Dissertation

Eigenvalue problems in the numerical analysis of electromagnetic fields

Dipl.-Ing. Christian Scheiber

Institute for Fundamentals and Theory in Electrical Engineering
Graz University of Technology

Supervisor: Univ.-Prof. Dipl.-Ing. Dr. techn. Oszkár Bíró

Graz, September 2011

Abstract

Finding solutions to eigenvalue problems is of great importance in computational electromagnetics. Many technical problems and numerical approaches lead to eigenvalue problems. This work deals with accurate and efficient algorithms to solve such problems arising from finite-element discretizations. In the first part electromagnetic waveguide structures are investigated. A formulation is presented to resolve the dispersion relation for these structures where the focus is on a vectorial finite-element formulation and on gauging strategies to find solutions free from spurious modes. The second part of this thesis is devoted to an analysis of photonic crystals, a class of material promising for interesting optical applications. The goal is an efficient way to compute so called band structures, which significantly characterize these materials. Based on conventional finite-element formulations, model-order reduction schemes are presented allowing for a considerable calculation speed-up. The parameter sweeps, required for the computation of band structures, can thereby be carried out much faster. The principle of the order-reduction techniques is to create a model of much lower dimension while keeping the necessary information needed for an accurate solution. The cost of setting up the reduced-order model is comparable to a conventional finite-element solution at selected parameter points. A multi-point model-order reduction technique is presented that uses the eigen-solution at a few parameter points to create a reduced basis. Band structures have been calculated using both 2d and 3d formulations. In the latter case special care has to be devoted to avoid spurious solutions produced by the reduced model. Finally, a further method has been presented that uses only one finite-element solution for creating a proper basis. Therein the respective quantities are expressed as Taylor series. By mode matching an equation system is set up whose solution, together with the eigenvector at the selected parameter, serve as the reduced basis.

Kurzfassung

Das Lösen von Eigenwertproblemen nimmt eine zentrale Rolle in der numerischen Simulation von elektromagnetischen Feldern ein. Viele technische Fragestellungen sowie numerische Zugänge führen zu Eigenwertproblemen. Diese Arbeit beschäftigt sich mit genauen und effizienten Algorithmen, um Lösungen von solchen Problemen zu finden, welche durch Finite-Elemente Diskretisierung entstehen. Im ersten Teil der Arbeit werden elektromagnetische Wellenleiter untersucht. Eine Formulierung wird vorgestellt, um Dispersionsrelationen für diese Strukturen zu berechnen, wobei der Fokus an einer vektoriellen Finite-Elemente Formulierung und an Eichstrategien liegt. Das Ziel dabei ist, Lösungen zu finden, die frei von unphysikalischen Moden sind. Der zweite Teil der Dissertation ist einer Untersuchung photonischer Kristalle gewidmet. Diese Materialklasse verspricht interessante optische Anwendungen. Das Ziel ist die Berechnung von Bandstrukturen, die eine wichtige Kenngröße dieser Materialen darstellen. Aufbauend auf konventionellen Finite-Elemente Formulierungen, werden Methoden zur Modellordnungsreduktion vorgestellt, mit deren Hilfe die Rechenzeit erheblich verringert wird. Das Prinzip dieser Ordnungsreduktionstechniken ist das Erstellen eines Modells mit wesentlich kleinerer Dimension, während die notwendige Information für eine genaue Lösung beibehalten wird. Der Aufwand für die Erzeugung der Projektionsbasis ist vergleichbar mit jenem einer konventionellen Finite-Elemente Lösung für ausgewählte Parameter. Ein Mehrpunktverfahren wird vorgestellt, das die Eigenlösungen an einiger weniger Parameterpunkte verwendet, mit deren Hilfe eine reduzierte Basis gebildet wird. Bandstrukturen werden sowohl für 2d als auch für 3d Formulierungen berechnet. Im letzteren Fall muss darauf geachtet werden, sodass vom reduzierten Modell keine unphysikalische Lösungen produziert werden. Schließlich wird eine weitere Methode vorgestellt, die nur eine Lösung des vollen Finite-Elemente Problems benötigt, um eine geeignete reduzierte Basis erzeugen zu können. In diesem Fall werden die entsprechenden Größen des Eigenwertproblems in Taylorreihen entwickelt und mit Hilfe eines Modenabgleichs wird ein Gleichungssystem erzeugt. Die Lösungen von letzterem System bilden hier das reduzierte Modell.

Eidesstattliche Erklärung

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Ort

Datum

Unterschrift

Table of Contents

List of Abbreviations and Symbols

FE	Finite-Element (Method)
BVP	Boundary-Value Problem
EVP	Eigenvalue Problem
JDA	Jacobi-Davidson Algorithm
MOR	Model-Order Reduction
ROM	Reduced-Order Model
$\mathbf{E}$	Electric Field Strength
$\mathbf{H}$	Magnetic Field Strength
$\mathbf{D}$	Electric Displacement Field
$\mathbf{B}$	Magnetic Flux Density
ρ	Free Charge Density
$\mathbf{J}$	Electric Current Density
ε	Electric Permettivity
μ	Magnetic Permeability
k	Wave Number
ω	Angular Frequency

$\mathbf{A}$	Magnetic Vector Potential
V	Electric Scalar Potential
$\mathbf{F}$	Electric Vector Potential
TE	Transverse Electric
TM	Transverse Magnetic
BCC	Body-Centered Cubic
FCC	Face-Centered Cubic
$\mathbf{n}$	Normal Vector of Unit Length
γ	Propagation Constant
β	Phase Constant
α	Attenuation Constant
$\mathbf{S}$	Stiffness Matrix
$\mathbf{T}$	Mass Matrix
PBC	Periodic Boundary Conditions

Part I

Introduction and Theory

1 Motivation and Outline

This work has been undertaken under a project of the doctoral school of Numerical Simulations in Technical Sciences whose goal is to bridge theoretical and mathematical studies with engineering disciplines. In this thesis the thematic umbrella is composed of eigenvalue problems arising in the simulations of electromagnetic fields. The goal is to find efficient algorithms to efficiently describe various problems arising from electromagnetic wave phenomena. So, speaking about the practical side of this work, the focus is on describing the transmission characteristics of electromagnetic waveguide structures on the one hand and on computing dispersion relations of photonic crystals on the other hand. Due to several complexities of the structure of these systems under investigation analytical solutions are rarely available. Therefore one has to rely on efficient numerical methods to solve these problems.

