\boldsymbol{k}}(\rho_1)(\boldsymbol{u}, \boldsymbol{u})}{m(\boldsymbol{u}, \boldsymbol{u})} \leq \frac{a_{\boldsymbol{k}}(\rho_2)(\boldsymbol{u}, \boldsymbol{u})}{m(\boldsymbol{u}, \boldsymbol{u})} \qquad \text{for all } \boldsymbol{u} \in \boldsymbol{V_k} \setminus \{\boldsymbol{0}\}.$$

Using the Min-Max Principle stated in Proposition 4.20 on page 70, the above inequality implies that $\lambda_j(\rho_1, \boldsymbol{k}) \leq \lambda_j(\rho_2, \boldsymbol{k})$. □

So far we considered the $j$-th smallest eigenvalue $\lambda_j(\,\cdot\,, \boldsymbol{k})$ as a functional on the set $\mathcal{D}$ for given $j \in \mathbb{N}$ and given $\boldsymbol{k} \in \mathbb{N}$. Next, we consider the restrictions of this functional to the function set

$$\overline{\mathcal{C}} := \big\{ \rho \in \mathcal{E} \;\big|\; \rho_{\min} \leq \rho \leq \rho_{\max} \text{ almost everywhere in } \Omega \big\}, \tag{5.4}$$

where the constants $\rho_{\min}$ and $\rho_{\max}$ coincide with those defining the admissible set $\mathcal{C}$ according to (5.2). We shall refer to the set $\overline{\mathcal{C}}$ as the *extended admissible set*. Functions, which belong to $\overline{\mathcal{C}}$, are also referred to as density functions. One easily verifies that $\mathcal{C} \subset \overline{\mathcal{C}} \subset \mathcal{D}$. With the extended admissible set $\overline{\mathcal{C}}$ at hand, we can state the following theorem.

**Theorem 5.4.** *For every index $j \in \mathbb{N}$ and every vector $\boldsymbol{k} \in \mathbb{B}$ the functional $\lambda_j(\,\cdot\,, \boldsymbol{k})$ is Lipschitz continuous on $\overline{\mathcal{C}}$.*

*Proof.* Clearly, every density function $\rho \in \rho_{\min}$ is positive and uniformly bounded away from zero by $\rho_{\min}$ almost everywhere in $\Omega$. Therefore, we have that $\alpha(\rho, \boldsymbol{k}) > \rho_{\min}/2$ for all $\rho \in \overline{\mathcal{C}}$. By the same line of arguments as in the proof of Theorem 5.2, we deduce that the function $\mu_j(\,\cdot\,, \boldsymbol{k})$ is Lipschitz continuous on $\overline{\mathcal{C}}$. By Proposition 5.3 the corresponding function $\lambda_j(\cdot, \boldsymbol{k})$ is monotonically increasing on $\mathcal{D}$. One easily verifies, that the same holds true for the function $\kappa(\,\cdot\,, \boldsymbol{k})$. Since $1/\mu_j(\,\cdot\,, \boldsymbol{k})$ coincides with $\lambda_j(\,\cdot\,, \boldsymbol{k}) + \kappa(\,\cdot\,, \boldsymbol{k})$ by (4.60) on page 70, we conclude that $\mu_j(\,\cdot\,, \boldsymbol{k})$ is monotonically decreasing on $\overline{\mathcal{C}}$. Hence, $\mu_j(\,\cdot\,\boldsymbol{k})$ is uniformly bounded from below by $\mu_j(\rho_{\min}, \boldsymbol{k}) > 0$ on $\overline{\mathcal{C}}$. Because of this, the functional $1/\mu_j(\,\cdot\,, \boldsymbol{k})$, and hence also the functional $\lambda_j(\,\cdot\,, \boldsymbol{k})$, are Lipschitz continuous on $\overline{\mathcal{C}}$. □

We now relate the result of Theorem 5.2 to the regularity of the goal functional $J$ in Problem 5.1. According to (4.74) on page 80, the functionals $\omega_j$ are given by $\omega_j := \sqrt{\lambda_j}$ for every index $j \in \mathbb{N}$. Using the Mean-Value Theorem, one can show that the function $x \mapsto \sqrt{x}$ is locally Lipschitz continuous on $\mathbb{R}_{>0}$. Since the composition of two locally Lipschitz continuous functions is locally Lipschitz continuous again (see Section 2.3), we deduce that the functions $\omega_j(\,\cdot\,, \boldsymbol{k})$ are locally Lipschitz continuous for every index $j \in \mathbb{N}$ and every vector $\boldsymbol{k} \in \mathbb{B}$. Since the functional $\Upsilon$, defined in (5.3), is also locally Lipschitz continuous by assumption, we deduce the following corollary to Theorem 5.2.

**Corollary 5.5.** *The goal functional $J : \mathcal{D} \to \mathbb{R}$ in Problem 5.1 is locally Lipschitz continuous.*

Finally, we remark that under further assumptions on the functional $\Upsilon$, the goal functional $J$ can be shown to be Lipschitz continuous on the set $\overline{\mathcal{C}}$. Consider, for example, the case where the optimization problem only involves the first photonic band, and where the functional $\Upsilon : C^0(\mathbb{B}, \mathbb{R}) \to \mathbb{R}$ is given by

$$\Upsilon\big(f\big) = f(\boldsymbol{k}_0) \qquad \text{for all } f \in C^0(\mathbb{B}, \mathbb{R}),$$

for some prescribed vector $\boldsymbol{k}_0 \in \mathbb{B} \setminus \{\boldsymbol{0}\}$. According to Proposition 4.20 on page 70 and Proposition 5.3 we that

$$\lambda_1(\rho, \boldsymbol{k}_0) \geq \lambda_1(\rho_{\min}, \boldsymbol{k}_0) > 0 \qquad \text{for all } \rho \in \overline{\mathcal{C}}.$$

Hence, the function $\lambda_1(\,\cdot\,, \boldsymbol{k}_0)$ is uniformly bounded away from zero on the extended admissible set $\overline{\mathcal{C}}$ by $\lambda_1(\rho_{\min}, \boldsymbol{k}_0)$. Since $\lambda_1(\,\cdot\,, \boldsymbol{k}_0)$ is Lipschitz continuous on $\overline{\mathcal{C}}$ according to Theorem 5.4, it follows that the function $\omega_1(\,\cdot\,, \boldsymbol{k}_0) = \sqrt{\lambda_1(\,\cdot\,, \boldsymbol{k}_0)}$, and hence also the goal functional $J = \omega_1(\,\cdot\,, \boldsymbol{k}_0)$, are Lipschitz continuous on $\overline{\mathcal{C}}$.

## 5.3 Existence of Optima in the TM Setting

So far, we considered photonic band structure optimization problems (PBSOPs) for three-dimensional photonic crystals. In this section we shift the focus of our investigations to PBSOPs, which involve TM band structures of two-dimensional photonic crystals.

Throughout this section, we assume that $\Lambda \subset \mathbb{R}^2$ is a Bravais lattice of rank 2, and that $\Omega$ is a primitive domain of $\Lambda$. By $\mathbb{B}$ we denote the first Brillouin zone of $\Lambda$. In Section 4.7 we defined $\lambda_j^{\mathrm{TM}}(\rho, \boldsymbol{k})$ as the $j$-th smallest eigenvalue of the eigenvalue equation (4.83) on page 85 for given $\rho \in \mathcal{D}$ and given $\boldsymbol{k} \in \mathbb{B}$. The sequence of the corresponding eigenfunctions $\{u_j^{\mathrm{TM}}(\rho, \boldsymbol{k})\}_{j \in \mathbb{N}}$ forms a complete system of $V$. Recall

that $V$ denotes the complex Sobolev space $H^1_{\mathrm{per}}(\Omega)$. The system is orthogonal with respect to the inner product $\langle \cdot\,,\,\cdot \rangle_\rho$, which was defined by (4.87) on page 86.

In analogy to the three-dimensional setting defined in the Sections 5.1 and Sections 5.2, we define the function set $\mathcal{C}$ as

$$\mathcal{C} := \big\{ \rho_{\min}(1 - \chi_S) + \rho_{\max}\chi_S \;\big|\; S \in \mathcal{M}(\Omega) \big\},$$

and the function set $\overline{\mathcal{C}}$ as

$$\overline{\mathcal{C}} := \big\{ \rho \in \mathcal{E} \;\big|\; \rho_{\min} \leq \rho \leq \rho_{\max} \text{ almost everywhere in } \Omega \big\}, \tag{5.5}$$

for some constants $\rho_{\min}, \rho_{\max} \in \mathbb{R}$ satisfying $0 < \rho_{\min} < \rho_{\max}$. As in the three-dimensional setting, the sets $\mathcal{C}$ is called the *admissible set*, and the set $\overline{\mathcal{C}}$ is called the *extended admissible set*. Clearly, every function $\rho \in \mathcal{C}$ represents the medium structure of a two-dimensional photonic crystal consisting of exactly two different materials. The $\Lambda$-periodic relative electric permittivity function $\varepsilon_r$ of such a crystal is determined by $\varepsilon_r|_\Omega = \rho$.

Given some fixed number $n \in \mathbb{N}$, a functional $J^{\mathrm{TM}} : \mathcal{D} \to \mathbb{R}$, which is given by

$$J^{\mathrm{TM}}(\rho) = \Upsilon^{\mathrm{TM}}\big(\omega_1^{\mathrm{TM}}(\rho,\,\cdot\,), \ldots, \omega_n^{\mathrm{TM}}(\rho,\,\cdot\,)\big) \qquad \text{for all } \rho \in \mathcal{D}, \tag{5.6}$$

for some locally Lipschitz continuous functional

$$\Upsilon^{\mathrm{TM}} : \underbrace{C^0(\mathbb{B},\mathbb{R}) \times \cdots \times C^0(\mathbb{B},\mathbb{R})}_{n\text{-times}} \to \mathbb{R},$$

is called a *TM goal functional*. In analogy to the three-dimensional setting (see Section 5.1), we assume that the TM goal functional $J^{\mathrm{TM}}$ attains its minimum at some density function $\rho_* \in \mathcal{C}$, if and only if the TM band structure of the two-dimensional photonic crystal, whose medium structure is represented by $\rho_*$, exhibits some desired property. We then consider the following problem scheme.

**Problem 5.6.** *Given a TM goal functional $J^{\mathrm{TM}} : \mathcal{D} \to \mathbb{R}$ of the form* (5.6)*, find an density function $\rho_* \in \mathcal{C}$, such that*

$$J^{\mathrm{TM}}(\rho_*) \leq J^{\mathrm{TM}}(\rho) \quad \textit{for all } \rho \in \mathcal{C}.$$

Clearly, Problem 5.6 is an analogue of Problem 5.1 for two-dimensional photonic crystals, and by a similar line of argument as in Section 5.2, one can show that every TM goal functional is locally Lipschitz continuous on $\mathcal{D}$.

In the remainder of this section we shall address the question as to whether or not the optimization problem stated in Problem 5.6 admits a solution. We remark

that similar optimization problems arise in the design of frequency-optimized, two-composite membranes. Such problems have been studied extensively by Cox and McLaughling (cf. [31], [32]). It turns out, that one has to study such problems on the extended admissible set $\overline{\mathcal{C}}$ to prove the existence of solutions.

In the following, we intoduce a topology on the function space $\mathcal{E}$ that is weaker than the norm topology. Recall that $L^1(\Omega,\mathbb{R})^*$, the normed dual of the real Banach space $L^1(\Omega,\mathbb{R})$, is isometrically linearly isomorphic to $\mathcal{E} = L^\infty(\Omega,\mathbb{R})$ (see e.g. Theorem 4.12 in [4]). The canonical isometric, linear isomorphism $\Phi : L^\infty(\Omega,\mathbb{R}) \to L^1(\Omega,\mathbb{R})^*$ is given by

$$\Phi(\xi)[f] := \int_\Omega \xi f \qquad \text{for all } \xi \in \mathcal{E},\ f \in L^1(\Omega,\mathbb{R}). \tag{5.7}$$

The *weak-∗-topology on* $L^1(\Omega,\mathbb{R})^*$ is defined as the weakest topology on $L^1(\Omega,\mathbb{R})^*$, with respect to which all continuous linear functionals $\varphi : L^1(\Omega,\mathbb{R})^* \to \mathbb{R}$, that are given by $\varphi(p) = p(f)$ for some $f \in L^1(\Omega,\mathbb{R})$ and for all $p \in L^1(\Omega,\mathbb{R})^*$, are continuous. By virtue of the isomorphism $\Phi$ every open set in that topology can be identified with a subset of $\mathcal{E}$. Thus, the weak-∗-topology on $L^1(\Omega,\mathbb{R})^*$ also defines a topology on $\mathcal{E}$ which is called the *weak-∗-topology on* $\mathcal{E}$. A sequence in $\mathcal{E}$, which is convergent with respect to this topology, is called *weak-∗-convergent* in $\mathcal{E}$. One easily verifies that a sequence $\{\xi^{(l)}\}_{l\in\mathbb{N}}$ in $\mathcal{E}$ is weak-∗-convergent in $\mathcal{E}$, if and only if there exists a function $\xi \in \mathcal{E}$, such that

$$\Phi\big(\xi^{(l)} - \xi\big)[f] = \int_\Omega \big(\xi^{(l)} - \xi\big) f \to 0 \quad \text{as } l \to \infty \qquad \text{for all } f \in L^1(\Omega,\mathbb{R}).$$

In this case we shall say that the sequence $\{\xi^{(l)}\}_{l\in\mathbb{N}}$ weak-∗-converges to $\xi$. Obviously, convergence in $\mathcal{E}$, i.e., convergence with respect to the norm topology on $\mathcal{E}$, also implies weak-∗-convergence in $\mathcal{E}$. The converse, however, does not hold in general. Hence, the norm topology on $\mathcal{E}$ is strictly stronger than the weak-∗-topology on $\mathcal{E}$.

In the following, terms like weak-∗-compactness or weak-∗-continuity always refer to the respective properties with respect to the weak-∗-topology on $\mathcal{E}$. The following proposition states in particular that the extended admissible set $\overline{\mathcal{C}}$ is compact with respect to that topology.

**Proposition 5.7.** *The set* $\overline{\mathcal{C}}$ *is convex, bounded, and weak-∗-compact.*

*Proof.* Convexity and boundedness directly follow from the definition of $\overline{\mathcal{C}}$ in (5.5). It hence remains to show the weak-∗-compactness of $\overline{\mathcal{C}}$. To this end, we denote by $\overline{\mathcal{B}_1(0)}$ the closed unit ball in $\mathcal{E}$ and define the operator $\Psi : \overline{\mathcal{C}} \to \overline{\mathcal{B}_1(0)}$ by

$$\Psi(\rho) := \frac{2\rho - \rho_{\max} - \rho_{\min}}{\rho_{\max} - \rho_{\min}} \qquad \text{for all } \rho \in \overline{\mathcal{C}}.$$

One easily verifies that $\Psi$ is a bijection. Choosing an arbitrary sequence $\{\rho^{(l)}\}_{l\in\mathbb{N}}$ in $\overline{\mathcal{C}}$, we define the sequence $\{\xi^{(l)}\}_{l\in\mathbb{N}}$ in $\overline{\mathcal{B}_1(0)}$ by $\xi^{(l)} := \Psi(\rho^{(l)})$ for all $l \in \mathbb{N}$, as well as the sequence $\{p^{(l)}\}_{l\in\mathbb{N}}$ in $L^1(\Omega,\mathbb{R})^*$ by $p^{(l)} := \Phi(\xi^{(l)})$ for all $l \in \mathbb{N}$. Here, $\Phi$ denotes the isometric linear isomorphism given by (5.7). Since

$$\|p^{(l)}\|_{L^1(\Omega,\mathbb{R})^*} = \|\Psi(\rho^{(l)})\|_{\Omega,\infty} \leq 1 \qquad \text{for all } l \in \mathbb{N},$$

we have that $\{p^{(l)}\}_{l\in\mathbb{N}}$ is a sequence in the the closed unit ball $\overline{B_1^*(0)}$ in $L^1(\Omega,\mathbb{R})^*$. According to Alaoglu's Theorem (see e.g. Theorem 13.9 [40]), $\overline{B_1^*(0)}$ is compact with respect to the weak-$*$-topology on $L^1(\Omega,\mathbb{R})^*$. Hence, there exists a subsequence $\{p^{(1,l)}\}_{l\in\mathbb{N}}$ of $\{p^{(l)}\}_{l\in\mathbb{N}}$ and a continuous linear functional $p \in L^1(\Omega,\mathbb{R})^*$ such that

$$(p^{(1,l)} - p)(f) \to 0 \quad \text{as } l \to \infty \qquad \text{for all } f \in L^1(\Omega,\mathbb{R}).$$

Letting $\xi := \Phi^{-1}(p)$ and $\rho := \Psi^{-1}(\xi)$, we have that

$$\int_\Omega (\rho^{(1,l)} - \rho) f = \frac{\rho_{\max} - \rho_{\min}}{2} \int_\Omega (\xi^{(1,l)} - \xi) f = (p^{(1,l)} - p)(f) \qquad \text{for all } l \in \mathbb{N}.$$

Hence, $\{\rho^{(1,l)}\}_{l\in\mathbb{N}}$ is a weak-$*$-convergent subsequence of $\{\rho^{(l)}\}_{l\in\mathbb{N}}$. Since $\{\rho^{(l)}\}_{l\in\mathbb{N}}$ was chosen arbitrarily, it follows that $\overline{\mathcal{C}}$ is compact with respect to the weak-$*$-topology on $\mathcal{E}$. □

Next, we study the relationship between the sets $\mathcal{C}$ and $\overline{\mathcal{C}}$ with respect to the weak-$*$-topology on $\mathcal{E}$. The following theorem is a particularization of Theorem 3 in [72].

**Theorem 5.8.** *Let $R$ be a non-empty subset of $\mathbb{R}$.*

(a) *Let $\{\xi^{(l)}\}_{l\in\mathbb{N}}$ be a sequence in $\mathcal{E}$, such that $\{\xi^{(l)}\}_{l\in\mathbb{N}}$ weak-$*$-converges to a function $\xi \in \mathcal{E}$, and such that $\xi^{(l)}(\boldsymbol{x}) \in R$ for almost all $\boldsymbol{x} \in \Omega$. Then, $\xi(\boldsymbol{x}) \in \overline{\text{conv}}(R)$ for almost all $\boldsymbol{x} \in \Omega$.*

(b) *Conversely, let $\xi \in \mathcal{E}$ be a function, such that $\xi(\boldsymbol{x}) \in \overline{\text{conv}}(R)$ for almost all $\boldsymbol{x} \in \Omega$. Then, there exists a sequence $\{\xi^{(l)}\}_{l\in\mathbb{N}}$ in $\mathcal{E}$, such that $\xi^{(l)}(\boldsymbol{x}) \in R$ for almost all $\boldsymbol{x} \in \Omega$ and all $l \in \mathbb{N}$, and such that $\{\xi^{(l)}\}_{l\in\mathbb{N}}$ weak-$*$-converges to $\xi$.*

Choosing $R = \{\rho_{\min}, \rho_{\max}\}$ in Theorem 5.8, we obtain that every weak-$*$-limit function of a weak-$*$-convergent sequence in the admissible set $\mathcal{C}$ takes function values in the closed interval $[\rho_{\min}, \rho_{\max}]$ almost everywhere on $\Omega$. Therefore, every such limit function belongs to the extende admissible set $\overline{\mathcal{C}}$. Conversely, for every

function in the extended admissible $\overline{\mathcal{C}}$ there exists a sequence in the admissible set $\mathcal{C}$, which weak-$*$-converges to that function. Hence, we have the following corollary to Theorem 5.8.

**Corollary 5.9.** *The set $\overline{\mathcal{C}}$ is the weak-$*$-closure of $\mathcal{C}$.*

By definition the TM goal functional $J^{\mathrm{TM}}$ is locally Lipschitz continuous and hence continuous on $\mathcal{D}$ with respect to the norm topology on $\mathcal{E}$. Since the norm topology is strictly stronger than the weak-$*$-topology on $\mathcal{E}$, the weak-$*$-compactness of the extended admissible set $\overline{\mathcal{C}}$ does not imply that the TM goal functional attains a minimum on that set. We first need to establish weak-$*$-continuity for the TM goal functional, in order to obtain such a result. The following lemma is an auxiliary result and follows from Lemma 4.2 in [31]. Recall that $Z$ denotes the complex Lebesgue space $L^2(\Omega)$.

**Lemma 5.10.** *Let $\{\rho^{(l)}\}_{l\in\mathbb{N}}$ be a sequence in $\overline{\mathcal{C}}$ that weak-$*$-converges to $\rho \in \overline{\mathcal{C}}$, and let $\{f^{(l)}\}_{l\in\mathbb{N}}$ and $\{g^{(l)}\}_{l\in\mathbb{N}}$ be sequences in $Z$ that converge in $Z$ to $f \in Z$ and $g \in Z$, respectively. Then, we have that*

$$m^{\mathrm{TM}}(\rho^{(l)})(f^{(l)}, g^{(l)}) \to m^{\mathrm{TM}}(\rho)(f, g) \qquad \text{as } l \to \infty.$$

The following theorem states that the functions $\lambda_j^{\mathrm{TM}}(\,\cdot\,, \boldsymbol{k})$ are weak-$*$-continuous.

**Theorem 5.11.** *Let $\{\rho^{(l)}\}_{l\in\mathbb{N}}$ be a sequence in $\overline{\mathcal{C}}$, which weak-$*$-converges to a coefficient $\rho \in \overline{\mathcal{C}}$. Then, we have for every index $j \in \mathbb{N}$, and every vector $\boldsymbol{k} \in \mathbb{B}$ that the sequence $\{\lambda_j^{\mathrm{TM}}(\rho^{(l)}, \boldsymbol{k})\}_{l\in\mathbb{N}}$ converges to $\lambda_j^{\mathrm{TM}}(\rho, \boldsymbol{k})$.*

*Proof.* This proof is an adaptation of the proof of Proposition 4.3(i) in [31]. Let $\boldsymbol{k} \in \mathbb{B}$ be fixed, and let $\{\rho^{(l)}\}_{l\in\mathbb{N}}$ be a sequence in $\overline{\mathcal{C}}$ that weak-$*$-converges to a density function $\rho \in \overline{\mathcal{C}}$. In order to simplify notations, we define ${\lambda_j}^{(l)} := \lambda_j(\rho^{(l)}, \boldsymbol{k})$ and ${u_j}^{(l)} := u_j^{\mathrm{TM}}(\rho^{(l)}, \boldsymbol{k})$, as well as $\lambda_j := \lambda_j^{\mathrm{TM}}(\rho, \boldsymbol{k})$ and $u_j := u_j^{\mathrm{TM}}(\rho, \boldsymbol{k})$ for all $j, l \in \mathbb{N}$. Without loss of generality we assume that the sequences $\{{u_j}^{(l)}\}_{j\in\mathbb{N}}$ and $\{u_j\}_{j\in\mathbb{N}}$ form a complete system of $V$, which is orthonormal with respect to the inner product $\langle\,\cdot\,,\,\cdot\,\rangle_\rho$.

Using the Min-Max Principle (4.88) on page 86, one can show that the mapping $\lambda_j^{\mathrm{TM}}(\,\cdot\,, \boldsymbol{k})$ is monotonically decreasing for all $j \in \mathbb{N}$ and all $\boldsymbol{k} \in \mathbb{B}$. Hence, for every index $j \in \mathbb{N}$, the sequence $\{{\lambda_j}^{(l)}\}_{l\in\mathbb{N}}$ is bounded from above by $\lambda_j^{\mathrm{TM}}(\rho_{\min}, \boldsymbol{k})$ and from below by $\lambda_j^{\mathrm{TM}}(\rho_{\max}, \boldsymbol{k})$. Since the sesquilinear forms $a_{\boldsymbol{k}}^{\mathrm{TM}}$ are coercive, we have that

$$\alpha^{\mathrm{TM}}(\boldsymbol{k})\, \|u_j^{(l)}\|_{1,\Omega}^2 \leq a_{\boldsymbol{k}}^{\mathrm{TM}}(u_j^{(l)}, u_j^{(l)}) + \kappa^{\mathrm{TM}}(\boldsymbol{k}) \|u_j^{(l)}\|_{\Omega}^2$$

$$\begin{aligned}
&\leq \lambda_j^{\mathrm{TM}}(\rho^{(l)}, \boldsymbol{k}) + \kappa^{\mathrm{TM}}(\boldsymbol{k}) \int_\Omega \frac{\rho^{(l)}}{\rho_{\min}} |u_j^{(l)}|^2 \\
&\leq \lambda_j^{\mathrm{TM}}(\rho_{\min}, \boldsymbol{k}) + \frac{\kappa^{\mathrm{TM}}(\boldsymbol{k}) \|u_j^{(l)}\|_\rho^2}{\rho_{\min}} \\
&= \lambda_j^{\mathrm{TM}}(\rho_{\min}, \boldsymbol{k}) + \frac{\kappa^{\mathrm{TM}}(\boldsymbol{k}) \|u_j^{(l)}\|_\rho^2}{\rho_{\min}} \qquad \text{for all } j, l \in \mathbb{N}.
\end{aligned}$$

Recall that the functions $\alpha^{\mathrm{TM}}$ and $\kappa^{\mathrm{TM}}$ were defined by (4.84) and (4.84) on page 85, respectively. From the above estimate we deduce that, for every index $j \in \mathbb{N}$, the sequence $\{u_j{}^{(l)}\}_{l\in\mathbb{N}}$ is bounded in $V$.

It follows that there exists a subsequence $\{\rho^{(1,l)}\}_{l\in\mathbb{N}}$ of $\{\rho^{(l)}\}_{l\in\mathbb{N}}$, such that the corresponding subsequence $\{\lambda_1{}^{(1,l)}\}_{l\in\mathbb{N}}$ of $\{\lambda_1{}^{(l)}\}_{l\in\mathbb{N}}$ converges to a non-negative, real number $\nu_1 \in \mathbb{R}_{\geq 0}$, and such that the subsequence $\{u_1{}^{(1,l)}\}_{l\in\mathbb{N}}$ of $\{u_1{}^{(l)}\}_{l\in\mathbb{N}}$ converges weakly in $V$ to a function $w_1 \in V$. The existence of the first subsequence follows from the Theorem of Bolzano–Weierstrass. The existence of the second subsequence follows from the fact that every bounded subset of a Hilbert space is weakly sequentially compact (see e.g. Theorem 6.9 in [4]). Since the identical embedding from $H^1(\Omega)$ into $Z$ is compact, we also have that the subsequence $\{u_1{}^{(1,l)}\}_{l\in\mathbb{N}}$ converges strongly in $Z$ to $w_1$ (see e.g. Lemma 8.2 in [4]). By the same reasoning, we can extract a second subsequence $\{\rho^{(2,l)}\}_{l\in\mathbb{N}}$ of $\{\rho^{(1,l)}\}_{l\in\mathbb{N}}$, such that $\{\lambda_2{}^{(2,l)}\}_{l\in\mathbb{N}}$ converges to a non-negative, real number $\nu_2 \in \mathbb{R}_{\geq 0}$, and such that $\{u_2{}^{(2,l)}\}_{l\in\mathbb{N}}$ converges weakly in $V$ and strongly in $Z$ to a function $w_2 \in V$. By repeating this procedure, we finally obtain a sequence $\{\{\rho^{(j,l)}\}_{l\in\mathbb{N}}\}_{j\in\mathbb{N}}$ of sequences in $\overline{\mathcal{C}}$ with the property that $\{\rho^{(j+1,l)}\}_{l\in\mathbb{N}}$ is a subsequence of $\{\rho^{(j,l)}\}_{l\in\mathbb{N}}$ for all $j \in \mathbb{N}$. Furthermore, we have for every index $j \in \mathbb{N}$ that the sequence $\{\lambda_j{}^{(j,l)}\}_{l\in\mathbb{N}}$ converges to a non-negative, real number $\nu_j \in \mathbb{R}_{\geq 0}$, and that the sequence $\{u_j{}^{(j,l)}\}_{j\in\mathbb{N}}$ converges weakly in $V$ and strongly in $Z$ to a function $w_j \in V$. We now consider the "diagonal" subsequence $\{\rho^{(l,l)}\}_{l\in\mathbb{N}}$ of $\{\rho^{(l)}\}_{l\in\mathbb{N}}$. It follows from the discussion above, that for every index $j \in \mathbb{N}$ the sequence $\{\lambda_j{}^{(l,l)}\}_{l\in\mathbb{N}}$ converges to $\nu_j$. Furthermore, we have that the sequence $\{u_j{}^{l,l}\}_{l\in\mathbb{N}}$ converges weakly in $V$ and strongly in $Z$ to $w_j$.

By Lemma 5.10, we have that the sequence $\{w_j\}_{j\in\mathbb{N}}$ forms an orthonormal system with respect to the inner product $\langle \cdot\ , \cdot \rangle_\rho$, since

$$\begin{aligned}
\delta_{ij} &= \langle u_i^{(l,l)}\ , u_j^{(l,l)} \rangle_{\rho^{(l)}} = m^{\mathrm{TM}}(\rho^{(l,l)})(u_i^{(l,l)}, u_j^{(l,l)}) \\
&\to m^{\mathrm{TM}}(\rho)(w_i, w_j) = \langle w_i\ , w_j \rangle_\rho \quad \text{as } l \to \infty \qquad \text{for all } i, j \in \mathbb{N}.
\end{aligned}$$

Furthermore, since $\{\lambda_j{}^{(l,l)}\}_{j\in\mathbb{N}}$ is an increasing sequence for all $l \in \mathbb{N}$, we also have that $\{\nu_j\}_{j\in\mathbb{N}}$ is an increasing sequence. Since every sequence $\{\lambda_j{}^{(l,l)}\}_{j\in\mathbb{N}}$ is unbounded, we also have that $\nu_j \to \infty$ as $j \to \infty$.

Our next aim is to show that for every index $j \in \mathbb{N}$ there exists an index $i \in \mathbb{N}$, such that $\nu_j = \lambda_i$. We choose $j \in \mathbb{N}$ arbitrarily. Due to the weak converge of the sequence $\{u_j^{(l,l)}\}_{l\in\mathbb{N}}$ in $V$, we have that

$$a_{\boldsymbol{k}}^{\mathrm{TM}}(u_j^{(l,l)}, v) \to a_{\boldsymbol{k}}^{\mathrm{TM}}(w_j, v) \quad \text{as } l \to \infty \qquad \text{for all } v \in V.$$

By Lemma 5.10, we also have that

$$m^{\mathrm{TM}}(\rho^{(l,l)})(u_j^{(l,l)}, v) \to m^{\mathrm{TM}}(\rho)(w_j, v) \quad \text{as } l \to \infty. \qquad \text{for all } v \in V.$$

Since $\lambda_j^{(l,l)} \to \nu_j$ as $l \to \infty$, we thus obtain

$$a_{\boldsymbol{k}}^{\mathrm{TM}}(\rho)(w_j, v) = \nu_j m(w_j, v) \qquad \text{for all } v \in V,$$

It follows that $\nu_j$ is an eigenvalue, and that $w_j$ is a corresponding eigenfunction. Hence, there exists an index $i \in \mathbb{N}$, such that $\nu_j = \lambda_i$.

Next, we show that for every index $i \in \mathbb{N}$ there exists an index $j \in \mathbb{N}$, such that $\lambda_i = \nu_j$. We proof this by contradiction. Suppose that there exists an index $i \in \mathbb{N}$, such that $\lambda_i \neq \nu_j$ for all $j \in \mathbb{N}$. Then, we define

$$\widetilde{\lambda} := \lambda_i, \quad \widetilde{u} := \frac{u_i}{\widetilde{\lambda} + \kappa^{\mathrm{TM}}(\boldsymbol{k})}.$$

Given an arbitrary index $j \in \mathbb{N}$, we find that

$$\nu_j \, m^{\mathrm{TM}}(\rho)(w_j, \widetilde{u}) = a_{\boldsymbol{k}}^{\mathrm{TM}}(w_j, \widetilde{u}) = \overline{a_{\boldsymbol{k}}^{\mathrm{TM}}(\widetilde{u}, w_j)} = \widetilde{\lambda}\, \overline{m^{\mathrm{TM}}(\rho)(\widetilde{u}, w_j)}.$$

Since $\widetilde{\lambda} \neq \nu_j$ for all $j \in \mathbb{N}$ by assumption, we conclude that $m^{\mathrm{TM}}(\rho)(\widetilde{u}, w_j) = \langle \widetilde{u}\,, w_j \rangle_\rho = 0$ for all $j \in \mathbb{N}$. Because the sequence $\{u_j{}^{(l,l)}\}_{l\in\mathbb{N}}$ converges strongly in $Z$ to $w_j$ for every index $j \in \mathbb{N}$, we have by Lemma 5.10 that

$$P_{\boldsymbol{k},j-1}^{\mathrm{TM}}(\rho^{(l,l)})[\widetilde{u}] = \sum_{i=1}^{j-1} \langle \widetilde{u}\,, u_i^{(l,l)} \rangle_{\rho^{(l,l)}} u_i^{(l,l)} \to \sum_{i=1}^{j-1} \langle \widetilde{u}\,, w_i \rangle_\rho w_i = 0 \quad \text{as } l \to \infty.$$

By Auchmuty's principle, stated in (4.91) on page 86, we have that

$$\frac{-1}{2\big(\lambda_j^{(l,l)} - \kappa^{\mathrm{TM}}(\boldsymbol{k})\big)} \leq \left( \frac{1}{2} \left( a_{\boldsymbol{k}}^{\mathrm{TM}}(\widetilde{u}, \widetilde{u}) + \kappa^{\mathrm{TM}}(\boldsymbol{k})\, m^{\mathrm{TM}}(\rho^{(l,l)})(\widetilde{u}, \widetilde{u}) \right) \right.$$
$$\left. - \left\| \widetilde{u} - P_{\boldsymbol{k},j-1}^{\mathrm{TM}}(\rho^{(l,l)}) \widetilde{u} \right\|_{\rho^{(l,l)}} \right) \qquad \text{for all } j, l \in \mathbb{N}.$$

Letting $l$ tend to infinity in the above inequality, we obtain

$$\frac{-1}{2\big(\nu_j - \kappa^{\mathrm{TM}}(\boldsymbol{k})\big)} \leq \left( \frac{1}{2} \left( a_{\boldsymbol{k}}^{\mathrm{TM}}(\widetilde{u}, \widetilde{u}) + \kappa^{\mathrm{TM}}(\boldsymbol{k})\, m^{\mathrm{TM}}(\rho)(\widetilde{u}, \widetilde{u}) \right) \right.$$

$$
\begin{aligned}
&\quad - \left\| \widetilde{u} - P^{\mathrm{TM}}_{\boldsymbol{k},j-1}(\rho)\widetilde{u} \right\|_\rho \Big) \\
&= \left( \frac{1}{2} \left( \widetilde{\lambda} + \kappa^{\mathrm{TM}}(\boldsymbol{k}) \right) \left\| \widetilde{u} \right\|_\rho^2 - \left\| \widetilde{u} \right\|_\rho \right) \\
&= \frac{-1}{2\big(\widetilde{\lambda} - \kappa^{\mathrm{TM}}(\boldsymbol{k})\big)} \qquad \text{for all } j \in \mathbb{N}.
\end{aligned}
$$

Notice that this implies $\nu_j \leq \widetilde{\lambda}$ for all $j \in \mathbb{N}$, which is a contradiction to the fact that $\nu_j \to \infty$ as $j \to \infty$. We thus have that for every $i \in \mathbb{N}$ there exists a $j \in \mathbb{N}$, such that $\nu_i = \lambda_j$. It follows that the sequences $\{\nu_j\}_{j\in\mathbb{N}}$ and $\{\lambda_j^{\mathrm{TM}}(\rho, \boldsymbol{k})\}_{j\in\mathbb{N}}$ coincide. □

Clearly, the weak-$*$-continuity of the functions $\lambda_j^{\mathrm{TM}}(\,\cdot\,, \boldsymbol{k})$ also implies the weak-$*$-continuity of the functions $\omega_j^{\mathrm{TM}}(\,\cdot\,, \boldsymbol{k})$ on $\overline{\mathcal{C}}$, where $j \in \mathbb{N}$. It thus follows from (5.6) that the TM goal functional $J^{\mathrm{TM}}$ is also weak-$*$-continuous on $\overline{\mathcal{C}}$. Since $\overline{\mathcal{C}}$ is weak-$*$-compact according to Proposition 5.7, we obtain the following corollary to Theorem 5.11.

**Corollary 5.12.** *The TM goal functional $J^{\mathrm{TM}}$ in Problem 5.6 attains a minimum on the extended admissible set $\overline{\mathcal{C}}$.*

It should be noted that Corollary 5.12 only guarantees the existence of optimal solutions in the extended admissible set $\overline{\mathcal{C}}$. In view of PBSOPs problems, it would be desirable to have optimal solutions in the admissible set $\mathcal{C}$. This is because only solutions which belong to the admissible set represent two-valued relative relectric permittivity functions. According to Proposition 5.7, the extended admissible set $\overline{\mathcal{C}}$ is a convex set. It can be shown that the set of extremal points of $\overline{\mathcal{C}}$ is precisely the admissible set $\mathcal{C}$ (see Proposition 2.5 in [31]). Therefore, every convex and weak-$*$-continuous functional on $\overline{\mathcal{C}}$ attains its minimum on $\mathcal{C}$. In Section 5.5 we introduce a special class of TM goal functionals. Unfortunately, we could not show that these functionals are convex. The numerical results we present in Chapter 9 indicate, however, that these functionals indeed attain their minima on the admissible set.

## 5.4 Existence of Optima in Other Settings

In the previous section, we showed that PBSOPs involving the TM band structure of a two-dimensional photonic crystal admit an optimal solution in the extended admissible set $\overline{\mathcal{C}}$. This set consists of all functions in $\mathcal{D}$, which are essentially bounded from above and from below by the constants $\rho_{\min}$ and $\rho_{\max}$, respectively.

In this section we comment on why an analogous result could not be shown so far either for PBSOPs involving the TE band structure of a two-dimensional photonic crystal, or for PBSOPs involving the band structure of a three-dimensional photonic crystal.

In the following, we assume that $\Omega$ is a primitive domain of a Bravais lattice $\Lambda \subset \mathbb{R}^2$ of rank 2, and the $\mathbb{B}$ is the first Brillouin zone of $\Lambda$. Recall that $\lambda_j^{\mathrm{TE}}(\rho, \boldsymbol{k})$ denotes the $j$-th smallest eigenvalue of the eigenvalue equation (4.94) on page 87 for every coefficient $\rho \in \mathcal{D}$ and every vector $\boldsymbol{k} \in \mathbb{B}$. If the coefficient $\rho$ belongs to the admissible set $\mathcal{C}$, which was defined in (5.3), then it represents the $\Lambda$-periodic relative electric permittivity function $\varepsilon_r$ of a two-dimensional photonic crystal according to $\rho = 1/\varepsilon_r|_\Omega$.

Given some fixed number $n \in \mathbb{N}$, we shall call a functional $J^{\mathrm{TE}} : \mathcal{D} \to \mathbb{R}$ a *TE goal functional* if it is of the form

$$J^{\mathrm{TE}}(\rho) = \Upsilon^{\mathrm{TE}}\big(\omega_1^{\mathrm{TE}}(\rho, \cdot), \ldots, \omega_n^{\mathrm{TE}}(\rho, \cdot)\big) \qquad \text{for all } \rho \in \mathcal{D}, \tag{5.8}$$

where

$$\Upsilon^{\mathrm{TE}} : \underbrace{C^0(\mathbb{B}, \mathbb{R}) \times \cdots \times C^0(\mathbb{B}, \mathbb{R})}_{n\text{-times}} \to \mathbb{R},$$

is a locally Lipschitz continuous functional. Supposing that a minimizer $\rho_* \in \mathcal{C}$ of the TE goal functional $J^{\mathrm{TE}}$ represents the relative electric permittivity function of a two-dimensional photonic crystal, whose TE band structure exhibits some desired property, we consider the following scheme of optimization problems.

**Problem 5.13.** *Given a TE goal functional $J^{\mathrm{TE}} : \mathcal{D} \to \mathbb{R}$ of the form (5.8), find an density function $\rho_* \in \mathcal{C}$, such that*

$$J^{\mathrm{TE}}(\rho_*) \leq J^{\mathrm{TE}}(\rho) \quad \textit{for all } \rho \in \mathcal{C}.$$

One can show that every TE goal functional is locally Lipschitz continuous on the extended admissible set $\overline{\mathcal{C}}$, which was defined by (5.5). The question is, whether or not one can show the existence of optimal solutions in $\overline{\mathcal{C}}$, as for Problem 5.13.

Recall that the existence of optimal solutions of Problem 5.13 followed from the weak-$*$-compactness of the extended admissible set $\overline{\mathcal{C}}$ and the weak-$*$-continuity of the TM eigenvalue functions $\lambda_j^{\mathrm{TM}}(\,\cdot\,, \boldsymbol{k})$. The latter result was established by Theorem 5.11 in the previous section. The proof of Theorem 5.11 relies on the result stated by Lemma 5.10, namely that for arbitrary sequences $\{\rho^{(l)}\}_{l\in\mathbb{N}}$, $\{f^{(l)}\}_{l\in\mathbb{N}}$ and $\{g^{(l)}\}_{l\in\mathbb{N}}$ in $\overline{\mathcal{C}}$, $Z$, and $Z$, the convergence of $m^{\mathrm{TM}}(\rho^{(l)})(f^{(l)}, g^{(l)})$ to $m^{\mathrm{TM}}(\rho)(f, g)$ as $l$ tends to infinity holds, if the sequence $\{\rho^{(l)}\}_{l\in\mathbb{N}}$ weak-$*$-converges in $\mathcal{E}$ to $\rho \in \overline{\mathcal{C}}$, and if the sequences $\{f^{(l)}\}_{l\in\mathbb{N}}$ and $\{g^{(l)}\}_{l\in\mathbb{N}}$ converge strongly in $\boldsymbol{Z}$ to $f \in Z$ and $g \in Z$, respectively.