One of the most powerful tool in handling electromagnetic problems or, in general, systems described by partial differential equations including boundary conditions, is the finite-element method, see e.g. [1]. Thereby the solution of a given differential equation is discretized and approximated by functions with local support, so called finite elements. This is in conceptual contrast to the finite-difference method, where the operator itself is approximated rather than its solution [2]. The latter method is commonly used in a time-domain context, since it allows for explicit time step procedures without the need of computing costly matrix factorizations. By the use of discontinuous Galerkin methods, however, also the finite-element performs well in the time-domain [3]. The formulations being presented in this thesis are given in the frequency domain, where the finite-element method is superior. Another advantage of the latter is that it allows for a higher geometric flexibility, since the implementation of unstructured meshes is straightforward there. Both methods have in common, that in the end the continuous boundary value problem is transformed into either an algebraic linear system of equations or, as it will be the case in this thesis, an algebraic eigenvalue problem. As already stated these methods belong to the class of routines based on differential equations, in contrast to methods based on integral equations, where the problem is formulated as an integral equation. The boundary-element method is an example of this class. These methods lead to algebraic linear systems of equations where the respective matrices are fully occupied.

On the contrary, one feature of the methods belonging to the class solving differential equations is the fact that the matrices describing the discretized systems are usually very sparsely occupied. As a consequence, iterative linear solvers and iterative eigenvalue solvers can be applied efficiently since these routines profit from a small occupation density of the matrices in question.

The topic of this thesis is the eigenvalue analysis of electromagnetic wave problems. In the first part, the theoretical and mathematical fundamentals are reviewed, whereas in the second part formulations and methods are presented to describe and analyze applications arising from numerical field simulations. It is the latter part which contains the scientific achievements of this thesis. Chapter 2 deals with an introduction into the electromagnetic field equations and explains the concepts behind wave propagation. In Chapter 3, the above mentioned numerical methods are outlined and explained in some detail. As the main method used in this thesis, the finite-element method is described and parts are highlighted that are implemented later on in this work. In the practical part, both scalar hierarchical and vectorial field formulations are used. For the first case, used in two dimensional formulations, a hierarchical basis up to third order is implemented. Vectorial elements are needed in the context of the electromagnetic field equation, since they guarantee for an operator conforming description. Historically, the finite-element method has been developed for scalar problems, and the natural component-wise extension to vector valued cases has proven to be unfeasible due to the creation of spurious modes. Using proper element basis functions, however, can avoid the latter. By using these formulations the given differential equations will be transformed into algebraic eigenvalue problems. A systematic overview of iterative eigenvalue solvers and a mathematical description is the content of the second part of Chapter 3. Krylov subspace iteration routines, like the Lanczos or Arnoldi method, can be applied in an effective manner to extract a few eigenvalues with largest magnitude. The desired spectral information can be adjusted upon the use of certain preconditioners. This comes at the price of having to solve an additional linear system of equations. Since the Krylov subspace relies on a certain orthogonalization structure, these solutions have to be computed very accurately. The Jacobi-Davidson method, on the other hand, only requires approximate solution for the subspace expansion. The latter is therefore preferable when the system size exceeds a certain value, where matrix factorizations get very time and memory consuming.

As a first application of the theoretical background outlined in the opening chapters, Chapter 4 deals with the analysis of electromagnetic waveguides structures. A quasi three dimensional finite-element formulation is presented where the cross section of the waveguide is meshed and the degree of freedom in the direction of the axes of propagation is taken into account in the differential equations. As mentioned above, special care has to be devoted to the use of proper operator conforming elements that neither produce spurious modes nor give approximated null solutions that can only hardly be distinguished from the physical ones, especially near the static limit. For the used $2 + 1$-d formulation different gauging strategies have been presented in [4]. Based on this work, one suggested gauging strategy is presented and analyzed in this thesis. For this purpose a code has been programmed, based on an already existing in-house software package, using rectangular second order finite elements.

The second major application is devoted to a band structure analysis of photonic crystals. Chapter 5 starts with a physical introduction into this class of materials, serving as a research motivation. The focus will be on numerical computations of band structure diagrams which pose an important characteristics of the material's optical properties. Certain photonic crystals have the interesting property that waves with certain wavelengths may be prevented from propagation through the material. To understand and quantify these so called band gaps a band structure diagram is beneficial. There are a few methods to numerically compute these quantities, but again the finite-element method will be the approach of choice in this thesis. One popular method is a plane-wave expansion, where the respective quantities are expressed as Fourier series and the problem is in the end transformed to an eigenvalue problem, which in general is fully occupied. In the finite-element method, the strategy is to model solely one unit cell and extend it via the use of periodic boundary conditions, relating each boundary to its counterpart. Again, an eigenvalue problem remains to be solved, but its matrices will be sparsely occupied in the FE case, since only entries from adjacent elements are different from zero.

First a two dimensional formulation is given where the incoming wave's electric field vector is set to be linearly polarized, perpendicular to the crystal plane. Obviously, with this H-plane formulation only singly and doubly periodic structures can be captured. Hence, a three dimensional full wave formulation is required to capture triply periodic examples which will also be outlined in this chapter. The focus and major scientific achievement, however, lie in model-order reduction techniques to efficiently compute these dispersion relations. Since a band structure calculation requires the evaluation of a generalized eigenvalue problem at many different parameter points, the whole computation tends to get quite time consuming. Hence a method will be presented that solves the full finite-element model for a few selected parameters and thence uses the obtained eigenvectors to create a reduced model then serving as a basis for the parameter sweep. The idea of this so called multi-point model-order reduction scheme is that the created model is of much lower dimension and therefore faster to solve. For the two dimensional case it will be shown that the reduction works straightforward, whereas for the full wave case one has to invest a bit more so that the model eventually does not produce spurious solutions. This is due to the special structure of the differential operator. Although this structure is correctly taken into account for the solutions at the selected parameters, it will be shown that this is not the case for the evaluation points in between. Therefore a strategy is suggested to overcome this problem by successfully filtering out the non-physical solutions when performing the parameter sweep.

Finally a single-point model-order reduction scheme is presented using solely the solution at one selected parameter for creating the reduced basis. Such techniques have been successfully applied to waveguiding structures. There the parameter dependence of the matrices is polynomial whereas it is exponential in the case of photonic crystals. The model is created by expanding the respective matrices of the generalized eigenvalue problems in Taylor series. Upon mode matching a system of equations is established. The solution of the original eigenvalue problem at the expansion point, along with the vectors solving the system, serve as a basis for the reduced model with which the parameter sweep can be carried out efficiently.

Results are presented in the end of the chapter showing the accuracy of the methods and demonstrating their efficiency.

2 Electromagnetic Theory

Let us start with reviewing the physics describing the phenomena used in this thesis. After a few introductory arguments about Maxwell's equations, the fundamentals of wave propagation are shortly outlined. It will set the physical fundament for the mathematical analysis outlined further on in the thesis.

2.1 Maxwell's Equations

The fundamentals of electromagnetic theory are described by Maxwell's system of equations [5]

$$\oint_{\partial\Gamma} \mathbf{H} \cdot d\mathbf{s} = \frac{d}{dt} \int_{\Gamma} \mathbf{D} \cdot d\Gamma + \int_{\Gamma} \mathbf{J} \cdot d\Gamma, \tag{2.1a}$$

$$\oint_{\partial\Gamma} \mathbf{E} \cdot d\mathbf{s} = -\frac{d}{dt} \int_{\Gamma} \mathbf{B} \cdot d\Gamma, \tag{2.1b}$$

$$\oint \mathbf{B} \cdot d\Gamma = 0, \tag{2.1c}$$

$$\oint \mathbf{D} \cdot d\Gamma = \int_{\Omega} \rho d\Omega. \tag{2.1d}$$

E and **H** denote the macroscopic electrical and magnetic field, respectively. **D** and **B** stand for the displacement and magnetic induction fields. Free charge and electrical current densities are represented by ρ and **J**.