When trying to adapt the proof of Theorem 5.11 to setting of Problem 5.13, one finds that the following statement needs to be true: Given a sequence $\{\rho^{(l)}\}_{l\in\mathbb{N}}$ in $\overline{\mathcal{C}}$ that weak-$*$-converges in $\mathcal{E}$ to a density function $\rho \in \overline{\mathcal{C}}$, as well as two sequences $\{u^{(l)}\}_{l\in\mathbb{N}}$ and $\{v^{(l)}\}_{l\in\mathbb{N}}$ in $V$ that converge strongly in $Z$ and weakly in $V$ to $u \in V$ and $v \in V$, respectively, it follows that $a_{\boldsymbol{k}}^{\mathrm{TE}}(\rho^{(l)})(u^{(l)}, v^{(l)})$ converges to $a_{\boldsymbol{k}}^{\mathrm{TE}}(\rho)(u, v)$ as $l$ tends to infinity. Unfortunately, this statement cannot be shown. In fact, for $\boldsymbol{k} = \boldsymbol{0}$ the statement is known to be false. In the following we shall make this notion more precise.

Suppose that $\{u^{(l)}\}_{l\in\mathbb{N}}$ is a sequence in $V$ that converges weakly in $V$ to a function $u \in V$. Then, the sequence of weakly defined gradients $\{\nabla u^{(l)}\}_{l\in\mathbb{N}}$ converges weakly in $Z$ (see e.g. Example 6.4(3) in [4]). Let $\{\rho^{(l)}\}_{l\in\mathbb{N}}$ be a sequence in $\overline{\mathcal{C}}$ that weak-$*$-converges in $\mathcal{E}$ to a function $\rho \in \overline{\mathcal{C}}$. Then, the sequence $\{\rho^{(l)}\nabla u^{(l)}\}_{l\in\mathbb{N}}$ does not converge weakly in $Z$ to $\rho\nabla u$, in general. A counter-example is constructed, for example, in Section 5.1 in [21]. Hence, we cannot expect $a_{\boldsymbol{0}}^{\mathrm{TE}}(\rho^{(l)})(u^{(l)}, v)$ to converge to $a_{\boldsymbol{0}}^{\mathrm{TE}}(\rho)(u, v)$ for every $v \in V$.

Sequences of the form $\{\rho^{(l)}\nabla u^{(l)}\}_{l\in\mathbb{N}}$ typically occur in problems, which are studied in homogenization theory. An important concept that is related to these sequences is the concept of *H-convergence* (see [56]), which we shall briefly introduce. In the following, we denote by $D \subset \mathbb{R}^r$ an open, simply connected, and bounded domain, where $r \in \mathbb{N}$ is an arbitrary space dimension. Then, we defined and for every two constants $\alpha_{\min}, \alpha_{\max} \in \mathbb{R}$, that satisfy $0 < \alpha_{\min} < \alpha_{\max}$, the set

$$M(D, \alpha_{\min}, \alpha_{\max}) := \big\{ \boldsymbol{A} \in L^\infty(D, \mathbb{R})^{r\times r} \;\big|\; \langle \boldsymbol{A}(\,\cdot\,)\boldsymbol{x}\,, \boldsymbol{x}\rangle \geq \alpha_{\min}|\boldsymbol{x}|^2, \; |\boldsymbol{A}(\,\cdot\,)\boldsymbol{x}| \leq \alpha_{\max}|\boldsymbol{x}| \text{ for all } \boldsymbol{x} \in \mathbb{R}^r \text{ a.e. on } D \big\}. \tag{5.9}$$

Clearly, the set $M(D, \alpha_{\min}, \alpha_{\max})$ consists of specific matrix-valued functions defined on $D$. A sequence $\{\boldsymbol{A}^{(l)}\}_{l\in\mathbb{N}}$ in $M(D, \alpha_{\min}, \alpha_{\max})$ is said to *H-converge* to a matrix-valued function $\boldsymbol{A} \in M(D, \beta_{\min}, \beta_{\max})$, if for every continuous, linear functional $f \in H^{-1}(D)$ the solution $u^{(l)} \in H_0^1(D)$ of the weakly formulated boundary value problem

$$\int_D \boldsymbol{A}^{(l)}\overline{\nabla u^{(l)}} \cdot \nabla v = f(v) \qquad \text{for all } v \in H_0^1(D),$$

converges weakly in $H_0^1(D)$ to the solution $u \in H_0^1(D)$ of the weakly formulated boundary value problem

$$\int_D \boldsymbol{A}\overline{\nabla u} \cdot \nabla v = f(v) \qquad \text{for all } v \in H_0^1(D),$$

and if $\boldsymbol{A}^{(l)}\nabla u^{(l)}$ converges weakly in $L^2(D)$ to the vector field $\boldsymbol{A}\nabla u$ as $l$ tends to $\infty$. It is known that for every sequence $\{\boldsymbol{S}^{(l)}\}_{l\in\mathbb{N}}$ of symmetric-matrix-valued

functions in $M(D, \alpha_{\min}, \alpha_{\max})$ there exists a symmetric-matrix-valued function $\boldsymbol{S} \in M(D, \alpha_{\min}, \alpha_{\max})$, such that a subsequence of $\{\boldsymbol{S}^{(l)}\}_{l\in\mathbb{N}}$ H-converges to $\boldsymbol{S}$ (see Theorem 13.2(iii) and Proposition 13.6 in [21]). This notion is expressed in stating that the set $M_2(\alpha_{\min}, \alpha_{\max})$ is *G-compact* (see Theorem 13.2 in [21]).

By definition, the concepts of H-convergence and G-compactness are linked to weakly formulated boundary value problems of the form

$$\int_D \boldsymbol{A}\overline{\nabla u} \cdot \nabla v = f(v) \qquad \text{for all } v \in H_0^1(\Omega),$$

where $u \in H_0^1(D)$ is the unknown function, where $f \in H^{-1}(D)$ is a given right-hand side, and where $\boldsymbol{A} \in M(D, \rho_{\min}, \rho_{\max})$ is a matrix-valued coefficient. Unfortunately, it seems that no attempts have been made so far to generalize these concepts. Anyway, we were not able to find conclusive evidence in the literature as to whether or not the results related to H-convergence also apply to weakly formulated boundary value problems of the form

$$\int_\Omega \boldsymbol{A}\overline{(\nabla + \mathrm{i}\boldsymbol{k})u} \cdot (\nabla + \mathrm{i}\boldsymbol{k})v = f(v) \qquad \text{for all } v \in V,$$

where $u \in V$ is the unknown function, $f \in V^*$ is a prescribed right-hand side, $\boldsymbol{k} \in \mathbb{B}$ is a given vector, and $\boldsymbol{A} \in M(\Omega, \rho_{\min}, \rho_{\max})$ is a given matrix-valued coefficient. Provided that the set $M(\Omega, \rho_{\min}, \rho_{\max})$ is also G-compact with respect to these problems, one could show that every TE goal functional attains a minimum on the set $M(\Omega, \rho_{\min}, \rho_{\max})$. The corresponding minimizer would then be a symmetric-matrix-valued function, representing the relative electric permittivity function of an anisotropic medium.

The situation is very similar for PBSOPs that involve band structures of three-dimensional photonic crystals. Henceforth, we assume that $\Lambda \subset \mathbb{R}^3$ is a Bravais lattice of rank 3 with primitive cell $\Omega$ and first Brillouin zone $\mathbb{B}$. We assume that the admissible set $\mathcal{C}$ and the extended admissible set $\overline{\mathcal{C}}$ are defined according to (5.2) and (5.4).

Under the assumptions of Section 5.1, one is not able to prove the existence of optimal solutions of Problem 5.1 in the extended admissible set $\overline{\mathcal{C}}$ by adapting the proof of Theorem 5.11. What is missing, again, is a suitable analogue of Lemma 5.10. More precisely, given a sequence $\{\rho^{(l)}\}_{l\in\mathbb{N}}$ in $\overline{\mathcal{C}}$ that weak-$*$-converges in $\mathcal{E}$ to $\rho \in \overline{\mathcal{C}}$, and two sequences $\{\boldsymbol{u}^{(l)}\}_{l\in\mathbb{N}}$ and $\{\boldsymbol{v}^{(l)}\}_{l\in\mathbb{N}}$ in $\boldsymbol{V_k}$ that converge weakly in $\boldsymbol{V_k}$ and strongly in $\boldsymbol{Z}$ to $\boldsymbol{u} \in \boldsymbol{V_k}$ and $\boldsymbol{v} \in \boldsymbol{V_k}$, respectively, one cannot conclude that $a_{\boldsymbol{k}}(\rho^{(l)})(\boldsymbol{u}^{(l)}, \boldsymbol{v}^{(l)})$ converges to $a_{\boldsymbol{k}}(\rho)(\boldsymbol{u}, \boldsymbol{v})$ as $l$ tends to infinity. As in the TE setting, one could prove the existence of optimal solutions in the set $M(\Omega, \rho_{\min}, \rho_{\max})$, if the results related to H-convergence also apply to weakly

formulated boundary value problems of the form

$$\int_\Omega \boldsymbol{A}\overline{(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}} \cdot (\nabla + \mathrm{i}\boldsymbol{k})\boldsymbol{v} = f(\boldsymbol{v}) \qquad \text{for all } v \in \boldsymbol{V_k},$$

where $\boldsymbol{u} \in \boldsymbol{V_k}$ is the unknown function, $f \in \boldsymbol{V_k}^*$ is a prescribed right-hand side, $\boldsymbol{k} \in \mathbb{B}$ a given vector and $\boldsymbol{A} \in M(\Omega, \rho_{\min}, \rho_{\max})$ a given coefficient. An optimal solution in $M(\Omega, \rho_{\min}, \rho_{\max})$ would then be a symmetric-matrix-valued function representing the relative electric permittivity function of an anisotropic medium. To the best of our knowledge, however, the question as to whether or not the concepts of H-convergence and G-compactness also apply to the above type of boundary value problem has not been considered in the literature so far.

## 5.5 Optimization Goals

In the previous sections we studied general aspects of photonic band structure optimization problems (PBSOPs). In this section we discuss several concrete optimization problems and introduce a specific class of model problems, which will be further investigated in this dissertation.

The most common PBSOPs, which are studied in the literature, are so-called *photonic band gap maximization problems* (PBGMPs). A photonic band gap maximization problem consists in finding a medium structure of a photonic crystal, which yields a band structure that exhibits a band gap as large as possible. The main reason for the popularity of PBGMPs is, that photonic crystals, whose band structures exhibit large band gaps, are expected to play and important role in the creation of certain nano-optical devices. Recall that photonic band gaps correspond to ranges of optical frequencies, at which the photonic crystal inhibits any light propagation. Therefore, photonic crystals with large band gaps are favorable for the design of mirrors or nano-scale optical wave guides.

In the following we present two PBSOPs, which were discussed in the literature and which fit into the problem scheme given by Problem 5.1 on page 92. It is easy to see that every PBSOP, which fits into that scheme is uniquely identified by the respective goal functional $J$.

In their work on the maximization of photonic band gaps in two-dimensional photonic crystals [29], the authors Cox and Dobson considered TM goal functionals $J^{\mathrm{TM}}_{j;\omega_0} : \mathcal{D} \to \mathbb{R}$, which were defined by

$$J^{\mathrm{TM}}_{j;\omega_0}(\rho) := -\min_{\boldsymbol{k}\in\mathbb{B}} \min\left\{\omega^{\mathrm{TM}}_{j+1}(\rho,\boldsymbol{k})^2 - \omega_0^2, \omega_0^2 - \omega^{\mathrm{TM}}_j(\rho,\boldsymbol{k})^2\right\} \tag{5.10}$$

for all $\rho \in \mathcal{D}$. Here, $j \in \mathbb{N}$ is a given index, and $\omega_0$ is a given real number. Under the assumption that there exists an initial density function $\rho^{(0)} \in \overline{\mathcal{C}}$, such that the

photonic crystal represented by $\rho^{(0)}$ exhibits a band gap between the $j$-th and the $(j+1)$-th TM band, the number $\omega_0$ can be chosen such that

$$\max_{\boldsymbol{k}\in\mathbb{B}} \omega_j^{\mathrm{TM}}(\rho^{(0)}, \boldsymbol{k}) < \omega_0 < \min_{\boldsymbol{k}\in\mathbb{B}} \omega_{j+1}^{\mathrm{TM}}(\rho^{(0)}, \boldsymbol{k}).$$

Such a number $\omega_0$ represents a so-called *gap frequency* in the TM band structure of the photonic crystal, which is represented by $\rho^{(0)}$. The minimization problem for the corresponding goal functional $J_{j;\omega}^{\mathrm{TM}}$ aims at increasing the minimal distance between the $j$-th and the $(j+1)$-th TM band to the gap frequency $\omega_0$. By this, Cox and Dobson were able to widen the TM band gaps of several two-dimensional photonic crystals.

In a second paper on band structure optimization of two-dimensional photonic crystals [30], Cox and Dobson used another type of TE goal functionals. These functionals $J_{j;\omega}^{\mathrm{TE}} : \mathcal{D} \to \mathbb{R}$ are given by

$$J_{j;\omega}^{\mathrm{TE}}(\rho) := -\min_{\boldsymbol{k}\in\mathbb{B}} \min\left\{\omega_{j+1}^{\mathrm{TE}}(\rho, \boldsymbol{k})^2 - \omega(\boldsymbol{k})^2, \omega(\boldsymbol{k})^2 - \omega_j^{\mathrm{TE}}(\rho, \boldsymbol{k})^2\right\} \tag{5.11}$$

for all $\rho \in \mathcal{D}$. Here, $\omega : \mathbb{B} \to \mathbb{R}$ denotes a prescribed, continuous function. Under the assumption that there exists an index $j \in \mathbb{N}$ and an initial density function $\rho^{(0)} \in \overline{\mathcal{C}^{\mathrm{TE}}}$, such that the corresponding $j$-th and $(j+1)$-th TE band are strictly separated (see (4.75) on page 80), the function $\omega$ can be chosen such that

$$\omega_j^{\mathrm{TE}}(\rho^{(0)}, \boldsymbol{k}) < \omega(\boldsymbol{k}) < \omega_{j+1}^{\mathrm{TE}}(\rho^{(0)}, \boldsymbol{k}) \qquad \text{for all } \boldsymbol{k} \in \mathbb{B}. \tag{5.12}$$

The minimization problem for the corresponding goal functional $J_{j;\omega}^{\mathrm{TM}}$ aims at increasing the minimal distance of the $j$-th and the $(j+1)$-th TE band to the graph of the function $\omega$. In an iterative optimization algorithm, one can modify the function $\omega$ slightly after each iteration and thus create a sequence of functions $\{\omega_l\}_{l\in\mathbb{N}}$, such that every element of that sequence satisfies (5.12). Cox and Dobson devised an iterative algorithm, in which the sequence $\{\omega_l\}_{l\in\mathbb{N}}$ converged to a constant function. By this, they were able to open band gaps in the TE band structures of some two-dimensional photonic crystals.

In the following sections, we consider PBSOPs for three-dimensional photonic crystals, which are given by goal functionals $J_j : \mathcal{D} \to \mathbb{R}$ of the form

$$J_j(\rho) := \max_{\boldsymbol{k}\in\mathbb{B}} \lambda_j(\rho, \boldsymbol{k}) - \min_{\boldsymbol{k}\in\mathbb{B}} \lambda_{j+1}(\rho, \boldsymbol{k}) \qquad \text{for all } \rho \in \mathcal{D}, \tag{5.13}$$

for some index $j \in \mathbb{N}$. We will refer to these functions as *gap width functionals*, since minimization problems for these functionals aim at widening the gap width of a band gap between the $j$-th and the $(j+1)$-th photonic band. Given an index $j \in \mathbb{N}$, an optimization algorithm for the goal functional $J_j$ can also be used to open up a band gap between the $j$-th and the $(j+1)$-th photonic band, provided

that these bands are strictly separated. In Chapter 9 we present some numerical examples, where band gaps between separated bands could be opened.

For completeness we also define for every index $j \in \mathbb{N}$ the TM and TE analogues $J_j^{\mathrm{TM}} : \mathcal{D} \to \mathbb{R}$ and $J_j^{\mathrm{TE}} : \mathcal{D} \to \mathbb{R}$ of $J_j$ by

$$J_j^{\mathrm{TM}}(\rho) := \max_{\boldsymbol{k}\in\mathbb{B}} \lambda_j^{\mathrm{TM}}(\rho, \boldsymbol{k}) - \min_{\boldsymbol{k}\in\mathbb{B}} \lambda_{j+1}^{\mathrm{TM}}(\rho, \boldsymbol{k}) \qquad \text{for all } \rho \in \mathcal{D}, \tag{5.14}$$

$$J_j^{\mathrm{TE}}(\rho) := \max_{\boldsymbol{k}\in\mathbb{B}} \lambda_j^{\mathrm{TE}}(\rho, \boldsymbol{k}) - \min_{\boldsymbol{k}\in\mathbb{B}} \lambda_{j+1}^{\mathrm{TE}}(\rho, \boldsymbol{k}) \qquad \text{for all } \rho \in \mathcal{D}. \tag{5.15}$$

These goal functionals were also used by Kao, Osher and Yablonovitch in their work on the maximization of band gaps in two-dimensional photonic crystals (cf. [44]).

We compared the goal functionals $J_{j;\omega_0}^{\mathrm{TM}}$ and $J_j^{\mathrm{TM}}$ by conducting some numerical experiments. In these experiments we found that solving a PBSOPs for the TM gap width functionals $J_j^{\mathrm{TM}}$ usually yields larger band gaps than solving such problems for the goal functionals $J_{j;\omega_0}^{\mathrm{TM}}$. This is mostly due to the fact, that an optimization algorithm for the TM gap width functionals $J_j^{\mathrm{TM}}$ can freely change the absolute position of the band gap in order to widen it. In contrast to this, optimization algorithms for the goal functionals $J_{j,\omega_0}^{\mathrm{TM}}$ are restricted in their ability to change the position of the band gap by the fixed gap frequency $\omega_0$.

For some applications, however, prescribing a fixed gap frequency $\omega_0$ is essential. Consider, for example, the problem of designing a photonic crystal that should inhibit light propagation around a prescribed center frequency. The center frequency could be determined, for example, by a given light source, such as a specific laser device. For this problem it is more adequate, of course, to choose a goal functional similar to $J_{j;\omega_0}^{\mathrm{TM}}$ for some index $j \in \mathbb{N}$, and to use to the prescribed center frequency as $\omega_0$. Another situation, where it can be useful to prescribe a certain gap frequency $\omega_0$, is when one is interested in finding two-dimensional photonic crystals, which exhibit maximal complete band gaps. As was stated in Section 4.7 a complete band gap consists of all frequencies lying in a TM band gap as well as in a TE band gap. Maximizing complete band gaps thus amounts to maximizing TM and TE band gaps, which contain a common frequency $\omega_0$.

In this work, we are mostly interested in finding photonic crystals with band gaps that are as large as possible. Therefore, we shall only consider PBSOPs for the gap width functionals $J_j$, $J_j^{\mathrm{TM}}$, and $J_j^{\mathrm{TE}}$ in the following chapters.

# Chapter 6

# Nonsmooth Analysis

In Section 5.2 we showed that the goal functionals, which arise in photonic band structure optimization problems (PBSOPs), are typically locally Lipschitz continuous. It was also emphasized that the goal functionals in general fail to be differentiable. However, in order to devise a local optimization algorithm for PBSOPs one needs some information about the local monotonicity of the goal functionals. For locally Lipschitz continuous functionals, this information can be obtained from so-called generalized differentials. Generalized differentials are a fundamental tool of nonsmooth analysis and were first introduced by Clarke (see e.g. [22]). In this chapter we list some important results concerning generalized differentials. In Section 6.1 we define generalized directional derivatives, as well as generalized differentials and discuss some of their properties. In Section 6.2 we provided some rules of calculus for generalized differentials. These rules are used in Section 6.3 in order to determined the generalized differentials of the gap width functionals for three-dimensional photonic crystals, which were introduced in Section 5.5. In Section 6.4 we give a characterization of the gap width functionals for two-dimensional photonic crystals.

## 6.1 Generalized Differentials

Throughout this section we denote by $X$ a real Banach space, which is endowed with a norm $\|\cdot\|_X$. By $X^*$ we denote its normed dual. Given a functional $f : X \to \mathbb{R}$, a point $x \in X$, and a vector $v \in X$, we define after Clarke (see Chapter 2.1 in [22]) the upper limit

$$f^\circ(x; v) := \limsup_{\substack{y \to x \\ t \to 0+}} \frac{f(y + tv) - f(y)}{t}.$$

For locally Lipschitz continuous functionals $f$ this upper limit is a real number, which is called the *generalized directional derivative of $f$ at $x$ in direction $v$*. For functionals, that are continuously differentiable in a neighbourhood of $x$, $f^\circ(x;v)$ coincides with the one-sided directional derivative of $f$ at $x$ in direction $v$. The following proposition summarizes important properties of generalized directional derivatives. The proposition corresponds to Proposition 1.1 in Chapter 2 in [23].

**Proposition 6.1.** *Let $f : X \to \mathbb{R}$ be a functional, which is $c$-Lipschitz continuous near a point $x \in X$ for some local Lipschitz constant $c > 0$. Then, the following statements are valid.*

(a) *$f^\circ(x;\cdot) : X \to \mathbb{R}$ is positively homogeneous, subadditive, and satisfies*

$$|f^\circ(x;v)| \leq c\|v\|_X \quad \text{for all } v \in X.$$

(b) *$f^\circ(x;\cdot) : X \to \mathbb{R}$ is $c$-Lipschitz continuous near $x$.*

(c) *$f^\circ : X \times X \to \mathbb{R}$ is upper semicontinuous.*

(d) *$f^\circ(x;-v) = (-f)^\circ(x;v)$ for all $v \in X$.*

To a certain extent the sign of the functional $f^\circ(x;\,\cdot\,) : X \to \mathbb{R}$ is determined by the local monotonicity of the functional $f$ at the point $x$. This notion is made precise by the following proposition. We denote by $B_r(x)$ the open ball in $X$ with radius $r > 0$ and center $x \in X$.

**Proposition 6.2.** *Let $f : X \to \mathbb{R}$ be a locally Lipschitz continuous functional, let $x \in X$ be a point, and let $V \subset X$ be a closed, convex cone with vertex at the origin.*

(a) *If $f$ is monotonically increasing at $x$ in all directions of $V$, i.e., if there exists a radius $r > 0$, such that*

$$f(x+w) \geq f(x) \quad \text{for all } w \in V \cap B_r(0),$$

*then $f^\circ(x;v) \geq 0$ for all $v \in V$.*

(b) *If $f$ is monotonically increasing in an open neighbourhood $U$ of $x$ in all directions of $V$, i.e., if there exists a radius $r > 0$, such that*

$$f(y+w) \geq f(y) \quad \text{for all } y \in U, w \in V \cap B_r(0),$$

*then $f^\circ(x;v) \geq 0$ and $f^\circ(x;-v) \leq 0$ for all $v \in V$.*

*Proof.* Let $v \in C$. Then, there exists a number $t_0 > 0$, such that $tv \in B_r(0)$ for all $t \in [0, t_0]$. Hence,

$$\begin{aligned} f^\circ(x; v) &= \limsup_{\substack{y \to x \\ t \to 0}} \frac{f(y + tv) - f(y)}{t} \\ &\geq \limsup_{t \to 0+} \frac{f(x + tv) - f(y)}{t} \\ &= \limsup_{t \to 0,\, t \in [0, t_0]} \frac{f(x + tv) - f(y)}{t} \\ &\geq 0, \end{aligned}$$

which proves the inequality $f^\circ(x; v) \geq 0$ in (a) and (b). Furthermore, we have

$$\begin{aligned} f^\circ(x; -v) &= (-f)^\circ(x; v) \\ &= \limsup_{\substack{y \to x \\ t \to 0}} \frac{f(y) - f(y + tv)}{t} \\ &= \limsup_{\substack{y \to x,\, y \in U \\ t \to 0,\, t \in [0, t_0]}} \frac{f(y) - f(y + tv)}{t} \\ &\leq 0, \end{aligned}$$

which establishes the inequality $f^\circ(x; -v) \leq 0$ in (b). □

In contrast to one-sided directional derivatives, the sign of the generalized directional derivative does not provide conclusive information about the local monotonicity of a locally Lipschitz continuous functional. We illustrate this by the following example.

**Example 6.3.** Consider the function $f : \mathbb{R} \to \mathbb{R}$, defined as $f(x) := \sqrt{1 - |x|} - 1$ for all $x \in \mathbb{R}$ (see Figure 6.1). This function is Lipschitz continuous on $\mathbb{R}$, the minimal Lipschitz constant being $1/2$. Moreover, $f$ is continuously differentiable at all points except at the origin. Explicit computation shows that the generalized directional derivative at the origin is given by $f^\circ(0; v) = |v|/2$. Hence, $f^\circ(0; v) > 0$ for all $v \in \mathbb{R}$, which intuitively corresponds well to the fact that $f$ increases monotonically at the origin in all directions. However, by Proposition 6.1(c) we also have $(-f)^\circ(0; v) = f^\circ(0; v) > 0$ for all $v \in \mathbb{R}$, even though the function $-f$ decreases monotonically at the origin in all directions.

By Proposition 6.1(a) the generalized directional derivative of a locally Lipschitz continuous functional constitutes a positively homogeneous, subadditive, bounded

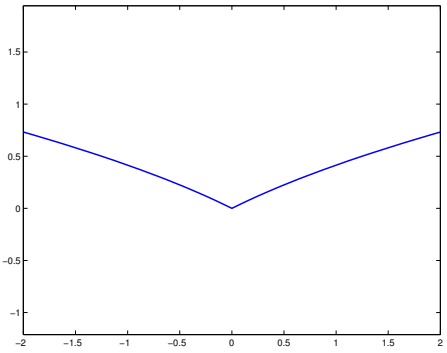



**Figure 6.1:** Graph of the function $f$ in Example 6.3 at the origin.

functional on $X$. According to the Hahn–Banach Theorem, the generalized directional derivative thus majorizes at least one bounded, linear functional on $X$. Given a locally Lipschitz functional $f : X \to \mathbb{R}$ and a point $x \in X$, we therefore have that the set

$$\partial f(x) := \big\{ g \in X^* \,\big|\, g(v) \leq f^\circ(x; v) \text{ for all } v \in X \big\}. \tag{6.1}$$

is non-empty. This set $\partial f(x)$ is called the *generalized differential of $f$ at $x$*. The following proposition summarizes important properties of generalized differentials. The assertions of the proposition correspond to those stated in the Propositions 2.1.2 and 2.3.1 in [22].

**Proposition 6.4.** *Let $f : X \to \mathbb{R}$ be $L$-Lipschitz continuous near $x \in X$. Then,*

(a) *$\partial f(x)$ is a non-empty, convex, weak-$*$-compact, and bounded subset of $X^*$. In particular,*

$$\|p\|_{X^*} \leq L \quad \text{for all } p \in \partial f(x).$$

(b) *$f^\circ(x; v) = \max\big\{ p(v) \,\big|\, p \in \partial f(x) \big\}$ for all $v \in X$.*

(c) *$\partial(-f)(x) = -\partial f(x)$.*

As the name already suggests, generalized differentials are generalizations of classical differentials or derivatives for locally Lipschitz continuous functionals. In particular, they provide necessary conditions for the existence of local extrema according to the following lemma.

**Lemma 6.5.** *If a locally Lipschitz continuous functional $f : X \to \mathbb{R}$ attains a local extremum at some point $x \in X$, then $0 \in \partial f(x)$.*

For a proof of Lemma 6.5, see Proposition 2.3.2 in [22]. There also exists an analog of the Mean-Value Theorem for generalized differentials, which is stated in the following Theorem of Lebourg (cf. [52]).

**Theorem 6.6** (Lebourg)**.** *Let $f : X \to \mathbb{R}$ be Lipschitz continuous on a convex, open subset $K$ of $X$. Then, for every two points $x, y \in K$ there exists a point $z \in \operatorname{conv}\{x, y\}$ and a continuous, linear functional $p \in \partial f(z)$, such that*

$$f(y) - f(x) = p(y - x).$$

The following proposition reveals the relationship between generalized differentials and other concepts of differentiability. The assertions of the proof are established in Section 2.2 in [22].

**Proposition 6.7.** *Let $f : X \to \mathbb{R}$ be a locally Lipschitz continuous functional, and let $U$ be an open neighbourhood of some point $x \in X$.*

(a) *If $f$ is convex on $U$, then $\partial f(x)$ coincides with the subdifferential of $f$ at $x$.*

(b) *If $f$ is Gâteaux differentiable or Fréchet differentiable at $x$ with derivative $\mathrm{D}f(x)$, then $\mathrm{D}f(x) \in \partial f(x)$.*

(c) *If $f$ is continuously Gâteaux differentiable on $U$ with derivative $\mathrm{D}f(x)$ at $x$, then $\{\mathrm{D}f(x)\} = \partial f(x)$.*

(d) *The functional $f$ is strictly differentiable in $U$ with strict derivative $\mathrm{D}_{\mathrm{s}}f(x)$ at $x$, if and only if $\{\mathrm{D}_{\mathrm{s}}f(x)\} = \partial f(x)$.*

Recall that the *subdifferential* of a convex functional $f : X \to \mathbb{R}$ at a point $x \in X$ is defined as set of all continuous linear functionals $p \in X^*$, for which

$$f(y) - f(x) \geq p(y - x) \quad \text{for all } y \in X.$$

In the literature the subdifferential of $f$ at a point $x$ is also commonly denoted by $\partial f(x)$.

The concept of strict differentiability is less well-known, which is why we briefly comment on it here. A functional $f : X \to \mathbb{R}$ is called *strictly differentiable* at a point $x \in X$, if there exists a continuous linear functional $p \in X^*$, such that

$$\lim_{\substack{y \to x \\ t \to 0+}} \frac{f(y + tv) - f(y)}{t} = p(v) \qquad \text{for all } v \in X.$$

The functional $p$ is then called the *strict differential of $f$ at $x$* and denoted by $\mathrm{D}_{\mathrm{s}}f(x)$. It can be shown that continuous Gâteaux differentiability at a point $x \in X$

implies strict differentiability at that point (see Corollary to Proposition 2.2.1 in [22]). Moreover, it is easy to see that strict differentiability implies Gâteaux differentiability.

It should be noted that Clarke, as well as other authors, call the set $\partial f(x)$ the "generalized gradient" of $f$ at $x$. We prefer the term "generalized differential", however, since it alludes to the fact, that $\partial f(x)$ is a subset of $X^*$. This corresponds to the notion of the differentials of a Gâteaux differentiable functional $f : X \to \mathbb{R}$. By definition, the differential $\mathrm{D}f(x)$ of such a functional at a given point $x \in X$ is an element of $X^*$. Now suppose that $Y$ is another real Banach space, and that $\langle \cdot\ ,\ \cdot \rangle : Y \times X \to \mathbb{R}$ is a bilinear form, such that $(Y, X, \langle \cdot\ ,\ \cdot \rangle)$ is a *dual pair* in the sense of functional analysis (see e.g. Definition VIII.3.1 in [75]). If there exists a uniquely defined element $u \in Y$, such that $\mathrm{D}\varphi(x) = \langle u\ , \cdot \rangle$ then $u$ is called the *gradient of $f$ at $x$* with respect to $\langle \cdot\ ,\ \cdot \rangle$. If $X$ is a Hilbert space, one usually chooses the inner product on $X$ as the bilinear form $\langle \cdot\ ,\ \cdot \rangle$. The gradient of $f$ at a point $x$ is then commonly denoted by $\nabla f(x)$. The connection between differentials and gradients in a Hilbert space is established by the Rieszian isomorphism $R : X \to X^*$ through $\nabla f(x) = R^{-1}(\mathrm{D}f(x))$.

According to the discussion above, it is quite natural to define generalized gradients as follows. Let $X$ and $Y$ be real Banach spaces, and let $\langle \cdot\, , \cdot \rangle : Y \times X \to \mathbb{R}$ be a bilinear form, such that $(Y, X, \langle \cdot\ ,\ \cdot \rangle)$ is a dual pair. Given a locally Lipschitz continuous functional $f : X \to \mathbb{R}$ and a point $x \in X$, the set

$$\big\{ u \in Y \;\big|\; \langle u\ , \cdot \rangle \in \partial f(x) \big\} \tag{6.2}$$

is called the *generalized gradient of $f$ at $x$* with respect to $\langle \cdot\ ,\ \cdot \rangle$. If $X$ is a Hilbert space coinciding with $Y$, and if $\langle \cdot\ ,\ \cdot \rangle$ is the inner product on $X$, then the set in (6.2) is often identified with the generalized differential of $f$ at $x$ and also denoted by $\partial f(x)$.

We conclude this section with the following example.

**Example 6.8.** Consider again the function $f$ defined in Example 6.3. The function is defined on $\mathbb{R}$. The dual space $\mathbb{R}^*$ of $\mathbb{R}$ consists of all functionals $p : \mathbb{R} \to \mathbb{R}$, which are of the form $p(v) = \alpha v$, where $\alpha \in \mathbb{R}$. By (6.1) a functional $p \in \mathbb{R}^*$ belongs to $\partial f(0)$, if and only if $p(v) \le f^\circ(x, v)$ for all $v \in \mathbb{R}$. Hence, one easily deduces that

$$\partial f(0) = \left\{ p \in \mathbb{R}^* \;\middle|\; p(v) = \alpha v,\ \alpha \in \left[ -\frac{1}{2}, \frac{1}{2} \right] \right\}.$$

In view of Lemma 6.5 it is not surprising, that the zero function on $\mathbb{R}$ is contained in $\partial f(0)$, since $f$ attains a local minimum at 0. Since $\mathbb{R}$ is a Hilbert space, one can also define a generalized gradient of $f$ at $x$ as a subset of $\mathbb{R}$. Clearly, this subset is given by the interval $[-1/2, 1/2]$.

## 6.2 Generalized Differential Calculus

In this section we list some rules of calculus for generalized differentiation. As we shall see below, most of these rules only establish certain inclusion relations. Therefore, these rules can only be used to determine supersets of generalized differentials. Still, these rules will prove useful in the following section, where we aim to characterize the generalized differentials of the gap width functionals introduced in Section 5.5.

As in the previous section, we denote by $X$ a real Banach space. Our first rule of calculus, which corresponds to Proposition 2.3.1 in [22], states that generalized differentiation is a homogeneous operation.

**Proposition 6.9.** *Let $f : X \to \mathbb{R}$ be a functional, which is Lipschitz continuous near a point $x \in X$, and let $\theta \in \mathbb{R}$ be a real number. Then we have*

$$\partial(\theta f)(x) = \theta \partial f(x).$$

The next proposition, which corresponds to Proposition 2.3.3 in [22], states that generalized differentials are subadditive in the sense of set inclusions.

**Proposition 6.10.** *Let $f_1, \dots, f_N : X \to \mathbb{R}$ be a finite number of functionals for some $N \in \mathbb{N}$, which are Lipschitz continuous near a point $x \in X$. Then, we have*

$$\partial(f_1 + \cdots + f_n)(x) \subseteq \partial f_1(x) + \cdots + \partial f_N(x).$$

The following theorem establishes a chain rule for certain continuously differentiable functionals with locally Lipschitz continuous functionals. This theorem is a particularization of Theorem 2.3.9 in [22].

**Theorem 6.11.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function, and let $g : X \to \mathbb{R}$ be a functional, which is Lipschitz continuous near a point $x \in X$. Then, the functional $f \circ g$ is Lipschitz continuous near $x$, and we have*

$$\partial(f \circ g)(x) = f'\big(g(x)\big)\partial g(x) := \big\{ f'\big(g(x)\big)q \;\big|\; q \in \partial g(x) \big\},$$

*where $f'(g(x)) \in \mathbb{R}$ denotes the derivative of $f$ at the point $g(x)$.*

The following theorem is a particularization of Theorem 2.3.10 in [22]. It establishes another chain rule for generalized differentials of functionals, which are given by compositions of locally Lipschitz continuous functions and continuously Gâteaux differentiable operators. Just like Proposition 6.10, the theorem only establishes an inclusion relation.

**Theorem 6.12.** *Let $Y$ be a second real Banach space, and let $G : X \to Y$ be an operator, which is continuously Gâteaux differentiable in a neighbourhood of a point $x \in X$. Furthermore, let $f : Y \to \mathbb{R}$ be a functional, which is Lipschitz continuous near $G(x)$. Then, the functional $f \circ G$ is Lipschitz continuous near $x$, and we have*

$$\partial(f \circ G)(x) \subseteq \big\{p \circ \mathrm{D}G(x) \,\big|\, p \in \partial f(G(x))\big\},$$

*where $\mathrm{D}G(x)$ denotes the Gâteaux derivative of $G$ at $x$.*

Before we can state the next rule of calculus for generalized differentiation, we need to introduce some concepts from measure theory. The concepts we present here are discussed in more detail in [67]. Suppose that $T$ is a topological Hausdorff space and that $\mathcal{T}$ is a $\sigma$-algebra on $T$ containing the open subsets of $T$. Then, a measure $\mu$ on $\mathcal{T}$ is called is called *locally finite*, if for every point $t \in T$ there exists an open neighbourhood $U$ of $t$, such that $\mu(U) < \infty$. The measure $\mu$ is called *inner regular*, if

$$\mu(S) = \sup\{\mu(K) \mid K \subset S,\ K \text{ compact}\} \qquad \text{for all } S \in \mathcal{T}.$$

Let $\mathcal{B}(T)$ be the Borel $\sigma$-algebra on $T$. Recall that $\mathcal{B}(T)$ is defined as the smallest $\sigma$-algebra on $T$ containing all open subsets of $T$. It follows from the definition of $\sigma$-algebras that $\mathcal{B}(T)$ is equivalently characterized as the smallest $\sigma$-algebra on $T$, which contains all closed subsets of $T$. Measures on $\mathcal{B}(T)$, which are locally finite and inner regular, are commonly referred to as *Radon measures*. To be more specific, we shall call a Radon measure on $\mathcal{B}(T)$ a *$T$-Radon measure*. A $T$-Radon measure $\mu$ is called a *probability $T$-Radon measure*, if $\mu(T) = 1$. The set $\mathrm{pr}_T$ consisting of all probability $T$-Radon measures is hence given by

$$\mathrm{pr}_T := \big\{\mu : \mathcal{B}(T) \to [0,\infty] \,\big|\, \mu \text{ locally finite, inner regular; } \mu(T) = 1\big\}. \tag{6.3}$$

We also define for every Borel set $S \in \mathcal{B}(T)$ the set $\mathrm{pr}_T(S)$, which consists of all probability $T$-Radon measures that are supported on $S$. More precisely, the set $\mathrm{pr}_T(S)$ is given by

$$\mathrm{pr}_T(S) := \big\{\mu \in \mathrm{pr}_T \,\big|\, \mu(S) = 1\big\}. \tag{6.4}$$

Finally, if $T$ is a sequentially compact topological space, we define for every continuous functional $g : T \to \mathbb{R}$ the set

$$\arg\min_T g := \Big\{t_* \in T \,\Big|\, g(t_*) = \min_{t \in T} g(t)\Big\}, \tag{6.5}$$

$$\arg\max_T g := \Big\{t_* \in T \,\Big|\, g(t_*) = \max_{t \in T} g(t)\Big\} \tag{6.6}$$

The set defined by (6.5) is called the *set of minimizers of $g$ on $T$*, and the set defined by (6.6) is called the *set of maximizers of $g$ on $T$*. Notice that neither of the two sets is empty due to the assumed compactness of $T$ and the assumed continuity of $g$.

With the above definitions at hand, we are able to state the following theorem, which corresponds to Theorem 2.8.2 in [22].

**Theorem 6.13.** *Let $T$ be a sequentially compact, separable topological Hausdorff space, and let $g : T \times X \to \mathbb{R}$ be a functional. Let $x \in X$ be a point, such that the image of $g(\,\cdot\,, x)$ is bounded. Furthermore, let $U$ be a neighbourhood of $x$, and let $c > 0$ be a positive constant, such that for each $y \in U$ the functional $g(\,\cdot\,, y)$ is continuous, and such that for each $t \in T$ the functional $g(t, \,\cdot\,)$ is $c$-Lipschitz continuous on $U$. Then, the functional $f : X \to \mathbb{R}$, defined by*

$$f := \max_{t \in T} g(t, \,\cdot\,)$$

*is $c$-Lipschitz continuous near $x$, and we have*

$$\partial f(x) \subseteq \left\{ \int_T q_t(\,\cdot\,) \,\mathrm{d}\mu(t) \;\middle|\; q_t \in \partial g(t, \,\cdot\,)(x),\ \mu \in \mathrm{pr}_T\left( \arg\max_T g(\,\cdot\,, x) \right) \right\}.$$

Now, suppose that $T$ is a compact subset of $\mathbb{R}^r$ for some $r \in \mathbb{R}$ endowed with the relative topology of $\mathbb{R}^r$. Given a point $x \in X$, let us further assume that the functional $g(\,\cdot\,, x)$ in Theorem 6.13 attains its maximum on a finite set of points, i.e.,

$$\arg\max_T g(\,\cdot\,, x) = \{t_{*1}, \ldots, t_{*N}\}$$

for some $N \in \mathbb{N}$. Then, it follows from (6.4) that the set of probability $T$-Radon measures on $T$, which are supported on the set of maximizers of $g(\,\cdot\,, x)$ on $T$, is given by

$$\mathrm{pr}_T\left( \arg\max_T g(\,\cdot\,, x) \right) = \overline{\mathrm{conv}}\{\delta_{t_{*1}}, \ldots, \delta_{t_{*N}}\}. \tag{6.7}$$

By $\delta_t$ we denote the so-called *Dirac measure* on $\mathcal{B}(T)$, which is centered on the point $t \in T$. This measure is defined by

$$\delta_t(S) = \begin{cases} 1 & \text{if } t \in S, \\ 0 & \text{else} \end{cases} \qquad \text{for all } S \in \mathcal{B}(T).$$

One easily verifies that $\delta_t$ is locally finite and inner regular, and hence a $T$-Radon measure, for all $t \in T$. The inner regularity follows from the fact that singletons, i.e. sets which contain exactly one point, are compact in $\mathbb{R}^r$. Furthermore, $\delta_t$ is

a probability $T$-Radon measure for all $t \in T$, since $\delta_t(T) = 1$. Since we also have $\delta_t(\{t\}) = 1$ for all $t \in T$, it follows that every measure in $\mathrm{pr}_T(\{t_{*1}, \ldots, t_{*N}\})$ is given by a convex combination of the Dirac measures, which are centered on $t_{*1}, \ldots, t_{*N}$. This is precisely the assertion of (6.7).

It is well-known, that for every continuous functional $h : T \to \mathbb{R}$ we have

$$\int_T h(t)\,\mathrm{d}\delta_{t_0}(t) = h(t_0) \qquad \text{for all } t_0 \in T. \tag{6.8}$$

From (6.8) and the discussion above we deduce the following corollary to Theorem 6.13.

**Corollary 6.14.** *Let $T$ be a compact subset of $\mathbb{R}^r$ for some $r \in \mathbb{N}$. Under the assumptions of Theorem 6.13, we further assume that*

$$\arg\max_T g(\,\cdot\,, x) = \big\{t_{*1}, \ldots, t_{*N}\big\},$$

*for some number $N \in \mathbb{N}$. Then, we have that the generalized differential of the functional $f$ at the point $x$ is given by*

$$\partial f(x) \subseteq \Big\{ \sum_{l=1}^{N} \theta_\ell\, q_\ell \;\Big|\; q_l \in \partial g(t_{*l},\,\cdot\,)(x),\ \theta_l \in [0,1],\ \sum_{l=1}^{N} \theta_l = 1 \Big\}.$$

Another result, which is related to generalized gradients of point-wise maxima, is stated by the following proposition. This proposition corresponds to Proposition 2.3.12 in [22].

**Proposition 6.15.** *Given a finite number of functionals $g_1, \ldots, g_N : X \to \mathbb{R}$ for some $n \in \mathbb{N}$, let $f : X \to \mathbb{R}$ be defined by*

$$f(x) := \max\big\{g_l(x) \;\big|\; l = 1, \ldots, N\big\}.$$

*If the functionals $g_1, \ldots, g_N$ are Lipschitz continuous near a point $x \in X$, then $f$ is Lipschitz continuous near $x$, and we have*

$$\partial f(x) \subseteq \Big\{ \sum_{l \in I(x)} \theta_l\, q_l \;\Big|\; q_l \in \partial g_l(x),\ \theta_l \in [0,1],\ \sum_{l \in I(x)} \theta_l = 1 \Big\},$$

*where the index set $I(x)$ is given by*

$$I(x) := \big\{l \in \{1, \ldots, N\} \;\big|\; g_l(x) = f(x)\big\}.$$

Finally, we remark that some of the inclusion relations in the propositions and theorems above can be replaced by identities under additional assumptions on the functionals. In Proposition 6.10, for example, equality holds between the two sets if all but at most one of the functionals $f_1, \ldots, f_N$ are strictly differentiable at $x$ (see Corollary 1 to Proposition 2.3.3 in [22]). According to Proposition 6.7(d) strict differentiability holds at a point $x \in X$, if and only if the generalized differentials at $x$ reduce to singletons. As we will see in the following section, this condition does not hold in the case of the gap width functionals. For this reason, we chose to present the above propositions and theorems in their most general form.

## 6.3 Differentiability of the Gap Width Functionals

In Section 6.1 and Section 6.2 we presented a number of general results related to generalized differentials of locally Lipschitz continuous functionals. In Section 5.2 we showed that the goal functionals of photonic band structure optimization problems are typically locally Lipschitz continuous. In Section 5.5 we defined in particular for every index $j \in \mathbb{N}$ the locally Lipschitz continuous gap width functional $J_j : \mathcal{D} \to \mathbb{R}$ by

$$J_j(\rho) := \max_{\boldsymbol{k} \in \mathbb{B}} \omega_j(\rho, \boldsymbol{k}) - \min_{\boldsymbol{k} \in \mathbb{B}} \omega_{j+1}(\rho, \boldsymbol{k}).$$

The aim of this section is to characterize the generalized differentials of these functionals.

We begin with some preliminary definitions. Given a coefficient $\rho \in \mathcal{D}$, a vector $\boldsymbol{k} \in \mathbb{B}$ and an index $j \in \mathbb{N}$ we define the function space

$$\boldsymbol{U}_j(\rho, \boldsymbol{k}) := \operatorname{span}\big\{ \boldsymbol{u}_i(\rho, \boldsymbol{k}) \;\big|\; \lambda_i(\rho, \boldsymbol{k}) = \lambda_j(\rho, \boldsymbol{k}) \big\}.$$

Recall that $\lambda_j(\rho, \boldsymbol{k})$ denotes the $j$-th smallest eigenvalue of the eigenvalue equation (4.42), and that $\boldsymbol{u}_j(\rho, \boldsymbol{k})$ denotes its corresponding function. Hence, the function space $\boldsymbol{U}_j(\rho, \boldsymbol{k})$ is the eigenspace corresponding to $\lambda_j(\rho, \boldsymbol{k})$. By Proposition 4.20, every eigenspace $\boldsymbol{U}_j(\rho, \boldsymbol{k})$ is finite-dimensional. Recall also that every eigenfunction $\boldsymbol{u}_j(\rho, \boldsymbol{k})$ is an element of the complex Hilbert space $\boldsymbol{V}_{\boldsymbol{k}}$, which we defined in Section 4.3. Hence, $\boldsymbol{U}_j(\rho, \boldsymbol{k})$ is a finite-dimensional, linear subspace of $\boldsymbol{V}_{\boldsymbol{k}}$ for all $\rho \in \mathcal{D}$, $\boldsymbol{k} \in \mathbb{B}$ and $j \in \mathbb{N}$.

Next, we define for given $\rho \in \mathcal{D}$, $\boldsymbol{k} \in \mathbb{B}$ and $j \in \mathbb{N}$ the set of functions

$$\boldsymbol{U}_j^1(\rho, \boldsymbol{k}) := \big\{ \boldsymbol{u} \in \boldsymbol{U}_j(\rho, \boldsymbol{k}) \;\big|\; \|\boldsymbol{u}\|_\Omega = 1 \big\}.$$

Clearly, the set $\boldsymbol{U}_j^1(\rho, \boldsymbol{k})$ is given by the intersection of the eigenspace $\boldsymbol{U}_j(\rho, \boldsymbol{k})$ with the unit sphere in the complex Hilbert space $\boldsymbol{Z}$. The space $\boldsymbol{Z}$ was defined in Section 4.3.

With the above definitions, we have the following fundamental result.

**Theorem 6.16.** *For every index $j \in \mathbb{N}$ and every vector $\boldsymbol{k} \in \mathbb{B}$ the generalized differentials of the functional $\lambda_j(\,\cdot\,,\boldsymbol{k}) : \mathcal{D} \to \mathbb{R}$ are given by*

$$\partial\lambda_j(\,\cdot\,,\boldsymbol{k})(\rho) = \left\{ p \in \mathcal{E}^* \;\middle|\; p(\eta) = \int_\Omega \eta|(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}|^2,\ \boldsymbol{u} \in \boldsymbol{U}_j^1(\rho,\boldsymbol{k}) \right\}$$

*for all $\rho \in \mathcal{D}$.*

*Proof.* According to Theorem 5.2 on page 93 the functional $\lambda_j(\,\cdot\,,\boldsymbol{k}) : \mathcal{D} \to \mathbb{R}$ is locally Lipschitz continuous for every index $j \in \mathbb{N}$ and every vector $\boldsymbol{k} \in \mathbb{B}$. Hence, the generalized differential of $\lambda_j(\,\cdot\,,\boldsymbol{k})$ exists at every point $\rho \in \mathcal{D}$.

Now let $\rho_0 \in \mathcal{D}$ be chosen arbitrarily. By definition of the set $\mathcal{D}$ (see (4.32) on page 59) we have that

$$\operatorname*{ess\,inf}_\Omega(\rho_0) \geq \delta > 0$$

for some positive number $\delta > 0$. Letting $\mathcal{B}_{\delta/2}(\rho_0)$ denote the open ball in $\mathcal{E}$ with radius $\delta/2$ and center $\rho_0$, we have that $\mathcal{B}_{\delta/2}(\rho_0) \subset \mathcal{D}$, as well as

$$\operatorname*{ess\,sup}_\Omega(\rho_0) + \delta \geq \operatorname*{ess\,sup}_\Omega(\rho) \geq \operatorname*{ess\,inf}_\Omega(\rho) \qquad \text{for all } \rho \in \mathcal{B}_{\delta/2}(\rho_0).$$

Next, we define the function $\widehat{\kappa} : \mathbb{B} \to \mathbb{R}_{\geq 0}$ by

$$\widetilde{\kappa}(\boldsymbol{k}) := \Big(\operatorname*{ess\,sup}_\Omega(\rho_0) + \delta\Big)\kappa_0(\boldsymbol{k}) \qquad \text{for all } \boldsymbol{k} \in \mathbb{B},$$

where $\kappa_0$ denotes the function defined in Lemma 4.4 on page 52. By defintion of the function $\kappa$ (see (4.40) on page 63) and the above estimate, we have that $\widetilde{\kappa}(\boldsymbol{k}) \geq \kappa(\rho,\boldsymbol{k})$ for all $\rho \in \mathcal{B}_{\delta/2}(\rho_0)$ and all $\boldsymbol{k} \in \mathbb{B}$.

For every $\rho \in B^\infty_{\delta/2}(\rho_0)$ and every vector $\boldsymbol{k} \in \mathbb{B}$ we define the sesquilinear form $\widetilde{l}_{\boldsymbol{k}}(\rho) : \boldsymbol{V}_{\boldsymbol{k}} \times \boldsymbol{V}_{\boldsymbol{k}} \to \mathbb{C}$ by $\widetilde{l}_{\boldsymbol{k}}(\rho) := a_{\boldsymbol{k}}(\rho) + \widetilde{\kappa}(\boldsymbol{k})m$. Recall that the sesquilinear forms $a_{\boldsymbol{k}}(\rho)$ and $m$ were defined by (4.33) and (4.35) on page 59. According to Proposition 4.14, for every $\rho \in B^\infty_{\delta/2}(\rho_0)$ and every $\boldsymbol{k} \in \mathbb{B}$ we have that

$$\widetilde{l}_{\boldsymbol{k}}(\rho)(\boldsymbol{u},\boldsymbol{u}) \geq a_{\boldsymbol{k}}(\rho)(\boldsymbol{u},\boldsymbol{u}) + \kappa(\rho,\boldsymbol{k})m(\boldsymbol{u},\boldsymbol{u}) \geq \alpha(\rho,\boldsymbol{k})\|\boldsymbol{u}\|^2_{\mathrm{curl},\Omega} \qquad \text{for all } \boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}}.$$

It follows that the sesquilinear form $\widetilde{l}_{\boldsymbol{k}}(\rho)$ is conjugate-symmetric, bounded and $\boldsymbol{V}_{\boldsymbol{k}}$-elliptic for all $\rho \in B^\infty_{\delta/2}(\rho_0)$ and all $\boldsymbol{k} \in \mathbb{B}$.

Given a vector $\boldsymbol{k} \in \mathbb{B}$, we denote by $\overline{\mathbb{L}}(\boldsymbol{V}_{\boldsymbol{k}})$ the linear space consisting of all continuous conjugate linear operators from $\boldsymbol{V}_{\boldsymbol{k}}$ into $\boldsymbol{V}_{\boldsymbol{k}}^*$. Furthermore, let $\mathbb{E}(\boldsymbol{V}_{\boldsymbol{k}})$ be the set of all operators $T \in \overline{\mathbb{L}}(\boldsymbol{V}_{\boldsymbol{k}})$, such that the mapping $(\boldsymbol{u},\boldsymbol{v}) \mapsto T\boldsymbol{u}[\boldsymbol{v}]$ is

a conjugate-symmetric, bounded, $\boldsymbol{V_k}$-elliptic sesquilinear form on $\boldsymbol{V_k} \times \boldsymbol{V_k}$. With this, we define for every $\boldsymbol{k} \in \mathbb{B}$ the operator-valued function $\widetilde{L}_{\boldsymbol{k}} : \mathcal{B}_{\delta/2}(\rho_0) \to \mathbb{E}(\boldsymbol{V_k})$ by

$$\widetilde{L}_{\boldsymbol{k}}(\rho)\boldsymbol{u}[\boldsymbol{v}] := \widetilde{l}_{\boldsymbol{k}}(\rho)(\boldsymbol{u}, \boldsymbol{v}) \qquad \text{for all } \rho \in \mathcal{B}_{\delta/2}(\rho_0),\ \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V_k}.$$

According to the definition in Section 4.4, we have that $\widetilde{L}_{\boldsymbol{k}}(\rho) = A_{\boldsymbol{k}}(\rho) + \widetilde{\kappa}(\boldsymbol{k}) M_{\boldsymbol{k}}|_{\boldsymbol{V_k}}$ for all $\rho \in \mathcal{B}_{\delta/2}(\rho_0)$ and all $\boldsymbol{k} \in \mathbb{B}$. It follows that the $j$-th smallest eigenvalue $\widetilde{\lambda}_j(\rho, \boldsymbol{k})$ of $\widetilde{L}_{\boldsymbol{k}}(\rho)$ coincides with $\lambda_j(\rho, \boldsymbol{k}) + \widetilde{\kappa}(\boldsymbol{k})$ for all $\rho \in \mathcal{B}_{\delta/2}(\rho_0)$ and all $\boldsymbol{k} \in \mathbb{B}$. Furthermore, we have that the eigenspaces of $\widetilde{\lambda}_j(\rho, \boldsymbol{k})$ and $\lambda_j(\rho, \boldsymbol{k}) + \widetilde{\kappa}(\boldsymbol{k})$ coincide. Since the function $\widetilde{\kappa}$ only depends on $\boldsymbol{k}$, we also have that $\partial\widetilde{\lambda}_j(\,\cdot\,, \boldsymbol{k})(\rho) = \partial\lambda_j(\,\cdot\,, \boldsymbol{k})(\rho)$ for all $\rho \in \mathcal{B}_{\delta/2}(\rho_0)$ and all $\boldsymbol{k} \in \mathbb{B}$.

Now let $\boldsymbol{k} \in \mathbb{B}$ be fixed. Note that for every coefficient $\rho \in \mathcal{B}_{\delta/2}(\rho_0)$ the operator $\widetilde{L}_{\boldsymbol{k}}(\rho)$ satisfies

$$\widetilde{L}_{\boldsymbol{k}}(\rho)\boldsymbol{u}[\boldsymbol{v}] = \int_\Omega \rho \overline{(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}} \cdot (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{v} + \widetilde{\kappa}(\boldsymbol{k}) \int_\Omega \overline{\boldsymbol{u}} \cdot \boldsymbol{v} \quad \text{for all } \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V_k}.$$

Given a function $\eta \in \mathcal{E}$, such that $\rho_0 + \eta \in \mathcal{B}_{\delta/2}(\rho_0)$, we thus find that

$$\begin{aligned}\big(\widetilde{L}_{\boldsymbol{k}}(\rho_0 + \eta) - \widetilde{L}_{\boldsymbol{k}}(\rho_0)\big)\boldsymbol{u}[\boldsymbol{v}] &= \int_\Omega \eta \overline{(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}} \cdot (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{v} \\ &= A_{\boldsymbol{k}}(\eta)\boldsymbol{u}[\boldsymbol{v}] \qquad \text{for all } \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V_k}.\end{aligned}$$

Clearly, the mapping $\eta \mapsto A_{\boldsymbol{k}}(\eta)$ is linear from $\mathcal{E}$ into $\overline{\mathbb{L}}(\boldsymbol{V_k})$. From this we deduce that the function $\widetilde{L}_{\boldsymbol{k}}$ is continuously Gâteaux differentiable at $\rho_0$ with Gâteaux derivative $\mathrm{D}\widetilde{L}_{\boldsymbol{k}}(\rho_0) = A_{\boldsymbol{k}}(\,\cdot\,)$.

We have thus shown that the operator-valued function $\widetilde{L}_{\boldsymbol{k}}$ satisfies the hypotheses in Section 1 in [28]. The assertion of this theorem then follows directly from Theorem 1 in [28]. □

According to (4.74) on page 80 the functions $\omega_j : \mathcal{D} \times \mathbb{B} \to \mathbb{R}$ defined as $\omega_j = \sqrt{\lambda_j}$ for all $j \in \mathbb{N}$. By Theorem 6.11 we have that the generalized differentials of these functionals are given by

$$\begin{aligned}\partial\omega_j(\,\cdot\,, \boldsymbol{k})(\rho) = \Bigg\{ p \in \mathcal{E}^* \;\Bigg|\; p(\eta) = \int_\Omega \eta \frac{|(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}|^2}{2\omega_j(\rho, \boldsymbol{k})}, \\ \boldsymbol{u} \in \boldsymbol{U}_j^1(\rho, \boldsymbol{k}) \Bigg\}\end{aligned} \tag{6.9}$$

for all $j \in \mathbb{N}$, $\rho \in \mathcal{D}$ and $\boldsymbol{k} \in \mathbb{B}$. For convenience, we define for every index $j \in \mathbb{N}$ the functionals $J_{\mathrm{up},j}, J_{\mathrm{lo},j} : \mathcal{D} \to \mathbb{R}$ by

$$J_{\mathrm{lo},j}(\rho) := \max_{\boldsymbol{k} \in \mathbb{B}} \omega_j(\rho, \boldsymbol{k}), \qquad \text{for all } j \in \mathbb{N},\ \rho \in \mathcal{D}, \tag{6.10}$$

$$J_{\mathrm{up},j}(\rho) := \min_{\boldsymbol{k}\in\mathbb{B}} \omega_{j+1}(\rho,\boldsymbol{k}), \qquad \text{for all } j\in\mathbb{N},\ \rho\in\mathcal{D}. \tag{6.11}$$

Given an index $j$, we call the functional $J_{\mathrm{up},j}$ the *$j$-th upper band edge functional*. Similarly, we refer to $J_{\mathrm{lo},j}$ as the *$j$-th lower band edge functional*. By Theorem 6.13 we obtain

$$\partial J_{\mathrm{lo},j}(\rho) \subseteq \Bigg\{ \int_{\mathbb{B}} q_{\boldsymbol{k}}(\,\cdot\,)\,\mathrm{d}\mu(\boldsymbol{k}) \;\Bigg|\; q_{\boldsymbol{k}} \in \partial\omega_j(\,\cdot\,,\boldsymbol{k})(\rho), \ \mu \in \mathrm{pr}_{\mathbb{B}}\Big(\arg\max_{\mathbb{B}} \omega_j(\rho,\,\cdot\,)\Big)\Bigg\}, \tag{6.12}$$

$$\partial J_{\mathrm{up},j}(\rho) \subseteq \Bigg\{ \int_{\mathbb{B}} q_{\boldsymbol{k}}(\,\cdot\,)\,\mathrm{d}\mu(\boldsymbol{k}) \;\Bigg|\; q_{\boldsymbol{k}} \in \partial\omega_{j+1}(\,\cdot\,,\boldsymbol{k})(\rho), \ \mu \in \mathrm{pr}_{\mathbb{B}}\Big(\arg\min_{\mathbb{B}} \omega_{j+1}(\rho,\,\cdot\,)\Big)\Bigg\} \tag{6.13}$$

for all $j\in\mathbb{N}$ and all $\rho\in\mathcal{D}$. Since $J_j = J_{\mathrm{lo},j} - J_{\mathrm{up},j}$, we have by Proposition 6.9 and Proposition 6.10 that the generalized differential of the $j$-th gap width functional satisfies the inclusion relation

$$\partial J_j(\rho) \subseteq \partial J_{\mathrm{lo},j}(\rho) - \partial J_{\mathrm{up},j}(\rho) \qquad \text{for all } j\in\mathbb{N},\ \rho\in\mathcal{D}. \tag{6.14}$$

In practice, one finds that the functionals $\omega_j(\rho,\,\cdot\,)$ typically attain their extrema at a small number of points in the first Brillouin zone $\mathbb{B}$. Hence, by assuming that the set of minimizers and maximizers are given by

$$\arg\min_{\mathbb{B}} \omega_j(\rho,\,\cdot\,) = \big\{\boldsymbol{k}_{\mathrm{lo},j}^{(1)},\dots,\boldsymbol{k}_{\mathrm{lo},j}^{(N_{\mathrm{lo},j})}\big\}, \tag{6.15}$$

$$\arg\max_{\mathbb{B}} \omega_{j+1}(\rho,\,\cdot\,) = \big\{\boldsymbol{k}_{\mathrm{up},j}^{(1)},\dots,\boldsymbol{k}_{\mathrm{up},j}^{(N_{\mathrm{up},j})}\big\} \tag{6.16}$$

for some numbers $N_{\mathrm{lo},j}, N_{\mathrm{up},j}\in\mathbb{N}$, we obtain by Corollary 6.14 the inclusion relations

$$\partial J_{\mathrm{lo},j}(\rho) \subseteq \Bigg\{ \sum_{l=1}^{N_{\mathrm{lo},j}} \theta_l q_l \;\Bigg|\; q_l \in \partial\omega_j(\,\cdot\,,\boldsymbol{k}_{\mathrm{lo},j}^{(l)})(\rho),\ \theta_l\in[0,1],\ \sum_{l=1}^{N_{\mathrm{lo},j}}\theta_l = 1\Bigg\}, \tag{6.17}$$

$$\partial J_{\mathrm{up},j}(\rho) \subseteq \Bigg\{ \sum_{l=1}^{N^{\mathrm{up}}} \theta_l q_l \;\Bigg|\; q_l \in \partial\omega_{j+1}(\cdot,\boldsymbol{k}_{\mathrm{up},l}^{(l)})(\rho),\ \theta_l\in[0,1],\ \sum_{l=1}^{N_{\mathrm{up},j}}\theta_l = 1\Bigg\} \tag{6.18}$$

for all $j\in\mathbb{N}$ and all $\rho\in\mathcal{D}$.

For our further discussion, we find it useful to define for every function $\boldsymbol{u}\in\boldsymbol{V}_{\boldsymbol{k}}$, every index $j\in\boldsymbol{V}_{\boldsymbol{k}}$ the functions $\gamma_{\mathrm{lo},j}^{(l)}(\boldsymbol{u}), \gamma_{\mathrm{up},j}^{(l)}(\boldsymbol{u}) : \Omega\to\mathbb{R}$ by

$$\gamma_{\mathrm{lo},j}^{(l)}(\boldsymbol{u}) := \frac{\big|(\nabla + \mathrm{i}\boldsymbol{k}_{\mathrm{lo},j}^{(l)})\times\boldsymbol{u}\big|^2}{2\,\omega_j(\rho,\boldsymbol{k}_{\mathrm{lo},j}^{(l)})} \qquad \text{for } l=1,\dots,N_{\mathrm{lo},j},$$

$$\gamma_{\mathrm{up},j}^{(l)}(\boldsymbol{u}) := \frac{\left|(\nabla + \mathrm{i}\boldsymbol{k}_{\mathrm{up},j}^{(l)}) \times \boldsymbol{u}\right|^2}{2\,\omega_{j+1}(\rho, \boldsymbol{k}_{\mathrm{up},j}^{(l)})} \qquad \text{for } l = 1, \ldots, N_{\mathrm{up},j}.$$

Notice that $(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}$ is a vector field in $\boldsymbol{Z} = L^2(\Omega)^3$ for every vector $\boldsymbol{k} \in \mathbb{B}$ and every vector field $\boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}}$ according to Proposition 4.3(a) on page 50. It thus follows, that $|(\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{u}|^2$ is a scalar field in $L^1(\Omega, \mathbb{R})$ for all $\boldsymbol{k} \in \mathbb{B}$ and all $\boldsymbol{u} \in \boldsymbol{V}_{\boldsymbol{k}}$, and the same also applies to the functions $\gamma_{\mathrm{lo},j}^{(l)}(\boldsymbol{u})$ and $\gamma_{\mathrm{up},j}^{(l)}(\boldsymbol{u})$.

According to Theorem 6.16 and (6.9), the inclusion relations (6.17) and (6.18) can be written equivalently as

$$\partial J_{\mathrm{lo},j}(\rho) \subseteq \left\{ p \in \mathcal{E}^* \;\middle|\; p(\eta) = \int_\Omega \eta\gamma,\ \gamma \in \Gamma_{\mathrm{lo},j}(\rho) \right\}, \tag{6.19}$$

$$\partial J_{\mathrm{up},j}(\rho) \subseteq \left\{ p \in \mathcal{E}^* \;\middle|\; p(\eta) = \int_\Omega \eta\gamma,\ \gamma \in \Gamma_{\mathrm{up},j}(\rho) \right\} \tag{6.20}$$

for all $j \in \mathbb{N}$ and all $\rho \in \mathcal{D}$, where the sets $\Gamma_{\mathrm{lo},j}(\rho)$ and $\Gamma_{\mathrm{up},j}(\rho)$ are defined by

$$\Gamma_{\mathrm{lo},j}(\rho) := \left\{ \sum_{l=1}^{N_{\mathrm{lo},j}} \theta_l\, \gamma_{\mathrm{lo},j}^{(l)}(\boldsymbol{u}^{(l)}) \;\middle|\; \boldsymbol{u}^{(l)} \in \boldsymbol{U}_j^1(\rho, \boldsymbol{k}_{\mathrm{lo},j}^{(l)}), \right.$$
$$\left. \theta_l \in [0,1],\ \sum_{l=1}^{N_{\mathrm{lo},j}} \theta_l = 1 \right\},$$

$$\Gamma_{\mathrm{up},j}(\rho) := \left\{ \sum_{l=1}^{N_{\mathrm{up},j}} \theta_l\, \gamma_{\mathrm{up},j}^{(l)}(\boldsymbol{u}^{(l)}) \;\middle|\; \boldsymbol{u}^{(l)} \in \boldsymbol{U}_{j+1}^1(\rho, \boldsymbol{k}_{\mathrm{up},j}^{(l)}), \right.$$
$$\left. \theta_l \in [0,1],\ \sum_{l=1}^{N_{\mathrm{up},j}} \theta_l = 1 \right\}.$$

Under the hypotheses (6.15) and (6.16), the inclusion relations (6.19) and (6.20) reveal some important aspects about the generalized differentials of the gap width functionals. First, we notice that $\Gamma_{\mathrm{lo},j}(\rho)$ and $\Gamma_{\mathrm{up},j}(\rho)$ are subsets of $L^1(\Omega, \mathbb{R})$. Hence, according to (6.14), every continuous, linear functional contained in $\partial J_j(\rho)$ can be expressed as a weighted integral over $\Omega$ with a weight function in $L^1(\Omega, \mathbb{R})$. The generalized differentials of the gap width functionals are hence weak-$*$-closed, convex, bounded subsets of

$$\mathcal{E}_{\mathrm{reg}}^* := \left\{ p \in \mathcal{E}^* \;\middle|\; p(\eta) = \int_\Omega \eta w,\ w \in L^1(\Omega, \mathbb{R}) \right\}.$$

Since the Banach space $\mathcal{E}$ is not reflexive, the space $\mathcal{E}_{\mathrm{reg}}^*$ is a proper linear subspace of $\mathcal{E}^*$. It is easy to see that the space $\mathcal{E}_{\mathrm{reg}}^*$ is isometrically, linearly isomorphic to $L^1(\Omega, \mathbb{R})$.

Concerning the differentiability of gap width functionals, we are able to make the following conclusive statement. Recall that by Proposition 6.7(d) a gap width functional $J_j$ is strictly differentiable at a point $\rho \in \mathcal{D}$, if and only if the generalized differential $\partial J_j(\rho)$ contains exactly one element, which is then the strict differential of $J_j$ at $\rho$. According to (6.19) and (6.20), the generalized differential $\partial J_j(\rho)$ is given by a singleton, if and only if each of the functions $\omega_j(\rho, \cdot\,)$ and $\omega_{j+1}(\rho, \cdot\,)$ attain their minimum and maximum at exactly one vector $\boldsymbol{k}_{\mathrm{lo},j} \in \mathbb{B}$ and $\boldsymbol{k}_{\mathrm{up},j} \in \mathbb{B}$, respectively, and if the corresponding eigenspaces $\boldsymbol{U}_j(\rho, \boldsymbol{k}_{\mathrm{lo},j})$ and $\boldsymbol{U}_{j+1}(\rho, \boldsymbol{k}_{\mathrm{lo},j})$ are one-dimensional.

Finally, we remark that the sets $\Gamma_{\mathrm{lo},j}(\rho)$ and $\Gamma_{\mathrm{up},j}(\rho)$ are completely determined by eigensolutions of the weakly formulated family of eigenvalue problems given by Problem 4.16. These eigensolutions also need to be computed in order to evaluate the gap width functional $J_j$ at a point $\rho \in \mathcal{D}$. Therefore, in a numerical algorithm the sets $\Gamma_{\mathrm{lo},j}(\rho)$ and $\Gamma_{\mathrm{up},j}(\rho)$ can be computed simultaneously with the goal value $J_j(\rho)$.

## 6.4 The Two-Dimensional Case

In the previous section we derived the generalized differentials of the gap width functionals $J_j$ for three-dimensional photonic crystals. In Section 5.5 we also defined for every index $j \in \mathbb{N}$ the TM and TE gap width functionals $J^{\mathrm{TE}} : \mathcal{D} \to \mathbb{R}$ and $J^{\mathrm{TE}} : \mathcal{D} \to \mathbb{R}$ by

$$\begin{aligned} J_j^{\mathrm{TM}}(\rho) &:= \max_{\boldsymbol{k}\in\mathbb{B}} \omega_j^{\mathrm{TM}}(\rho, \boldsymbol{k}) - \min_{\boldsymbol{k}\in\mathbb{B}} \omega_{j+1}^{\mathrm{TM}}(\rho, \boldsymbol{k}) \qquad \text{for all } \rho \in \mathcal{D}, \\ J_j^{\mathrm{TE}}(\rho) &:= \max_{\boldsymbol{k}\in\mathbb{B}} \omega_j^{\mathrm{TE}}(\rho, \boldsymbol{k}) - \min_{\boldsymbol{k}\in\mathbb{B}} \omega_{j+1}^{\mathrm{TE}}(\rho, \boldsymbol{k}) \qquad \text{for all } \rho \in \mathcal{D}. \end{aligned}$$

For completeness, we give a characterization of the generalized differentials of these functionals in this section.

In the following we assume that $\Omega$ is a primitive domain of a Bravais lattice in $\mathbb{R}^2$ of rank 2. By $\mathbb{B}$ we denote the lattice's first Brillouin zone. In analogy to the three-dimensional setting, we define the eigenspaces

$$\begin{aligned} U_j^{\mathrm{TM}}(\rho, \boldsymbol{k}) &:= \operatorname{span}\big\{ u_i^{\mathrm{TM}}(\rho, \boldsymbol{k}) \;\big|\; \lambda_i(\rho, \boldsymbol{k}) = \lambda_j(\rho, \boldsymbol{k}) \big\}, \\ U_j^{\mathrm{TE}}(\rho, \boldsymbol{k}) &:= \operatorname{span}\big\{ u_i^{\mathrm{TE}}(\rho, \boldsymbol{k}) \;\big|\; \lambda_i^{\mathrm{TE}}(\rho, \boldsymbol{k}) = \lambda_j^{\mathrm{TE}}(\rho, \boldsymbol{k}) \big\}, \end{aligned}$$

as well as the function sets

$$\begin{aligned} U_j^{\mathrm{TM}1}(\rho, \boldsymbol{k}) &:= \big\{ u \in U_j^{\mathrm{TM}}(\rho, \boldsymbol{k}) \;\big|\; \|u\|_\rho = 1 \big\}, \\ U_j^{\mathrm{TE}1}(\rho, \boldsymbol{k}) &:= \big\{ u \in U_j^{\mathrm{TE}}(\rho, \boldsymbol{k}) \;\big|\; \|u\|_\Omega = 1 \big\}. \end{aligned}$$

for all $j \in \mathbb{N}$, $\rho \in \mathcal{D}$, $\rho \in \mathcal{D}$ and all $\boldsymbol{k} \in \mathbb{B}$. Recall that the norm $\|\cdot\|_\rho$ was defined by $\|u\|_\rho^2 = m^{\mathrm{TM}}(\rho)(u,u)$ for all $u \in L^2(\Omega)$ (see (4.87) on page 86). With the above definition, we can give the following characterization of the generalized differentials of the TM and TE eigenvalues.

**Theorem 6.17.** *For every index $j \in \mathbb{N}$ and every vector $\boldsymbol{k} \in \mathbb{B}$ the generalized differentials of the functionals $\lambda_j^{\mathrm{TM}}(\,\cdot\,,\boldsymbol{k}) : \mathcal{D} \to \mathbb{R}$ and $\lambda_j^{\mathrm{TE}}(\,\cdot\,,\boldsymbol{k}) : \mathcal{D} \to \mathbb{R}$ are given by*

$$\partial\lambda_j^{\mathrm{TM}}(\,\cdot\,,\boldsymbol{k})(\rho) = \left\{ p \in \mathcal{E}^* \;\middle|\; p(\eta) = -\lambda_j^{\mathrm{TM}}(\rho,\boldsymbol{k}) \int_\Omega \eta|u|^2,\ u \in U_j^{\mathrm{TM}^1}(\rho,\boldsymbol{k}) \right\},$$

$$\partial\lambda_j^{\mathrm{TE}}(\,\cdot\,,\boldsymbol{k})(\rho) = \left\{ p \in \mathcal{E}^* \;\middle|\; p(\eta) = \int_\Omega \eta|(\nabla + \mathrm{i}\boldsymbol{k})u|^2,\ u \in U_j^{\mathrm{TE}^1}(\rho,\boldsymbol{k}) \right\}$$

*for all $\rho \in \mathcal{D}$ and all $\rho \in \mathcal{D}$.*

A proof of the first identity in Theorem 6.17 is given in Section 3 in [29]. The second identity can be proved by a similar line of argument as given in the proof of Theorem 6.16. According to Theorem 6.11, the generalized differentials of the functions $\omega_j^{\mathrm{TM}} : \mathcal{D} \to \mathbb{R}$ and $\omega_j^{\mathrm{TE}} : \mathcal{D} \to \mathbb{R}$ are given by

$$\partial\omega_j^{\mathrm{TM}}(\,\cdot\,,\boldsymbol{k})(\rho) = \left\{ p \in \mathcal{E}^* \;\middle|\; p(\eta) = -\omega_j^{\mathrm{TM}}(\rho,\boldsymbol{k}) \int_\Omega \eta|u|^2, \right.$$
$$\left. u \in U_j^{\mathrm{TM}^1}(\rho,\boldsymbol{k}) \right\},$$

$$\partial\omega_j^{\mathrm{TE}}(\,\cdot\,,\boldsymbol{k})(\rho) = \left\{ p \in \mathcal{E}^* \;\middle|\; p(\eta) = \int_\Omega \eta \frac{|(\nabla + \mathrm{i}\boldsymbol{k})u|^2}{2\omega_j^{\mathrm{TE}}(\rho,\boldsymbol{k})}, \right.$$
$$\left. u \in U_j^{\mathrm{TE}^1}(\rho,\boldsymbol{k}) \right\}$$

for all $j \in \mathbb{N}$, $\rho \in \mathcal{D}$, and all $\boldsymbol{k} \in \mathbb{B}$. From here, one can proceed exactly as in the previous section and characterize the generalized differentials of $J_j^{\mathrm{TM}}$ and $J_j^{\mathrm{TE}}$ in terms of integrals with respect to the variable $\boldsymbol{k}$ over subsets of $\mathbb{B}$, where the functions $\omega_j^{\mathrm{TM}}$ and $\omega_j^{\mathrm{TE}}$ attain their minima or maxima. By defining the *TM band edge functionals* $J_{\mathrm{lo},j}^{\mathrm{TM}}, J_{\mathrm{up},j}^{\mathrm{TM}} : \mathcal{D} \to \mathbb{R}$, as well as the *TE band edge functionals* $J_{\mathrm{lo},j}^{\mathrm{TE}}, J_{\mathrm{up},j}^{\mathrm{TE}} : \mathcal{D} \to \mathbb{R}$ in analogy to (6.10) and (6.11), one obtains inclusion relations for their generalized differentials, which are analogous to (6.12) and (6.13). By assuming that the minima and maxima are attained at a finite number of quasimomentum vectors $\boldsymbol{k}$, one obtains inclusion relations analogous to (6.17) and (6.18). The generalized gradients of the TM gap width functionals $J_j^{\mathrm{TM}}$ satisfy the inclusion relations $\partial J_j^{\mathrm{TM}}(\rho) \subseteq \partial J_{\mathrm{lo},j}^{\mathrm{TM}}(\rho) - J_{\mathrm{up},j}^{\mathrm{TM}}(\rho)$ for all $j \in \mathbb{N}$ and all $\rho \in \mathcal{D}$, and the analogous result also holds for the generalized gradients of the TE gap width functionals $J_j^{\mathrm{TE}}$.

# Chapter 7

# A Generalized Gradient Method

In this section, we develop an optimization algorithm for discretized photonic band structure optimization problems based on the concept of generalized gradient methods. In Section 7.1 we outline the basic idea of generalized gradient methods and comment on their mathematical justification and practical applicability. In Section 7.2 we present a discretization scheme for photonic band structure optimization problems. In Section 7.3 and Section 7.4 we develop two approaches to incorporate certain optimization constraints into a generalized gradient method for photonic band structure optimization problems. An iterative optimization algorithm based on these approaches is presented in Section 7.5. A crucial step in each iteration of this algorithm is the choice of a so-called descent direction. In Section 7.6 we present a strategy for the choice of descent directions, which can be applied to photonic band gap maximization problems. In Section 7.7 we make some final remarks on the generalized gradient method.

## 7.1 Basic Concepts

We consider a minimization problem of the form

$$\underset{x \in X}{\text{minimize}} \; f(x), \tag{7.1}$$

where $f$ is a locally Lipschitz continuous goal functional defined on a finite-dimensional Hilbert space $X$ over $\mathbb{R}$ with inner product $\langle \cdot \, , \, \cdot \rangle_X$. A minimization problem of this type is commonly referred to as a *locally Lipschitz minimization problem*. Due to its local Lipschitz continuity, the functional $f$ admits a generalized differential $\partial f(x)$ at every point $x \in X$. Since $X$ is a Hilbert space, the generalized differential $\partial f(x)$ can be identified with a subset $G_f(x)$ of $X$ via the Rieszian isomorphism. This subset is called the generalized gradient of $f$ at $x$. It follows from Proposition 6.4 on page 113 that $G_f(x)$ is a closed, convex and bounded

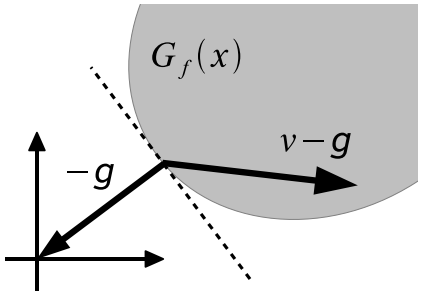



**Figure 7.1:** The geometric meaning of the variational inequality $\langle -g, v-g \rangle \leq 0$, where $v$ is an arbitrary vector in $G_f(x)$, and where $g$ is the best approximation in $G_f(x)$ of the zero element in $X$.

subset of $X$ for every $x \in X$. The closedness of $G_f(x)$ follows from the fact that $\partial f(x)$ is weak-$*$-compact, which implies that $G_f(x)$ is weakly compact. Since $X$ is a finite-dimensional Hilbert space, weak compactness implies compactness, which in turn implies closedness according to the Heine–Borel Theorem.

A *generalized gradient method* is an optimization algorithm, which can be used to find local minima of locally Lipschitz continuous functionals defined on a finite-dimensional, real Hilbert space. A generalized gradient method for the minimization problem (7.1) attempts to construct a minimizing sequence $x_0, x_1, x_2, \ldots$, etc. in $X$ for the goal functional $f$. The *initial point* $x_0 \in X$ is chosen arbitrarily. The points $x_1, x_2, \ldots$ etc., are determined iteratively according to

$$x_l := x_{l-1} - s_l \, g_{l-1} \quad \text{for } l = 1, 2, \ldots,$$

where $s_l > 0$ is a suitable *step size*, and where $g_{l-1}$ is a suitable element of $G_f(x_{l-1})$ for every $l \in \mathbb{N}$. Here, the variable $l \in \mathbb{N}_0$ enumerates the iterations of the generalized gradient method. In each iteration, the generalized gradient element $g_{l-1}$ is chosen such that $-g_{l-1}$ is a *descent direction* of $f$ at the point $x_{l-1}$, if possible. More precisely, $g_{l-1} \in G_f(x_{l-1})$ is chosen such that there exists a positive number $t_{\max} > 0$, such that

$$f(x_{l-1} - t \, g_{l-1}) \leq f(x_{l-1}) \qquad \text{for all } t \in [0, t_{\max}).$$

The following lemma informs us on how an element $g_{l-1} \in G_f(x)$ can be chosen, such that $-g_{l-1}$ is a descent direction.

**Lemma 7.1.** *Suppose that $0 \notin G_f(x)$. Let $g \in G_f(x)$, such that $\|g\|_X \leq \|v\|_X$ for all $v \in G_f(x)$. Then, $-g$ is a descent direction of $f$ at $x$.*

*Proof.* The following proof is a particularization of the proof of Proposition 6.2.4 in [22]. Since $G_f(x)$ is convex and closed, there exists a uniquely defined element $g \in G_f(x)$, such that $\|g\|_X \leq \|v\|_X$ for all $v \in G_f(x)$. Note that the element $g$ is the best approximation in the convex set $G_f(x)$ of the zero element in $X$. Hence, we have the variational inequality

$$\langle -g \, , v - g \rangle \leq 0 \qquad \text{for all } v \in G_f(x).$$

It should be noted that this variational inequality has a simple geometric meaning, namely that the angle between the vector $-g$ and every other vector, that connects the point $g$ to another point in the set $G_f(x)$, is greater than or equal to $\pi/2$ (see also Figure 7.1).

$$\langle -g \, , v \rangle_X \leq \langle -g \, , g \rangle_X = -\|g\|_X^2.$$

By Proposition 2.1.5 in [22], there exists for every given number $\varepsilon > 0$ a number $\delta > 0$, such that

$$G_f(y) \subset G_f(x) + B_\varepsilon(0) \qquad \text{for all } y \in X \text{ with } \|x - y\|_X < \delta,$$

where $B_\varepsilon(0)$ denotes the open ball in $X$ with radius $\varepsilon$ centered at zero. Here, we choose $\delta$, such that the above inclusion relation holds for $\varepsilon := \|g\|_X/2$. We also set $t_{\max} := \delta/\|g\|_X$. Choosing an arbitrary number $t \in [0, t_{\max})$, we have by Lebourg's theorem (see Theorem 6.6 on page 114) that there exists an element $z \in \operatorname{conv}\{x, x - \delta g\}$ and an element $u \in G_f(z)$, such that

$$f(x - tg) - f(x) = -t\langle u \, , g \rangle.$$

Since $\|x - z\|_X \leq t\|g\|_X < \delta$, we have that $G_f(z) \subset G_f(x) + B_\varepsilon(0)$. Hence, there exists an element $v \in G_f(z)$ and an element $w \in X$ with $\|w\|_X < \delta$, such that $u = v + w$. Therefore, we have that

$$\begin{aligned} f(x - tg) - f(x) &= -t\langle v + w \, , g \rangle_X = t\langle -g \, , v \rangle_X - t\langle w \, , g \rangle_X \\ &\leq -t\|g\|_X^2 + t\|w\|_X\|g\|_X \leq -t\|g\|_X^2 + t\varepsilon\|g\|_X \\ &= -t\frac{\|g\|_X^2}{2}, \end{aligned}$$

which completes this proof. □

Notice that one needs to solve a convex minimization problem in order to find the element $g$ proposed by Lemma 7.1. If the element $g$ turns out to be zero, then a

necessary condition for a local extremum of $f$ is satisfied at the point $x$ according to Lemma 6.5 on page 113.