With the help of the theorems of Gauß and Stokes one can derive the differential form of

Maxwell's equations:

$$\nabla \times \mathbf{H} = \frac{\partial}{\partial t}\mathbf{D} + \mathbf{J}, \tag{2.2a}$$

$$\nabla \times \mathbf{E} = -\frac{\partial}{\partial t}\mathbf{B}, \tag{2.2b}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{2.2c}$$

$$\nabla \cdot \mathbf{D} = \rho. \tag{2.2d}$$

If not explicitly stated differently in this work, materials are considered to be linear and there are no free charges or currents. In other words, ρ and $\mathbf{J}$ are set to zero. With these assumptions and additionally assuming isotropic materials the following relations are achieved for the field quantities:

$$\mathbf{B}(\mathbf{r}) = \mu(\mathbf{r})\mathbf{H}(\mathbf{r}), \tag{2.3a}$$

$$\mathbf{D}(\mathbf{r}) = \varepsilon(\mathbf{r})\mathbf{E}(\mathbf{r}), \tag{2.3b}$$

with ε and μ denoting the electrical permittivity and magnetic permeability, respectively. Under these assumptions Maxwell's equations finally read

$$\nabla \times \mathbf{H} - \varepsilon(\mathbf{r})\frac{\partial \mathbf{E}(\mathbf{r},t)}{\partial t} = 0, \tag{2.4a}$$

$$\nabla \times \mathbf{E} + \mu(\mathbf{r})\frac{\partial \mathbf{H}(\mathbf{r},t)}{\partial t} = 0, \tag{2.4b}$$

$$\nabla \times \mathbf{H}(\mathbf{r},t) = 0, \tag{2.4c}$$

$$\nabla \times \mathbf{E}(\mathbf{r},t) = 0. \tag{2.4d}$$

Due to the linearity of these equations it is possible to split off the time dependence from the spatial terms using an expansion into harmonic modes, see e.g. [6] pages 8f:

$$\mathbf{H}(\mathbf{r},t) = \mathbf{H}(\mathbf{r})e^{j\omega t}, \tag{2.5a}$$

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}(\mathbf{r})e^{j\omega t}. \tag{2.5b}$$

Inserting the last relation into the differential forms (2.2b)-(2.2d) yields Maxwell's equations in the Fourier domain:

$$\nabla \times \mathbf{H}(\mathbf{r}) = j\omega\varepsilon(\mathbf{r})\mathbf{E}(\mathbf{r}), \tag{2.6a}$$

$$\nabla \times \mathbf{E}(\mathbf{r}) = -j\omega\mu(\mathbf{r})\mathbf{H}(\mathbf{r}), \tag{2.6b}$$

$$\nabla \cdot [\mu(\mathbf{r})]\, \mathbf{H}(\mathbf{r}) = 0, \tag{2.6c}$$

$$\nabla \cdot [\varepsilon(\mathbf{r})\mathbf{E}(\mathbf{r})] = 0. \tag{2.6d}$$

These can further be simplified combining the first two equations. Depending on which field quantity is removed one obtains one of the following curl-curl-relations:

$$\nabla \times \left(\frac{1}{\varepsilon(\mathbf{r})}\nabla \times \mathbf{H}(\mathbf{r})\right) = \omega^2\mu(\mathbf{r})\mathbf{H}(\mathbf{r}), \tag{2.7a}$$

$$\nabla \times \left(\frac{1}{\mu(\mathbf{r})}\nabla \times \mathbf{E}(\mathbf{r})\right) = \omega^2\varepsilon(\mathbf{r})\mathbf{E}(\mathbf{r}). \tag{2.7b}$$

One of the last equations, along with the divergence conditions (2.6c)-(2.6d) provide all the information needed to describe the electromagnetic field quantities. For a given geometrical configuration the procedure will be to solve one of the major curl-curl-equations (2.7a) or (2.7b) to get the respective field modes and the corresponding frequencies.

2.2 Wave Propagation

This section deals with a presentation of some fundamental concepts regarding wave propagation. Thereby some specific concepts are outlined which are used later in Chapters 4-5 where practical applications are discussed. When analyzing the propagation behaviour of electromagnetic waves, one is often interested in the dispersion relation, i.e. the dependency of the wavenumber on the frequency of the wave. This will also be the case in the applications here, where first the propagation behaviour of waveguides is discussed, followed by an analysis of propagating behaviour in periodic structures such as photonic crystals.

2.2.1 Waveguides

The first practical application of this thesis will be an investigation of dispersion relations of waveguides in Chapter 4. Hence a few introductory words about waveguides are presented in this subsection. Generally one speaks about waveguides when the transversal dimensions are comparable to the wavelength, i.e. when electrically large systems are under investigation.

Under the assumption of homogeneous systems electromagnetic fields can always be considered as a superposition of two classes of wave types. The first class of modes, called transverse-electric (TE), is composed of waves whose electrical field is purely transversal and whose magnetic field has components in every direction. The opposite is the case for TM-modes, where the magnetic field lies in the plane perpendicular to the axis of propagation. With the help of single component vector potentials, the electromagnetic fields can be described by two scalar potentials. Under the additional assumption of a lossless and a charge-free medium, a magnetic vector potential $\mathbf{A}$ and an electric scalar potential V can be introduced such that $\mathbf{H} = \frac{1}{\mu}\nabla \times \mathbf{A}$ and $\mathbf{E} + j\omega\mathbf{A} = -\nabla V$. Inserting into Maxwell's equations (2.1) yields

$$-\Delta\mathbf{A} - k^2\mathbf{A} = 0, \tag{2.8a}$$

$$-\Delta V - k^2 V = 0, \tag{2.8b}$$

with $k = \omega\sqrt(\mu\varepsilon)$.