If the functional $f$ is convex in a neighbourhood $U$ of the point $x$, then $-g$ is a descent direction for every element $g \in G_f(x) \setminus \{0\}$. This is due to Proposition 6.7(a) on page 114, which states that $\partial f(x)$ coincides with the subdifferential of $f$ at $x$ under the above assumption. Hence, $G_f(x)$ corresponds to the so-called *subgradient* of $f$ at $x$. This subgradient is defined as the set of all elements $u \in X$, such that

$$f(y) - f(x) \geq \langle u \,,\, y - x \rangle_X \qquad \text{for all } y \in U.$$

Finally, recall that $G_f(x)$ is a singleton, if and only if $f$ is strictly differentiable at $x$. Then, the single element contained in $G_f(x)$ is the strict gradient $\nabla_{\mathrm{s}} f(x)$ of $f$ at $x$. If the strict gradient does not vanish, it easily follows that $-\nabla_{\mathrm{s}} f(x)$ is a descent direction of $f$ at the point $x$.

It should be noted, that in many cases the generalized gradient $G_f(x)$ is not known explicitly. Often, one can only determine a superset of $G_f(x)$ using the rules of calculus presented in Section 6.2. In these cases, specific *choice strategies* for the descent directions have to be devised. These choice strategies have a significant impact on the performance of a generalized gradient method. We shall comment on this in Section 7.7.

So far, results concerning the convergence of generalized gradient methods could only be established for specific types of locally Lipschitz minimization problems (see e.g. [79], [19]). As far as can be judged from the literature, a general convergence theory for generalized gradient methods is not available yet.

## 7.2 Discretization

The proof of Lemma 7.1 relies on the fact that $X$ is a finite-dimensional Hilbert space. In particular, the proof invokes Proposition 2.1.5 in [22], which also assumes that $X$ is finite-dimensional. So far, it is not known whether the results of Lemma 7.1 also apply to infinite-dimensional Hilbert spaces. If $X$ is a Banach space, the additional problem arises that the generalized differential $\partial f(x) \subset X^*$ cannot be identified with a set $G_f(x) \subset X$, in general. In particular, such an identification is generally not possible for $X = \mathcal{E}$. Recall, for example, that the generalized differentials of the gap width functionals can only be identified with subsets of $L^1(\Omega, \mathbb{R})$ (see Section 6.3). Due to this, we were not able to develop a generalized gradient method for the minimization problem

$$\underset{\rho \in \mathcal{C}}{\text{minimize}} \;\; J(\rho).$$

However, we could develop a optimization algorithm that can be applied to the corresponding discretized problem

$$\underset{\rho_h \in \mathcal{C}_h}{\text{minimize}} \; J_h(\rho_h), \tag{7.2}$$

where $\mathcal{C}_h$ denotes a discretization of the admissible set $\mathcal{C}$, and where $J_h$ denotes the corresponding discretization of the gap width functional $J$. In this section we explain the underlying discretization scheme

Given a primitive domain $\Omega$ of a Bravais lattice $\Lambda \subset \mathbb{R}^3$ of rank 3, we suppose that

$$\mathcal{T}_h = \{T_1, T_2, \ldots, T_N\}$$

is a *finite element mesh* on $\Omega$ for some $N \in \mathbb{N}$, where the $h > 0$ indicates the maximal diameter of the elements $T_i$ in the mesh. We suppose that the Hilbert spaces $\boldsymbol{W}$ and $Q$, as defined by (4.28) and (4.29) on page 59, are discretized by conforming finite element spaces $\boldsymbol{W}_h$ and $Q_h$. Given a coefficient $\rho \in \mathcal{D}$ and a finite set of quasimomentum vectors $K_h \subset \mathbb{B}$, an approximate band structure is computed by solving for each $\boldsymbol{k} \in K_h$ the constrained matrix eigenvalue problem

$$\begin{cases} \mathbf{A}_{\boldsymbol{k},h}(\rho)\mathbf{u} = \lambda_h \mathbf{M}_{\boldsymbol{k},h}\mathbf{u}, \\ \quad \mathbf{B}_{\boldsymbol{k},h}\mathbf{u} = \mathbf{0}. \end{cases} \tag{7.3}$$

Here, $\boldsymbol{A}_{\boldsymbol{k},h}(\rho_h)$, $\boldsymbol{M}_{\boldsymbol{k},h}$, and $\boldsymbol{B}_{\boldsymbol{k},h}$ are the finite element matrices that realize the operators $A_{\boldsymbol{k}}(\rho)$, $M_{\boldsymbol{k}}$ and $B_{\boldsymbol{k}}$, as defined by (4.43) and (4.44) on page 65, and by (4.70) on page 78, on the respective finite element spaces.

Typically, finite element matrices are assembled by applying specific quadrature rules. These quadrature rules are realized by evaluating the element shape functions at a finite set of quadrature points. During the matrix assemblation, coefficients such as $\rho$ are also evaluated at these quadrature points only. For convenience, we define the set

$$\mathcal{X}_h := \big\{ \boldsymbol{x} \in \mathbb{R}^3 \,\big|\, \boldsymbol{x} \text{ is a quadrature point of } \mathcal{T}_h \big\}. \tag{7.4}$$

Since the coefficient $\rho$ is only evaluated at the points in $\mathcal{X}_h$ during the assemblation of $\boldsymbol{A}_{h,\boldsymbol{k}}(\rho)$, we devise the following discretization schemes for the sets $\mathcal{C}, \overline{\mathcal{C}}$, and $\mathcal{D}$. First, we define the finite-dimensional, linear function space

$$\mathcal{E}_h := \{\xi_h \mid \xi_h : \mathcal{X}_h \to \mathbb{R}\}. \tag{7.5}$$

Clearly, the space $\mathcal{E}_h$ consists of all real-valued functions, which are defined on the set of quadrature points of the finite element mesh $\mathcal{T}_h$. In the following, we shall use $\mathcal{E}_h$ as a discretization of $\mathcal{E}$. An interpolation from the infinite-dimensional

Banach space $\mathcal{E}$ into the discrete space $\mathcal{E}_h$ can be defined as follows: Suppose that for every element $T_i$ in the finite element mesh $\mathcal{T}_h$ we are given an associated linear function space $P_i$ that consists of real-valued polynomial functions defined on $T_i$. Then, we define the *interpolation operator* $\Pi_h : \mathcal{E} \to \mathcal{E}_h$ by

$$\Pi_h \xi := \pi(\xi)|_{\mathcal{X}_h} \qquad \text{for all } \xi \in \mathcal{E}, \tag{7.6}$$

where $\pi(\xi)$ is the element-wise polynomial function that satisfies $\pi(\xi)|_{T_i} \in P_i$, as well as

$$\|\pi(\xi)|_{T_i} - \xi\|_{T_i,\infty} \leq \|p - \xi\|_{T_i,\infty} \qquad \text{for all } p \in P_i,\ i = 1, \ldots, N.$$

Note that $\Pi_h$ is a *Clément*-type interpolation operator (cf. [24]). In analogy to the function sets $\mathcal{D}$, $\mathcal{C}$, and $\overline{\mathcal{C}}$, we also define the discrete function sets

$$\mathcal{D}_h := \{\rho_h \in \mathcal{E}_h \mid \rho_h(\boldsymbol{x}) > 0 \text{ for all } \boldsymbol{x} \in \mathcal{X}_h\}, \tag{7.7}$$

$$\mathcal{C}_h := \{\rho_h \in \mathcal{E}_h \mid \rho_h(\boldsymbol{x}) \in \{\rho_{\min}, \rho_{\max}\} \text{ for all } \boldsymbol{x} \in \mathcal{X}_h\}, \tag{7.8}$$

$$\overline{\mathcal{C}}_h := \{\rho_h \in \mathcal{E}_h \mid \rho_{\min} \leq \rho_h(\boldsymbol{x}) \leq \rho_{\max} \text{ for all } \boldsymbol{x} \in \mathcal{X}_h\}. \tag{7.9}$$

We suppose that, for every index $j \in \mathbb{N}$, we have a function $\lambda_h : \mathcal{D}_h \times K_h \to \mathbb{R}$, such that for every coefficient $\rho \in \mathcal{D}$ and every vector $\boldsymbol{k} \in K_h$, the function value $\lambda_h(\Pi_h \rho, \boldsymbol{k})$ coincides with the $j$-th smallest eigenvalue of (7.3).

Given a function $\xi_h \in \mathcal{E}_h$, we define the integral of this function over $\Omega$ as

$$\int_{\Omega,h} \xi_h := \sum_{\boldsymbol{x} \in \mathcal{X}_h} w_{\boldsymbol{x}}\, \xi_h(\boldsymbol{x}), \tag{7.10}$$

where $w_{\boldsymbol{x}}$ denotes the quadrature weight corresponding to the quadrature point $\boldsymbol{x} \in \mathcal{X}_h$. Now, suppose that for every coefficient $\rho_h \in \mathcal{D}_h$ a subset $\Gamma_{J_h}(\rho_h)$ of $\mathcal{E}_h$ can be computed, such that

$$\partial J_h(\rho_h) \subseteq \left\{ p \in \mathcal{E}_h^* \;\middle|\; p(\eta_h) = \int_{\Omega,h} \gamma_h \eta_h,\ \gamma_h \in \Gamma_{J_h}(\rho_h) \right\}. \tag{7.11}$$

Then, our aim is to construct a minimizing sequence $\rho_h^{(0)}, \rho_h^{(1)}, \rho_h^{(2)}, \ldots$, etc. in $\mathcal{C}_h$ for the discretized goal functional $J_h$. The initial coefficient $\rho^{(0)} \in \mathcal{C}_h$ is chosen arbitrarily. The coefficients $\rho_h^{(1)}, \rho_h^{(2)}, \ldots$, are constructed iteratively according to

$$\rho_n^{(l)} := \rho_h^{(l-1)} - s_l\, \gamma_h^{(l-1)} \quad \text{for } l = 1, 2, \ldots, \tag{7.12}$$

where $s_l > 0$ is a suitable step size, and where $\gamma_h^{(l-1)}$ is an element of the set $\Gamma_{J_h}(\rho_h{}^{(l-1)})$ for every $l \in \mathbb{N}$. Clearly, this approach corresponds to a generalized gradient method for minimization problem of the form (7.1) for $X = \mathcal{E}_h$, $f = J_h$, and for the inner product $\langle \cdot\ , \cdot \rangle_{\mathcal{E}_h}$ on $\mathcal{E}_h$, which is given by

$$\langle \xi_h\ , \eta_h \rangle_{\mathcal{E}_h} := \int_{\Omega,h} \xi_h \eta_h \qquad \text{for all } \xi_h, \eta_h \in \mathcal{E}_h.$$

## 7.3 Incorporating Optimization Constraints

In the previous section we presented a general approach for an optimization algorithm than can be applied to discretized photonic band structure optimization problems (PBSOPs). In essence, this approach consists in constructing a minimizing sequence for the goal functional $J_h$ by means of a generalized gradient method. The minimizing sequence is constructed iteratively according to (7.12). In each iteration a function in $\mathcal{E}_h$ is added to the respective discretized coefficient.

So far, our approach does not account for certain optimization constraints which are inherent to PBSOPs. Recall that the discretized minimization problem (7.2) is posed on the admissible set $\mathcal{C}_h$, which consists of two-valued functions only. In order find an admissible, optimal density function, we have to impose an optimization constraint on the minimizing sequence ${\rho_h}^{(0)}, {\rho_h}^{(1)}, {\rho_h}^{(2)}, \ldots$, etc., which ensures that every density function of that sequence is an admissible density function. Clearly, such an optimization constraint is given by requiring that

$$\rho_h^{(l)}(\boldsymbol{x}) \in \{\rho_{\min}, \rho_{\max}\} \qquad \text{for all } \boldsymbol{x} \in \mathcal{X}_h,\ l = 0, 1, 2, \ldots\ . \tag{7.13}$$

Notice however, that by adding a function in $\mathcal{E}_h$ to an admissible density function in $\mathcal{C}_h$, one generally obtains a function in $\mathcal{E}_h \setminus \mathcal{C}_h$. Therefore, we cannot expect that the optimization constraint (7.13) is automatically fulfilled by a standard generalized gradient method.

In the following we shall relax the optimization constraint on the minimizing sequence by considering the minimization problem

$$\underset{\rho_h \in \overline{\mathcal{C}}_h}{\text{minimize}}\ J_h(\rho_h).$$

The extended admissible set $\overline{\mathcal{C}}_h$ imposes the optimization constraint

$$\rho_{\min} \leq \rho_h^{(l)}(\boldsymbol{x}) \leq \rho_{\max} \qquad \text{for all } \boldsymbol{x} \in \mathcal{X}_h,\ l = 0, 1, 2, \ldots \tag{7.14}$$

on every minimizing sequence ${\rho_h}^{(0)}, {\rho_h}^{(1)}, {\rho_h}^{(2)}, \ldots$, etc. that is constructed by a generalized gradient algorithm. Optimization constraints of the above form (7.14) are often referred to as *box constraints*. Again, one observes that adding a function in $\mathcal{E}_h$ to a density function in $\overline{\mathcal{C}}_h$ generally yields a function in $\mathcal{E}_h \setminus \overline{\mathcal{C}}_h$. Therefore, one cannot expect that the general construction principle (7.12) yields a sequence in $\overline{\mathcal{C}}_h$.

In the following we present a method by which the box constraint (7.14) can be ensured. To this end, we define an unconstrained minimization problem on the finite-dimensional, linear function space $\mathcal{E}_h$. Suppose that $Y_h : \mathcal{E}_h \to \overline{\mathcal{C}}_h$ is an operator, which is given by

$$Y_h(\xi_h) := y \circ \xi_h \qquad \text{for all } \xi_h \in \mathcal{E}_h, \tag{7.15}$$

where $y$ is a strictly increasing, continuously differentiable function from $\mathbb{R}$ onto the open interval $(\rho_{\min}, \rho_{\max})$. In the following, we shall refer to such an operator as a *funnel operator*, alluding to the fact that $Y_h$ maps a function in $\xi_h \in \mathcal{E}_h$ to a function $Y_h(\xi_h) \in \overline{\mathcal{C}}_h$ by "funnelling" the function values of $\xi_h$ from the "wider" interval $\mathbb{R} = (-\infty, \infty)$ into the "narrower" interval $(\rho_{\min}, \rho_{\max})$. Since the function $y$ is assumed to be strictly increasing, we have that every funnel operator is a bijection.

Given a goal functional $J_h : \mathcal{D}_h \to \mathbb{R}$ and a specific funnel operator $Y_h : \mathcal{E}_h \to \mathcal{C}_h$, we consider the minimization problem

$$\underset{\xi_h \in \mathcal{E}_h}{\text{minimize}} \; (J_h \circ Y_h)(\xi_h). \tag{7.16}$$

By constructing a minimizing sequence ${\xi_h}^{(0)}, {\xi_h}^{(1)}, {\xi_h}^{(2)}, \ldots$, etc. in the function space $\mathcal{E}_h$ for $J_h \circ Y_h$, we also obtain a minimizing sequence ${\rho_h}^{(0)}, {\rho_h}^{(1)}, {\rho_h}^{(2)}, \ldots$, etc. in the extended admissible set $\overline{\mathcal{C}}_h$ for $J_h$. This minimizing sequence is given by

$$\rho_h^{(l)} := Y_h\big(\xi_h^{(l)}\big) \qquad l = 0, 1, 2, \ldots \; . \tag{7.17}$$

Next, we consider the generalized gradients of the functional $J_h \circ Y_h$. One can easily show that the funnel operator $Y_h$ is continuously Gâteaux differentiable in $\mathcal{E}_h$. At every point $\xi_h \in \mathcal{E}_h$, the corresponding Gâteaux differential $\mathrm{D}Y_h(\xi_h)$ of $Y_h$ is given by

$$\mathrm{D}Y_h(\xi_h)[\eta_h] := (y' \circ \xi_h)\eta_h \qquad \text{for all } \eta_h \in \mathcal{E}_h,$$

where $y'$ denotes the derivative of the function $y$ in (7.15). By Theorem 6.12 on page 117, we obtain the following inclusion relation for the generalized differentials of $J_h \circ Y_h$,

$$\partial(J_h \circ Y_h)(\xi_h) \subseteq \big\{ p \circ \mathrm{D}Y_h(\xi_h) \;\big|\; p \in \partial J_h\big(Y_h(\xi_h)\big) \big\} \qquad \text{for all } \xi_h \in \mathcal{E}_h.$$

Furthermore, we obtain from (7.11) the inclusion relation

$$\partial(J_h \circ Y_h)(\xi_h) \subseteq \left\{ p \in \mathcal{E}_h^* \;\middle|\; p(\eta_h) = \int_{\Omega,h} \gamma_h (y' \circ \xi_h)\eta_h, \; \gamma_h \in \Gamma_{J_h}\big(Y_h(\xi_h)\big) \right\}. \tag{7.18}$$

For convenience, we define for every function $\xi_h \in \mathcal{E}_h$ the set

$$\Gamma_{J_h \circ Y_h}(\xi_h) := \big\{ (y' \circ \xi_h)\gamma_h \;\big|\; \gamma_h \in \Gamma_{J_h}\big(Y_h(\xi_h)\big) \big\}.$$

With this, the inclusion relation (7.18) can be equivalently written as

$$\partial(J_h \circ Y_h)(\xi_h) \subseteq \left\{ p \in \mathcal{E}_h^* \;\middle|\; p(\eta_h) = \int_{\Omega,h} \zeta_h \eta_h, \; \zeta_h \in \Gamma_{J_h \circ Y_h}(\xi_h) \right\}. \tag{7.19}$$

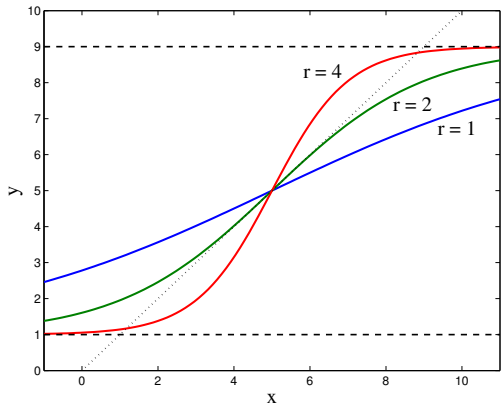



**Figure 7.2:** Graphs of the function $y$ as defined by (7.20) for $\rho_{\min} = 1$, $\rho_{\max} = 9$ and various values for $r$. The dotted line indicates the diagonal $x = y$.

Notice that for every function $\xi_h \in \mathcal{E}_h$ the functions in the set $\Gamma_{J_h \circ Y_h}(\xi_h)$ are obtained from the functions in the set $\Gamma_{J_h}\big(Y_h(\xi_h)\big)$ through a point-wise multiplication by the function $y' \circ \xi_h$. Therefore, computing the set $\Gamma_{J_h \circ Y_h}(\xi_h)$ is no more difficult than computing the set $\Gamma_{J_h}\big(Y_h(\xi_h)\big)$. It should also be noted that $y' \circ \xi_h$ is a strictly positive function, since the function $y$ is strictly increasing by assumption.

A minimizing sequence ${\xi_h}^{(0)}, {\xi_h}^{(1)}, {\xi_h}^{(2)}, \ldots$, etc. in $\mathcal{E}_h$ for the goal functional $J_h \circ Y_h$ can now be constructed using a generalized gradient method. After choosing an initial function $\xi_h^{(0)} \in \mathcal{E}_h$, the functions ${\xi_h}^{(1)}, {\xi_h}^{(2)}, \ldots$, etc. are constructed according to

$$\xi_h^{(l)} := \xi_h^{(l-1)} + s_l \, \zeta_h^{(l-1)} \qquad \text{for } l = 1, 2, \ldots,$$

where $s_l > 0$ is a suitable step size, and where $\zeta_h^{(l-1)}$ is an element of $\Gamma_{J_h \circ Y_h}({\xi_h}^{(l-1)})$, such that $-{\zeta_h}^{(l-1)}$ is a descent direction of $J_h \circ Y_h$ at the point ${\xi_h}^{(l-1)}$ for $l = 1, 2, \ldots$, etc. From this sequence, one obtains through (7.17) a minimizing sequence ${\rho_h}^{(0)}, {\rho_h}^{(1)}, {\rho_h}^{(2)}, \ldots$, etc. in the extended admissible set $\overline{\mathcal{C}}$ for the goal functional $J_h$.

We remark that the function $y : \mathbb{R} \to (\rho_{\min}, \rho_{\max})$ in (7.15) can be defined as

$$\begin{cases} y(x) := \dfrac{\rho_{\max} - \rho_{\min}}{1 - \exp(-t_r(x))} + \rho_{\min} & \text{for all } x \in \mathbb{R}, \\ t_r(x) := \dfrac{2r}{\rho_{\max} - \rho_{\min}}(x - \rho_{\min}) - r & \text{for all } x \in \mathbb{R} \end{cases} \tag{7.20}$$

for some number $r > 0$. In Figure 7.2 we depicted the function graphs of $y$ as defined by (7.20) for various values of the parameter $r$. Notice that, for every $r > 0$,

the graph of $y$ is a sigmoid curve, whose horizontal asymtotes are given by $\rho_{\min}$ and $\rho_{\max}$. All function graphs intersect at the point $((\rho_{\min}+\rho_{\max})/2, (\rho_{\min}+\rho_{\max})/2)$, which is also the point of inflection of every graph. The larger the value of $r$, the steeper the function graph is at that point, and the faster it tends to the horizontal asymptodes. Finally, we remark that the function $y$ defined in (7.20) is a so-called logistic function. It is also possible, however, to construct similar functions based on the inverse tangent function or the error function.

## 7.4 Preserving Symmetries

In each iteration of a generalized gradient method for the unconstrained minimization problem (7.16) the goal functional $J_h \circ Y_h$ is evaluated by computing an approximate photonic band structure. Typically, the band structure is computed only at a finite set of quasimomentum vectors $\boldsymbol{k}_1, \ldots, \boldsymbol{k}_M$ that belong to the first Brillouin zone $\mathbb{B}$. It was discussed in in Section 4.6, that under certain assumptions on the photonic crystal's point group the band structure is completely determined by its restriction to a so-called irreducible zone $K \subset \mathbb{B}$. The set of quasimomentum vectors $\boldsymbol{k}_1, \ldots, \boldsymbol{k}_M$ is hence usually chosen from within this irreducible zone $K$. Typically, these quasimomentum vectors lie on the critical path in $\mathbb{B}$ (see Figure 4.1(a) in Section 4.6).

It follows that a generalized gradient method has to guarantee, that the density functions of the minimizing sequence ${\rho_h}^{(0)}, {\rho_h}^{(1)}, {\rho_h}^{(2)}, \ldots$, etc. defined by (7.17) all possess the same symmetry. This imposes a so-called *symmetry constraint* on the sequence. Provided that the point group of the initial photonic crystal's medium structure is isomorphic to a group $G$ consisting of orthogonal transformations, the symmetry constraint is given by

$$\rho_h^{(l)} \circ \boldsymbol{\theta} = \rho_h^{(l)} \qquad \text{for all } \boldsymbol{\theta} \in G,\ l = 0, 1, 2, \ldots \ . \tag{7.21}$$

In Section 2.4, we introduced the symmetrization operation $f \mapsto f^{(G)}$, where $G$ is a finite group of isometries. This operation acts as an identity on the set of $G$-symmetric functions and maps each function $f$ to a $G$-symmetric function $f^{(G)}$. We constitute that the symmetrization operation is realized on the finite-dimensional, linear function space $\mathcal{E}_h$ by a so-called *symmetrization operator* $S_G : \mathcal{E}_h \to \mathcal{E}_h$ that is defined as

$$S_G(\xi_h) := \xi_h^{(G)} \qquad \text{for all } \xi_h \in \mathcal{E}_h. \tag{7.22}$$

Here as in the following, we assume that the set of quadrature points $\mathcal{X}_h$ is $G$-symmetric. From (2.11) on page 18 one easily deduces that $S_G$ is a linear operator for every group $G$ of orthogonal transformations. The action of $S_G$ on the set $\mathcal{E}_h$ can be realized as follows: Suppose that the set of quadrature points is given

by $\mathcal{X}_h = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$, where $N := |\mathcal{X}_h|$. Then, each function $\xi_h \in \mathcal{E}_h$ can be represented by a corresponding vector $\underline{\xi_h} \in \mathbb{R}^N$ by virtue of

$$\underline{\xi_h}_i = \xi_h(\boldsymbol{x}^{(i)}) \qquad \text{for all } i = 1, \ldots, N,$$

where $\underline{\xi_h}_i$ denotes the $i$-th component of the vector $\underline{\xi_h}$. Given a group $G$ of orthogonal transformations, such that $\mathcal{X}_h$ is $G$-symmetric, one can show that there exists a symmetric matrix $\boldsymbol{S} \in \mathbb{R}^{n \times n}$, such that

$$\underline{S_G(\xi_h)} = \boldsymbol{S}\underline{\xi_h} \qquad \text{for all } \xi_h \in \mathcal{E}_h.$$

As with the optimization constraint imposed by the extended admissible set $\overline{\mathcal{C}}_h$, we incorporate the symmetry constraint (7.21) into a generalized gradient method by considering yet another minimization problem, which is given by

$$\underset{\xi_h \in \mathcal{E}_h}{\text{minimize}}\ (J_h \circ Y_h \circ S_G)(\xi_h)$$

for some group $G$ of orthogonal transformations. Since $S_G$ is a linear operator, we have by Theorem 6.12 on page 117 that

$$\partial(J_h \circ Y_h \circ S_G)(\xi_h) \subseteq \left\{ p \circ S_G \;\middle|\; p \in \partial(J_h \circ Y_h)\big(S_G(\xi_h)\big) \right\} \qquad \text{for all } \xi_h \in \mathcal{E}_h.$$

From (7.18), we obtain the inclusion relation

$$\partial(J_h \circ Y_h \circ S_G)(\xi_h) \subseteq \left\{ p \in \mathcal{E}_h^* \;\middle|\; p(\eta_h) = \int_{\Omega,h} \gamma_h(y' \circ \xi_h) S_G(\eta_h), \right. \\ \left. \gamma_h \in \Gamma_{J_h}\big((Y_h \circ S_G)(\xi_h)\big) \right\}. \tag{7.23}$$

Recall that the integral over $\Omega$ in (7.18) is understood as the application of a quadrature rule on the set of quadrature point $\mathcal{X}_h$. Suppose that $\boldsymbol{W} \in \mathbb{R}^{N \times N}$ denotes the diagonal matrix with the corresponding quadrature weights on the diagonal. Given some arbitrary functions $\xi_h \in \mathcal{E}_h$ and $\gamma_h \in \Gamma_{J_h}(\xi_h)$, we find that

$$\begin{aligned} \int_{\Omega,h} \gamma_h(y' \circ \xi_h) S_G(\eta_h) &= \big(\underline{\gamma_h} \odot \underline{(y' \circ \xi_h)}\big) \cdot \boldsymbol{W}\boldsymbol{S}\underline{\eta_h} \\ &= \Big(\boldsymbol{S}\big(\underline{\gamma_h} \odot \underline{(y' \circ \xi_h)}\big)\Big) \cdot \boldsymbol{W}\underline{\eta_h} \\ &= \int_{\Omega,h} S_G\big(\gamma_h(y' \circ \xi_h)\big)\eta_h \qquad \text{for all } \eta_h \in \mathcal{E}_h. \end{aligned}$$

Here, $\odot$ denotes the *component-wise vector product*, which also known as the *Hadamard product* on $\mathbb{R}^N$. The above identity affords an alternative formulation

of the inclusion relation (7.23). In order to establish this formulation, we define for every function $\xi_h \in \mathcal{E}_h$ the set

$$\Gamma_{J_h \circ Y_h \circ S_G}(\xi_h) := \big\{ S_G\big((y' \circ \xi_h)\gamma_h\big) \;\big|\; \gamma_h \in \Gamma_{J_h}\big((Y_h \circ S_G)(\xi_h)\big) \big\}.$$

With this, the inclusion relation (7.23) can be equivalently written as

$$\partial(J_h \circ Y_h \circ S_G)(\xi_h) \subseteq \big\{ p \in \mathcal{E}_h^* \;\big|\; p(\eta_h) = \int_{\Omega,h} \upsilon_h \eta_h,\ \upsilon_h \in \Gamma_{J_h \circ Y_h \circ S_G}(\xi_h) \big\}.$$

From this we deduce that the set $\Gamma_{J_h \circ Y_h \circ S_G}(\xi_h)$ is a superset of the generalized gradient of $J_h \circ Y_h \circ S_G$ at $\xi_h$ for every function $\xi_h \in \mathcal{E}_h$.

We can now proceed as in the previous section and define a generalized gradient method, which constructs a minimizing sequence ${\xi_h}^{(0)}$, ${\xi_h}^{(1)}$, ${\xi_h}^{(2)}, \ldots$, etc. in $\mathcal{E}_h$ for the goal functional $J_h \circ Y_h \circ S_G$. In the $l$-th iteration of this method, the descent direction $-\upsilon_h^{(l-1)}$ is chosen such that $\upsilon_h^{(l-1)} \in \Gamma_{J_h \circ Y_h \circ S_G}({\xi_h}^{(l-1)})$ for $l = 1, 2, \ldots$, etc. By construction, every function in the set $\Gamma_{J_h \circ Y_h \circ S_G}({\xi_h}^{(l-1)})$ is $G$-symmetric. Provided that the initial function ${\xi_h}^{(0)} \in \mathcal{E}_h$ is also $G$-symmetric, we thus obtain a minimizing sequence, which consists of $G$-symmetric functions in $\mathcal{E}_h$. In particular, we have that

$$S_G\big({\xi_h}^{(l)}\big) = {\xi_h}^{(l)} \qquad \text{for } l = 0, 1, 2, \ldots \ .$$

From the sequence ${\xi_h}^{(0)}$, ${\xi_h}^{(1)}$, ${\xi_h}^{(2)}, \ldots$, we obtain by (7.17) a minimizing sequence ${\rho_h}^{(0)}$, ${\rho_h}^{(1)}$, ${\rho_h}^{(2)}, \ldots$, etc. in $\overline{\mathcal{C}}_h$ for the goal functional $J_h$, such that every density function of that sequence is $G$-symmetric.

## 7.5 The Algorithm

Based on the approaches that were introduced in the previous Section 7.3 and Section 7.4, we implemented an optimization algorithm for band structure optimization problems. In this section we describe this algorithm and comment on some implementation details. The algorithm is listed in pseudo-code as Algorithm 7.1. It was implemented in MATLAB as well as in C++ using the parallel finite element library M++ (see [76], [77]). In the following we shall refer to the functions $\xi_h^{(l)}$ in Algorithm 7.1 as *free density functions*.

The optimization algorithm constitutes a generalized gradient method for the locally Lipschitz minimization problem

$$\underset{\xi_h \in \mathcal{E}_h}{\text{minimize}}\ (J_h \circ Y_h \circ S_G)(\xi_h),$$

as introduced in Section 7.1. A minimizing sequence ${\xi_h}^{(0)}$, ${\xi_h}^{(1)}$, ${\xi_h}^{(2)}, \ldots$, etc. of free density functions in $\mathcal{E}_h$ is iteratively constructed from a given initial free

**Algorithm 7.1** The Generalized Gradient Algorithm

**Input:**

| | |
|---|---|
| $J_h : \mathcal{D}_h \to \mathbb{R}$ | *Goal functional* |
| $G$ | *Group of orthogonal transformations (Point group)* |
| ${\xi_h}^{(0)}$ | *Initial free density* |

**Parameters:**

| | |
|---|---|
| $s_{\min} > 0$ | *Minimal step size* |
| $s_{\max} > s_{\min}$ | *Maximal step size* |
| $\Delta_{\min} > 0$ | *Minimal decrease* |

**Algorithm:**

```
ρ_h^(0) := Y_h(ξ_h^(0))
compute z_0 := J_h(ρ_h^(0))
s_0 := s_max
for l = 1, 2, ... do
  compute Γ_{J_h∘Y_h∘S_G}(ξ_h^(l−1))
  choose a descent direction −ṽ_h^(l−1) ∈ −Γ_{J_h∘Y_h∘S_G}(ξ_h^(l−1))
  υ_h^(l−1) := ṽ_h^(l−1) / ‖ṽ^(l−1)‖_∞
  s_{l,1} := min{2s_{l−1}, s_max}
  for p = 1, 2, ... do
    ξ_h^(l,p) := ξ_h^(l−1) − s_{l,p} υ_h^(l−1)
    ρ_h^(l,p) := Y_h(ξ_h^(l,p))
    compute z_{l,p} := J_h(ρ_h^(l,p))
    if z_{l−1} − z_{l,p} > Δ_min then
      ξ_h^(l) := ξ_h^(l,p)
      z_l := z_{l,p}
      t_l := s_{l,p}
      break
    end if
    s_{l,p+1} := s_{l,p}/2
    if s_{l,p+1} < s_min then
      return ρ_h^(l−1)
    end if
  end for
end for
```

function ${\xi_h}^{(0)} \in \mathcal{E}_h$ according to

$$ {\xi_h}^{(l)} := {\xi_h}^{(l-1)} - s_l \, {\upsilon_h}^{(l-1)} \qquad \text{for } l = 1, 2, \dots \, . \tag{7.24} $$

In every iteration, the function ${\upsilon_h}^{(l-1)}$ is chosen from the set $\Gamma_{J_h \circ Y_h \circ S_G}({\xi_h}^{(l-1)})$, such

that $-v_h{}^{(l-1)}$ is a descent direction of $J_h \circ Y_h \circ S_G$ at $\xi_h{}^{(l-1)}$, where $l = 1, 2, \ldots$. The choice of the descent direction is accomplished by an appropriate choice strategy. In general, one has to devise a separate choice strategy for every goal functional $J_h$. In the following section we briefly discuss this issue and present a choice strategy, which is appropriate for the gap width functionals defined in Section 5.5.

The step sizes $s_l > 0$ are chosen adaptively using a trial-and-error scheme. Starting with a given maximimal step size $s_{\max}$ the step size is divided by 2, each time an update of the current free density function according to (7.24) fails to yield a prescribed minimal decrease $\Delta_{\min}$ in the goal value. In order to prevent excessively small step sizes, the algorithm always doubles the step size before attempting an update by a new descent direction. The algorithms terminates, when the step size becomes smaller than a prescribed minimal step size $s_{\min} > 0$. In this case, we shall say that the algorithm has *converged*. The *final free density function* is given by the last free density function $\xi_h{}^{(L)} \in \mathcal{E}_h$, for which a descrease of the goal value could be obtained. The final free density defines a corresponding *final density function* $\rho_h{}^{(L)} \in \overline{\mathcal{C}}_h$, which is given by

$$\rho_h{}^{(L)} := Y_h\big(\xi_h{}^{(L)}\big).$$

This final density function $\rho_h{}^{(L)}$ is returned by the algorithm.

By construction the final density function $\rho_h{}^{(L)}$ is $G$-symmetric. Due to the definition of the funnel operator $Y_h$, however, $\rho_h{}^{(L)}$ is always a function in $\overline{\mathcal{C}}_h \setminus \mathcal{C}_h$. Hence, the optimization algorithm is not capable to return admissible density functions. In numerical experiments we found, however, that the final density function was almost two-valued, attaining function values close to $\rho_{\min}$ and $\rho_{\max}$ at most quadrature points. We remark that an admissible function can always be obtained from a final density function $\rho_h{}^{(L)}$, by applying the so-called *threshold operator* $H_h : \overline{\mathcal{C}}_h \to \mathcal{C}_h$ to it. The threshold operator is defined by

$$H_h(\rho_h)(\boldsymbol{x}) := \begin{cases} \rho_{\min} & \text{if } \rho_h(\boldsymbol{x}) < \dfrac{\rho_{\min} + \rho_{\max}}{2}, \\ \rho_{\max} & \text{else} \end{cases} \qquad \text{for all } \boldsymbol{x} \in \mathcal{X}_h. \tag{7.25}$$

and for all $\rho_h \in \overline{\mathcal{C}}_h$. We used Algorithm 7.1 to solve several photonic band gap maximization problems. Some of the results are presented in Chapter 9.

## 7.6 Choosing Descent Directions

A crucial step each iteration of the generalized gradient method, which is given by Algorithm 7.1, is the choice of the descent directions $-v_h{}^{(l-1)}$. As we mentioned in the previous section, this choice is usually accomplished through a specific choice

strategy. In this section we briefly comment on possible choice strategies and develop a specific choice strategy for the band gap functionals defined in Section 5.5.

In certain situations, the choice of a descent direction is trivial. Suppose, for example, that the set $\Gamma_{J_h}(\xi_h)$ is a singleton. Then, the set $\Gamma_{J_h \circ Y_h \circ S_G}({\xi_h}^{(l-1)})$ is also a singleton and the functional $J_h \circ Y_h \circ S_G$ is hence strictly differentiable at ${\xi_h}^{(l-1)}$. In this case the set $-\Gamma_{J_h \circ Y_h \circ S_G}({\xi_h}^{(l-1)})$ contains the negative strict gradient of $J_h \circ Y_h \circ S_G$ at ${\xi_h}^{(l-1)}$ as the only element, which constitutes a descent direction. Furthermore, when $-\Gamma_{J_h \circ Y_h \circ S_G}({\xi_h}^{(l-1)})$ is known to be equal to the negative of the generalized gradient of $J_h \circ Y_h \circ S_G$ at $\xi_h^{(l-1)}$, one can choose the element with minimal norm as a descent direction according to Lemma 7.1. In the general case, however, the set $\Gamma_{J_h \circ Y_h \circ S_G}({\xi_h}^{(l-1)})$ is strictly larger than the generalized gradient.

We now turn to the specific case, where the goal functional $J_h$ is given by a *discretized gap width functional* $J_{j,h} : \mathcal{D}_h \to \mathbb{R}$ for some fixed index $j \in \mathbb{N}$. We assume that the $J_{j,h}$ is given by

$$J_{j,h}(\rho_h) := \max_{\boldsymbol{k} \in K_h} \omega_{j,h}(\rho_{j,h}, \boldsymbol{k}) - \min_{\boldsymbol{k} \in K_h} \omega_{j+1,h}(\rho_{j,h}, \boldsymbol{k}) \qquad \text{for all } \rho_h \in \mathcal{D}_h,$$

where $K_h$ denotes a finite set of quasimomentum vectors in the first Brillouin zone $\mathbb{B}$, and where $\omega_{j,h}(\rho_h, \boldsymbol{k})$ denotes the square-root of the $j$-th smallest eigenvalue $\lambda_{j,h}(\rho_h, \boldsymbol{k})$ of the generalized matrix eigenvalue problem (7.3) on page 131 for all $\rho_h \in \mathcal{D}_h$ and all $\boldsymbol{k} \in K_h$. We also define the *discretized band edge functionals* $J_{\mathrm{lo},j,h}, J_{\mathrm{up},j,h} : \mathcal{D}_h \to \mathbb{R}$ by

$$\begin{aligned} J_{\mathrm{lo},j,h}(\rho_h) &:= \max_{\boldsymbol{k} \in K_h} \omega_{j,h}(\rho_h, \boldsymbol{k}) && \text{for all } \rho_h \in \mathcal{D}_h, \\ J_{\mathrm{up},j,h}(\rho_h) &:= \min_{\boldsymbol{k} \in K_h} \omega_{j+1,h}(\rho_h, \boldsymbol{k}), && \text{for all } \rho_h \in \mathcal{D}_h. \end{aligned}$$

In the following we assume that every eigenvalue $\lambda_{j,h}(\rho_h, \boldsymbol{k})$ is simple. We remark that this assumption is usually fulfilled due to numerical errors. Then, by a similar result to that of Theorem 6.16 on page 121, we then have that

$$\partial \lambda_{j,h}(\,\cdot\,, \boldsymbol{k})(\rho_h) = \left\{ p \in \mathcal{E}_h^* \;\middle|\; p(\eta_h) := \int_{\Omega,h} \eta \big| \mathcal{I}_h\big[(\nabla + \mathrm{i}\boldsymbol{k})_h \times \boldsymbol{u}_{j,h}(\rho_h, \boldsymbol{k})\big]\big|^2 \right\},$$

where $\boldsymbol{u}_{j,h}(\rho_h, \boldsymbol{k})$ denotes an eigenvalue corresponding to $\lambda_{j,h}(\rho_h, \boldsymbol{k})$, and where $(\nabla + \mathrm{i}\boldsymbol{k})_h \times$ denotes a modified curl operator, which is defined appropriately on the finite element space $\boldsymbol{W}_h$. Recall that $\mathcal{I}_h$ denotes the interpolation operator with image space $\mathcal{E}_h$ (see Section 7.2).

Since the generalized differential $\partial \lambda_{j,h}(\,\cdot\,, \boldsymbol{k})(\rho_h)$ is a singleton, we have by Proposition 6.7(d) on page 114 that the functional $\lambda_{j,h}(\,\cdot\,, \boldsymbol{k})$ is strictly differentiable in a neighbourhood of $\rho_h$ for every vector $\boldsymbol{k} \in K_h$.