Since the divergence $\nabla \cdot \mathbf{A}$ is arbitrary, one can apply the Lorenz-gauge

$$\nabla \cdot \mathbf{A} + j\omega\varepsilon\mu V = 0. \tag{2.9}$$

The electrical and magnetic field strength can then be expressed in terms of potentials as

$$\mathbf{H} = \frac{1}{\mu}\nabla \times \mathbf{A}, \tag{2.10a}$$

$$\mathbf{E} = -\frac{1}{j\omega\mu\varepsilon}(k^2\mathbf{A} + \nabla(\nabla \cdot \mathbf{A})). \tag{2.10b}$$

In the TM-case one can choose a single-component vector potential $\mathbf{A}(\mathbf{r}) = A(\mathbf{r}) \cdot \mathbf{e}_z$. This ensures the magnetic field strength to be transverse, i.e. the component (H_z) into the direction of propagation $\mathbf{e}_z$ to be zero, since in Cartesian coordinates the field quantities are

$$\mathbf{H} = \frac{1}{\mu}\left[\frac{\partial A}{\partial y}\mathbf{e}_x - \frac{\partial A}{\partial x}\mathbf{e}_y\right], \tag{2.11a}$$

$$\mathbf{E} = \frac{1}{j\omega\mu\varepsilon}\left[\frac{\partial^2 A}{\partial x \partial z}\mathbf{e}_x + \frac{\partial^2 A}{\partial y \partial z}\mathbf{e}_y + \left(\frac{\partial^2 A}{\partial z^2} + k^2 A\right)\mathbf{e}_z\right]. \tag{2.11b}$$

Similarly, one can in general introduce an electric vector potential $\mathbf{F}$ and a magnetic scalar potential ψ, such that $\mathbf{D} = \nabla \times \mathbf{F}$ and $\mathbf{H} = j\omega\mathbf{F} - \nabla\psi$. In the TE-case a single-component vector potential $\mathbf{F}(\mathbf{r}) = F(\mathbf{r}) \cdot \mathbf{e}_z$ can be chosen. Inserting into Maxwell's equation and applying the Lorenz gauge leads to a transverse electrical field. The TE-mode field quantities can be expressed as

$$\mathbf{E} = \frac{1}{\varepsilon}\left[\frac{\partial F}{\partial y}\mathbf{e}_x - \frac{\partial F}{\partial x}\mathbf{e}_y\right], \tag{2.12a}$$

$$\mathbf{H} = -\frac{1}{j\omega\mu\varepsilon}\left[\frac{\partial^2 F}{\partial x \partial z}\mathbf{e}_x + \frac{\partial^2 F}{\partial y \partial z}\mathbf{e}_y + \left(\frac{\partial^2 F}{\partial z^2} + k^2 F\right)\mathbf{e}_z\right]. \tag{2.12b}$$

For the case of a rectangular waveguide, whose geometry is sketched in Fig. 2.1, an exact solution can be derived.

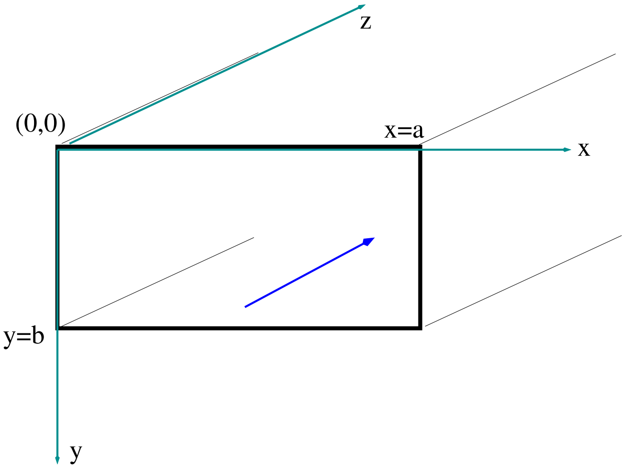


Figure 2.1: Geometry of the rectangular waveguide with axis of propagation into z-direction

With waves propagating into z-direction the following ansatz is made for the single-component vector potential in the TM-case

$$\mathbf{A}(x,y,z) = A(x,y,z)\mathbf{e}_z = A(x,y)\exp(-j\beta z)\mathbf{e}_z. \tag{2.13}$$

Together with the condition of vanishing tangential components of the electrical field at the walls, i.e. $\mathbf{n}\times\mathbf{E} = 0$, the ansatz (2.13) is inserted into (2.8a). The resulting differential equation can be easily solved in Cartesian coordinates by separating variables. The solution for the single component vector potential is then given by

$$A(x,y,z) = C\sin\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)\exp(-j\beta z), \tag{2.14}$$

with C a constant and m, n integers characterizing the respective modes.
The eigenvalues of the respective modes, i.e. the wavenumbers $k_x = \frac{m\pi}{a}$ and $k_y = \frac{n\pi}{b}$, are then related to the propagation constant β by the following relation

$$k_x^2 + k_y^2 + \beta^2 = \omega^2\mu\varepsilon. \tag{2.15}$$

In practical application one can either prescribe the propagation constant β and solve for the frequency, or fix the frequency and solve for the propagation constant. The latter case is practically more relevant and will be investigated in Chapter 4. Looking in some greater detail at (2.15) on sees that for certain frequencies β^2 might get negative. As a consequence, in these cases there is no propagation possible for the respective frequencies, the wave will be fully damped, since for $\beta^2 < 0$ one gets

$$\beta = \pm j\alpha \quad \Rightarrow \quad \exp(\mp\alpha z). \tag{2.16}$$

α is termed attenuation constant and the frequency where the transition from propagation to damping occurs is termed cut-off frequency.

With the same ansatz a similar solution is obtained for the single component vector potential describing TE-modes

$$F(x,y,z) = C\cos\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)\exp(-j\beta z). \tag{2.17}$$

From these potentials all the field quantities can be derived using (2.11) and (2.12), respectively. Of course, these rather simple analytical solutions are only possible for specific geometry configurations and under the assumption of homogeneous material parameters. For heterogeneous materials and when losses are included, the general solution is no longer given as a superposition of the transverse magnetic and transverse electrical modes. In order to compute the propagation behaviour in a general setting, one has to rely on efficient numerical methods. The finite-element method will be described in Chapter 3 and applied to waveguides in Chapter 4.

2.2.2 Periodic Structures

In chapter 5 of this thesis the focus will be on the analysis of wave propagation in periodic structures such as photonic crystals. Some theoretical concepts from solid state physics and the methodology of describing lattice structures are necessary to understand the formulations presented later on. This section serves as an introduction into the concept of reciprocal lattices and Brillouin zones used to describe dispersion relations of photonic crystals.

Due to the translational invariance of photonic crystals, such structures can be described by a periodic repetition of unit cells. The mathematical standard of solid state physics to characterize this unit cell is by a real lattice $\mathcal{L}$ and a reciprocal lattice $\mathcal{K}$ [7]. The first is mathematically defined by its three axis vectors $\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3$ so that any node of the lattice can