We further assume that

$$\arg\max_{K_h} \lambda_{j,h}(\rho_{j,h}, \cdot) = \big\{\boldsymbol{k}_{\mathrm{lo}}^{(1)}, \dots, \boldsymbol{k}_{\mathrm{lo}}^{(N_{\mathrm{lo}})}\big\}, \tag{7.26}$$

$$\arg\min_{K_h} \lambda_{j+1,h}(\rho_{j,h}, \cdot) = \big\{\boldsymbol{k}_{\mathrm{up}}^{(1)}, \dots, \boldsymbol{k}_{\mathrm{up}}^{(N_{\mathrm{up}})}\big\} \tag{7.27}$$

for small numbers $N_{\mathrm{lo}}, N_{\mathrm{up}} \in \mathbb{N}$. For convenience, we define the functionals $\widetilde{\omega}_{\mathrm{lo}}^{(i)}, \widetilde{\omega}_{\mathrm{up}}^{(i)} : \mathcal{E}_h \to \mathbb{R}$ by

$$\begin{aligned}
\widetilde{\omega}_{\mathrm{lo}}^{(i)} &:= \omega_{j,h}\big(\,\cdot\,, \boldsymbol{k}_{\mathrm{lo}}^{(i)}\big) \circ Y_h \circ S_G \qquad && i = 1, \dots, N_{\mathrm{lo}}, \\
\widetilde{\omega}_{\mathrm{up}}^{(i)} &:= \omega_{j+1,h}\big(\,\cdot\,, \boldsymbol{k}_{\mathrm{up}}^{(i)}\big) \circ Y_h \circ S_G \qquad && i = 1, \dots, N_{\mathrm{up}},
\end{aligned}$$

where $Y_h$ and $S_G$ are the funnel operator (see (7.15) in Section 7.3) and the symmetrization operator (see (7.22) in Section 7.4) in Algorithm 7.1. Given a fixed function $\xi_h \in \mathcal{E}_h$, we define the functions $\upsilon_{\mathrm{lo}}^{(i)}, \upsilon_{\mathrm{up}}^{(i)} \in \mathcal{E}_h$ by

$$\begin{aligned}
\upsilon_{\mathrm{lo}}^{(i)} &:= \frac{S_G\Big((y' \circ \xi_h)\,\big|\mathcal{I}_h\big[(\nabla + \mathrm{i}\boldsymbol{k}_{\mathrm{lo}}^{(i)})_h \times \boldsymbol{u}_{j,h}(\rho_h, \boldsymbol{k}_{\mathrm{lo}}^{(i)})\big]\big|^2\Big)}{2\omega_{j,h}(\rho_h, \kappa_{\mathrm{lo}}^{(i)})} \qquad && i = 1, \dots, N_{\mathrm{lo}}, \\
\upsilon_{\mathrm{up}}^{(i)} &:= \frac{S_G\Big((y' \circ \xi_h)\big|\mathcal{I}_h\big[(\nabla + \mathrm{i}\boldsymbol{k}_{\mathrm{up}}^{(i)})_h \times \boldsymbol{u}_{j+1,h}(\rho_h, \boldsymbol{k})\big]\big|^2\Big)}{2\omega_{j+1,h}(\rho_h, \kappa_{\mathrm{up}}^{(i)})} \qquad && i = 1, \dots, N_{\mathrm{up}}.
\end{aligned}$$

Then, we have that generalized differentials of the functionals $\widetilde{\omega}_{\mathrm{lo}}^{(i)}$ and $\widetilde{\omega}_{\mathrm{up}}^{(i)}$ at $\xi_h$ are given by

$$\begin{aligned}
\partial\widetilde{\omega}_{\mathrm{lo}}^{(i)}(\xi_h) &= \left\{ p \in \mathcal{E}_h^* \,\middle|\, p(\eta_h) := \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(i)} \eta_h \right\} \qquad && i = 1, \dots, N_{\mathrm{lo}}, \\
\partial\widetilde{\omega}_{\mathrm{up}}^{(i)}(\xi_h) &= \left\{ p \in \mathcal{E}_h^* \,\middle|\, p(\eta_h) := \int_{\Omega,h} \upsilon_{\mathrm{up}}^{(i)} \eta_h \right\} \qquad && i = 1, \dots, N_{\mathrm{up}}.
\end{aligned}$$

Since the generalized differentials $\partial\widetilde{\omega}_{\mathrm{lo}}^{(i)}(\xi_h)$ are singletons, we have that the functional functionals $\widetilde{\omega}_{\mathrm{lo}}^{(i)}$ are strictly differentiable at $\xi_h$ for all $i = 1, \dots, N_{\mathrm{lo}}$. By the same argument, we have that the functionals $\widetilde{\omega}_{\mathrm{up}}^{(i)}$ are also strictly differentiable at $\xi_h$ for all $i = 1, \dots, N_{\mathrm{up}}$. In particular, we have that

$$\widetilde{\omega}_{\mathrm{lo}}^{(i)}(\xi_h + \eta_h) = \widetilde{\omega}_{\mathrm{lo}}^{(i)}(\xi_h) + \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(i)} \eta_h + o(\|\eta\|_{\mathcal{E}_h}) \quad \text{as } \eta_h \to 0 \qquad i = 1, \dots, N_{\mathrm{lo}}, \tag{7.28}$$

$$\widetilde{\omega}_{\mathrm{up}}^{(i)}(\xi_h + \eta_h) = \widetilde{\omega}_{\mathrm{up}}^{(i)}(\xi_h) + \int_{\Omega,h} \upsilon_{\mathrm{up}}^{(i)} \eta_h + o(\|\eta\|_{\mathcal{E}_h}) \quad \text{as } \eta_h \to 0 \qquad i = 1, \dots, N_{\mathrm{up}}. \tag{7.29}$$

For convience we define the functionals $\widetilde{J}, \widetilde{J}_{\mathrm{lo}}, \widetilde{J}_{\mathrm{up}} : \mathcal{E}_h \to \mathbb{R}$ by

$$\begin{aligned}\widetilde{J} &:= J_{j,h} \circ Y_h \circ S_G,\\ \widetilde{J}_{\mathrm{lo}} &:= J_{\mathrm{lo},j,h} \circ Y_h \circ S_G,\\ \widetilde{J}_{\mathrm{up}} &:= J_{\mathrm{up},j,h} \circ Y_h \circ S_G.\end{aligned}$$

Clearly, we then have that

$$\widetilde{J} = \widetilde{J}_{\mathrm{lo}} - \widetilde{J}_{\mathrm{up}}.$$

We also define the sets $\widetilde{\Gamma}_{\mathrm{lo}}, \widetilde{\Gamma}_{\mathrm{up}}, \widetilde{\Gamma} \in \mathcal{E}_h$ as

$$\begin{aligned}\widetilde{\Gamma}_{\mathrm{lo}} &:= \overline{\mathrm{conv}}\big\{\upsilon_{\mathrm{up}}^{(1)}, \ldots, \upsilon_{\mathrm{up}}^{(N_{\mathrm{lo}})}\big\},\\ \widetilde{\Gamma}_{\mathrm{up}} &:= \overline{\mathrm{conv}}\big\{\upsilon_{\mathrm{up}}^{(1)}, \ldots, \upsilon_{\mathrm{up}}^{(N_{\mathrm{up}})}\big\},\\ \widetilde{\Gamma} &:= \widetilde{\Gamma}_{\mathrm{lo}} - \widetilde{\Gamma}_{\mathrm{up}}.\end{aligned}$$

According to (7.26)–(7.27) and Proposition 6.15 on page 119, we have that

$$\begin{aligned}\partial\widetilde{J}_{\mathrm{lo}}(\xi_h) &\subseteq \left\{ p \in \mathcal{E}_h^* \,\middle|\, p(\eta_h) = \int_{\Omega,h} \upsilon_h \eta_h,\ \upsilon_h \in \widetilde{\Gamma}_{\mathrm{lo}} \right\}\\ \partial\widetilde{J}_{\mathrm{up}}(\xi_h) &\subseteq \left\{ p \in \mathcal{E}_h^* \,\middle|\, p(\eta_h) = \int_{\Omega,h} \upsilon_h \eta_h,\ \upsilon_h \in \widetilde{\Gamma}_{\mathrm{up}} \right\}.\end{aligned}$$

Hence, we obtain by Proposition 6.9 and Proposition 6.10 on page 116 the following inclusion relation for the generalized differential of $\widetilde{J}$ at $\xi_h$,

$$\partial\widetilde{J}(\xi_h) \subseteq \left\{ p \in \mathcal{E}_h^* \,\middle|\, p(\eta_h) = \int_{\Omega,h} \upsilon_h \eta_h,\ \upsilon_h \in \widetilde{\Gamma} \right\}.$$

From this, we deduce that the set $\widetilde{\Gamma}$ is a superset of the generalized gradient of $\widetilde{J}$ at $\xi_h$. We remark that the set $\widetilde{\Gamma}$ reduces to a singleton, if and only if $N_{\mathrm{lo}} = N_{\mathrm{up}} = 1$.

Our aim now is to indentify a descent direction $-\upsilon_h \in -\widetilde{\Gamma}$ for the goal functional $\widetilde{J}$ at $\xi_h$. To this end we derive a linear program, whose solution determines a good candidate for such a descent direction. Note that by definition of the functionals $\widetilde{\omega}_{\mathrm{lo}}^{(i)}$ and $\widetilde{\omega}_{\mathrm{up}}^{(i)}$ we have that

$$\begin{aligned}\widetilde{J}_{\mathrm{lo}}(\xi_h) &= \widetilde{\omega}_{\mathrm{lo}}^{(1)}(\xi_h) = \cdots = \widetilde{\omega}_{\mathrm{lo}}^{(N_{\mathrm{lo}})}(\xi_h),\\ \widetilde{J}_{\mathrm{up}}(\xi_h) &= \widetilde{\omega}_{\mathrm{up}}^{(1)}(\xi_h) = \cdots = \widetilde{\omega}_{\mathrm{up}}^{(N_{\mathrm{up}})}(\xi_h).\end{aligned}$$

Under the assumption that the function identities

$$\widetilde{J}_{\mathrm{lo}} = \max\big\{\widetilde{\omega}_{\mathrm{lo}}^{(1)}, \cdots, \widetilde{\omega}_{\mathrm{lo}}^{(N_{\mathrm{lo}})}\big\},$$

$$\widetilde{J}_{\mathrm{up}} = \min\{\widetilde{\omega}_{\mathrm{up}}^{(1)}, \cdots, \widetilde{\omega}_{\mathrm{up}}^{(N_{\mathrm{up}})}\}$$

hold in some neighbourhood of $\xi_h$, we obtain by (7.28) and (7.29) the following asymptotic behaviour of $\widetilde{J}$ at $\xi_h$,

$$\begin{aligned}
&\widetilde{J}(\xi_h + \eta_h) - \widetilde{J}(\xi_h) \\
&= \widetilde{J}_{\mathrm{lo}}(\xi_h + \eta_h) - \widetilde{J}_{\mathrm{lo}}\big(P(\xi)\big) - \widetilde{J}_{\mathrm{up}}(\xi_h + \eta_h) + \widetilde{J}_{\mathrm{up}}(\xi_h) \\
&= \max\{\widetilde{\omega}_{\mathrm{lo}}^{(1)}(\xi_h + \eta_h) - \widetilde{\omega}_{\mathrm{lo}}^{(1)}(\xi_h), \ldots, \widetilde{\omega}_{\mathrm{lo}}^{(N_{\mathrm{lo}})}(\xi_h + \eta_h) - \widetilde{\omega}_{\mathrm{lo}}^{(N_{\mathrm{lo}})}(\xi_h)\} \\
&\qquad - \min\{\widetilde{\omega}_{\mathrm{up}}^{(1)}(\xi_h + \eta_h) - \widetilde{\omega}_{\mathrm{up}}^{(1)}(\xi_h), \ldots, \widetilde{\omega}_{\mathrm{up}}^{(N_{\mathrm{up}})}(\xi_h + \eta_h) - \widetilde{\omega}_{\mathrm{up}}^{(N_{\mathrm{up}})}(\xi_h)\} \\
&= \max\left\{\int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(1)}\eta_h, \ldots, \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(N_{\mathrm{lo}})}\eta_h\right\} - \min\left\{\int_{\Omega,h} \upsilon_{\mathrm{up}}^{(1)}\eta_h, \ldots, \int_{\Omega,h} \upsilon_{\mathrm{up}}^{(N_{\mathrm{up}})}\eta_h\right\} \\
&\qquad + o(\|\eta_h\|_{\mathcal{E}_h}) \quad \text{as } \eta_h \to 0.
\end{aligned}$$

This asymptotic behaviour implies, that a direction $\eta_h \in \mathcal{E}_h$, which minimizes

$$\max\left\{\int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(1)}\eta_h, \ldots, \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(N_{\mathrm{lo}})}\eta_h\right\} - \min\left\{\int_{\Omega,h} \upsilon_{\mathrm{up}}^{(1)}\eta_h, \ldots, \int_{\Omega,h} \upsilon_{\mathrm{up}}^{(N_{\mathrm{up}})}\eta_h\right\}$$

is a good candidate for a descent direction of $\widetilde{J}$ at $\xi_h$. Now, suppose that the direction $\eta_h$ is chosen from the set $-\widetilde{\Gamma}$. Then, by definition of the sets $\widetilde{\Gamma}$, $\widetilde{\Gamma}_{\mathrm{lo}}$ and $\widetilde{\Gamma}_{\mathrm{up}}$, there exists non-negative numbers $\theta_{\mathrm{lo}}^{(1)}, \ldots, \theta_{\mathrm{lo}}^{(N_{\mathrm{lo}})}, \theta_{\mathrm{up}}^{(1)}, \ldots, \theta_{\mathrm{up}}^{(N_{\mathrm{up}})} \in \mathbb{R}$, such that

$$0 \le \theta_{\mathrm{lo}}^{(i)} \le 1 \qquad i = 1, \ldots, N_{\mathrm{lo}}, \tag{7.30}$$

$$0 \le \theta_{\mathrm{up}}^{(i)} \le 1 \qquad i = 1, \ldots, N_{\mathrm{up}}, \tag{7.31}$$

$$\sum_{i=1}^{N_{\mathrm{lo}}} \theta_{\mathrm{lo}}^{(i)} = 1, \tag{7.32}$$

$$\sum_{i=1}^{N_{\mathrm{up}}} \theta_{\mathrm{up}}^{(i)} = 1 \tag{7.33}$$

and such that

$$\eta_h = -\sum_{\ell=1}^{N_{\mathrm{lo}}} \theta_{\mathrm{lo}}^{(\ell)} \upsilon_{\mathrm{lo}}^{(\ell)} + \sum_{\ell=1}^{N_{\mathrm{up}}} \theta_{\mathrm{up}}^{(\ell)} \upsilon_{\mathrm{up}}^{(\ell)}. \tag{7.34}$$

It follows that

$$\int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(i)} \eta_h = -\sum_{i=1}^{N_{\mathrm{lo}}} \theta_{\mathrm{lo}}^{(\ell)} \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(i)} \upsilon_{\mathrm{lo}}^{(\ell)} + \sum_{i=1}^{N_{\mathrm{up}}} \theta_{\mathrm{up}}^{(\ell)} \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(i)} \upsilon_{\mathrm{up}}^{(\ell)}, \qquad i = 1, \ldots, N_{\mathrm{lo}}, \tag{7.35}$$

$$\int_{\Omega,h} \upsilon_{\mathrm{up}}^{(i)} \eta_h = -\sum_{i=1}^{N_{\mathrm{lo}}} \theta_{\mathrm{lo}}^{(\ell)} \int_{\Omega,h} \upsilon_{\mathrm{up}}^{(i)} \upsilon_{\mathrm{lo}}^{(\ell)} + \sum_{i=1}^{N_{\mathrm{up}}} \theta_{\mathrm{up}}^{(\ell)} \int_{\Omega,h} \upsilon_{\mathrm{up}}^{(i)} \upsilon_{\mathrm{up}}^{(\ell)}, \qquad i = 1, \ldots, N_{\mathrm{up}}. \tag{7.36}$$

We can now formulate a minimization problem, which is given by a linear program. To this end we define the functions $f_{\mathrm{lo}}, f_{\mathrm{up}} : -\widetilde{\Gamma} \to \mathbb{R}$ by

$$\begin{aligned} f_{\mathrm{lo}}(\eta_h) &:= \max\left\{ \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(1)} \eta_h, \ldots, \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(N_{\mathrm{lo}})} \eta_h \right\} \qquad \text{for all } \eta_h \in -\widetilde{\Gamma}, \\ f_{\mathrm{up}}(\eta_h) &:= \min\left\{ \int_{\Omega,h} \upsilon_{\mathrm{up}}^{(1)} \eta_h, \ldots, \int_{\Omega,h} \upsilon_{\mathrm{up}}^{(N_{\mathrm{up}})} \eta_h \right\} \qquad \text{for all } \eta_h \in -\widetilde{\Gamma} \end{aligned}$$

and consider the minimization problem

$$\underset{\eta_h \in -\widetilde{\Gamma}}{\text{minimize}} \, \big( f_{\mathrm{lo}}(\eta_h) - f_{\mathrm{up}}(\eta_h) \big). \tag{7.37}$$

By introducing the formal variables $\zeta_{\mathrm{lo}}$ and $\zeta_{\mathrm{up}}$ for $f_{\mathrm{lo}}(\eta_h)$ and for $f_{\mathrm{up}}(\eta_h)$, we obtain from (7.35)–(7.36) the contraints

$$\begin{aligned} -\sum_{\ell=1}^{N_{\mathrm{lo}}} \theta_{\mathrm{lo}}^{(\ell)} \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(i)} \upsilon_{\mathrm{lo}}^{(\ell)} + \sum_{\ell=1}^{N_{\mathrm{up}}} \theta_{\mathrm{up}}^{(\ell)} \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(i)} \upsilon_{\mathrm{up}}^{(\ell)} - \zeta_{\mathrm{lo}} \le 0 \qquad \ell = 1, \ldots, N_{\mathrm{lo}}, \\ \sum_{\ell=1}^{N_{\mathrm{lo}}} \theta_{\mathrm{lo}}^{(\ell)} \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(i)} \upsilon_{\mathrm{lo}}^{(\ell)} - \sum_{\ell=1}^{N_{\mathrm{up}}} \theta_{\mathrm{up}}^{(\ell)} \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(i)} \upsilon_{\mathrm{up}}^{(\ell)} + \zeta_{\mathrm{up}} \le 0 \qquad \ell = 1, \ldots, N_{\mathrm{up}}. \end{aligned}$$

Recall that the coefficients $\theta_{\mathrm{lo}}^{(i)}$ and $\theta_{\mathrm{up}}^{(i)}$ have to obey the constraints (7.30)–(7.33). By defining the unknown vector $\boldsymbol{x} \in \mathbb{R}^{N_{\mathrm{lo}}+N_{\mathrm{up}}+2}$ as

$$\boldsymbol{x} := \Big( \theta_{\mathrm{lo}}^{(1)}, \ldots, \theta_{\mathrm{lo}}^{(N_{\mathrm{lo}})}, \theta_{\mathrm{up}}^{(1)}, \ldots, \theta_{\mathrm{up}}^{(N_{\mathrm{up}})}, \zeta_{\mathrm{lo}}, \zeta_{\mathrm{up}} \Big)^{\mathrm{T}}, \tag{7.38}$$

the minimization problem (7.37) can be formulated as a linear program according to

$$\begin{aligned} &\underset{\boldsymbol{x}}{\text{minimize}} \; \boldsymbol{f} \cdot \boldsymbol{x} \\ &\text{subject to } \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{0}, \quad \boldsymbol{C}\boldsymbol{x} = \boldsymbol{d}, \quad \boldsymbol{l} \le \boldsymbol{x} \le \boldsymbol{u}. \end{aligned}$$

The vector $\boldsymbol{f} \in \mathbb{R}^{N^{\mathrm{lo}}+N^{\mathrm{up}}+2}$ and the vectors $\boldsymbol{l}, \boldsymbol{u} \in \overline{\mathbb{R}}^{N^{\mathrm{lo}}+N^{\mathrm{up}}+2}$ are given by

$$\boldsymbol{f} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ -1 \end{pmatrix}, \quad \boldsymbol{l} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ -\infty \\ -\infty \end{pmatrix}, \quad \boldsymbol{u} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \\ \infty \\ \infty \end{pmatrix}.$$

The matrix $\boldsymbol{A} \in \mathbb{R}^{(N^{\mathrm{lo}}+N^{\mathrm{up}})\times(N^{\mathrm{lo}}+N^{\mathrm{up}}+2)}$ is given as a block matrix by

$$\boldsymbol{A} := \begin{pmatrix} -\boldsymbol{A}_{\mathrm{lo,lo}} & \boldsymbol{A}_{\mathrm{lo,up}} & -\mathbf{1}_{N_{\mathrm{lo}},1} & \mathbf{0}_{N_{\mathrm{lo}},1} \\ \boldsymbol{A}_{\mathrm{up,lo}} & -\boldsymbol{A}_{\mathrm{up,up}} & \mathbf{0}_{N_{\mathrm{up}},1} & \mathbf{1}_{N_{\mathrm{up}},1} \end{pmatrix}.$$

Here, we denote by $\mathbf{0}_{m,n}$ and $\mathbf{1}_{m,n}$ for given numbers $m, n \in \mathbb{N}$ those $m \times n$ matrices with components equal to zero and one, respectively. The matrices $\boldsymbol{A}_{\mathrm{lo,lo}} \in \mathbb{R}_{N_{\mathrm{up}}\times N_{\mathrm{up}}}$, $\boldsymbol{A}_{\mathrm{lo,up}} \in \mathbb{R}^{N_{\mathrm{up}}\times N_{\mathrm{up}}}$, $\boldsymbol{A}_{\mathrm{up,lo}} \in \mathbb{R}^{N_{\mathrm{up}}\times N_{\mathrm{up}}}$ and $\boldsymbol{A}_{\mathrm{up,up}} \in \mathbb{R}^{N_{\mathrm{up}}\times N_{\mathrm{up}}}$ are given by

$$\begin{aligned}
[\boldsymbol{A}_{\mathrm{lo,lo}}]_{i,\ell} &= \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(i)} \upsilon_{\mathrm{lo}}^{(\ell)} && i = 1, \ldots, N_{\mathrm{lo}},\ \ell = 1, \ldots, N_{\mathrm{lo}}, \\
[\boldsymbol{A}_{\mathrm{lo,up}}]_{i,\ell} &= \int_{\Omega,h} \upsilon_{\mathrm{lo}}^{(i)} \upsilon_{\mathrm{up}}^{(\ell)} && i = 1, \ldots, N_{\mathrm{lo}},\ \ell = 1, \ldots, N_{\mathrm{up}}, \\
[\boldsymbol{A}_{\mathrm{up,lo}}]_{i,\ell} &= \int_{\Omega,h} \upsilon_{\mathrm{up}}^{(i)} \upsilon_{\mathrm{lo}}^{(\ell)} && i = 1, \ldots, N_{\mathrm{up}},\ \ell = 1, \ldots, N_{\mathrm{lo}}, \\
[\boldsymbol{A}_{\mathrm{up,up}}]_{i,\ell} &= \int_{\Omega,h} \upsilon_{\mathrm{up}}^{(i)} \upsilon_{\mathrm{up}}^{(\ell)} && i = 1, \ldots, N_{\mathrm{up}},\ \ell = 1, \ldots, N_{\mathrm{up}}.
\end{aligned}$$

Finally, the matrix $\boldsymbol{C} \in \mathbb{R}^{2\times(N^{\mathrm{lo}}+N^{\mathrm{up}}+2)}$ and the vector $\boldsymbol{d} \in \mathbb{R}^2$ are given by

$$\boldsymbol{C} := \begin{pmatrix} \mathbf{1}_{1,N_{\mathrm{lo}}} & \mathbf{0}_{1,N_{\mathrm{up}}} & 0 & 0 \\ \mathbf{0}_{1,N_{\mathrm{lo}}} & \mathbf{1}_{1,N_{\mathrm{up}}} & 0 & 0 \end{pmatrix}, \quad \boldsymbol{d} := \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

By solving the linear program (7.37) we obtain an optimal vector $\boldsymbol{x}_*$, which defines a corresponding direction $\eta_{h,*} \in -\widetilde{\Gamma}$ via (7.34) and (7.38). The element $\eta_{h,*}$ is expected to define a descent direction for the goal functional $\widetilde{J} = J_{h,j} \circ Y_h \circ S_G$ at $\xi_h$. The choice strategy hence consists in choosing $-\upsilon_h := \eta_{h,*}$ for the update (7.24) of Algorithm 7.1.

We remark that linear programs of the form (7.37) can be solved e.g. by the MATLAB function `linprog`, which is provided by MATLAB's Optimization Toolbox. We also used the C++ function library lp_solve, which is available under a GNU Lesser General Public License (cf. lpsolve.sourceforge.net/5.5/).

## 7.7 Remarks

In Section 7.3 and Section 7.4 we discussed, that a photonic band structure optimization problems is, in fact, a constrained minimization problem. The so-called box constraint (7.14) arises from the definition of the extended admissible set $\overline{\mathcal{C}}_h$, and the symmetry constraint (7.21) arises from the assumption that approximate band structure computations can be restricted to quasimomentum vectors lying inside an irreducible zone of the first Brillouin zone. Algorithm 7.1 ensures both constraints by defining an alternative minimization, which is posed on the unconstrained space $\mathcal{E}_h$. In essence, this was accomplished through a parametrization of a subset of $\overline{\mathcal{C}}$ by virtue of the operator $Y_h \circ S_G$.

We also tested other methods for to ensure the optimization constraints. A simple way for ensuring the box constraint is to use a so-called *projective generalized gradient method*. Such a method was used e.g. by Cox and Dobson (cf. [29], [30]). A projective generalized gradient method constructs a minimizing sequence ${\rho_h}^{(0)}, {\rho_h}^{(1)}, {\rho_h}^{(2)}, \ldots$, etc. in $\overline{\mathcal{C}}_h$ by choosing in each iteration a descent direction $-{\upsilon_h}^{(l-1)} \in -\Gamma_{J_h}(\xi^{(l-1)})$ and by applying the update rule

$$ {\rho_h}^{(l)} := P_h\big({\rho_h}^{(l-1)} - s_l\, {\upsilon_h}^{(l-1)}\big) \qquad \text{for } l = 1, 2, \ldots \,, $$

where the projection operator $P_h : \mathcal{E}_h \to \mathcal{C}_h$ is given by

$$ P_h\big(\xi_h\big)(\boldsymbol{x}) := \begin{cases} \rho_{\min} & \text{if } \xi_h(\boldsymbol{x}) < \rho_{\min}, \\ \rho_{\max} & \text{if } \xi_h(\boldsymbol{x}) > \rho_{\max}, \\ \xi_h(\boldsymbol{x}) & \text{else} \end{cases} \qquad \text{for all } \xi_h \in \mathcal{E}_h,\ \boldsymbol{x} \in \mathcal{X}_h. $$

We found that certain problems can occur for this approach: The choice strategies for the descent directions are usually based on a linear program similar to that in Section 7.6. Such linear programs, in essence, stem from a local linearization of the goal functional $J_h$. However, these linearizations cannot take into accout the effect of the non-linear operator $P_h$. Therefore, it can happen that a descent direction for the functional $J_h$ is chosen, which is not a descent direction for $J_h \circ P_h$. In such cases the optimization algorithm stops prematurely.

We also tested a so-called *penalty approach*, which consists in devising a strictly convex functional $g : \mathcal{E}_h \to [0, \infty)$, such that $g(\xi_h) = 0$ is satisfied if and only if $\xi_i \in \mathcal{C}$. Choosing a penalty parameter $a > 0$, one then considers the minimization problem for the functional $J_h + \alpha g$. We tested with different choices for $\alpha$ and $g$ but could not obtain good results with this approach, so far.

# Chapter 8

# A Level-Set Method

In the previous chapter we developed an optimization algorithm for photonic band structure optimization problems. The algorithm was based on a so-called gradient method. We discussed that this method is not guaranteed to converge to an admissible solution. In this section we describe another algorithm, which is based on a so-called level set method. In Section 8.1 we introduce some preliminary concepts. In Section 8.1 we motivate an approach for a level set method for photonic band structure optimization problems. In Section 8.3 we describe a concrete optimization algorithm for such problems. In Section 8.4 we make some general remarks about this algorithm.

## 8.1 Basic Concepts

In this section we introduce some fundamental concepts of shape optimization, such as shape functionals, shape derivatives and level set methods. Here, we only give a brief overview of the most important concepts leaving out any mathematically rigorous derivations. For more details on shape optimization the reader can refer to the book of Sokołowski and Zolesio [70]. Level set methods are discussed in detail in the standard textbook of Osher and Fedkiw [58].

Suppose that $\Omega$ is a simply-connected, open domain in $\mathbb{R}^n$ for some space dimension $n \in \mathbb{N}$, which is bounded by a Lipschitz smooth boundary curve $\partial\Omega$. In order to simplify notations, we define the linear function spaces

$$\mathcal{H} := H^1_{\mathrm{per}}(\Omega, \mathbb{R}), \tag{8.1}$$

$$\mathcal{V} := W^{1,\infty}(\Omega, \mathbb{R}). \tag{8.2}$$

Here, $W^{1,\infty}(\Omega, \mathbb{R})$ denotes the Sobolev space that consist of all functions in $L^\infty(\Omega, \mathbb{R})$, which admit weak first partial derivatives in $L^\infty(\Omega, \mathbb{R})$.

Let $\mathcal{J} : \mathcal{H} \to \mathcal{P}(\Omega)$ be the mapping, which is densely defined by

$$\mathcal{J}(\varphi) := \{\boldsymbol{x} \in \Omega \,|\, \varphi(\boldsymbol{x}) < 0\} \qquad \text{for all } \varphi \in C^\infty(\Omega, \mathbb{R}).$$

Here, $\mathcal{P}(\Omega)$ denotes the power set of $\Omega$. We then define the system

$$\mathcal{S} := \{S \subset \Omega \,|\, \exists\ \varphi \in \mathcal{H} \text{ such that } S = \mathcal{J}(\varphi)\}. \tag{8.3}$$

A set $S \in \mathcal{S}$ is called an *admissible shape* in $\Omega$. By definition, every admissible shape is an open subset of $\Omega$. Furthermore, the boundary $\partial S$ of each admissible shape $S \in \mathcal{S}$ coincides with the zero level set of a function in $\mathcal{H}$.

A function $f : \mathcal{S} \to \mathbb{R}$ is called a *shape functional*, and a minimization problem of the form

$$\underset{S \in \mathcal{S}}{\text{minimize}}\ f(S) \tag{8.4}$$

is called a *shape optimization problem*. Problems of this type typically arise in structural optimization (see e.g. [3], [2]).

A *level set method* is an optimization method for shape optimization problems. The basic idea of a level set method is to represent an initial, admissible shape $S^{(0)} \in \mathcal{S}$ by a function $\varphi^{(0)} \in \mathcal{H}$ through

$$S^{(0)} = \mathcal{J}(\varphi^{(0)}).$$

By definition of $\mathcal{S}$, such a function $\varphi^{(0)}$ always exists. Now, suppose that $\psi \in C^1(\mathbb{R}, \mathcal{H})$ is a function that satisfies the first-order, parabolic initial boundary value problem

$$\begin{cases} \dot{\psi}(t, \cdot\,) + \boldsymbol{v}(t, \cdot\,) \cdot \nabla\psi(t, \cdot\,) = 0 & \text{in } \Omega \text{ for all } t \in (0, T], \\ \psi(0, \cdot\,) = \varphi^{(0)}. \end{cases} \tag{8.5}$$

for a given function $\boldsymbol{v} \in C^1(\mathbb{R}, \mathcal{V}^n)$ and a given positive number $T > 0$. Here, $\dot{\psi}(t)$ denotes the derivative of $\psi$ with respect to the first variable, and $\nabla\psi(t, \cdot\,)$ denotes the gradient of the function $\psi(t, \cdot\,) \in \mathcal{H}$ for all $t \in \mathbb{R}$. Note that the initial boundary value problem is a diffusion-convection-type problem. Therefore, the function $\boldsymbol{v}$ in (8.5) is commonly referred to as a *velocity field*. With the solution $\psi$ of (8.5) at hand, one can define a family $\{S^{(t)}\}_{t \in [0,T]}$ of admissible shapes by

$$S^{(t)} := \mathcal{J}\big(\psi(t, \cdot\,)\big), \qquad t \in [0, T]. \tag{8.6}$$

For every "point in time" $t \in [0, T]$ the function $\varphi^{(t)} := \psi(t, \cdot\,)$ is called the *level set function* of $S^{(t)}$. The initial boundary value problem (8.5) is commonly known as the *level set equation* for the family of level set functions $\{\varphi^{(t)}\}_{t \in [0,T]}$.

For a sufficiently small $T > 0$ the velocity field $\boldsymbol{v}$ in (8.5) uniquely defines a vector field $\boldsymbol{T}_{\boldsymbol{v}} \in C^1([0,T], \mathcal{V}^n)$ as the solution of the initial value problem

$$\begin{cases} \boldsymbol{T}_{\boldsymbol{v}}(t,\boldsymbol{x}) = \boldsymbol{v}\big(t, \boldsymbol{T}_{\boldsymbol{v}}(t,\boldsymbol{x})\big) & \text{for all } t \in I, \boldsymbol{x} \in \Omega, \\ \boldsymbol{T}_{\boldsymbol{v}}(0,\boldsymbol{x}) = \boldsymbol{x} & \text{for all } \boldsymbol{x} \in \Omega, \end{cases}$$

such that $\boldsymbol{T}_{\boldsymbol{v}}(t,\,\cdot\,)$ ist a $W^{1,\infty}$-diffeomorphism for all $t \in [0,T]$ (see Theorem 2.16 in [70]). Furthermore, one can show (see Section 2.9 in [70]) that

$$S^{(t)} = \boldsymbol{T}_{\boldsymbol{v}}(t, S^{(0)}) \qquad \text{for all } t \in [0,T],$$

where $S^{(t)}$ coincides with the admissible shape defined by (8.6).

Given a shape optimization problem of the form (8.4) and an initial admissible shape $S^{(0)} \in \mathcal{S}$, the basic idea of a level set method is to construct a velocity field $\boldsymbol{v} \in C^1(\mathbb{R}, \mathcal{V}^n)$, such that $f(S^{(t)}) = f\big(\boldsymbol{T}_{\boldsymbol{v}}(t, S^{(0)})\big)$ converges to a local minimum of $f$ as $t$ tends to $\infty$. In order to construct such a velocity field, one needs to obtain some information about the local behaviour of $f(S^{(t)})$ for a given point in time $t \in \mathbb{R}$ with respect to small perturbations in $t$. Such information is provided by the concept of *shape differentiability*.

Given a shape functional $f : \mathcal{S} \to \mathbb{R}$ and a velocity field $\boldsymbol{v} \in C^1(\mathbb{R}, \mathcal{V}^n)$, the shape functional $f$ is said to be *directionally shape differentiable* at a shape $S \in \mathcal{S}$ in direction $\boldsymbol{v}$, if the limit

$$f'(S;\boldsymbol{v}) := \lim_{t \to \infty} \frac{f\big(\boldsymbol{T}_{\boldsymbol{v}}(t,S)\big) - f(S)}{t}$$

exists. The limit $f'(S;\boldsymbol{v})$ is then called the *directional shape derivative* of $f$ at $S$ in direction $\boldsymbol{v}$. If $f'(S;\,\cdot\,)$ is a continuous, linear functional on $C^1(\mathbb{R}, \mathcal{V}^n)$, then $f$ is called *shape differentiable* at $S$. The *shape differential*

$$\mathrm{D}f(S) \in \big[C^1(\mathbb{R}, \mathcal{V}^n)\big]^*$$

of $f$ at $S$ is then defined by

$$\mathrm{D}f(S)[\boldsymbol{v}] := f'(S;\boldsymbol{v}) \qquad \text{for all } \boldsymbol{v} \in C^1(\mathbb{R}, \mathcal{V}^n).$$

We remark that $\mathrm{D}f(S)[\boldsymbol{v}]$ is often referred to as the *shape derivative* of $f$ at $S$.

## 8.2 Motivation

In this section we motivate a level set method for photonic band structure optimization problems, which involve the TE band structures of two-dimensional

photonic crystals. The motivation we present here is by no means mathematically rigorous. It is only meant to provide an intuitive understanding of how level set methods for photonic band structure optimization problems can be constructed.

In the following, we assume that $\Omega$ is a primitive domain of a Bravais lattice $\Lambda \subset \mathbb{R}^2$ of rank 2. By $\mathbb{B}$ we denote the first Brillouin zone of $\Lambda$. Given an index $j \in \mathbb{N}$, a coefficient $\rho \in \mathcal{D}$ and a vector $\boldsymbol{k} \in \mathbb{B}$, the non-negative real number $\lambda_j^{\mathrm{TE}}(\rho, \boldsymbol{k})$ was defined as the $j$-th smallest eigenvalue of the corresponding eigenvalue problem (4.94) on page 87. The function $u_j^{\mathrm{TE}}(\rho, \boldsymbol{k}) \in V$ was defined as the corresponding eigenfunction. Without loss of generality, we assume that

$$m^{\mathrm{TE}}\big(u_j^{\mathrm{TE}}(\rho, \boldsymbol{k}), u_j^{\mathrm{TE}}(\rho, \boldsymbol{k})\big) = 1 \qquad \text{for all } j \in \mathbb{N},\ \rho \in \mathcal{D},\ \boldsymbol{k} \in \mathbb{B}.$$

Then, the $j$-th smallest TE eigenvalue $\lambda_j^{\mathrm{TE}}(\rho, \boldsymbol{k})$ can be characterized as

$$\begin{aligned} \lambda_j^{\mathrm{TE}}(\rho, \boldsymbol{k}) &= a_{\boldsymbol{k}}^{\mathrm{TE}}(\rho)\big(u_j^{\mathrm{TE}}(\rho, \boldsymbol{k}), u_j^{\mathrm{TE}}(\rho, \boldsymbol{k})\big) \\ &= \int_\Omega \rho \big|(\nabla + \mathrm{i}\boldsymbol{k}) u_j^{\mathrm{TE}}(\rho, \boldsymbol{k})\big|^2 \qquad \text{for all } j \in \mathbb{N},\ \rho \in \mathcal{D},\ \boldsymbol{k} \in \mathbb{B}. \end{aligned}$$

For the purpose of this section we assume that every TE eigenvalue $\lambda_j^{\mathrm{TE}}(\rho, \boldsymbol{k})$ is a geometrically simple eigenvalue, i.e., we assume that the eigenspace of every TE eigenvalue is one-dimensional.

We assume that the system $\mathcal{S}$ of admissible shapes in $\Omega$ is defined according to (8.3). Then, we define the function set

$$\widetilde{\mathcal{C}} := \big\{ \rho_{\min} + (\rho_{\max} - \rho_{\min}) \chi_S \;\big|\; S \in \mathcal{S} \big\}.$$

Clearly, $\widetilde{\mathcal{C}}$ is a subset of the admissible set $\mathcal{C}$. Next, we define the mapping $\mathcal{R} : \mathcal{S} \to \widetilde{\mathcal{C}}$ by

$$\mathcal{R}(S) := \rho_{\min} + (\rho_{\max} - \rho_{\min}) \chi_S \qquad \text{for all } S \in \mathcal{S}.$$

Gien an index $j \in \mathbb{N}$ and a quasimomentum vector $\boldsymbol{k} \in \mathbb{B}$, we define the shape functional $\widetilde{\lambda}_j^{\mathrm{TE}}(\,\cdot\,, \boldsymbol{k}) : \mathcal{S} \to \mathbb{R}$ and the mapping $\widetilde{u}_j^{\mathrm{TE}}(\,\cdot\,, \boldsymbol{k}) : \mathcal{S} \to V$ by

$$\begin{aligned} \widetilde{\lambda}_j^{\mathrm{TE}}(\,\cdot\,, \boldsymbol{k}) &:= \lambda_j^{\mathrm{TE}}(\,\cdot\,, \boldsymbol{k}) \circ \mathcal{R}, \\ \widetilde{u}_j^{\mathrm{TE}}(\,\cdot\,, \boldsymbol{k}) &:= u_j^{\mathrm{TE}}(\,\cdot\,, \boldsymbol{k}) \circ \mathcal{R}. \end{aligned}$$

One easily verifies that the shape functional $\widetilde{\lambda}_j^{\mathrm{TE}}(\,\cdot\,, \boldsymbol{k})$ is given by

$$\begin{aligned} \widetilde{\lambda}_j^{\mathrm{TE}}(S, \boldsymbol{k}) = \rho_{\min} &\int_\Omega \big|(\nabla + \mathrm{i}\boldsymbol{k}) \widetilde{u}_j^{\mathrm{TE}}(S, \boldsymbol{k})\big|^2 \\ &+ (\rho_{\max} - \rho_{\min}) \int_S \big|(\nabla + \mathrm{i}\boldsymbol{k}) \widetilde{u}_j^{\mathrm{TE}}(S, \boldsymbol{k})\big|^2 \qquad \text{for all } S \in \mathcal{S}. \end{aligned}$$

It can be shown that $\widetilde{\lambda}_j^{\mathrm{TE}}(\,\cdot\,,\boldsymbol{k})$ is shape differentiable at an admissible shape $S \in \mathcal{S}$, provided that $\lambda_j^{\mathrm{TE}}(\widetilde{\rho},\boldsymbol{k})$ is a geometrically simple eigenvalue for all coefficients $\widetilde{\rho}$ in a neighbourhood of $\mathcal{R}(S)$. The shape differential is given by

$$\mathrm{D}\widetilde{\lambda}_j^{\mathrm{TE}}(\,\cdot\,,\boldsymbol{k})(S)[\boldsymbol{v}] = (\rho_{\max}-\rho_{\min})\int_{\partial S} \boldsymbol{\nu}_S \cdot \boldsymbol{v}(0,\,\cdot\,)\big|(\nabla+\mathrm{i}\boldsymbol{k})\widetilde{u}_j^{\mathrm{TE}}(S,\boldsymbol{k})\big|^2$$

for all $\boldsymbol{v} \in \mathcal{V}$, where $\boldsymbol{\nu}_S : \partial S \to \mathbb{S}^1$ denotes the outer unit normal field on the boundary of $S$.

In the following, we aim at constructing a minimizer $S_* \in \mathcal{S}$ for $\widetilde{\lambda}_j^{\mathrm{TE}}(\,\cdot\,,\boldsymbol{k})$. We want to construct the minimizer with a level set method. Therefore, we choose an arbitrary initial admissible shape $S_0 \in \mathcal{S}$, which is given by $S_0 = \mathcal{J}(\varphi^{(0)})$ for some initial function level set function $\varphi^{(0)} \in \mathcal{H}$. Given a velocity field $\boldsymbol{v} \in C^1(\mathbb{R},\mathcal{V}^2)$, we can construct a family $\{S^{(t)}\}_{t\in[0,T]}$ of admissible shapes by solving the level set equation

$$\begin{cases} \dot{\psi}(t,\,\cdot\,) + \boldsymbol{v}(t)\cdot\nabla\psi(t,\,\cdot\,) = 0 & \text{in } \Omega \text{ for all } t \in (0,T], \\ \psi(0,\,\cdot\,) = \varphi^{(0)}. \end{cases} \tag{8.7}$$

for $\psi \in C^1([0,T],\mathcal{H})$, and by setting $S^{(t)} := \mathcal{J}(\psi(t,\,\cdot\,))$ for all $t \in [0,T]$. Clearly, this specific family of admissible shapes is completely determined by the initial level set function $\varphi^{(0)}$ and by the velocity field $\boldsymbol{v}$. Furthermore, one can show that the shape differential of $\widetilde{\lambda}_j^{\mathrm{TE}}(\,\cdot\,,\boldsymbol{k})$ satisfies

$$\begin{aligned}\mathrm{D}\widetilde{\lambda}_j^{\mathrm{TE}}(\,\cdot\,,\boldsymbol{k})(S^{(t)})[\boldsymbol{v}] = (\rho_{\max}-\rho_{\min})\int_{\partial S^{(t)}} \boldsymbol{\nu}_{S^{(t)}} \cdot \boldsymbol{v}(t,\,\cdot\,)\big|(\nabla+\mathrm{i}\boldsymbol{k})\widetilde{u}_j^{\mathrm{TE}}(S^{(t)},\boldsymbol{k})\big|^2 \\ + o(t^2).\end{aligned} \tag{8.8}$$

Recall that every shape $S^{(t)}$ is determined by its corresponding level set function $\varphi^{(t)} = \psi(t,\,\cdot\,)$. Provided that the gradient $\nabla\varphi^{(t)}$ does not vanish on $\partial S^{(t)}$ for any $t \in [0,T]$, we can represent the outer unit normal fields $\boldsymbol{\nu}_{S^{(t)}} : \partial S^{(t)} \to \mathbb{S}^1$ on the boundaries of the admissible shapes $S^{(t)}$ by

$$\boldsymbol{\nu}_{S^{(t)}} = \frac{\nabla\varphi^{(t)}}{|\nabla\varphi^{(t)}|}\bigg|_{\partial S^{(t)}} \qquad \text{for all } t \in [0,T].$$

Under the further assumption that $\nabla\varphi^{(t)}$ does not vanish almost everywhere on $\Omega$ for any $t \in [0,T]$, we can define continuous extensions $\boldsymbol{\nu}^{(t)} : \Omega \to \mathbb{S}^1$ of the outer unit normal fields $\boldsymbol{\nu}_{S^{(t)}}$ by

$$\boldsymbol{\nu}^{(t)} = \frac{\nabla\varphi^{(t)}}{|\nabla\varphi^{(t)}|} \qquad \text{for all } t \in [0,T].$$

Now, suppose that for every $t \in [0,T]$ the velocity field $\boldsymbol{v}(t,\,\cdot\,)$ is normal to $\partial S^{(t)}$, i.e., that $\boldsymbol{v}(t,\,\cdot\,)$ is given by

$$\boldsymbol{v}(t,\,\cdot\,) = c(t,\,\cdot\,)\boldsymbol{\nu}^{(t)} \qquad \text{for all } t \in [0,T] \tag{8.9}$$

for some function $c \in C^1(\mathbb{R},\mathcal{V})$. Then,(8.8) can be rewritten as

$$\begin{aligned}\mathrm{D}\widetilde{\lambda}_j^{\mathrm{TE}}(\,\cdot\,,\boldsymbol{k})(S^{(t)})[\boldsymbol{v}] = (\rho_{\max}-\rho_{\min})\int_{\partial S^{(t)}} c(t,\,\cdot\,)\big|(\nabla+\mathrm{i}\boldsymbol{k})\widetilde{u}_j^{\mathrm{TE}}(S^{(t)},\boldsymbol{k})\big|^2 \\ + o(t^2).\end{aligned} \tag{8.10}$$

Furthermore, the level set equation (8.7) reduces to a so-called *Hamilton–Jacobi*-type system, which reads

$$\begin{cases} \dot{\psi}(t,\,\cdot\,) + c(t,\,\cdot\,)|\nabla\psi(t,\,\cdot\,)| = 0 & \text{in } \Omega \text{ for all } t \in (0,T], \\ \psi(0,\,\cdot\,) = \varphi^{(0)}. \end{cases} \tag{8.11}$$

The function $c$ in (8.11) is commonly referred to as the *speed field*.

Recall that, for every $t \in [0,T]$, the shape differential $\mathrm{D}\widetilde{\lambda}_j^{\mathrm{TE}}(\,\cdot\,,\boldsymbol{k})(S^{(t)})$ is a linearization of the shape functional $\widetilde{\lambda}_j^{\mathrm{TE}}(\,\cdot\,,\boldsymbol{k})$ at the admissible shape $S^{(t)}$ with respect to perturbations of $S^{(t)}$ that are generated by velocity fields $\boldsymbol{v} \in C^1(\mathbb{R},\mathcal{V}^2)$. Hence, a negative sign of the integral in (8.10) indicates, that the value of $\widetilde{\lambda}_j^{\mathrm{TE}}(S^{(t)},\boldsymbol{k})$ is likely to decrease, if the shape $S^{(t)}$ is perturbed by the velocity $\boldsymbol{v}$, which is defined in (8.9). Clearly, a negative sign of the integral can be guaranteed by setting

$$c(t,\,\cdot\,) := -(\rho_{\max}-\rho_{\min})\big|(\nabla+\mathrm{i}\boldsymbol{k})\widetilde{u}_j^{\mathrm{TE}}(S^{(t)},\boldsymbol{k})\big|^2 \qquad \text{for all } t \in [0,T].$$

Now, let $J^{\mathrm{TE}} : \mathcal{D} \to \mathbb{R}$ be a given TE goal functional. We then define a corresponding shape functional $\widetilde{J}^{\mathrm{TE}} : \mathcal{S} \to \mathbb{R}$ by $\widetilde{J}^{\mathrm{TE}} := J^{\mathrm{TE}} \circ \mathcal{R}$ and consider the minimization problem

$$\underset{S\in\mathcal{S}}{\text{minimize}}\ \widetilde{J}^{\mathrm{TE}}(S).$$

Suppose that the goal functional $J^{\mathrm{TE}}$ is strictly differentiable at some density function $\rho = \mathcal{R}(S)$, where $S \in \mathcal{S}$ is an admissible shape, and that the strict differential of $J^{\mathrm{TE}}$ at $\rho$ is given by

$$\mathrm{D}_{\mathrm{s}}J^{\mathrm{TE}}(\rho)[\eta] = \int_\Omega \gamma\eta \qquad \text{for all } \eta \in \mathcal{E} \tag{8.12}$$

for some uniquely defined function $\gamma \in L^1(\Omega)$. Then, one can show that the shape differential of $\widetilde{J}^{\mathrm{TE}}$ at the admissible shape $S$ is given by

$$\mathrm{D}\widetilde{J}^{\mathrm{TE}}(S)[\boldsymbol{v}] = (\rho_{\max}-\rho_{\min})\int_{\partial S} \boldsymbol{\nu}_S \cdot \boldsymbol{v}(0,\,\cdot\,)\gamma \qquad \text{for all } \boldsymbol{v} \in C^1(\mathbb{R},\mathcal{V}^2).$$

Starting off with an initial admissible shape $S^{(0)} \in \mathcal{S}$, one can attempt to construct a family $\{S^{(t)}\}_{t\in[0,T]}$ of admissible shapes, such that $\widetilde{J}^{\mathrm{TE}}(S^{(T)}) < J^{\mathrm{TE}}(S^{(0)})$ for some $T > 0$. Using a level set method, this family is constructed by choosing an initial level set function $\varphi^{(0)}$, such that $S^{(0)} = \mathcal{J}(\varphi^{(0)})$. The level set function is then evolved by solving the Hamilton–Jacobi equation (8.11). According to the discussion above, it is reasonable to choose the speed field $c \in C^1(\mathbb{R}, \mathcal{V})$, such that

$$c(t, \cdot\,) = -(\rho_{\max} - \rho_{\min})\, \gamma^{(t)} \qquad \text{for all } t \in [0, T],$$

where $\gamma^{(t)} \in L^1(\Omega, \mathbb{R})$ is the function that determines the strict derivative of $J^{\mathrm{TE}}$ at $\rho^{(t)} := \mathcal{R}(S^{(t)})$ according to (8.12).

If the goal functional $J^{\mathrm{TE}}$ is not strictly differentiable but only locally Lipschitz continuous, it is necessary to adapt the construction principle for the level set method presented above. We hence construct the family $\{S^{(t)}\}_{t\in[0,T]}$ of admissible shapes as follows: Given a family member $S^{(t)}$ and a corresponding density function $\rho^{(t)} := \mathcal{R}(S^{(t)})$, we compute a convex set of functions $\Gamma(\rho^{(t)}) \subset L^1(\Omega, \mathbb{R})$, such that

$$\partial J^{\mathrm{TE}}(\rho^{(t)}) \subset \left\{ p \in \mathcal{E}^* \;\middle|\; p(\eta) = \int_\Omega \gamma\eta,\ \gamma \in \Gamma(\rho^{(t)}) \right\}.$$

The speed field $c \in C^1(\mathbb{R}, \mathcal{V})$ is then chosen, such that

$$c(t, \cdot\,) \in -(\rho_{\max} - \rho_{\min})\Gamma(\rho^{(t)}) \qquad \text{for all } t \in [0, T].$$

To the best of our knowledge there is no conclusive evidence that this approach is justified in all cases. Nevertheless, we were able to develop an optimization algorithm based on this approach, which performed quite well. We decribed this algorithm in the next section.

## 8.3 The Algorithm

Based on the assumptions in the previous section, we implemented a level set algorithm for photonic band structure optimization problems. In this section we describe this algorithm, which is also listed in pseudo-code as Algorithm 8.1.

Under the assumptions of Section 7.2, we further assume that the Hilbert space $\mathcal{H}$ is discretized by a conforming finite element ansatz on the underlying mesh $\mathcal{T}_h$. We denote the corresponding finite element space by $\mathcal{H}_h$. We then define the so-called *interpretation operator* $\Theta_h : \mathcal{H} \to \mathcal{C}_h$ according to

$$\Theta_h(\varphi_h)(\boldsymbol{x}) := \begin{cases} \rho_{\min} & \text{if } \varphi_h(\boldsymbol{x}) > 0, \\ \rho_{\max} & \text{if } \varphi_h(\boldsymbol{x}) < 0, \end{cases} \qquad \text{for all } \boldsymbol{x} \in \mathcal{X}_h.$$

**Algorithm 8.1** Level-Set Method

**Input:**

| | |
|---|---|
| $J_h : \mathcal{D}_h \to \mathbb{R}$ | *Goal functional* |
| $G$ | *Group of orthogonal transformations (Point group)* |
| ${\varphi_h}^{(0)} \in \mathcal{H}_h$ | *Initial level set function* |

**Parameters:**

| | |
|---|---|
| $t_{\min} > 0$ | *Minimal time step* |
| $t_{\max} > t_{\min}$ | *Maximal step size* |
| $\Delta_{\min} > 0$ | *Minimal decrease* |

**Algorithm:**

```
ρ_h^(0) := Θ_h(φ_h^(0))
compute z_0 := J_h(ρ_h^(0))
t_0 := t_max
for l = 1, 2, ... do
  compute Γ_{J_h}(ρ_h^(l-1))
  choose a descent direction −γ_h^(l-1) ∈ −(ρ_max − ρ_min)Γ_{J_h}(ρ_h^(l-1))
  γ_h^(l-1) := γ̃_h^(l-1) / ‖γ̃_h^(l-1)‖_∞
  t_{l,1} := min{2t_{l-1}, t_max}
  for p = 1, 2, ... do
    φ_h^(l,p) := Ψ_h(φ_h^(l-1), −γ_h^(l-1), t_{l,p})
    ρ_h^(l,p) := Θ_h(φ_h^(l,p))
    compute z_{l,p} := J_h(ρ_h^(l))
    if z_{l-1} − z_{l,p} > Δ_min then
      φ_h^(l) := φ_h^(l,p)
      ρ_h^(l) := ρ_h^(l,p)
      z_l := z_{l,p}
      t_l := t_{l,p}
      break
    end if
    t_{l,p+1} := t_{l,p}/2
    if t_{l,p+1} < t_min then
      return ρ_h^(l-1)
    end if
  end for
end for
```

Here, the understanding is that the function $\varphi_h$ represents the level set function of a shape $S \in \mathcal{S}$. The shape $S$ itself is understood to represent the area inside the primitive domain of a photonic crystal with optical density $\rho_{\max}$.

For convenience, we define the so-called *evolution operator* $\Psi : \mathcal{H} \times \mathcal{V} \times \mathbb{R} \to \mathcal{H}$ as follows. Given a level set function $\varphi \in \mathcal{H}$, a speed field $c \in \mathcal{V}$, and a positive number $T > 0$, let $\psi \in C^1([0,T], \mathcal{H})$ be the solution of the initial boundary value problem

$$\begin{cases} \dot{\psi}(t,\,\cdot\,) + c|\nabla\psi(t,\,\cdot\,)| = 0 & \text{in } \Omega \text{ for all } t \in (0,T], \\ \psi(0,\,\cdot\,) = \varphi. \end{cases}$$

Then, we define $\Psi(\varphi, c, T) := \psi(t,\,\cdot\,)$. In practice, the evolution operator $\Psi$ is realized by a *discretized evolution operator* $\Psi_h : \mathcal{H}_h \times \mathcal{E}_h \times \mathbb{R} \to \mathcal{H}_h$. Such a discretized evolution operator is evaluated by solving the above initial boundary value problem numerically. In two space dimensions we implemented a second-order, semi-discrete central evolution scheme for Hamilton-Jacobi equations, which was proposed by Kurganov and Tadmor (see [51]).

Given a discretized band structure optimization problem of the form

$$\underset{\rho_h \in \mathcal{C}_h}{\text{minimize}} \; J_h(\rho_h),$$

we assume that the inital density function $\rho_h^{(0)}$ is determined by an *initial level set function* ${\varphi_h}^{(0)}$, which we require as an input parameter for the algorithm. In every iteration, a superset $\Gamma_{J_h}({\rho_h}^{(l-1)})$ of the generalized gradient of $J_h$ at ${\rho_h}^{(l-1)}$ is computed. A descent direction $-{\gamma_h}^{(l-1)}$ is then chosen from the set $-\Gamma_{J_h}({\rho_h}^{(l-1)})$. The choice is accomplished through a suitable choice strategy. We found that a choice strategy similar to that developed in Section 7.6 is suitable, if the goal functional $J_h$ is given by a gap width functional as defined in Section 5.5. A new discretized level set function ${\varphi_h}^{(l)}$ is obtained as the result of $\Psi_h({\varphi_h}^{(l-1)}, -{\gamma_h}^{(l-1)}, t_{l,p})$.

The "time step" $t_{l,p}$ of the level set evolution is determined adaptively by a trial-and-error approach. Whenever an evolution by a given time step fails to produce a level set function that yields a prescribed minimal decrease $\Delta_{\min}$ of the goal value, the step size is divided by two. In order to avoid excessively small step sizes, the step size is doubled at the beginning of each iteration, without exceeding a prescribed maximal time step $t_{\min}$, however. If the time step drops below a prescribed minimal time step $t_{\min}$, the algorithm terminates. In this case the discretized level set function of the previous time step is chosen as the *final level set function* ${\phi_h}^{(L)}$. Through ${\rho_h}^{(L)} := \Theta_h({\varphi_h}^{(L)})$ a final density function ${\rho_h}^{(L)}$ is determined, which is then returned by the algorithm. By construction, this final density function belongs to the admissible set $\mathcal{C}_h$.

We conducted a series of numerical experiments on photonic band gap maximization problems for two-dimensional photonic crystals using Algorithm 8.1. Some of the results are presented in Chapter 9.

## 8.4 Remarks

From a algorithmic point of view the generalized gradient method as given by Algorithm 7.1 and the level set method as given by Algorithm 8.1 are very similar. In both methods descent directions are chosen from a superset of the goal functional's generalized gradient. The methods only differ in how they use this descent direction to construct new density functions. In the generalized gradient method the descent directions are simply added to the current density functions. In contrast to this, the level method uses the descent direction as a speed function in a Hamilton–Jacobi equation by which the level set is evolved.

In the literature, it is often discussed that the level set functions can develop slopes, which are either very steep or very flat. Both cases are known to cause problems for the convergence of a level set method. It is often proposed to restart the level set method periodically. Another method consists in regularizing the level set function after several iterations by solving the problem

$$\begin{aligned} \dot{\psi}(t,\,\cdot\,) + \operatorname{sgn}(\varphi)\big(|\nabla\psi(t,\,\cdot\,)| - 1\big) = 0 \quad &\text{in } \Omega \text{ for all } t > 0, \\ \psi(0,\,\cdot\,) = \varphi. \end{aligned}$$

This problem is known to converge to a so-called signed distance function for the zero level set of $\phi$ (see e.g. Section 7.4 in [58]). In our numerical experiments we did not implement either of the above techniques as convergence was not found to be a problem.

# Chapter 9

# Numerical Results

In this chapter we demonstrate the performance of the optimization algorithms developed in Chapter 7 and Chapter 8. In Section 9.1 and Section 9.2 we present numerical experiments on photonic band gap maximization problems (PBGMPs) for two-dimensional photonic crystals. Numerical experiment on PBGMPs for three-dimensional photonic crystals are presented in Section 9.3.

## 9.1 Maximizing TM Band Gaps

In this section we present some numerical solutions of photonic band gap maximization problems (PBGMPs) involving TE band gaps of two-dimensional photonic crystals. Each problem was given by a minimization problem of the form

$$\underset{\rho\in\mathcal{C}}{\text{minimize}}\ J_j^{\mathrm{TM}}(\rho),$$

where $J_j^{\mathrm{TM}}$ denotes the TM band gap functional defined by (5.14) for a given index $j \in \mathbb{N}$. Throughout this section, we only consider photonic crystals, whose medium structures are periodic with respect to a square lattice $a\mathbb{Z}^2$, where $a > 0$ denotes the lattice constant (see Section 3.2). Furthermore, we assume that the point group of each crystal is isomorphic to the orthogonal group $\mathrm{O}_2(\mathbb{Z})$. In all numerical experiments the computational domain $\Omega$ was given by the square $(-a/2, a/2)^2$.

The photonic crystals that were considered consisted of exactly two different materials with relative electric permittivities $\varepsilon_{r,\min} = 1$ and $\varepsilon_{r,\max} = 9$. Hence, the admissible set was given by

$$\mathcal{C} := \big\{\rho \in \mathcal{E} \,\big|\, \rho = 1 + 8\chi_S,\ S \in \mathcal{M}(\Omega)\big\}.$$

We discretized the computational domain $\Omega$ by a $100 \times 100$, rectangular, structured mesh. The approximate band structures were computed through an $H^1$-conforming, bilinear finite element ansatz. The density functions were discretized

quadrature-point-wise as described in Section 7.2. We used a custom-built MATLAB toolbox to implement the finite element ansatz. This toolbox is briefly described in the Appendix Chapter A. The toolbox uses Lobatto-type shape functions (cf. in [73]) and Gauss-type quadrature rules. We used a rule of consistency order 7, which gave 12 quadrature points on each mesh element (cf. Table 4.40 in [73]). The finite element mesh as well as the quadrature rule were chosen such that such that the set of quadrature points $\mathcal{X}_h$ was $\mathrm{O}_2(\mathbb{Z})$-symmetric, which facilitated the implementation of symmetrization operations. All eigenvalue problems were solved using MATLAB's `eigs` function.

### 9.1.1 Results of the Generalized Gradient Method

In the following examples we applied the generalized gradient method as given by Algorithm 7.1 to various PBGMPs. The parameters of Algorithm 7.1 were chosen as $s_{\min} = 10^{-6}$, $s_{\max} = 1$, and $\Delta_{\min} = 10^{-8}$. The funnel operator $Y_h$ was given as defined by (7.15) and (7.20) with $r = 4$. When applied to an admissible density function in $\rho_h \in \mathcal{C}_h$, this funnel operator returns a two-valued function with function values of approximately 1.1439 and 7.8679. The descent directions where chosen by solving linear programs, as explained in Section 7.6. The linear programs were solved using the MATLAB function `linprog`, which is provided by the MATLAB Optimization Toolbox.

Since the generalized gradient method is not guaranteed to converge to an admissible, two-valued solution, we always comment on whether or not a solution can be considered to be close to an admissible solution. To make this notion precise, define for every positive number $\delta > 0$ the set

$$\Xi(\delta) := [\varepsilon_{r,\min}, \varepsilon_{r,\min} + \delta) \cup (\varepsilon_{r,\max} - \delta, \varepsilon_{r,\max}].$$

With this, a discretized permittivity function $\varepsilon_h \in \mathcal{E}_h$ with function values in the interval $[\varepsilon_{r,\min}, \varepsilon_{r,\max}]$ is considered to be close to an admissible permittivity function, if there exists a small number $\delta > 0$, such that $\varepsilon_h$ takes function values in $\Xi(\delta)$ at most quadrature points. Recall that the discrete function space $\mathcal{E}_h$ was defined by (7.5) on page 131. In order to obtain admissible permittivity functions, we subjected each solution of Algorithm 7.1 to the so-called *threshold operator* $\widetilde{H}_h$, which is defined by

$$\widetilde{H}_h\varepsilon_h(\boldsymbol{x}) := \begin{cases} \varepsilon_{r,\min} & \text{if } \rho(\boldsymbol{x}) < \dfrac{\varepsilon_{r,\min} + \varepsilon_{r,\max}}{2}, \\ \varepsilon_{r,\max} & \text{else .} \end{cases} \qquad \text{for all } \varepsilon_h \in \mathcal{E}_h, \boldsymbol{x} \in \mathcal{X}_h, \quad (9.1)$$

We also remind the reader that $\Pi_h$ was defined as the interpolation operator, which projects suitable functions onto the discrete function space $\mathcal{E}_h$ (see Section 7.2).

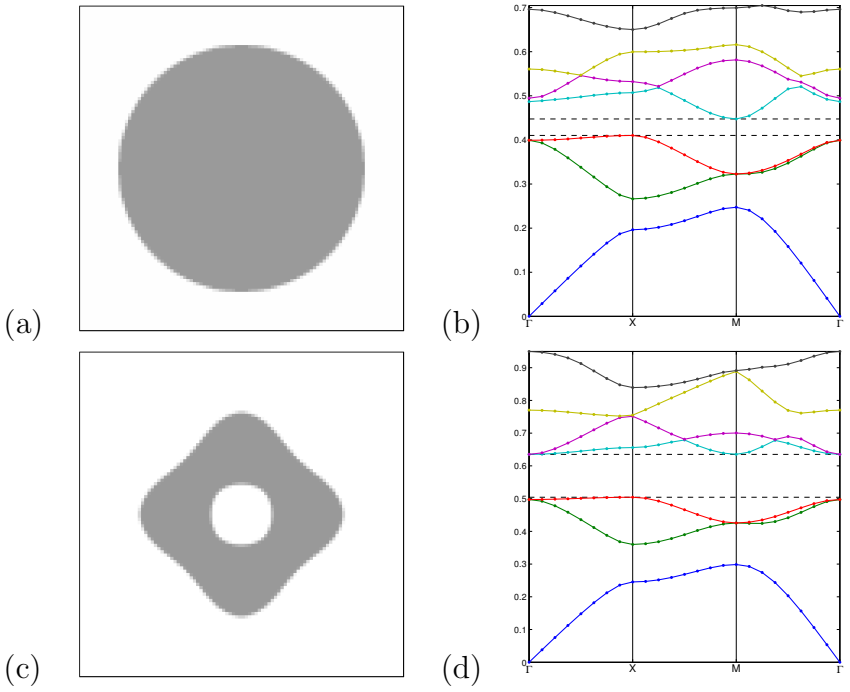



**Figure 9.1:** Results of Example 9.1. (a) Initial permittivity. (b) Initial band structure. (c) Final permittivity after 117 iterations. (d) Final band structure.

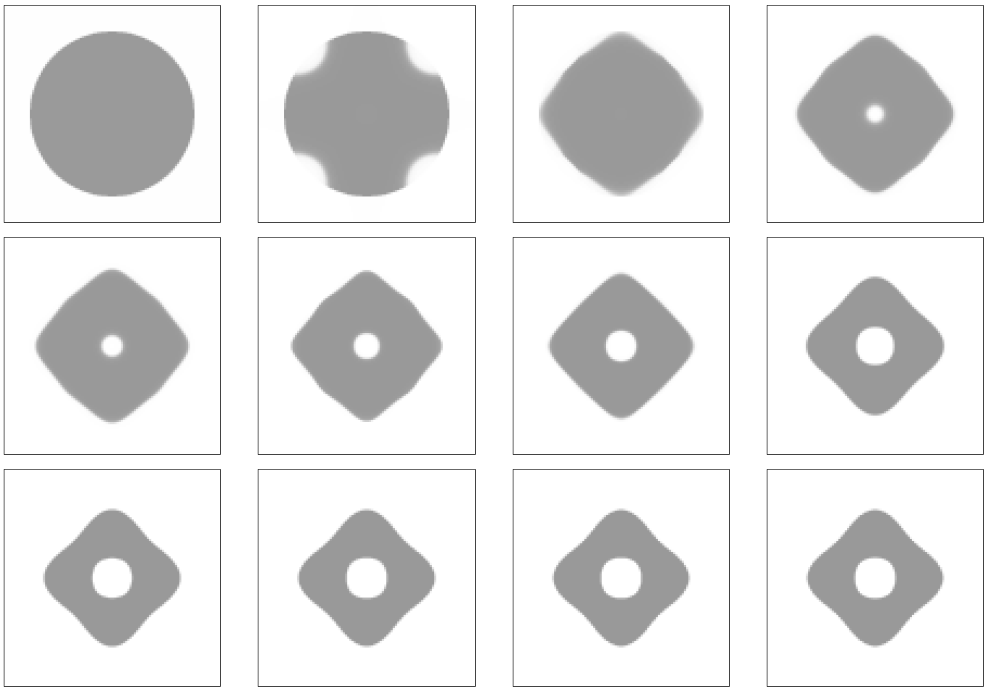
**Example 9.1.** In our first example we attempted to maximize a band gap in the TM band structure of a two-dimensional photonic crystal. The crystal consisted of a periodic arrangement of cylindrical rods. The radius of each rod was given by $0.38a$, where $a$ denotes the lattice constant. The relative electric permittivity of the rods was given by $\varepsilon_{r,\max} = 9$, that of the ambient medium by $\varepsilon_{r,\min} = 1$. This medium structure is represented by the relative electric permittivity function $\varepsilon_r$, whose restriction to $\Omega$ is given by

$$\varepsilon_r|_\Omega = 1 + 8\chi_{B_{0.38a}(\mathbf{0})},$$

where $B_r(\boldsymbol{x})$ denotes the open ball in $\mathbb{R}^2$ with radius $r > 0$ and center point $\boldsymbol{x} \in \mathbb{R}^2$. The TM band structure of the photonic crystal exhibited a band gap between the third and the fourth band. We computed an approximate gap width of $w_{3,h} = 0.0444 \cdot 2\pi c_0/a$, where $c_0$ denotes the speed of light.

Choosing $\xi_h^{(0)} := \mathcal{I}_h \varepsilon_r|_\Omega$ as the initial free density function, we obtained an initial permittivity function ${\varepsilon_r}_h^{(0)}$, which yielded a smaller gap width of approximately ${w_{3,h}}^{(0)} = 0.0374 \cdot 2\pi c_0/a$ (see Figure 9.1(a) and (b)). After 117 iterations, however, the algorithm converged to a final permittivity function ${\varepsilon_r}_h^{(117)}$,

(a)



(b)

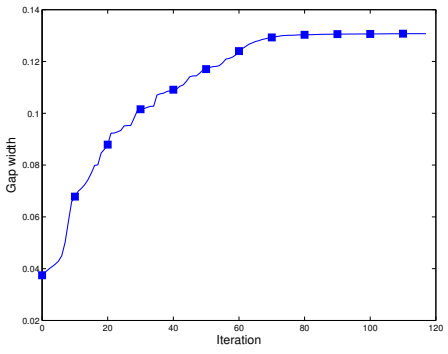



**Figure 9.2:** The evolution of the discretized permittivity function (a) and the corresponding gap width (b) in Example 9.1.

which represented a periodic arrangement of hollow rods with flower-shaped cross-sections (see Figure 9.1(c)). The width of the band gap increased to $w_{3,h}{}^{(117)} = 0.1307 \cdot 2\pi c_0/a$ (see Figure 9.1(d)).

We found that the minimal and maximal function values of $\varepsilon_{r,h}{}^{(117)}$ were given by 1.0000 and 8.9998, respectively. Moreover, $\varepsilon_{r,h}{}^{(117)}$ took function values in $\Xi(10^{-4})$ at 54.% of all quadrature points. Function values in $\Xi(10^{-3})$ and in $\Xi(10^{-2})$ were attained by $\varepsilon_{r,h}{}^{(117)}$ at 93.5% and 99.3% of all quadrature points,

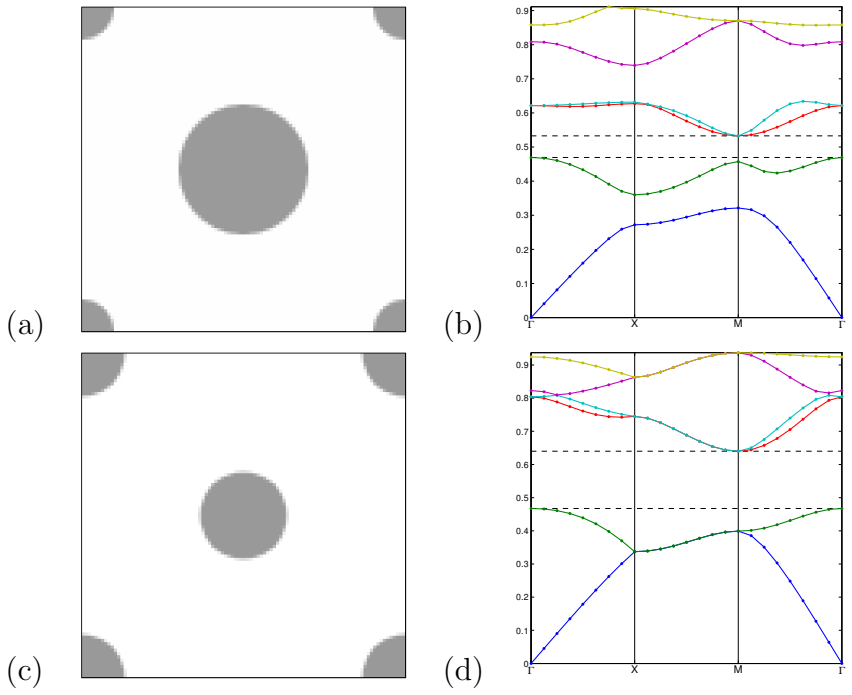



**Figure 9.3:** Results of Example 9.2. (a) Initial permittivity. (b) Initial band structure. (c) Final permittivity after 133 iterations. (d) Final band structure.

respectively. The results indicate that the final permittivity function in reasonably close to an admissible, two-valued function. By applying the threshold operator $\widetilde{H}_h$ to ${\varepsilon_{r,h}}^{(117)}$, we obtained such a function $\widetilde{\varepsilon}_{r,h}$, which yielded a slightly larger gap width of $\widetilde{w}_{3,h} = 0.1308 \cdot 2\pi c_0/a$.

Figure 9.2(a) shows the evolution of the permittivity function. From left to right and from top to bottom the images in Figure 9.2(a) show the discretized permittivity functions after every tenth iteration, starting with the initial permittivity function. Figure 9.2(b) shows the corresponding evolution of the gap width. The solid squares indicate points on the graph, which correspond to images in Figure 9.2(a).

We remark that the results in this example correspond well to results published by Cox and Dobson (cf. Figure 5.1 in [29]). Using a different generalized gradient method and a different goal functional, Cox and Dobson found a final permittivity function similar to ours, which yielded an approximate gap width of $0.128 \cdot 2\pi c_0/a$. In contrast to our generalized gradient method, the algorithm by Cox and Dobson needed a total of 1620 iterations to converge.

**Example 9.2.** In this example we maximize the band gap between the second and the third TM band of a two-dimensional photonic crystal. The crystal we start off with is composed of a staggered arrangement of cylindrical rods with two different radii. Each rod with a larger radius of $0.2a$ is surrounded by four rods with a smaller radius of $0.1a$. This medium structure is represented by a relative electric permittivity function, which satisfies

$$\varepsilon_r|_\Omega = 1 + 8\chi_{B_{0.2a}(\mathbf{0})} + \sum_{x_1,x_2\in\{-a/2,a/2\}} \chi_{B_{0.1a}(x_1\boldsymbol{e}^{(1)}+x_2\boldsymbol{e}^{(2)})\cap\Omega},$$

where $\boldsymbol{e}^{(1)}$ and $\boldsymbol{e}^{(2)}$ denote the standard basis vectors in $\mathbb{R}^2$. The width of the band gap between the second and third TM band is approximately equal to $w_{2,h} = 0.0732 \cdot 2\pi c_0/a$.

As in the previous example, we chose $\mathcal{I}_h\varepsilon_r|_\Omega$ as the initial free density $\xi^{(0)}$. The corresponding initial gap width was given by ${w_{2,h}}^{(0)} = 0.0634 \cdot 2\pi c_0/a$ (see Figure 9.3(a) and (b)). After 133 iterations Algorithm 7.1 converged to a final permittivity function $\varepsilon_{r,h}^{(133)}$ representing a two-dimensional photonic crystal consisting of a periodic arrangement of identical cylindrical rods. The radius of each rod was approximately equal to $0.125a$. The gap width increased to ${w_{2,h}}^{(133)} = 0.1725 \cdot 2\pi c_0/a$ (see Figure 9.3(c) and (d)). This result corresponds well to a result published by Koa, Osher, and Santosa (cf. Figure 3 in [59]), who used a level set method and a different value for $\varepsilon_{r,\max}$.

The minimal and maximal function values of the final relative permittivity function ${\varepsilon_{r,h}}^{(133)}$ were 1.0000 and 9.0000. Remarkably, the function ${\varepsilon_{r,h}}^{(133)}$ was virtually two-valued in the sense that it took function values in $\Xi(10^{-4})$ at 99.75% of all quadrature points. We applied the threshold operator $\widetilde{H}_h$ to ${\varepsilon_{r,h}}^{(133)}$ and found that the resulting two-valued permittivity function yielded a gap width $\widetilde{w}_{2,h}$, which satisfied $|\widetilde{w}_{2,h} - {w_{2,h}}^{(133)}|/\widetilde{w}_{2,h} < 10^{-6}$.

### 9.1.2 Results of the Level Set Method

Here, we present results obtained by the level set method as given by Algorithm 8.1. The parameters for this algorithm were chosen as $t_{\min} = 0.01$, $t_{\max} = 0.1$, $\Delta_{\min} = 10^{-8}$. Descent direction were chosen by a choice strategy similar to that described in Section 7.6.

**Example 9.3.** In this example we maximize the band gap between the first and the second TM band of a two-dimensional photonic crystal, which is composed of periodically placed, square rods. The cross-section of each square rod has an side length of $0.6a$. The TM band structure of this photonic crystal exhibits a band

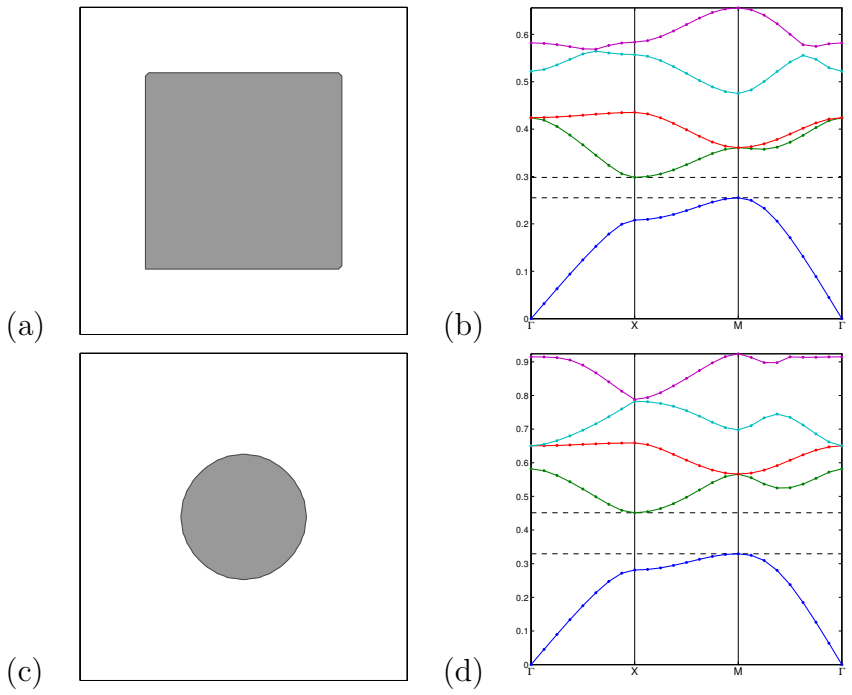



**Figure 9.4:** Results of Example 9.3. (a) Initial permittivity. (b) Initial band structure. (c) Final permittivity after 117 iterations. (d) Final band structure.

gap with a width of approximately $w_{1,h} = 0.0145 \cdot 2\pi c_0/a$ between the first and the second band (see Figure 9.4(a) and (b)).

We represented the crystal's medium structure in the primitive domain $\Omega$ by an inital level set function $\varphi^{(0)}$, which was given by

$$\varphi^{(0)}(\boldsymbol{x}) := 0.3a - |\boldsymbol{x}|_\infty \qquad \text{for all } \boldsymbol{x} \in \Omega.$$

Here, $|\cdot|_\infty$ denotes the surpremum norm in $\mathbb{R}^2$. We remark that $\varphi^{(0)}$ is a signed distance function for the boundary of the rod's cross-section. After only 14 iterations Algorithm 8.1 converged to a final level set function $\varphi^{(14)}$, which represented the medium structure of a photonic crystal consisting of a periodic arrangement of cylindrical rods. The radius of each rod was approximately equal to $0.19a$. We computed an approximate gap width of $w_{1,h}^{(14)} = 0.1219 \cdot 2\pi c_0/a$ for this final medium structure (see Figure 9.4(c) and (d)). A similar optimal medium structure was also found by Kao, Osher, and Yablonovitch, though for a different relative electric permittivity of the rods (cf. Figure 2 in [44]).

In Figure 9.5(a) we depicted the permittivity functions after every second iteration of Algorithm 8.1. Figure 9.5 shows the evolution of the gap width. Each solid square corresponds to a permittivity function in Figure 9.5(a). Notice how

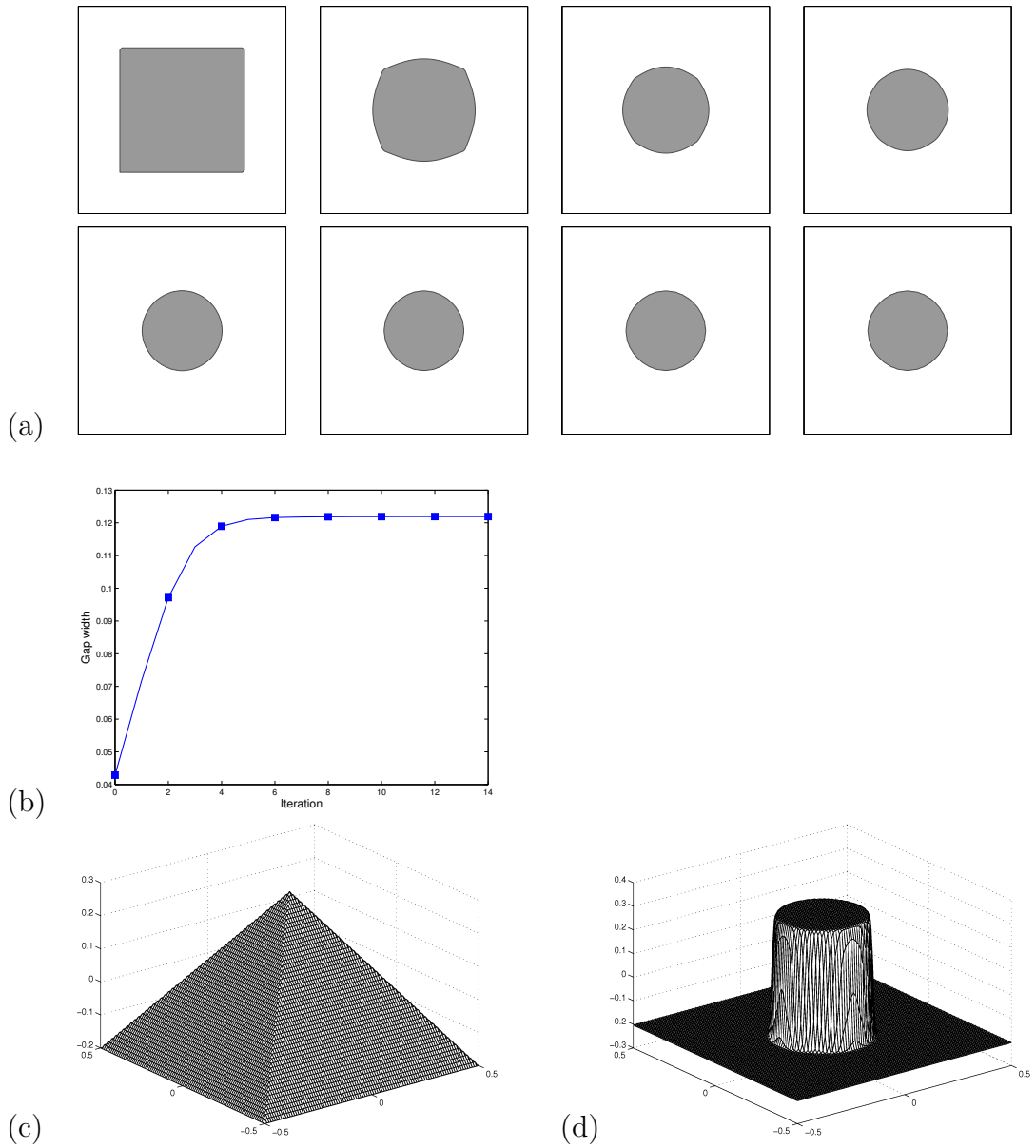



**Figure 9.5:** The evolution of the permittivity function (a) and the corresponding gap width (b) in Example 9.3.

fast the level set algorithm changes the rod's cross-section to a circle-like shape in the first few iterations. We compared the performance of Algorithm 8.1 to that of Algorithm 7.1, the generalized gradient method, and found that the latter needed 141 iterations to converge to the same final permittivity function.