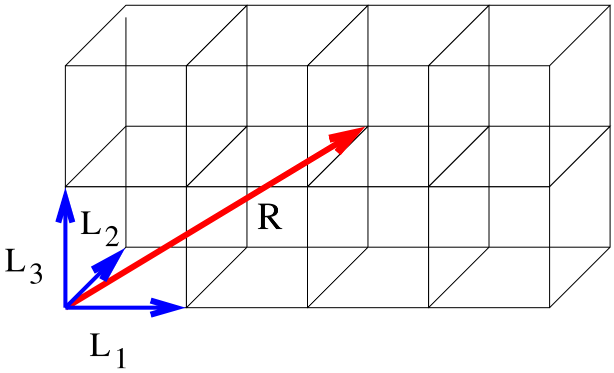


Figure 2.2: Lattice being composed of the basis vectors $\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3$

be described by a vector $\mathbf{R}$ pointing to it

$$\mathbf{R} = n_1\mathbf{L}_1 + n_2\mathbf{L}_2 + n_3\mathbf{L}_3, \tag{2.18}$$

for all integers n_1, n_2, n_3. The standard definition of the reciprocal lattice $\mathcal{K}$ is such that any vector $\mathbf{K}$ from the reciprocal lattice fulfills the following condition:

$$\exp(j\mathbf{R} \cdot \mathbf{K}) = 1. \tag{2.19}$$

With the help of (2.18) and (2.19) the basis vectors $(\mathbf{K}_1, \mathbf{K}_2, \mathbf{K}_3)$ spanning the reciprocal space are obtained as

$$\mathbf{K}_1 = 2\pi \frac{\mathbf{L}_2 \times \mathbf{L}_3}{\mathbf{L}_1 \cdot \mathbf{L}_2 \times \mathbf{L}_3}, \tag{2.20a}$$

$$\mathbf{K}_2 = 2\pi \frac{\mathbf{L}_3 \times \mathbf{L}_1}{\mathbf{L}_1 \cdot \mathbf{L}_2 \times \mathbf{L}_3}, \tag{2.20b}$$

$$\mathbf{K}_3 = 2\pi \frac{\mathbf{L}_1 \times \mathbf{L}_2}{\mathbf{L}_1 \cdot \mathbf{L}_2 \times \mathbf{L}_3}. \tag{2.20c}$$

Thus any vector from the reciprocal space is then given as $\mathbf{K} = m_1\mathbf{K}_1 + m_2\mathbf{K}_2 + m_3\mathbf{K}_3$ for some integers m_1, m_2, m_3.

The two spaces, real and reciprocal, are connected via a Fourier transform

$$f(\mathbf{r}) = \sum_{\mathbf{K} \in \mathcal{K}} \widetilde{f}(\mathbf{K}) \exp(j\mathbf{K}\mathbf{r}) \tag{2.21}$$

with

$$\widetilde{f}(\mathbf{k}) = \frac{1}{V} \int_{cell} f(\mathbf{r}) \exp(-j\mathbf{K} \cdot \mathbf{r})\, dV. \tag{2.22}$$

In a translationally invariant system electromagnetic modes can be written as *Bloch* or *Floquet* modes, i.e. they are in the form of

$$\mathbf{E}_{\mathbf{k}}(\mathbf{r}) = \exp(j\mathbf{k} \cdot \mathbf{r})\mathbf{u}_{\mathbf{k}}(\mathbf{r}) = \exp(j\mathbf{k} \cdot \mathbf{r})\mathbf{u}_{\mathbf{k}}(\mathbf{r} + \mathbf{R}), \tag{2.23}$$

with $\mathbf{u_k}$ being a periodic function. The exponential term can be interpreted as a phase shift over one lattice cell. An important characteristics of such a *Bloch* mode is the fact that adding a vector $\mathbf{K}$ from the reciprocal lattice to the wave vector $\mathbf{k}$ leads to the same mode, since the phase increment is then precisely given by (2.19) which is one. This nice property allows one to reduce the reciprocal space to its unit cell, termed *Brillouin*-zone, since every point in the reciprocal space can be related to one of the Brillouin zone by adding multiplies of the reciprocal unit vectors [6] p. 235.

For the simple cubic lattice, whose unit cell has dimension a, the reciprocal space is given as sketched in Fig. 2.3. The volume surrounded by the dashed line represents the reduced *Brillouin* zone. For the analysis of periodic structures it is sufficient to solely cover the reciprocal space within one reduced *Brillouin* zone, since all other points in the reciprocal space can be reached by mirror operations to the reduced *Brillouin* zone and by translations with reciprocal lattice vectors.

The different symmetry points are labelled by letters whose coordinates are listed below:

$$\Gamma : \frac{\pi}{a}\begin{pmatrix}0\\0\\0\end{pmatrix} \quad X : \frac{\pi}{a}\begin{pmatrix}1\\0\\0\end{pmatrix}, \tag{2.24a}$$

$$R : \frac{\pi}{a}\begin{pmatrix}1\\1\\1\end{pmatrix} \quad M : \frac{\pi}{a}\begin{pmatrix}1\\1\\0\end{pmatrix}. \tag{2.24b}$$

Another important class of lattices is the one of body-centered cubic (bcc) and face-centered cubic (fcc) lattices. In the bcc case an additional lattice point is placed in the interior of the unit cell of the simple cubic lattice, whereas the fcc lattice is created by putting a lattice point in the middle of each face of the simple cubic lattice. The *Brillouin* zones of the two lattices are illustrated in Fig. 2.4.

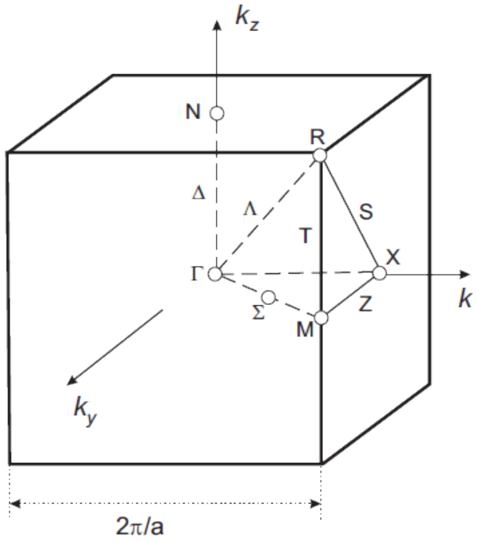


Figure 2.3: Reciprocal lattice of simple cubic lattice [8]

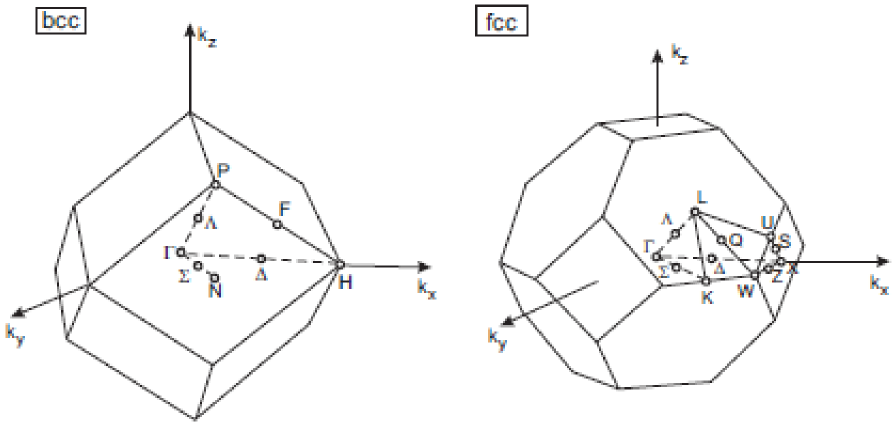


Figure 2.4: Reciprocal lattice of body-centered (bcc) and face-centered (fcc) cubic lattice [8]

3 Numerical Methods

This introductory chapter reviews the fundamental concepts of numerical methods used in this thesis. The arising differential equations are treated by the finite-element method where special focus is devoted to an appropriate representation of the respective differential operators. The fundamentals of this method are shortly reviewed in the first part of this chapter, followed by an overview of algebraic eigenvalue solvers. These methods will be used later on in this thesis when applications are considered.