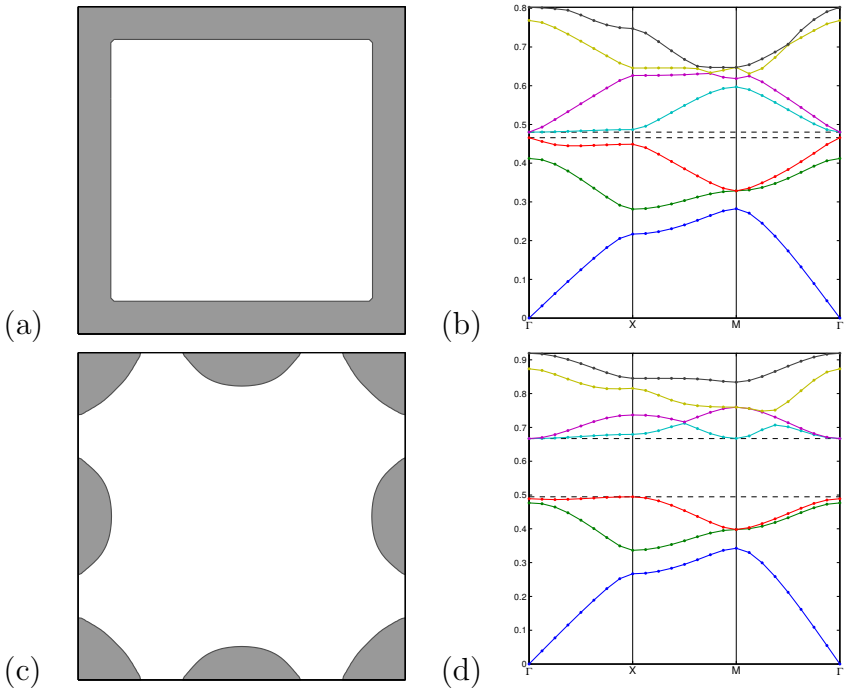



**Figure 9.6:** Results of Example 9.4. (a) Initial permittivity. (b) Initial band structure. (c) Final permittivity after 6 iterations. (d) Final band structure.

**Example 9.4.** In this example we start off with a photonic crystal, whose medium structure is given by a square lattice framework. The width of the framework is given by $0.2a$. The crystal's TM band structure exhibits a band gap between the third and the fourth band with an approximate gap width of $w_{3,h} = 0.0145 \cdot 2\pi c_0/a$ (see Figure 9.6(a) and (b)). We represented the medium structure by an inital level set function $\varphi^{(0)}$, given by

$$\varphi^{(0)}(\boldsymbol{x}) := |\boldsymbol{x}|_\infty - 0.4a \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^2.$$

Only 6 iterations were needed to widen the band gap to an approximate gap width of ${w_{3,h}}^{(6)} = 0.1723 \cdot 2\pi c_0/a$ (see Figure 9.6(c) and (d)). This gap width is even larger than the one found in Example 9.1. We remark that our final medium structure corresponds well to optimal structures found by Kao, Osher, and Yablonovitch (cf. Figure 4 in [44]), as well as by Cox and Dobson (cf. Figure 3 in [29]).

**Example 9.5.** In this example we demonstrate that the level set method given by Algorithm 8.1 can also be used to open band gaps between separated photonic

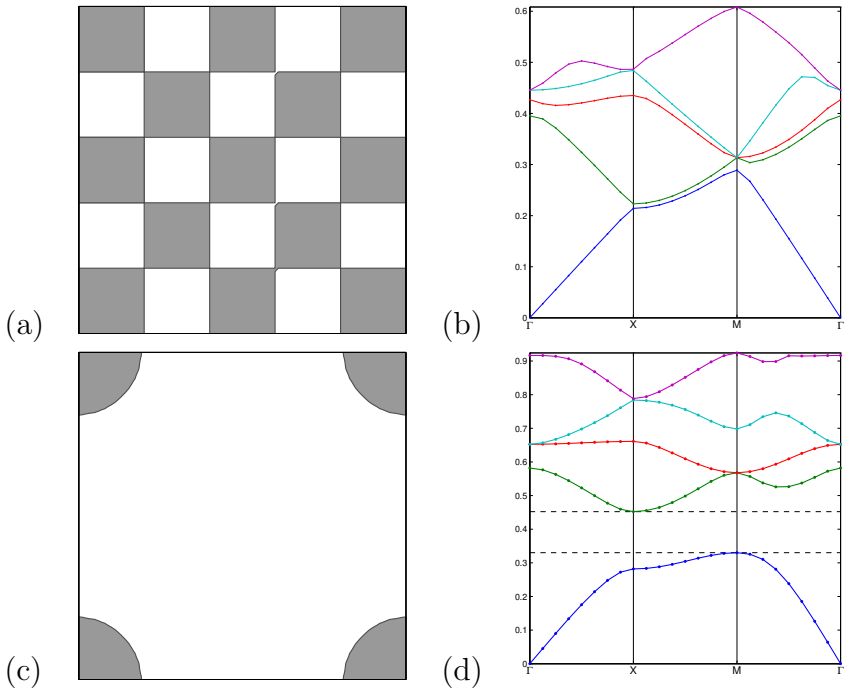



**Figure 9.7:** Results of Example 9.5. (a) Initial permittivity. (b) Initial band structure. (c) Final permittivity after 17 iterations. (d) Final band structure.

bands. By separated we mean that the bands do not intersect in any point. We considered a two-dimensional photonic crystal with a checkerboard-like medium structure. We chose the computational domain $\Omega$, such that it contained a $5 \times 5$ portion of the checkerboard pattern (see Figure 9.7(a)). The TM band structure of the photonic crystal does not exhibit any band gap betwenn the first 5 bands. One notices, however, that the first and the second band are separated (see Figure 9.7). We therefore applied Algorithm 8.1 with the gap width functional $J_1^{\mathrm{TM}}$. We represented the initial medium structure by a level set function $\varphi^{(0)}$, which was given by

$$\varphi^{(0)}(\boldsymbol{x}) := \cos\left(\frac{5\pi \boldsymbol{x}_1}{a}\right)\cos\left(\frac{5\pi \boldsymbol{x}_2}{a}\right) \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^2.$$

In the course of 17 iterations Algorithm 8.1 opened up a band gap between the first and the second TM band. The final level set function $\varepsilon^{(18)}$ represented the same optimized medium structure as in Example 9.3. It is therefore not surprising the final width of the band gap was approximately equal to ${w_{1,h}}^{(17)} = 0.1219 \cdot 2\pi c_0/a$.

In Figure 9.8 we depicted the change of the crystal's permittivity function during the first 7 iterations of the level set method. As one can see, the level set method had no trouble changing the medium structure's topology. The final

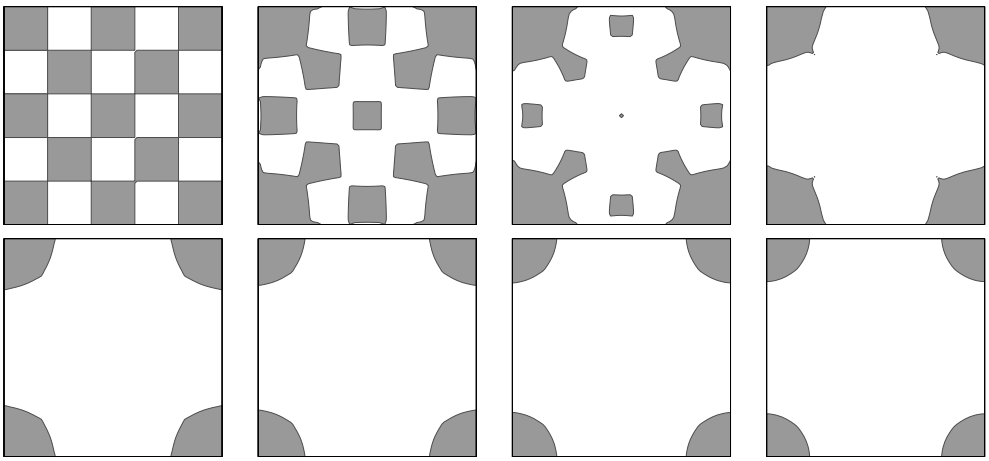


**Figure 9.8:** The evolution of the permittivity over the first 7 iterations in Example 9.5.

topology is already attained after the fourth iteration.

## 9.2 Maximizing TE Band Gaps

Here, we present numerical solutions of photonic band gap maximization problems involving the TE band structures of two-dimensional photonic crystals. Each such problem is given by a minimization problem

$$\underset{\rho\in\mathcal{C}^{\mathrm{TE}}}{\text{minimize}}\ J_j^{\mathrm{TE}}(\rho),$$

where $J_j^{\mathrm{TE}}$ denotes the TM band gap functional defined by (5.14) for a given index $j \in \mathbb{N}$. In the following experiments the same type of photonic crystals were considered. With the two permittivities $\varepsilon_{r,\min} = 1$ and $\varepsilon_{r,\max} = 9$, the admissible is given by

$$\mathcal{C}^{\mathrm{TE}} := \left\{ \rho \in \mathcal{E} \;\middle|\; \rho = \frac{1}{9} + \frac{8}{9}\chi_S,\ S \in \mathcal{M}(\Omega) \right\}.$$

The problems were discretized as described in the previous section.

### 9.2.1 Results of the Generalized Gradient Method

The results in the following examples were obtain by the Algorithm 7.1. The algorithm parameters where chosen as $s_{\min} = 10^{-6}$, $s_{\max} = 10^{-1}$, and $\Delta_{\min} = 10^{-8}$.

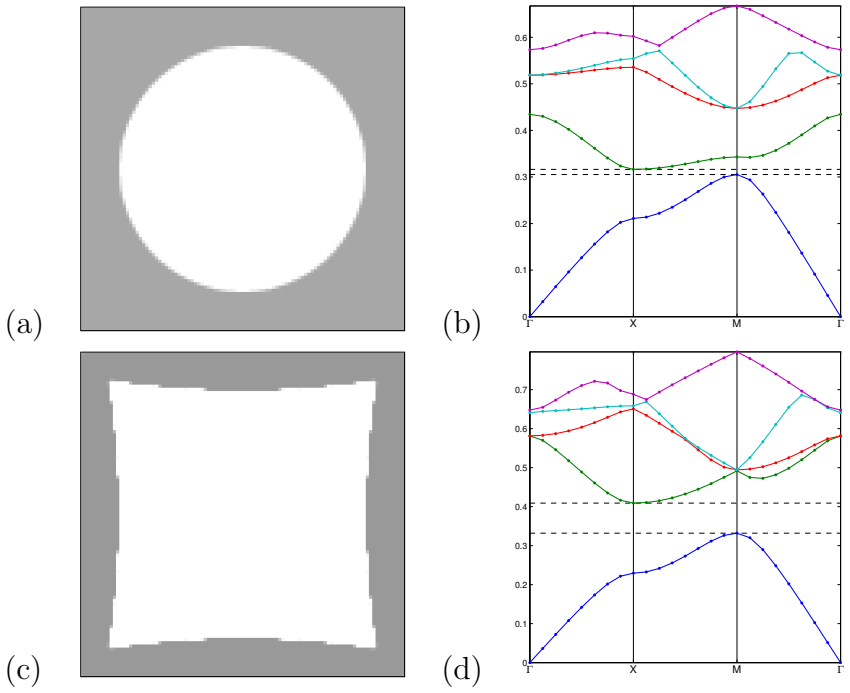



**Figure 9.9:** Results of Example 9.6. (a) Initial permittivity. (b) Initial band structure. (c) Final permittivity after 58 iterations. (d) Final band structure.

The funnel operator $Y_h$ was given as defined by (7.15) and (7.20) with $r = 4$. When applied to an admissible density function in $\mathcal{C}^{\mathrm{TE}}$, this funnel operator yields a two-valued density function, which represents a two-value permittivity function with function values of approximately 7.8679 and 1.0162. The descent directions where chosen as described in Section 7.6.

**Example 9.6.** In this example we maximize the band gap between the first and second TE band of a two-dimensional photonic crystal consisting of a periodic arrangement of cylindrical holes drilled into a material with relative electric permittivity $\varepsilon_{r,\max} = 9$. The medium structure is represented by the relative electric permittivity function $\varepsilon_r$, which satisfies

$$\varepsilon_r|_\Omega = 9 - 8B_{0.38a}(\mathbf{0}),$$

where $a$ denotes the lattice constant, and we computed an approximate gap width of $w_{1,h} = 0.0158 \cdot 2\pi c_0/a$ for this medium structure.

Choosing $\xi_h^{(0)} := \mathcal{I}_h(1/\varepsilon_r|_\Omega)$, the approximate initial gap width was given by ${w_{1,h}}^{(0)} = 0.0110 \cdot 2\pi c_0/a$ (see Figure 9.9(a) and (b)). Over the course of 58 iterations the gap width increased to a value of ${w_{1,h}}^{(58)} = 0.0771 \cdot 2\pi c_0/a$. The final

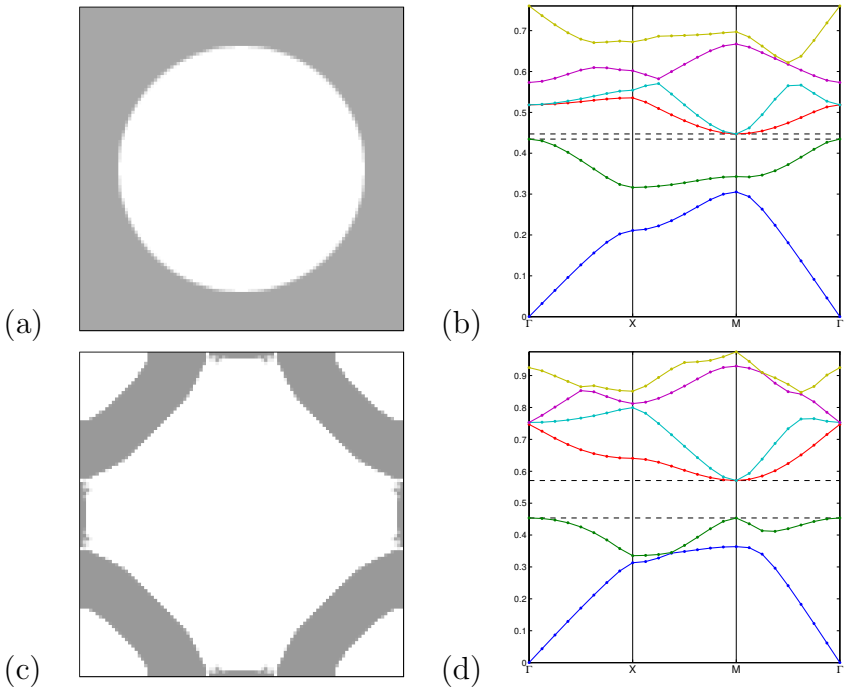



**Figure 9.10:** Results of Example 9.7. (a) Initial permittivity. (b) Initial band structure. (c) Final permittivity after 87 iterations. (d) Final band structure.

discrete permittivity ${\varepsilon_{r,h}}^{(58)}$ and TE band structure are depicted in Figure 9.9(c) and (d). The maximal and minimal function values of the final discrete permittivity ${\varepsilon_{r,h}}^{(58)}$ were equal to 9 and 1, respectively. We also found that the range of ${\varepsilon_{r,h}}^{(58)}$ was a subset of $\Xi(10^{-5})$, and that ${\varepsilon_{r,h}}^{(58)}$ took values in $\Xi(10^{-6})$ at 99.9% of all quadrature points. Compared to the Examples 1 and 2, this result is significantly better. Moreover, we found that $|\widetilde{w}_{1,h} - {w_{1,h}}^{(58)}|/\widetilde{w}_1 < 10^{-9}$, where $\widetilde{w}_1$ denotes the approximate gap width for the two-valued relative electric permittivity $\widetilde{\varepsilon}_{r,h} = \widetilde{H}_h({\varepsilon_{r,h}}^{(58)})$.

**Example 9.7.** As can be seen in Figure 9.9, the initial TE band structure of the photonic crystal studied in Example 9.6 also exhibits a band gap between the second and third band with an approximate gap width of $w_{2,h} = 0.0243 \cdot 2\pi c_0/a$. By initializing Algorithm 7.1 exactly as in Example 9.6, we obtain a reduced initial gap width of ${w_{2,h}}^{(0)} = 0.0126 \cdot 2\pi c_0/a$ (see Figure 9.10(a) and (b)). After only 87 iterations the algorithm converged to a final discrete permittivity ${\varepsilon_{r,h}}^{(87)}$, which yielded a gap width of ${w_{2,h}}^{(87)} = 0.1174 \cdot 2\pi c_0/a$ (see Figure 9.10(c) and (d)). We found the maximal and minimal function values of ${\varepsilon_{r,h}}^{(87)}$ to be 9 and 1, respectively. Moreover, ${\varepsilon_{r,h}}^{(87)}$ took function values in $\Xi(10^{-4})$ at 99.9% of

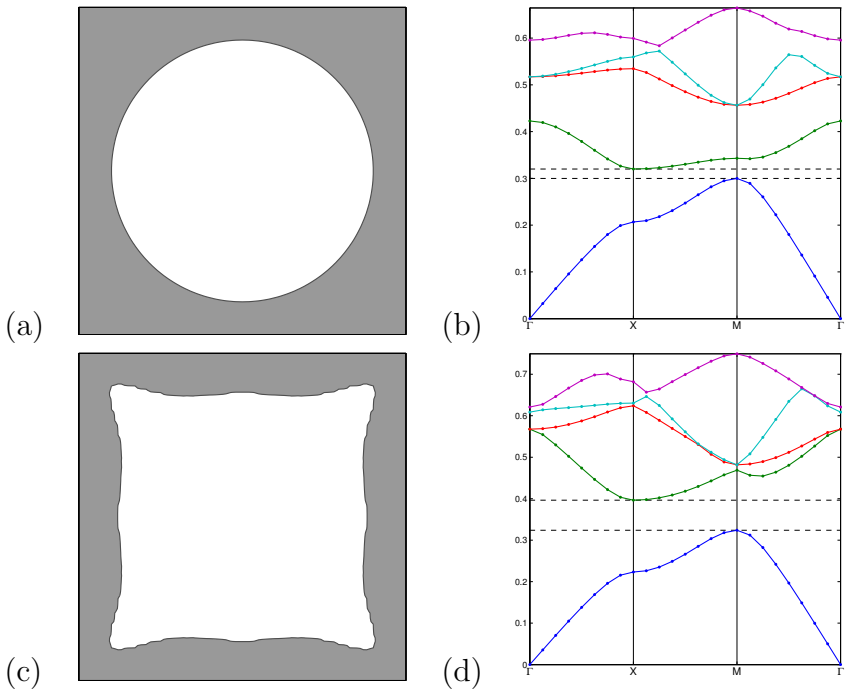



**Figure 9.11:** Results of Example 9.8. (a) Initial permittivity. (b) Initial band structure. (c) Final permittivity after 4 iterations. (d) Final band structure.

all quadrature points. In view of these results, one is surprised to find that the two-valued permittivity $\widetilde{\varepsilon}_{r,h} = \widetilde{H}_h(\varepsilon_{r,h}{}^{(87)})$ yields a smaller gap width of $\widetilde{w}_2 = 0.1170 \cdot 2\pi c_0/a$. We attribute this rather large deviation to fact that the final photonic crystal structure in some parts shows cusp-like medium interfaces, which cannot be resolved properly by a uniform finite element mesh.

### 9.2.2 Results of the Level Set Method

Here we also present some results obtained by Algorithm 8.1. The algorithm parameters were chosen as $t_{\min} = 0.01$, $t_{\max} = 0.1$, $\Delta_{\min} = 10^{-8}$. Descent directions were chosen by a choice strategy similar to that described in Section 7.6.

**Example 9.8.** In our first example we start off with an initial photonic crystal which is similar to that in Example 9.6. The medium structure of this crystal is described by the initial level set function

$$\varphi^{(0)}(\boldsymbol{x}) := |\boldsymbol{x}| - 0.4a,$$

where $a$ denotes the lattice constant. Our aim is to maximize the band gap between the first and the second TE band. We computed an approximate initial gap

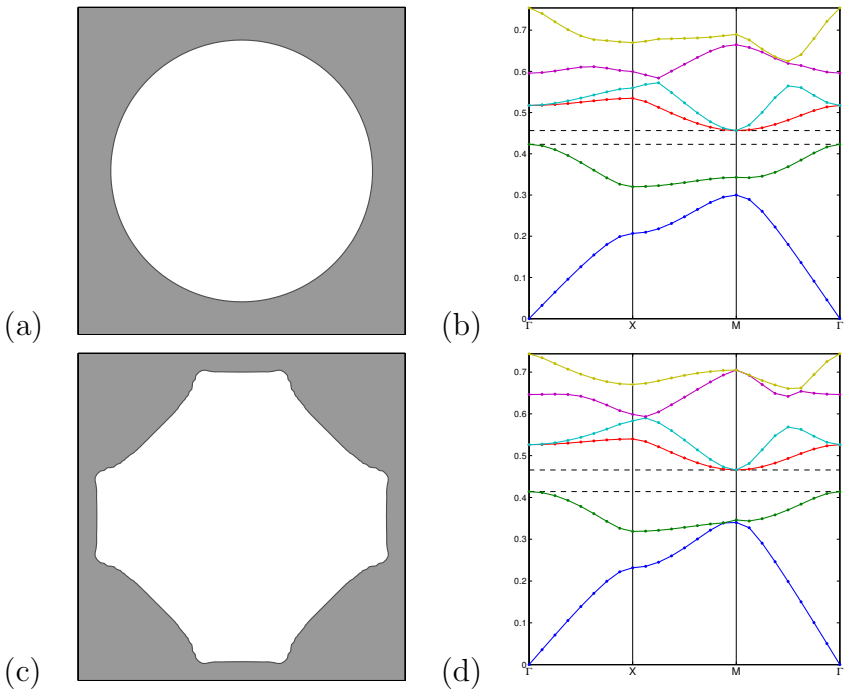



**Figure 9.12:** Results of Example 9.9. (a) Initial permittivity. (b) Initial band structure. (c) Final permittivity after 4 iterations. (d) Final band structure.

width of ${w_{1,h}}^{(0)} = 0.0201 \cdot 2\pi c_0/a$ (see Figure 9.11(a) and (b)). The level set method stopped after only 4 iterations yielding a band gap on ${w_{1,h}}^{(4)} = 0.0726 \cdot 2\pi c_0/a$ (see Figure 9.11(c) and (d)) . When we compare this result to the result found by the generalized gradient method in Example 9.6, where a gapwidth of $0.0771 \cdot 2\pi c_0/a$ was achieved, we are lead to conclude that the level set method stopped prematurely. By comparing the final medium structures, we notice that the optimal structure exhibits cusp-shaped material interfaces. We belief that these interfaces present a problem to the level set method.

**Example 9.9.** In this last example of a band gap maximization problem for two dimensional crystals, we start with the same initial photonic crystal as in the previous example, but aim at maximizing the band gap between the second and the third TE band. The initial gap width is given by ${w_{2,h}}^{(0)} = 0.0334 \cdot 2\pi c_0/a$ (see Figure 9.12(a) and (b)). As in the previous example the level set method stopped after 4 iterations failing to converge to an optimal solution. The gap is only widened to a final gap width of ${w_{2,h}}^{(4)} = 0.0517 \cdot 2\pi c_0/a$ (see Figure 9.12(c) and (d)). Compared to the result in Example 9.7, where a final gap width of $0.1174 \cdot 2\pi c_0/a$ could be achieved, the result of the level set method is far from

optimal. When comparing the final medium structures, one notices that they obviously have different topologies. Apparenty, the level set method was not able to change the medium structure's topology and thus converged to a less optimal result.

## 9.3 Maximizing Band Gaps of a 3D Photonic Crystal

In this section we present numerical solutions of photonic band gap maximization problems for three-dimensional photonic crystals. The results we present here are novel and, to the best of our knowledge, no similar results have been published yet. The results present in this section were obtained by the generalized gradient method as given by Algorithm 7.1.

The band gap maximization problems in the examples below are given by the minimization problem

$$\underset{\rho\in\mathcal{C}}{\text{minimize}}\ J_2(\rho),$$

i.e., we aim at maximizing the band gap between the second and the third photonic bands. In both examples the initial medium structure of the photonic crystal is given by a scaffold-like structure. The crystallographic space group of this structure is isomorphic to the semidirect product $\mathbb{Z}^3 \rtimes \mathrm{O}_3(\mathbb{Z})$. The crystal's Bravais lattice is the simple cubic lattice $a\mathbb{Z}^3$, where $a$ denotes the lattice constant. The width of the scaffold beams is given by $0.25a$.

The relative electric permittivity of the beams was given by $\varepsilon_{r,\max} = 13$, that of the ambient medium by $\varepsilon_{r,\min} = 1$. The admissible set was hence given by

$$\mathcal{C} := \left\{ \rho \in \mathcal{E} \;\middle|\; \rho = \frac{1}{13} + \frac{12}{13}\chi_S,\ S \in \mathcal{M}(\Omega) \right\}.$$

The problems were discretized by an $\boldsymbol{H}(\mathrm{curl})$-conformating finite element ansatz on a hexahedral mesh. We used a quadrature rule with 9 quadrature points on each element. The finite element discretization was implemented in C++ using the parallel finite element library M++ developed by Wieners (cf. [76], [77]). The approximate band structures were computed using a parallel, iterative eigenvalue solver developed by A. Buloyatov (cf. [17]). We used the C++ library lp_solve to solve the linear programs arising from the choice strategy for the descent directions (see Section 7.6).

**Example 9.10.** In this first example the primitive domain $\Omega$ was discretized by a $32 \times 32 \times 32$ mesh. We computed an initial gap width of ${w_{2,h}}^{(0)} = 0.0953 \cdot 2\pi c_0/a$ (see Figure 9.13(a) and (b)). The generalized gradient method stopped after 26

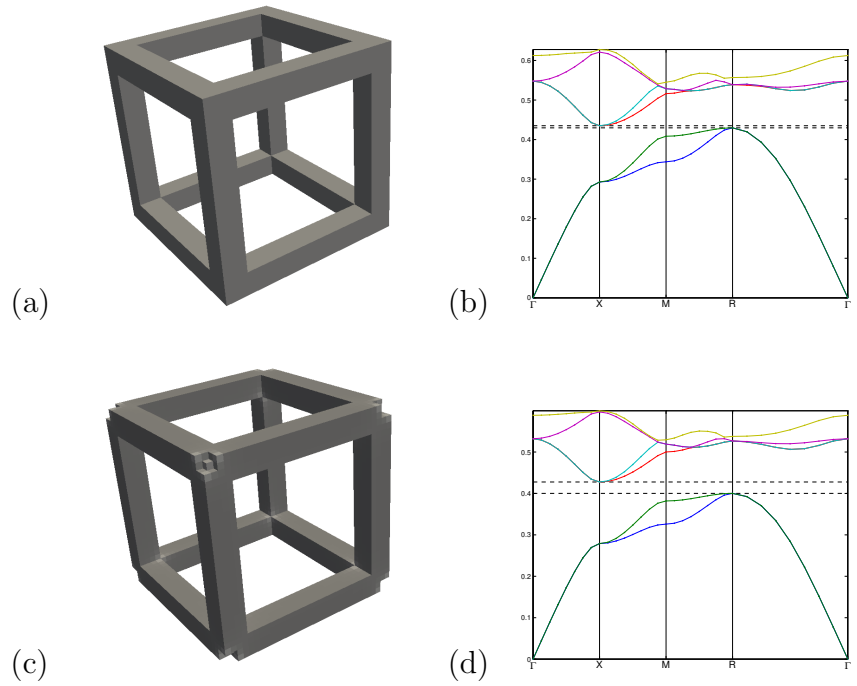



**Figure 9.13:** Results of Example 9.10. (a) Initial permittivity. (b) Initial band structure. (c) Final permittivity after 26 iterations. (d) Final band structure.

iterations. The final gap width was given by ${w_{2,h}}^{(26)} = 0.1721 \cdot 2\pi c_0/a$. The minimal and maximal function values of the final discretized relative electric permittivity ${\varepsilon_{r,h}}^{(26)}$ were given by 1.0137 and 12.8821. However, ${\varepsilon_{r,h}}^{(26)}$ was not not close to a two-valued, admissible permittivity function.

**Example 9.11.** In this second example the primitive domain $\Omega$ was discretized by a $64 \times 64 \times 64$ mesh. We computed an approximate initial gap width of ${w_{2,h}}^{(0)} = 0.0907 \cdot 2\pi c_0/a$ (see Figure 9.14(a) and (b)). After 32 iterations, the generalized gradient method had widened the band gap to a final width of ${w_{2,h}}^{(0)} = 0.1674 \cdot 2\pi c_0/a$. The minimal and maximal function values of the final discretized relative electric permittivity ${\varepsilon_{r,h}}^{(32)}$ were given by 1.0137 and 12.8794. As in the previous example, however, ${\varepsilon_{r,h}}^{(32)}$ was not close to an admissible permittivity function.

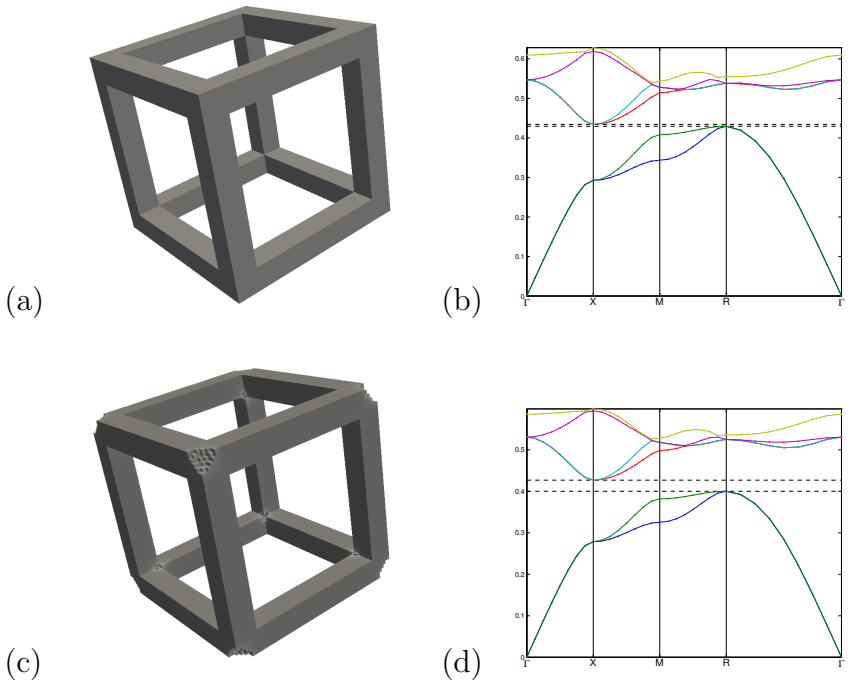



**Figure 9.14:** Results of Example 9.11. (a) Initial permittivity. (b) Initial band structure. (c) Final permittivity after 32 iterations. (d) Final band structure.

# Chapter 10

# Summary, Conclusions and Outlook

## 10.1 Summary

In this dissertation we studied mathematical aspects related to photonic band structure optimization problems and developed optimization algorithms for these problems. The performance of the optimization algorithms was demonstrated by several numerical experiments, included a band gap maximization problem for a three-dimensional photonic crystal.

Based on Maxwell's equations and the Bloch ansatz we derived a family of constrained eigenvalue problems, which determine the existence of so-called Bloch modes in a three-dimensional photonic crystal. Each eigenvalue problem in this family is identified by the so-called quasimomentum vector of the corresponding Bloch modes. The medium structure of the photonic crystal is represented by a coefficient in the eigenvalue equations, which is typically discontinuous. For this reason, we established a weak formulation of the eigenvalue problems. This weak formulation was posed for functions, which belong to certain Sobolev spaces, and which satisfy so-called periodic boundary conditions. The corresponding spectral theory showed that for every photonic crystal and for every quasimomentum vector there exists an increasing, unbounded sequence of non-negative, real eigenvalues. The sequence of corresponding eigenfunctions forms a complete system of the underlying Sobolev space, which is orthogonal in the $L^2$-sense.

Given a specific quasimomentum vector, each eigensolution of the corresponding eigenvalue problem determines a Bloch mode, which is able to propagate in the photonic crystal. The eigenvalue determines the phase frequency of such a Bloch mode. We showed that the Bloch mode frequencies depend continuously on the quasimomentum vector. The graphs of the Bloch mode frequencies with respect to

the quasimomentum vector are called the photonic bands of the photonic crystals. The union of all photonic bands is referred to as the photonic band structure of the photonic crystal. We discussed how symmetries in the crystal's medium structure relate to symmetries in its band structure. For certain medium structures, the band structure can exhibit so-called band gaps. In essence, a band gap is a range of frequencies with the property that there exists no Bloch mode, whose phase frequency is in that range.

For certain applications it is desirable that the width of a band gap is large. From this the question arises, how the medium structure of a photonic crystal should be designed, such that its band structure exhibits a band gap with maximal width. The emerging optimization problem is called a photonic band gap maximization problem (PBGMP). More generally, one can formulate so-called photonic band structure optimization problems (PBSOPs), which consist in finding optimal medium structures of a photonic crystal in view of certain features of its band structure. We devised a formal setting for a certain class of PBSOPs. This setting allowed us to treat PBSOPs as minimization problems over an admissible set of two-valued functions. Each function of this admissible set represents a specific medium structure of a photonic crystal. The PBSOP is characterized by a goal functional, which is to be minimized. We showed that the goal functionals are locally Lipschitz continuous. The existence of optimal solutions could only be proved for PBSOPs involving so-called TM band structures of two-dimensional photonic crystals.

In order to devise optimization algorithms for PBSOPs we reviewed the concept of generalized differentials, which extend the concept of differentiability to functionals, which are only locally Lipschitz continuous. By employing known results from nonsmooth analysis, we were able to establish inclusion relations for the generalized differentials of those goal functionals, which define PBGMPs. We were also able to identify supersets of the corresponding generalized gradients.

After commenting on suitable discretization schemes for PBSOPs we developed an iterative optimization algorithm based on the generalized gradients of the discretized goal functionals. The algorithm was designed such that certain optimization constraints, which are inherent to all PBSOPs, are satisfied in each iteration. This was achieved by solving a modified minimization problem on an unconstrained, open admissible set. The performance of the algorithm was demonstrated by several numerical experiments on PBGMPs for two- and three-dimensional photonic crystals. We described a second optimization method, which is based on a level set method. In contrast to the generalized gradient algorithm, the level set algorithm is guaranteed to converge to an admissible, optimal medium structure. We also conducted numerical experiments with the level set algorithm, though only on PBGMPs for two-dimensional photonic crystals.

## 10.2 Comparison of the Optimization Algorithms

In Chapter 7 and Chapter 8 we developed two optimization algorithms for photonic band structure optimization problems (PBSOPs). One algorithm was based on a generalized gradient method, the other algorithm was based on a level set method. The performance of both algorithms was demonstrated by several numerical experiments on photonic band gap maximization problems (PBGMPs) in Chapter 9. Both algorithms were capable to find optimal medium structures, which maximized certain band gaps in the TM and TE band structures of two-dimensional photonic crystals. The generalized gradient method could also maximize a band gap in the band structure of a three-dimensional photonic crystal. In our experiments we found that the level set method has some significant advantages over the generalized gradient method. One advantage is, that the level set method is by construction guaranteed to converge to a two-valued, optimal solution. In general, the generalized gradient method does not converge to a two-valued solution. Nevertheless, we observed in many experiments that this method converged to an almost two-valued solution. The second clear advantage of the level set method is, that it usually needs less iterations than the generalized gradient method in order to converge to an optimal solution. The computation time needed for a single iteration is about the same for both algorithms. Although level set methods generally cannot be expected to change to topology of a medium structure, we observed that topology changes did occur in some examples. We noticed, however, that the level set method was not able to converge to medium structures with cusp-shaped material interfaces. In that respect, the generalized gradient method performed slightly better.

## 10.3 Open Problems

In this work we already mentioned some problems, which are still open to the best of our knowledge. Let us briefly review these problems. In Section 5.3, we showed that photonic band structure optimization problems (PBSOPs) involving the TM band structure of a two-dimensional photonic crystal admit an optimal solution within an extended admissible set of functions, which are essentially bounded from above and below. Our numerical experiments suggest that the optimal solutions of photonic band gap maximization problems (PBGMPs) are indeed two-valued functions. However, we could not prove that this is generally the case. We mention here the works of Cox and MacLaughlin [31] and [32], which might be helpful to establish a proof, that optimal solutions of TM PBGMPs are always attained for two-valued relative electric permittivity functions.

For the TE setting, as well as in for three-dimensional setting the existence of optimal solutions could not be proved, either. As we discussed in Section 5.4, one would have to consider the optimization problems on admissible sets, which also contain anisotropic density functions, in order to prove the existence of optimal solutions by standard homogenization methods. The concepts of H-convergence and G-compactness would be central to such a proof. However, we could find conclusive evidence in the literature that these concepts can be transferred to PBSOPs. The question as to whether or not there exist isotropic or two-valued optimal solutions also remained unanswered so far.

Furthermore, we could not verify that the concept of generalized gradient methods can be extended to infinite-dimensional Hilbert spaces, or even to certain Banach spaces such as $L^\infty$-spaces. The numerical algorithms developed in Chapter 7 mainly rely on the fact that the optimization problems are, in essence, posed on $\mathbb{R}^N$ for some $N \in \mathbb{N}$. This is due to the discretization of PBSOPs. Notice however, that the discretization acts as a regularization operation. Therefore, the generalized gradient method in Chapter 7 only solves regularized versions of PBSOPs. We do not know to what extent the solution of a regularized PBSOP may differ from the solution of an unregularized PBSOP.

## 10.4 Final Remarks and Outlook

One of the aims of this dissertation was to solve photonic band gap maximization problems for three-dimensional photonic crystal. As was shown by the numerical experiments presented in Section 9.3, the generalized gradient method is able to widen the band gap in a three-dimensional photonic crystal. The main problem with this method is, however, that it does not converge to two-valued density functions in general. In contrast to this, the level set method is guaranteed to converge to two-valued solutions. Therefore, we plan to implement and test this method also for three-dimensional photonic crystals.

# Appendix A

# A FEM Toolbox for MATLAB

In the course of our research project we developed a finite element toolbox for MATLAB. This toolbox, which we shall refer to as the *FEM Toolbox* hereafter, was used to discretize all photonic band gap maximization problems (PBGMPs) involving TM and TE band structures of two-dimensional photonic crystals. In this appendix chapter we give a brief overview over this toolbox. In Section A.1 we comment on general aspects related to the FEM Toolbox, such as design principles and intended use. In Section A.2 we briefly describe some fundamental data structures and algorithms which are provided by the FEM Toolbox. In Section A.3 we present some features of the FEM Toolbox, which were designed specifically for the numerical solution of PBGMPs. The chapter is completed with an implementation example in Section A.4.

## A.1 General Remarks

The FEM Toolbox was developed with the intent to provide an easy-to-use and easy-to-adapt MATLAB function library, which can be used to implement finite element methods in one and two space dimensions. The FEM Toolbox can be used either as a tool for numerical computations or as a basic framework for research and education related to finite element methods. Since MATLAB provides a wide range of standard numerical algorithms and advanced visualization capabilities, we chose to implement the FEM Toolbox entirely in the MATLAB programming language. Wherever possible, mathematical entities are represented by MATLAB arrays or by MATLAB structure arrays. Object-oriented programming concepts, such as MATLAB classes, are not used due to their slower performance.

The development of the FEM Toolbox was mostly motivated by our research on photonic crystals. In order to compute band structures of one and two-dimensional photonic crystals, we needed a finite element software, which supported $H^1$-con-

forming one-dimensional and preferably quadrilateral elements, and which could handle periodic boundary conditions. We also required that the software could easily be adapted to the algorithmic requirements of band structure optimization algorithms.

In its current version the FEM Toolbox is capable to discretize boundary values problems, which are of the form

$$\begin{cases} -\nabla \cdot (\boldsymbol{\kappa} \nabla u) + \boldsymbol{\gamma} \cdot \nabla u + \mu u = f & \text{in } \Omega, \\ u - \sigma(u \circ \boldsymbol{\iota}) = r & \text{on } \Gamma_0, \\ -\boldsymbol{\nu} \cdot (\boldsymbol{\kappa} \nabla u) + \eta u = g & \text{on } \Gamma_1 \end{cases} \tag{A.1}$$

for the unknown function $u \in H^1(\Omega)$. Here, $\Omega$ denotes a computational domain in $\mathbb{R}^n$, where $n \in \{1, 2\}$, which is bounded by a Lipschitz smooth curve $\Gamma = \partial\Omega$. The boundary $\Gamma$ is partitioned into two part $\Gamma_0$ and $\Gamma_1$, where either boundary part can be empty. In the above system of equations, $\boldsymbol{\nu}$ denotes the outer unit normal field on $\Gamma_1$. The mapping $\boldsymbol{\iota} : \Gamma_0 \to \Gamma_0$ identifies certain points on $\Gamma_0$.

In the partial differential equation $\boldsymbol{\kappa} : \Omega \to \mathbb{R}^{n \times n}$, $\boldsymbol{\gamma} : \Omega \to \mathbb{R}^n$ and $\mu : \Omega \to \mathbb{R}$ are given coefficients. The right-hand side of the equation is given by a function $f : \Omega \to \mathbb{R}$. The equation posed on $\Gamma_0$ is referred to as the *essential boundary condition* of the problem. Depending on the choice of $\sigma : \Gamma_0 \to \mathbb{R}$ and $r : \Gamma_0 \to \mathbb{R}$, the essential boundary condition is a Dirichlet-type condition (for $\sigma = 0$), periodic boundary condition (for $\sigma = 1$ and $r = 0$) or a quasi-periodic boundary condition (for $r = 0$), provided that $\boldsymbol{\iota}$ is chosen appropriately. The equation posed on $\Gamma_1$ is referred to as the *natural boundary condition* of the problem. Depending on the choice of $\eta : \Gamma_1 \to \mathbb{R}$ and $g : \Gamma_1 \to \mathbb{R}$, this condition is a Neumann-type condition (for $\eta = 0$) or a Robin-type condition (for $\eta \neq 0$). In the next section we briefly describe, how the different entities are represented by the FEM Toolbox, and how the discretization of a boundary value problem of the above form (A.1) is realized. We remark that the FEM Toolbox was solely used to produce the numerical results presented in Chapter 9 in this work, as well as the numerical results presented in [36].