The FE method is the dominating instrument in scientific and computational treatment of engineering problems since it allows for a very flexible geometric modeling of complicated structures. It is widely used not only in the electrical engineering community but also in disciplines of material sciences, fluid dynamics, thermodynamics or Geo sciences, to name a few. Even the financial industry relies on the powers of these methods to correctly price complicated financial instruments arising from derivative structured products.

The drawback of the FE method is the fact that, depending on the actual problem, it can get computationally very challenging to obtain solutions. The fundamental principle of the FE method is to transform a continuously posed problem stated as differential equations into a discrete set of algebraic equations or algebraic eigenvalue problems, respectively. Depending on the condition number of the matrix describing this set of equations a given method to compute solutions may converge or not. Direct solutions are rare since the considered applications are usually of sizes far too big to be handled with these methods. The finding of efficient algebraic solving routines, along with an appropriate formulation of FE-methods, are research topics currently pursued. In this work, the FE method will be used to compute dispersion relations for waveguiding structures and photonic crystals. These problems are stated in such a way that the discretizations will lead to eigenvalue problems

3.1 Finite-Element Method

This section is based on the respective chapters of [1] and [9]. The basic idea of the FE is based on the principle that a given boundary value problem is divided into a large number of smaller sets and to approximate the solutions in these *elements* by polynomials of lower dimension. Since these basis solutions have a common property of local support, the method is called finite-elements. The basis functions are allocated to a specific degree of freedom, e.g. a node or an edge of the mesh, and differ from zero only in domains within the respective or adjacent elements.

In order to outline the fundamentals of the FE method, let us start with a typical differential equation over a domain Ω

$$\mathcal{L}u = f. \tag{3.1}$$

Together with appropriate conditions on the domain's enclosure Γ, (3.1) is termed a boundary value problem. Here, $\mathcal{L}$ is a the differential operator, f the source and u the physical quantity that is sought.

In most applications, the finding of an analytical solution of (3.1) is impossible. Therefore a numerical method, like the FE method, has to be applied. It is stated that FE is just one choice of methods but it stands out for its geometric flexibility and is the method used in this work. The basic step of a FE formulation is the finding of a variational scheme, i.e. the solution of (3.1) formulated as the result of minimizing a certain functional. The functional space $V(\Omega)$ has to contain functions of a definite level of smoothness and has to satisfy the prescribed boundary conditions on Γ.

The differential operator involved is said to be self-adjoint if the following relation holds for all functions of V:

$$\langle \mathcal{L}u, v \rangle = \langle \mathcal{L}v, u \rangle, \quad \forall v \in V, \tag{3.2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. This is a symmetric bilinear form. In the following only self-adjoint operators are considered and the notation can be simplified to

$$\mathcal{L}(u, v) \equiv \langle \mathcal{L}u, v \rangle = \langle \mathcal{L}v, u \rangle, \quad \forall v \in V. \tag{3.3}$$

By multiplying (3.1) with a test function v and integrating over the domain Ω, one obtains the weak form

$$\mathcal{L}(u, v) = \langle f, v \rangle, \quad \forall v \in V. \tag{3.4}$$

One way to discretize the weak form is to apply the Galerkin method by restricting the functions u and v to a finite-dimensional subspace $V_h \subset V$. The weak form then transforms to the problem of finding $u_h \in V_h$ such that

$$\mathcal{L}(u_h, v_h) = (f, v_h)\,, \quad \forall v_h \in V_h. \tag{3.5}$$

Assuming that a proper discretization of the space V to a set with dimensionality n has been carried out, i.e. the domain has been meshed, the approximated solution u_h can then be described as a linear combination of the discrete set's basis functions $\psi_i (i = 1, \dots, n)$:

$$u_h = \sum_{i=1}^{n} u_i \psi_i = \mathbf{u}^T \boldsymbol{\psi}, \tag{3.6}$$

where $\mathbf{u}$ represents the Eucledian vector of coefficients generally to be in $\mathbb{C}$. Thus the original problem has been transformed into the the algebraic task of finding proper coefficients $\mathbf{u}$.

In the same spirit a relation is established between the differential operator $\mathcal{L}$ and the Euclidean vectors $\mathbf{u}$ and $\mathbf{v}$ [9].

$$\langle \mathbf{S}\mathbf{u}, \mathbf{v} \rangle = \mathcal{L}(u_h, v_h). \tag{3.7}$$

This relation holds for any two functions $u_h, v_h \in V_h$ and their respective Euclidean vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. Here, the left hand side is the typical Euclidean scalar product with the square matrix $\mathbf{S}$ given by

$$S_{ij} = \langle \mathbf{S}\mathbf{e}_i, \mathbf{e}_j \rangle = \mathcal{L}(\psi_i, \psi_j), \tag{3.8}$$

where the vectors $\mathbf{e}_i$ represents the i-th column of the identity matrix. This means that in the end, after defining proper basis functions and computing the respective matrices, the continuous differential equation has been transformed into an algebraic system of equations. This equivalence is the main property of Galerkin methods and FEM in particular. The matrix S_{ij} is usually termed *stiffness matrix* since the method was originally developed for problems in structural mechanics.

Since the Galerkin formulation (3.5) is only a restriction of the weak continuous formulation to a finite-dimensional subspace, the numerical bilinear form has the same algebraic properties as the continuous one [9], pp. 79f. For elliptic differential operators $\mathcal{L}$, i.e. if

$$\mathcal{L}(u, u) \geq c(u, u), \quad \forall u \in V, \tag{3.9}$$

for some constant $c > 0$, the matrix $\mathbf{S}$ is strictly positive definite and, in addition, the following relation holds:

$$(\mathbf{Su}, \mathbf{u}) \geq c(\mathbf{Tu}, \mathbf{u}), \quad \forall \mathbf{u} \in \mathbb{R}^n, \tag{3.10}$$

where $\mathbf{T}$ is such that the Euclidean form $(\mathbf{Tu}, \mathbf{v})$ corresponds to the L_2 inner product of the respective functions:

$$(\mathbf{Tu}, \mathbf{v}) = (u_h, v_h). \tag{3.11}$$

This defines the so called *mass matrix*

$$T_{ij} = (\psi_i, \psi_j) \tag{3.12}$$

and is in complete analogy to the expressions (3.7) and (3.8).

Thus the continuous differential equation has to be transformed into an algebraic system or into an eigenvalue problem, depending on the problem setting. A difficulty lies in choosing the correct space for the basis functions ψ_i. This will be discussed in the next subsections where the focus is on formulations regarding electromagnetic problems.

A general recipe for solving differential equations of the form (3.1) with the FE method, is the following [10]. First the computational domain is divided into parts or elements, e.g. triangles, quadrilaterals for 2D cases or tetrahedra, cuboids for 3D problems. As a second step, the solution is expressed by a finite number of low order polynomial basis functions W_i, i.e. $u(\mathbf{r}) \approx \sum_{i=1}^{n} u_i W_i(\mathbf{r})$. Then a residual quantity $r = \mathcal{L}[u] - f$ is computed, which is required to be as small as possible in the weak sense. This is followed by choosing as many test functions ω_i as there are unknown functions for weighting the residual r. When test and basis functions coincide one speaks from Galerkin's method. Finally the weighted residual is set to zero and one solves for the unknown coefficients u_i:

$$\langle W_i, r \rangle = \int_{\Omega} W_i r \, d\Omega = 0. \tag{3.13}$$

To be mathematically correct, the term finite element contains the element together with a polynomial space defined in the element and a set of degrees of freedom defined on this space. The first can be a triangle, whereas the second stands for a certain space of polynomial functions and the third represents the values on the polynomial functions in the corners of the element [10].

3.1.1 Scalar Valued Quantities

For an analysis of 2d systems, the scalar Helmholtz equation quite often serves as a model describing the problem setting at hand. This is the case in the first part of Chapter 5. For a scalar potential $\mathbf{A} = A\mathbf{e}_z$ the scalar Helmholtz equation reads

$$-\nabla \cdot \nabla A - k^2 A = 0. \tag{3.14}$$

Since A is a scalar quantity, node based elements can be used to approximate a solution. The component of the potential pointing into the direction of propagation can thus be expressed as

$$A(\mathbf{r}) = \sum_{i=1}^{n_n} A_i N_i(\mathbf{r}), \tag{3.15}$$

with the vector $\mathbf{r}$ lying in the plane perpendicular to the axis of propagation and n_n denoting the number of nodes. The vector $\mathbf{r}_i$ is pointing to the i-th node. Choosing triangular elements with linear basis functions, the nodal basis functions N_i have the property $N_i(\mathbf{r}_i) = 1$ and $N_i(\mathbf{r}_j) = 0$ for $i \neq j$. Thus there is one basis function associated with each node. Inserting these trial functions into the weak form (3.5) of (3.14) and applying Galerkin's method leads to the following algebraic eigenvalue problem:

$$\left[\mathbf{S} - k^2\mathbf{T}\right] \mathbf{v} = 0. \tag{3.16}$$

Here $\mathbf{S}$ and $\mathbf{T}$ are termed stiffness and mass matrix, respectively. They are calculated by the following relations

$$S_{ij} = \int_S \nabla N_i \cdot \nabla N_j \, dS, \tag{3.17a}$$

$$T_{ij} = \int_S N_i \cdot N_j \, dS. \tag{3.17b}$$

In the first part of Chapter 5 the scalar Helmholtz equation will reappear when discussing a 2d-formulation for photonic crystals. There the electrical field is considered to lie in a plane perpendicular to the axis of propagation.

3.1.2 Vector Valued Quantities

When having vector valued differential equations at hand, like the vectorial Helmholtz equation, the obvious extension of the FE method would be to simply describe each component separately by node based elements. This procedure, however, leads to approximation errors, involving the occurrence of spurious modes. As an outcome the basis functions have to be chosen from a well-defined function space, taking care of the properties of the differential operator. A list of conventional function spaces, used in electromagnetic problems is presented in subsection 3.1.3. There is one class of elements, called edge-elements, fulfilling the continuity properties of the curl-operator [11]. There the degrees of freedom correspond to the edges of the elements, rather than to the nodes. In addition to a numbering of edges in the mesh, a reference direction of each edges has to be defined in this case. The electrical field strength **E** can then be expressed in terms of edge element basis functions $\mathbf{N}_i$ as

$$\mathbf{E}(\mathbf{r}) = \sum_{i=1}^{N_e} E_i \mathbf{N}_i(\mathbf{r}). \tag{3.18}$$

Theses basis functions are formed such that they have continuous tangential components across element interfaces, whereas they allow their normal components to jump. This property insures that an electrical field expanded in terms of edge elements has a curl that is square integrable, since their continuous tangential components imply that the curl of an edge element does not contain jumps at element interfaces [10]. The most important property, however, is the fact that the gradient functions ∇N_i are contained in the discrete space **N** spanned by the edge element basis functions $\mathbf{N}_k$:

$$\nabla N_i \subset \mathbf{N}. \tag{3.19}$$

This ensures the avoidance of spurious modes when using edge elements to approximate the solution of the electromagnetic wave equation.

In the remainder of the thesis, boundary value problems of the following type play an important role, whose solutions describe the propagation properties of electromagnetic waves

$$\nabla \times \nabla \times \mathbf{E} - k^2 \mathbf{E} = 0 \quad \text{in} \quad S \tag{3.20a}$$

$$\mathbf{n} \times \mathbf{E} = 0 \quad \text{on} \quad \partial S. \tag{3.20b}$$

In order to apply the FE method to the above BVP, the problem has to be transformed into the weak form given by

$$\int_S (\nabla \times \mathbf{W}_i) \cdot (\nabla \times \mathbf{E})\, dS = k^2 \int_S \mathbf{W}_i \cdot \mathbf{E}\, dS. \tag{3.21}$$

The electrical field is then approximated by edge element basis functions $\mathbf{N}_i$ tested with $\mathbf{W}_i = \mathbf{N}_i$, i.e. applying Galerkin's method. Thus one again arrives at a generalized eigenvalue problem

$$\left[\mathbf{S} - k^2 \mathbf{T}\right] \mathbf{v} = 0, \tag{3.22}$$

with the entries of the stiffness $\mathbf{S}$ and the mass matrix $\mathbf{T}$ now given by

$$S_{ij} = \int_S (\nabla \times \mathbf{N}_i) \cdot (\nabla \times \mathbf{N}_j)\, dS \tag{3.23a}$$

$$T_{ij} = \int_S \mathbf{N}_i \cdot \mathbf{N}_j\, dS. \tag{3.23b}$$

The reason for node based elements to fail for this kind of differential operators is the fact that they do not contain the proper null-space of the curl-operator, which is given by the electrostatic modes $\mathbf{E} = -\nabla V$. Node based elements can only approximate these modes and therefore pollute the spectrum between zero and the values of the physical modes. On the other hand, edge-based elements yield exactly one zero eigenvalue for each interior node. This is the case, since the electrostatic modes $\mathbf{E} = -\nabla V$, with piecewise bilinear V, belong to the set of edge elements according to the property of (3.19).