It should be noted that MATLAB's *Partial Differential Equations Toolbox* also provides data structures and algorithms for the implementation of finite element methods. In particular, this toolbox provides sophisticated concepts to describe the geometry of computational domains, as well as a mesh generator. Moreover, it features a graphical user interface which can be used to define and solve various types of elliptic boundary value problems. The reason why we did not use this toolbox was that it only provides data structures and algorithms for $H^1$-conforming, linear finite elements on triangular meshes. Furthermore, the Partial Differential Equations Toolbox is not able to handle periodic boundary conditions.

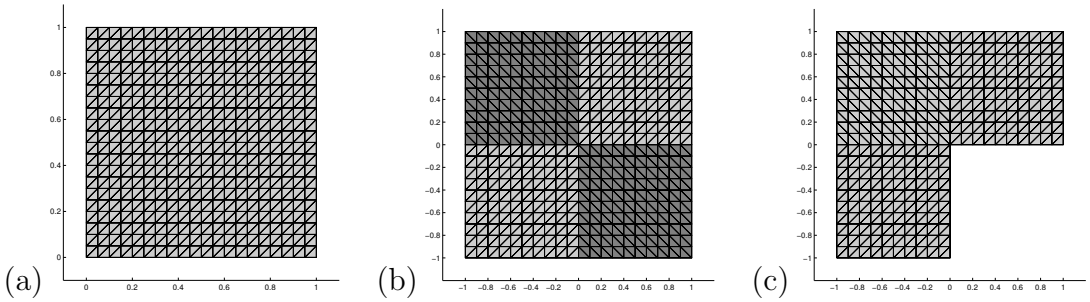



**Figure A.1:** Finite element meshes returned by the functions `femsquarem` (a), `femcheckerm` (b), and `femlshapem` (c). Different shades of gray indicate different subdomains.

## A.2 Data Structures and Algorithms

In this section we describe some important data structures and algorithms, which are provided by the FEM Toolbox. The section does not give a complete account of the toolbox's functionalities and only introduces the most basic concepts.

The fundamental data structure of the FEM Toolbox is a MATLAB structure array, which describes the mesh of a finite element discretization. This *mesh structure* is capable to represent one-dimensional, triangular, quadrilateral or hybrid, triangular-quadrilateral meshes. The mesh structure stores information about the mesh's vertices, edges and faces. In addition to that, the *mesh structure* also stores information about boundary vertices and boundary edges. In particular, the mesh structure stores information about boundary identifications, which were represented by the mapping $\boldsymbol{\iota}$ in the previous section. Furthermore, the mesh structure stores a *subdomain index* for every mesh element, and a *boundary part index* for every boundary edge or vertex. With this, a computational domain can be subdivided into several *subdomains*, and its boundary can be subdivided into several *boundary parts*.

The FEM Toolbox provides a collection of MATLAB functions that create mesh structures representing specific finite element meshes. Here, we only mention the functions `femintervm`, `femsquarem`, `femcheckerm` and `femlshapem`, which create mesh structures that represent finite element meshes on an interval, on a square, on a $2 \times 2$ checkerboard, and on an L-shaped domain, respectively (see Figure A.1). The different "fields" of the checkerboard are distinguished by the subdomain indices mentioned above. As a convention, all MATLAB functions of the FEM Toolbox begin with the letters "`fem`". The FEM Toolbox provides a number of functions, which can be used to retrieve informations about a given mesh. Given a mesh structure `msh`, the function call `femnelems(msh)`, for example, re-

turns the number of elements in mesh. Other functions, such as `femnsubdoms`, `femnbparts`, or `femhmax` return the number of different subdomains, the number of different boundary parts, or the maximal element diameter in a mesh. One-dimensional and triangular meshes can also be refined, either globally or locally, by the function `femrefine`.

Finite element spaces are represented by so-called *space structures*. A space structure is a MATLAB structure array, which stores the degrees of freedoms on every *node* in a finite element mesh. Here, the term "node" refers to a vertex, an edge, or a face in a finite element mesh. The space structure also stores informations about the shape functions of the finite element space. In the current version of the FEM Toolbox, space structures are used to represent higher-order $H^1$- and $L^2$-conforming finite element spaces. We implemented Lobatto-type finitie element shape functions, as defined in [73].

Finite element spaces can be created on given meshes by specific MATLAB functions, which are provided by the FEM Toolbox. Given a mesh structure `msh`, the function call `femh1(msh)`, for example creates a space structure, which represents a linear, $H^1$-conforming finite element space on the mesh, which is represented by `msh`. Other MATLAB functions provided by the FEM Toolbox retrieve informations about finite element spaces. Given a space structure `spc`, the function call `femndofs(spc)`, for example, returns the number of degrees of freedom in the finite element space represented by `spc`. Finite element spaces can be h-refined or p-refined by the functions `femhrefine`, `femprefine`. Both function can be used to perform either global or local refinements. The function `feminterp` interpolates a MATLAB function on a finite element space and returns the corresponding *coefficient vector*.

The partial differential equation

$$-\nabla \cdot (\boldsymbol{\kappa} \nabla u) + \boldsymbol{\gamma} \cdot \nabla u + \mu u = f \quad \text{on } \Omega$$

of a boudary value problem of the form (A.1) is represented by a MATLAB structure array, which is called the *PDE structure*. A PDE structure is typically generated by the FEM Toolbox function `fempde`, whose arguments define the coefficients $\boldsymbol{\kappa}$, $\boldsymbol{\gamma}$, $\mu$, as well as the right-hand side $f$ of an elliptic partial differential equation. The coefficients and the right-hand side can each be specified subdomain-wise, element-wise or quadrature-point-wise.

Boundary conditions of the form

$$\begin{aligned} u - \sigma(u \circ \boldsymbol{\iota}) &= r \quad \text{on } \Gamma_0, \\ -\boldsymbol{\nu} \cdot (\boldsymbol{\kappa} \nabla u) + \eta u &= g \quad \text{on } \Gamma_1 \end{aligned}$$

are represented by so-called *boundary condition structures*. These MATLAB structure arrays, in essence, store informations about the type of boundary condition,

as well as the boundary data. The boundary data $r$, $\sigma$, $r$, $\eta$, and $g$ can be specified boundary-part-wise. Several standard boundary conditions can be defined for a given finite element space by MATLAB functions, which are also provided by the FEM Toolbox. Given a mesh structure `msh` and a space structure `spc`, such that `spc` represents a finite element space on the mesh which is represented by `msh`, the function call `femperiodic(msh,spc)` generates a boundary condition structure representing periodic boundary conditions. Other functions, which generate boundary condition structures are `femdirichlet` or `femneumann`.

After defining a mesh, a finite element space on the mesh, and a partical differential equation, one can assemble the corresponding finite element matrices using the function `femassemble`. This function returns sparse matrices `K`, `C`, and `M`, as well as a vector `f`, which are the *stiffness matrix*, the *damping matrix*, the *mass matrix*, and the *load vector*.

In order to incorporate the boundary conditions, one also needs to assemble the matrices `E` and `H`, as well as the vectors `r` and `g` using the function `femassembcs`. The matrix `E` and the vector `r` represent essential boundary conditions. Recall that essential boundary conditions reduce the number of degrees of freedom in a finite element space and thus give rise to a constrained finite element space. The matrix `E` realizes the identical embedding from the constrained finite element space into the unconstrained finite element space. The vector `r` represents the boundary data. Given a coefficient vector `u` in the constrained finite element space, one obtains the corresponding coefficient vector `v` in the unconstrained space by

```
v = E * u + r;
```

The matrix `H` and the vector `g` represent natural boundary conditions. Natural boundary conditions modify the left-hand side, as well as the right-hand side in the linear system associated with the boundary value problem. In order to obtain this linear system, one needs to compute *system matrix* `A` and the *right-hand side vector* `b` according to

```
A = E' * (A + H) * E;
b = E' * (f + g - (A + H) * r);
```

The matrices `A` and `E`, as well as the vectors `b` and `r` can also be computed directly by the FEM Toolbox function `femassembvp`. The coefficient vector `u` of an approximate solution of the boundary value problem can then be computed using MATLAB's standard solver routines such as the MATLAB backslash `\`, `pcg`, or `bicg`.

Eigenvalue problems can be discretized in a similar way. After assembling the matrices `K`, `M`, and `E` mentioned above one obtains the left-hand side matrix `A` and the right-hand side matrix `B` of the associated generalized matrix eigenvalue problem by

```
A = E' * K * E;
B = E' * M * E;
```

Alternatively, one can assemble the matrices `A`, `B`, and `E` directly by using the FEM toolbox function `femassemevp`. In order to obtain approximate eigenvalues as well as the coefficient vectors of approximate eigenfunctions, one can solve the generalized matrix eigenvalue problem by using MATLAB's `eigs` function.

Most of the FEM Toolbox's functionalities are illustrated by a number of demo programs. These demo programs are MATLAB scripts which can be started through the commands `femdemo0`, `femdemo1`, ..., `femdemo12` (in the current version). In particular, the demo program `femdemo0` provides a "crash course" for users, who simply want to use the FEM Toolbox in order to solve a standard elliptic boundary value problem numerically. All functions in the FEM Toolbox are documented according to MATLAB conventions. A user manual is not available yet.

## A.3 Some Customized Features

In this section we present some features of the FEM Toolbox, which were developed specifically for the optimization algorithms introduced in Chapter 7 and Chapter 8.

In the generalized gradient method, as well as in the level set method, one often needs to evaluate the approximate eigenfunctions at the quadrature points in the mesh. Clearly, point evaluations are linear operations and can therefore be represented by matrices. We hence added a function named `femqdelta` to the FEM Toolbox. This function assembles a sparse matrix `D`, which evaluates finite element functions at all quadrature points in the corresponding finite element mesh. Given a coefficient vector `u`, one obtains by

```
y = D * u;
```

a vector `y`, which contains the function values of the finite element function represented by `u` at the quadrature points in the mesh. Thus, point evaluations at quadrature points can be performed by a single matrix-vector product. As long as the finite element mesh remains unchanged, which was the case for the above mentioned optimization algorithms, the matrix `D` does not need to be reassembled.

We found that the matrix `D` could also be used to speed up reassemblations of finite element matrices. Recall that the discretized density function $\rho$ changed in each iteration of the optimization algorithms for photonic band structure optimization problems. As a consequence, certain finite element matrices needed to be reassembled in order to compute the band structure for the new medium structure. In the TM setting, the mass matrix `M` depends on the coefficient $\rho$ and needs to

be reassembled after every iteration. In order to speed up this reassemblation, we devised the function `fempreassem`, which in particular returns a sparse diagonal matrix `W`, such that the mass matrix for the coefficient equal to 1 can be computed by

```
M = D' * W * D;
```

The matrix product on the right-hand side realizes the assemblation process for the mass matrix `M`. The elements on the diagonal of `W` are precisely the weights of the quadrature rule, which is used during the assemblation. A fast reassemblation of the mass matrix can now be performed as follows. Suppose that `rho` is a vector, which contains the function values of the discretized density function $\rho$ at all quadrature points in the mesh. One can then construct a sparse diagonal matrix `Rho` with the elements of `rho` on the diagonal. By

```
Mrho = D' * W * Rho * D;
```

one obtains the mass matrix `Mrho` for the coefficient $\rho$. As we expected, evaluating the above matrix product was significantly faster than a standard matrix assemblation. The function `fempreassem` returns further matrices, which can be used to assemble the stiffness and the damping matrix also by evaluating specific matrix products.

## A.4 Implementation Examples

In this final section, we demonstrate the use of the FEM Toolbox through small implementation examples. In the first example, our aim is to compute approximations of the four smallest eigenvalues $\lambda_1, \ldots, \lambda_4$ of the Laplace eigenvalue problem

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where $\Omega$ denotes the square $(0, \pi)^2$. Clearly, the eigenvalues of the above problem are given by $\lambda^{(m,n)} := m^2 + n^2$, where $m, n \in \mathbb{N}$. Hence, the four smallest eigenvalues are given by $\lambda_1 := \lambda^{(1,1)} = 2$, $\lambda_2 := \lambda^{(2,1)} = 5$, $\lambda_3 := \lambda^{(1,2)} = 5$, and $\lambda_4 := \lambda^{(2,2)} = 8$. Notice that the number 5 is a double eigenvalue.

The MATLAB code listed as Program A.1 demonstrates, how this problem can be discretized with the FEM Toolbox. At first, a mesh structure `msh` is created. The mesh structure represents a $20 \times 20$ structured, quadrilateral mesh on $\Omega$. Next, an $H^1$-conforming, bilinear finite element space is created on the mesh. The finite element space is represented by the space structure `spc`. The PDE structures `lsh`

**Program A.1** Approximation of the four smallest Laplace eigenvalues

```
% Create a 20-by-20 rectangular structured mesh
% on a square with side lengths equal to 2*pi.
n = 20;
msh = femsquarem(n,0,0,pi,'rectangular');

% Create an H1-conforming, linear finite element
% space on the above mesh.
p = 1;
spc = femh1(msh,p);

% Define the left-hand side and the right-hand side
% operators of the standard Laplace eigenvalue equation.
lhs = fempde(msh,spc,1);
rhs = fempde(msh,spc,0,0,1);

% Define homogeneous Dirichlet conditions.
bcs = femdirichlet(msh,spc);

% Assemble the finite element matrices for the generalized
% matrix eigenvalue problem A * v = lambda * M * v.
[A,B,E] = femassemevp(msh,spc,bcs,lhs,rhs);

% Compute the four smallest eigenvalues
k = 4;
s = 0;
[V,Lambda] = eigs(A,B,k,s);

% Realize the boundary conditions
U = E * V;

% Plot the eigenfunction
for i = 1:4
    figure(i);
    femplot(msh,spc,U(:,5-i));
end
```

and `rhs` represent the operator $-\Delta$ ($\boldsymbol{\kappa} = 1$, $\boldsymbol{\gamma} = 0$, $\mu = 0$) and the identity operator ($\boldsymbol{\kappa} = 0$, $\boldsymbol{\gamma} = 0$, $\mu = 1$), respectively. The boundary condition structure `bcs` represents the homogeneous Dirichlet boundary conditions imposed on $\partial\Omega$. The

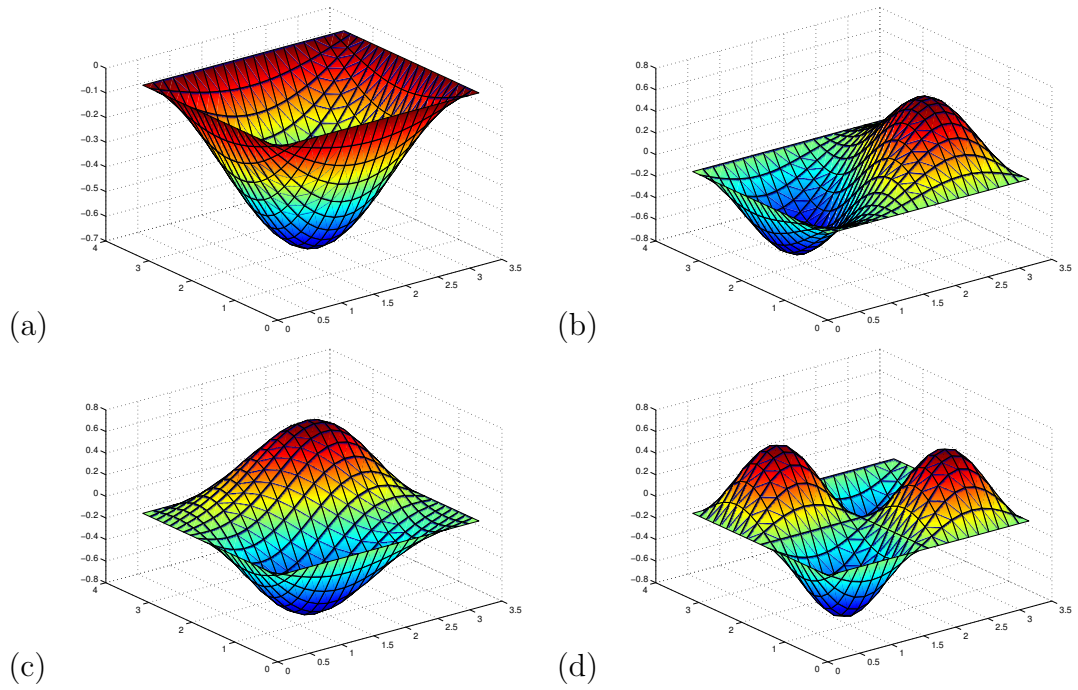



**Figure A.2:** Eigenfunctions corresponding to the four smallest Laplace eigenvalues. The plots (b) and (c) show eigenfunctions that correspond to the double eigenvalue $\lambda_2$.

finite element matrices are assembled using the function `femassemevp`. Approximations for the four smallest eigenvalues are then computed by MATLAB's `eigs` function. The function `eigs` also returns the coefficient vectors of the corresponding approximate eigenfunctions. However, these coefficent vectors only determine the eigenfunction's interior degrees of freedoms. The degrees of freedom at the boundary are realized by virtue of the boundary conditions. In the code, this is realization is accomplished by applying the embedding matrix `E` to the coefficient vectors. Finally, the eigenfunctions are plotted. The plots are shown in Figure A.2.

In our second example we investigate the convergence properties of the eigenvalue approximations with respect to the mesh width $h_{\max}$, i.e., with respect to the largest element diameter in the finite element mesh. Consider the MATLAB code listed as Program A.2. At first, a coarse triangular mesh on the computational domain $\Omega$ is created. An $H^1$-conforming finite element space on the mesh is created next. Notice that this space consists of linear finite elements, only (`p = 1;`). In the following loop the mesh, as well as finite element space are repeatedly globally refined. The refinement is accomplished by a newest-node bisection algorithm (see e.g. Chapter 3 in [68]). In Figure A.3(a) we depicted the first three refined

**Program A.2** Convergence analysis of the four smallest Laplace eigenvalues

```
% Define a coarse finite element discretization
p = 1;
msh = femsquarem(4,0,0,pi,'criss');
spc = femh1(msh,p);

% Allocate memory
hmax = zeros(1,6);
error = zeros(4,6);

for j = 1:8
    % Refine the finite element space
    [msh,spc] = femhrefine(msh,spc);

    % Mesh witdh and number of DOFs
    hmax(j) = femhmax(msh);

    % Solve the boundary value problem
    lhs = fempde(msh,spc,1);
    rhs = fempde(msh,spc,0,0,1);
    bcs = femdirichlet(msh,spc);

    % Compute the four smallest eigenvalues
    [A,B] = femassemevp(msh,spc,bcs,lhs,rhs);
    lambda = eigs(A,B,4,0);

    % Determine the errors
    error(:,j) = abs([8; 5; 5; 2] - lambda);
end
```

meshes. After every refinement, the current mesh size $h_{\max}$ is stored in the array `hmax`. Then, approximations of the smallest four Laplace eigenvalues on $\Omega$ are computed. The absolute approximation errors are then stored in the columns of the matrix `error`.

First, we ran Program A.2 as listed with linear finite element spaces. Afterwards, we ran the same program with quadratic finite element spaces, i.e., we replaced the statement `p = 1;` by `p = 2;`. The results of both program runs are displayed in Figure A.3(b) and (c). As expected, we observed an experimental order of convergence (EOC) of approximately 2 for linear finite elements, and an EOC of approximately 4 for quadratic finite elements. We remark that the EOCs

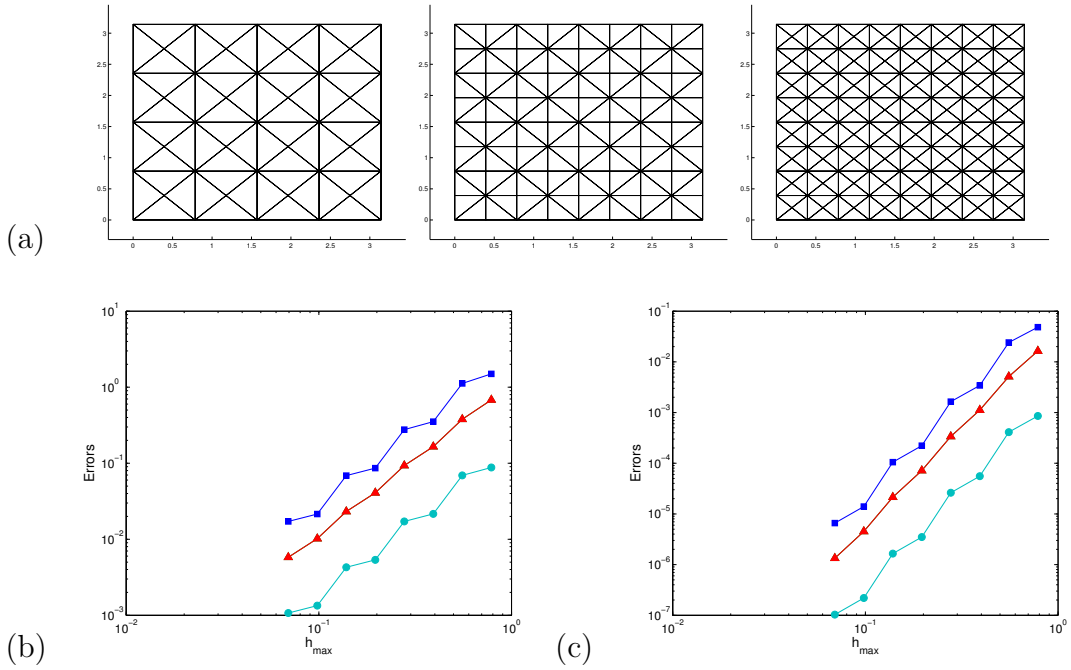



**Figure A.3:** The first three refined mesh generated in Program A.2 (a) and the absolute approximation errors for linear elements (b) and for quadratic elements (c). Round markers (•) indicate errors for the eigenvalue $\lambda_1$, triangular markers (▲) indicate errors for the eigenvalue $\lambda_2 = \lambda_3$, and square markers (■) indicate errors for the eigenvalue $\lambda_4$.

can be computed with the function `femheoc`, wich is also provided by the FEM Toolbox.

# Frequently Used Symbols

| | |
|---|---|
| $A_{\boldsymbol{k}}(\rho)$ | The conjugate-linear operator induced by $a_{\boldsymbol{k}}(\rho)$, see (4.43) on page 65 |
| $a_{\boldsymbol{k}}(\rho)$ | Sesquilinear form on $\boldsymbol{W} \times \boldsymbol{W}$, see (4.33) on page 60 |
| $a_{\boldsymbol{k}}^{\mathrm{TE}}(\rho)$ | Sesquilinear form on $V \times V$, see (4.92) on page 87 |
| $a_{\boldsymbol{k}}^{\mathrm{TM}}$ | Sesquilinear form on $V \times V$, see (4.81) on page 85 |
| $\mathbb{B}$ | The first Brillouin zone of $\Lambda$, see (3.34) on page 38 |
| $\mathcal{C}$ | The admissible set, see (5.2) on page 91 |
| $\overline{\mathcal{C}}$ | The extended admissible set, see (5.4) on page 95 |
| $\mathcal{C}_h$ | Discretization of $\mathcal{C}$, see (7.8) on page 132 |
| $\overline{\mathcal{C}}_h$ | Discretization of $\overline{\mathcal{C}}$, see (7.9) on page 132 |
| $\mathcal{D}$ | The subset of essentially positive functions in $\mathcal{E}$, see (4.32) on page 59 |
| $\mathcal{D}_h$ | Discretization of $\mathcal{D}$, see (7.7) on page 132 |
| $\mathcal{E}$ | The real Banach space $L^\infty(\Omega, \mathbb{R})$, see (4.31) on page 59 |
| $\mathcal{E}_h$ | Discretization of $\mathcal{E}$, see (7.5) on page 131 |
| $G_{\boldsymbol{k}}(\rho)$ | The Green solution operator for $L_{\boldsymbol{k}}(\rho)$, see (4.51) on page 66 |
| $\mathcal{H}$ | The real Hilbert space $H^1_{\mathrm{per}}(\Omega, \mathbb{R})$, see (8.1) on page 148 |
| $J$ | The generic goal functional, see (5.3) on page 92 |
| $J^{\mathrm{TE}}$ | The generic TE goal functional, see (5.8) on page 104 |
| $J^{\mathrm{TM}}$ | The generic TM goal functional, see (5.6) on page 97 |
| $M_{\boldsymbol{k}}$ | The conjugate-linear operator induced by $m$, see (4.44) on page 65 |

| | |
|---|---|
| $m$ | Sesquilinear form on $\boldsymbol{Z} \times \boldsymbol{Z}$, see (4.35) on page 60 |
| $m^{\mathrm{TE}}$ | Sesquilinear form on $Z \times Z$, see (4.93) on page 87 |
| $m^{\mathrm{TM}}(\rho)$ | Sesquilinear form on $Z \times Z$, see (4.82) on page 85 |
| $Q$ | The complex Hilbert space $H^1_{\mathrm{per}}(\Omega)$, see (4.29) on page 59 |
| $\boldsymbol{u}_j(\rho, \boldsymbol{k})$ | The $j$-th eigenfunction, see (4.53) on page 68 |
| $\mathcal{H}$ | The real Banach space $W^{1,\infty}(\Omega, \mathbb{R})$, see (8.2) on page 148 |
| $V$ | The complex Hilbert space $H^1_{\mathrm{per}}(\Omega)$, see (4.79) on page 84 |
| $\boldsymbol{V_k}$ | Linear subspace of $\boldsymbol{W}$, see (4.36) on page 60 |
| $\boldsymbol{W}$ | The complex Hilbert space $\boldsymbol{H}_{\mathrm{per}}(\mathrm{curl}; \Omega)$, see (4.28) on page 59 |
| $\mathcal{X}_h$ | The set of quadrature points, see (7.4) on page 131 |
| $Z$ | The complex Hilbert space $L^2(\Omega)$, see (4.80) on page 84 |
| $\boldsymbol{Z}$ | The complex Hilbert space $L^2(\Omega)^3$, see (4.30) on page 59 |
| $\Pi_h$ | The interpolation operator from $\mathcal{E}$ to $\mathcal{E}_h$, see (7.6) on page 132 |
| $\Lambda$ | Bravais lattice, see (3.32) on page 38 |
| $\lambda_j(\rho, \boldsymbol{k})$ | The $j$-th smallest eigenvalue of $A_{\boldsymbol{k}}(\rho)$, see (4.60) on page 70 |
| $\mu_j(\rho, \boldsymbol{k})$ | The $j$-th largest eigenvalue of $G_{\boldsymbol{k}}(\rho)$, see (4.52) on page 68 |
| $\Omega$ | Primitive cell of $\Lambda$, see (3.36) on page 39 |
| $\omega_j(\rho, \boldsymbol{k})$ | The $j$-th smallest eigenfrequency, see (4.74) on page 80 |
| $\omega_j^{\mathrm{TE}}(\rho, \boldsymbol{k})$ | The $j$-th smallest TE eigenfrequency, see (4.96) on page 87 |
| $\omega_j^{\mathrm{TM}}(\rho, \boldsymbol{k})$ | The $j$-th smallest TM eigenfrequency, see (4.95) on page 87 |
| $\rho_{\max}$ | Essential upper bound for the functions in $\mathcal{C}$, see (5.1) on page 91 |
| $\rho_{\min}$ | Essential lower bound for the functions in $\mathcal{C}$, see (5.1) on page 91 |

# About The Author

The author, Markus Thobias Richter, was born on July 8th 1980 in Frankfurt am Main, Germany. After living in Darmstadt–Eberstadt for a few years, his family moved to Alsbach–Hähnlein, a municipality near Darmstadt. There, he attended the local primary school from 1987 onwards. In 1991 he was admitted at the grammar school Altes Kurfürstliches Gymnasium (AKG) in Bensheim, where he received his Abitur in 2000. After fulfilling his Civilian Service as an emergency medical technician, he commenced his undergraduate studies in applied mathematics at the University of Karlsruhe in 2001. There, he moved on to graduate studies in 2003, specializing in the numerical analysis of finite element methods and mathematical models in optical communications technology. After writing his diploma thesis under the supervision of Prof. Dr. Willy Dörfler and Prof. Dr. Christian Wieners on the numerical computation of photonic band structures for dispersive media, he graduated from the University of Karlsruhe in 2006 with the academic degree Diplom–Technomathematiker. In the same year he began his doctoral studies as an associated member of the DFG Research Training Group 1294 at the Department of Mathematics of the University of Karlsruhe. His doctoral advisors were again Prof. Dr. Willy Dörfler and Prof. Dr. Christian Wieners. During his time as a doctoral candidate the author also worked as a research assistant at the Institute of Applied Mathematics and Numerical Analysis at the University of Karlsruhe.

# Bibliography


[1] Robert A. Adams and John J. F. Fournier, *Sobolev Spaces. 2nd Edition*, Pure and Applied Mathematics, no. 140, Academic Press, New York, 2003.

[2] Grégoire Allaire and François Jouve, *A level-set method for vibration and multiple loads structural optimization*, Comput. Methods Appl. Mech. Eng. **194** (2005), no. 30-33, 3269–3290.

[3] Grégoire Allaire and Robert V. Kohn, *Optimal design for minimum weight and compliance in plane stress using extremal microstructures*, Eur. J. Mech., A **12** (1993), no. 6, 839–878.

[4] Hans W. Alt, *Lineare Funktionalanalysis: Eine anwendungsorientierte Einführung. 4. Auflage*, Springer, Berlin, 2002.

[5] Angelo Alvino, Pierre-Louis Lions, and Guido Trombetti, *On optimization problems with prescribed rearrangements*, Nonlinear Anal., Theory Methods Appl. **13** (1989), no. 2, 185–220.

[6] Mark A. Armstrong, *Groups and Symmetry*, Springer, Berlin, 1988.

[7] Douglas N. Arnold, Richard S. Falk, and Ragnar Winther, *Finite element exterior calculus, homological techniques, and applications*, Acta Numerica **15** (2006), 1–155.

[8] Giles Auchmuty, *Duality for non-convex variational principles*, J. Differ. Equations **50** (1983), 80–145.

[9] ———, *Dual variational principles for eigenvalue problems*, Nonlinear Functional Analysis and its Applications (Felix E. Browder, ed.), AMS, Providence (RI), 1986, pp. 136–211.

[10] George Bachman, Lawrence Narici, and Edward Beckenstein, *Fourier and Wavelet Analysis*, Springer, Berlin, 2000.

[11] Martin P. Bendsøe and Ole Sigmund, *Topology Optimization: Theory Methods and Applications*, Springer, Berlin, 2003.

[12] Serge Berthier, *Iridescence: The Physical Color of Insects*, Springer, Berlin, 2007.

[13] Felix Bloch, *Über die Quantenmechanik der Elektronen in Kristallgittern*, Z. Phys. **52** (1928), 555–600.

[14] Daniele Boffi, Matteo Conforti, and Lucia Gastaldi, *Modified edge finite elements for photonic crystals*, Numer. Math. **105** (2006), no. 2, 249–266.

[15] Walter Borchardt-Ott, *Crystallography. 2nd edition*, Springer, Berlin, 1995.

[16] Raoul Bott and Loring W. Tu, *Differential Forms in Algebraic Topology*, Graduate texts in mathematics, no. 82, Springer, Berlin, 1995.

[17] Alexander Bulovyatov, *A Parallel Multigrid Method for Band Structure Computation of 3D Photonic Crystals with Higher Order Finite Elements*, Ph.D. thesis, Karlsruhe Institute of Technology (KIT), 2010.

[18] Bureau international des poids et mesures, Sévre, *Le Système international d'unités*, 8 ed., 2006.

[19] Dan Butnariu and Abraham Mehrez, *Convergence criteria for generalized gradient methods of solving locally lipschitz feasibility problems*, Comput. Optim. Appl. **1** (1992), no. 3, 307–326.

[20] Michel Cessenat, *Mathematical Methods in Electromagnetism: Linear Theory and Applications*, World Scientific Publishing, Singapore, 1996.

[21] Doina Cioranescu and Patrizia Donato, *An Introduction to Homogenization*, Oxford University Press, 1999.

[22] Frank H. Clarke, *Optimization and Nonsmooth Analysis*, Canadian Mathematical Society Series of Monographs and Advanced Texts, Wiley, New York, 1983.

[23] Frank H. Clarke, Yuri S. Ledyaev, Ron J. Stern, and Peter R. Wolenski, *Nonsmooth Analysis Optimization and Control Theory*, Graduate Texts in Mathematics, no. 178, Springer, Berlin, 1998.

[24] Philippe Clément, *Approximation by finite element functions using local regularization*, Rev. Franc. Automat. Inform. Rech. Operat. **9** (1975), no. R-2, 77–84.

[25] Gary C. Cohen, *Higher-Order Numerical Methods for Transient Wave Equations*, Scientific Computation, Springer, Berlin, 2002.

[26] Carlos Conca, Rajesh Mahadevan, and León Sanz, *An extremal eigenvalue problem for a two-phase conductor in a ball*, Appl. Math. Optim. **60** (2009), no. 2, 173–184.

[27] Steven Cox and Robert Lipton, *Extremal eigenvalue problems for two-phase conductors*, Arch. Ration. Mech. Anal. **136** (1996), no. 2, 101–117.

[28] Steven J. Cox, *The generalized gradient at a multiple eigenvalue*, J. Funct. Anal. **133** (1995), no. 1, 30–40.

[29] Steven J. Cox and David C. Dobson, *Maximizing band gaps in two-dimensional photonic crystals*, SIAM J. Appl. Math. **59** (1999), no. 6, 2108–2120.

[30] ———, *Band structure optimization of two-dimensional photonic crystals in h-polarization*, J. Comput. Phys. **158** (2000), no. 2, 214–224.

[31] Steven J. Cox and Joyce R. McLaughlin, *Extremal eigenvalue problems for composite membranes. I*, Appl. Math. Optimization **22** (1990), no. 2, 153–167.

[32] ———, *Extremal eigenvalue problems for composite membranes. II*, Appl. Math. Optimization **22** (1990), no. 2, 169–187.

[33] David C. Dobson and Joseph E. Pasciak, *Analysis of an algorithm for computing electromagnetic bloch modes using nedelec spaces*, Comput. Methods Appl. Math. **1** (2001), no. 2, 138–153.

[34] Georges Duvaut and Jacques-Louis Lions, *Les inéquations en mécanique et en physique*, Dunod, Paris, 1972.

[35] Ivar Ekeland and Roger Temam, *Analyse Convexe et Probléme Variationelles*, Dunod, Paris, 1974.

[36] Christian Engström and Markus Richter, *On the spectrum of an operator pencil with applications to wave propagation in periodic and frequency dependent materials*, SIAM J. Appl. Math. **70** (2009), no. 1, 231–247.

[37] Gaston Floquet, *Sur les équations différentielles linéaires à coefficients périodiques*, Ann. de l'Éc. N. **12** (1883), 47–89.

[38] Vivette Girault and Pierre-Arnaud Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Lecture Notes in Mathematics, no. 749, Springer, Berlin, 1979.

[39] Ralf Hiptmair, *Finite elements in computational electromagnetism*, Acta Numerica **11** (2002), 237–339.

[40] Friedrich Hirzebruch and Winfried Scharlau, *Einführung in die Funktionalanalysis*, Spektrum Akademischer Verlag, Heidelberg, 1971.

[41] John D. Jackson, *Classical Electrodynamics. 3rd Edition*, John Wiley & Sons, Hoboken (NJ), 1999.

[42] John D. Joannopoulos, Steven G. Johnson, Joshua N. Winn, and Robert D. Meade, *Photonic Crystals: Molding the Flow of Light. 2nd Edition*, Princeton University Press, Princeton (NJ), 2008.

[43] Sajeev John, *Strong localization of photons in certain disordered dielectric superlattices*, Phys. Rev. Lett. **58** (1987), no. 23, 2486–2489.

[44] C. Y. Kao, Stanley Osher, and Eli Yablonovitch, *Maximizing band gaps in two-dimensional photonic crystals by using level set methods*, Appl. Phys. B **81** (2005), 235–244.

[45] Tosio Kato, *Perturbation Theory for Linear Operators. 2nd Edition*, Grundlehren der mathematischen Wissenschaften, no. 132, Springer, Berlin, 1976.

[46] Shuichi Kinoshita, *Structural Colors in the Realm of Nature*, World Scientific Publishing, Singapore, 2008.

[47] Rolf Klein, *Concrete and Abstract Voronoi Diagrams*, Lecture Notes in Computer Science, no. 400, Springer, Berlin, 1989.

[48] Mark G. Krein, *On certain problems on the maximum and minimum of characteristic values and on the Lyapunov zones of stability*, AMS Transl. Ser. 2 **1** (1955), 163–187.

[49] Peter Kuchment, *Floquet Theory for Partial Differential Equations*, Birkhäuser, Basel, 1993.

[50] ———, *Mathematics of photonic crystals*, Mathematical Modeling in Optical Science (Gang Bao, ed.), SIAM, Philadelphia (PA), 2001, pp. 207–272.

[51] Alexander Kurganov and Eitan Tadmor, *New high-resolution semi-discrete central schemes for Hamilton-Jacobi equations*, J. Comput. Phys. **160** (2000), no. 2, 720–742.

[52] Gerard Lebourg, *Valeur moyenne pour gradient généralisé*, C. R. Acad. Sci., Paris, Sér. A **281** (1975), 795–797.

[53] Jerrold E. Marsden and Thomas J.R. Hughes, *Mathematical foundations of elasticity.*, Prentice-Hall, Englewood Cliffs (NJ), 1983.

[54] Clyde W. Mason, *Structural colors in feathers. II*, J. Phys. Chem. **27** (1923), no. 5, 401–448.

[55] James C. Maxwell, *A Treatise on Electricity and Magnetism*, Clarendon Press, Oxford, 1873.

[56] François Murat and Luc Tartar, *H-convergence*, Topics in the Mathematical Modelling of Composite Materials (Andrej Cherkaev and Robert Kohn, eds.), Birkhäuser, Basel, 1997, pp. 21–43.

[57] Jean-Claude Nédélec, *A new family of mixed finite elements in* $\mathbb{R}^3$, Numer. Math. **50** (1986), 57–81.

[58] Stanley Osher and Ronald Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Springer, Berlin, 2003.

[59] Stanley Osher and Fadil Santosa, *Level set methods for optimization problems involving geometry and constraints. I. Frequency of a two-density inhomogeneous drum*, J. Comput. Phys. **171** (2001), 272–288.

[60] Dennis W. Prather, Shouyuan Shi, Ahmed Sharkawy, Janusz Murakowski, and Garrett J. Schneider, *Photonic Crystals: Theory, Applications, and Fabrication*, John Wiley & Sons, Hoboken (NJ), 2009.

[61] Stefan Preble, Michal Lipson, and Hod Lipson, *Two-dimensional photonic crystals designed by evolutionary algorithms*, Appl. Phys. Lett. **86** (2005), 061111.

[62] Erhard Quaisser, *Geometrie: Einführung, Probleme, Übungen.*, Spektrum Akademischer Verlag, Heidelberg, 1994.

[63] Lord Rayleigh, *On the dynamical theory of gratings*, Lond. R. S. Proc. (A) **79** (1907), 399–416.

[64] Michael Reed and Barry Simon, *Methods of Modern Mathematical Physics. IV: Analysis of Operators*, Academic Press, New York, 1978.

[65] R. Tyrrell Rockafellar, *Convex Analysis*, Princeton University Press, Princeton (NJ), 1970.

[66] Kazuaki Sakoda, *Optical Properties of Photonic Crystals. 2nd Edition*, Springer, Berlin, 2005.

[67] Laurent Schwartz, *Radon Measures on Arbitrary Topological Spaces and Cylindrical Measures*, Oxford University Press, Oxford, 1973.

[68] Edward G. Sewell, *Automatic Generation of Triangulations for Piecewise Polynomial Approximations*, Ph.D. thesis, Purdue University, 1973.

[69] Ole Sigmund and Jensen Søndergaard, *Systematic design of photonic bandgap materials and structures by topology optimization*, Phil. Trans. R. Soc. Lond. A **361** (2003), 1001–1019.

[70] Jan Sokołowski and Jean-Paul Zolesio, *Introduction to Shape Optimization: Shape Sensitivity Analysis*, Springer-Verlag, Berlin, 1992.

[71] Krister Svanberg, *The method of moving asymptotes: A new method for structural optimization*, Int. J. Numer. Methods Eng. **24** (1987), 359–373.

[72] Luc Tartar, *Compensated compactness and applications to partial differential equations*, Nonlinear Analysis and Mechanics, Heriot-Watt Symposium IV (Robin J. Knops, ed.), Pitman, London, 1979, pp. 136–211.

[73] Pavel Šolín, Karel Segeth, and Ivo Doležel, *Higher-Order Finite Element Methods*, CRC Press, Boca Raton (FL), 2004.

[74] Christian Weber, *A local compactness theorem for maxwell's equations*, Math. Methods Appl. Sci. **2** (1980), 12–25.

[75] Dirk Werner, *Funktionalanalysis. 6. Auflage*, Springer, Berlin, 2007.

[76] Christian Wieners, *Distributed point objects: A new concept for parallel finite elements*, Domain Decomposition Methods in Science and Engineering (Berlin) (Ralf Kornhuber, Ronald Hoppe, Jacques Périaux, Olivier Pironneau, Olof Widlund, and Jinchao Xu, eds.), Lecture Notes in Computational Science and Engineering, vol. 40, Springer, 2004, pp. 175–183.

[77] ———, *A geometric data structure for parallel finite elements and the application to multigrid methods with block smoothing*, Comput. Vis. Sci. **13** (2010), 161–175.

[78] Joseph Wloka, *Partial Differential Equations*, Cambridge University Press, Cambridge, 1987.

[79] Rob S. Womersley, *Local properties of algorithms for minimizing nonsmooth composite functions*, Math. Program. **32** (1985), 69–89.

[80] Eli Yablonovitch, *Inhibited spontaneous emission in solid-state physics and electronics*, Phys. Rev. Lett. **58** (1987), no. 20, 2059–2062.