3.1.3 Function spaces

In this subsection, referring to [12], some theoretical concepts about function spaces of finite-element basis functions are discussed which are needed for an analysis of electromagnetic problems posed by the Maxwell's equations. There are four important function spaces, each describing a particular electromagnetic quantity:

$$H^1(\Omega, \Gamma_\phi) := \{\phi \in L^2(\Omega) | \nabla\phi \in L^2(\Omega) \wedge \phi = 0 \text{ on } \Gamma_\phi\}, \tag{3.24a}$$

$$H(curl; \Omega, \Gamma_E) := \{\mathbf{E} \in L^2(\Omega) | \nabla \times \mathbf{E} \in L^2(\Omega) \wedge \mathbf{n} \times (\mathbf{E} \times \mathbf{n}) = 0 \text{ on } \Gamma_E\}, \tag{3.24b}$$

$$H(div; \Omega, \Gamma_B) := \{\mathbf{B} \in L^2(\Omega) | \nabla \cdot \mathbf{B} \in L^2(\Omega) \wedge \mathbf{n} \cdot \mathbf{B} = 0 \text{ on } \Gamma_B\}, \tag{3.24c}$$

$$H^0(\Omega) := \{\rho \in L^2(\Omega)\} \tag{3.24d}$$

Here Ω stands for the whole domain and Γ_i for the respective boundary. L^2 denotes the space of Lebesgue integrable functions

$$L^2(\Omega) := \{\mathbf{f}(\mathbf{x}) | \|\mathbf{f}(\mathbf{x})\|_{L^2} < \infty\}, \tag{3.25}$$

with $\|\cdot\|$ standing for the norm induced by the scalar product

$$(\mathbf{f}_1, \mathbf{f}_2) = \int_\Omega \mathbf{f}_2^H(\mathbf{x}) \mathbf{f}_1(\mathbf{x})\, d\Omega. \tag{3.26}$$

In order to make the above function spaces to Hilbert spaces, i.e. complete under the scalar product, the latter (3.26) has to be slightly modified for each function space. With the following definitions of scalar products, the so called *Sobolev* spaces are obtained:

$$(\phi, \phi)_{H^1(\Omega)} := (\phi_1, \phi_2)_L^2(\Omega) + (\nabla\phi_1, \nabla\phi_2)_{L^2(\Omega)}, \tag{3.27a}$$

$$(\mathbf{E}_1, \mathbf{E}_2)_{H(curl;\Omega)} := (\mathbf{E}_1, \mathbf{E}_2)_{L^2(\Omega)} + (\nabla \times \mathbf{E}_1, \nabla \times \mathbf{E}_2)_{L^2(\Omega)}, \tag{3.27b}$$

$$(\mathbf{B}_1, \mathbf{B}_2) := (\mathbf{B}_1, \mathbf{B}_2)_{L^2(\Omega)} + (\nabla \cdot \mathbf{B}_1, \nabla \cdot \mathbf{B}_2)_{L^2(\Omega)}, \tag{3.27c}$$

$$(\rho_1, \rho_2)_{H^0(\Omega)} := (\rho_1, \rho_2)_{L^2(\Omega)}. \tag{3.27d}$$

The clue about the presented function spaces is such that elements form these spaces ensures the physically correct continuity properties. This means that an electrical scalar potential has to be an element from H^1 since it is a continuous spatial function. The electromagnetic field quantities **E** and **H**, on the other hand, are tangentially continuous so that they belong to the space $H(curl;\Omega)$. The latter is tangentially continuous due to the L^2-integrability of the curl. Since the flux quantities **D** and **B** are normally continuous they have to be members of the third function space $H(div;\Omega)$. There the property of normal continuity is due to the L^2-integrability of the divergence. Finally volume charges do not posses any continuity properties. Therefore they belong to the last space $H^0(\Omega)$. In the FE method one approximates these function spaces by discrete functions. Depending on the quantity a certain subspace is chosen. The space spanned by node based elements corresponds to a discrete subspace of H^1 of (3.24a), whereas edge elements are constructed so that they are members of a discretized subspace of $H(curl)$ of (3.24b).

3.2 Review of Eigenvalue Solvers

The definition of an eigenvalue problem is to find eigenvectors $\mathbf{v}_i$ and eigenvalues λ_i to a matrix $\mathbf{A}$ such that the following equation holds:

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i. \tag{3.28}$$

The set of eigenvalues λ_i is termed eigenspectrum. Generalized eigenvalue problems, like in (3.16) and (3.22), are defined for the matrix pencil $(\mathbf{A}, \mathbf{B})$ according to the following relation:

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{B}\mathbf{v}_i. \tag{3.29}$$

This section deals with a systematic overview of routines to solve algebraic eigenvalue problems. FE-formulations as described in the last section lead to generalized eigenvalue problems. The matrices obtained from the discretizations of the differential operators are in general far too big to be fully diagonalized by a direct application of conventional eigenvalue routines. In other words, the full eigenspectrum cannot be computed in practical applications. In many cases, however, it is sufficient to resolve the spectrum for extremal eigenvalues or for eigenvalues, located around a selected target. This is achieved by the use of so called iterative methods, that transform the original matrices into certain structures for which direct methods can be applied. These iterative methods work the more efficiently the lower the occupation number of the matrices in question is. The number of matrix entries different from zero defines the occupation number. Since one characteristic of matrices stemming from FE-discretizations is that they are sparsely occupied, iterative methods will be well suited in our context. (Subspace) iteration methods have in common, that first a subspace of much lower dimension is created which is then solved by direct methods. All one has to know about the matrix in question is its action on a vector. In the latter expression the matrix is understood as an algebraic representation of a linear tensorial operator, where the matrix-vector multiplication represents the action of the operator. This matrix-vector multiplication can be efficiently implemented and is treated as a black-box by the algorithm. Whereas the information about the full spectrum cannot be cheaply obtained by iterative methods, the computed solutions are more or less accurate approximations of the eigenvalues and eigenvectors near a selected target value.

It is noted that the above categorization of eigenvalue routines is somewhat misleading, since also direct methods rely on an iterative procedure to compute the eigensolutions. In [13] it is stated that any eigenvalue solver has to be iterative. This is to understand in comparison to solving linear systems, where certain methods are capable of producing exact solutions in a finite number of steps. In the context of eigenvalue problems, however, sequences are computed with the goal of a rapid convergence towards the true eigenvalues. Although this means that computing eigenvalues is in principle an *unsolvable* problem, the convergence of this sequence is very rapid and in practice differs from the solution of linear systems only by a small fraction. To close the picture, direct methods will be stated as such in this thesis, although they imply an intrinsic iterative behaviour. Iterative methods in the context of the thesis refer to methods, where a subspace is created. So, *subspace iteration methods* would be a more appropriate term, this class will be often denoted by the name iterative methods in the context of this thesis.

This chapter starts with a short description of direct routines, which are implicitly used to solve the transformed eigenproblems when applying the iterative schemes, followed by a description of the most traditional iterative schemes. Finally, it is outlined how these concepts can be extended, for the computation of generalized eigenvalue problems, defined in (3.29).

3.2.1 Direct Methods - Complete Spectrum

When speaking about direct routines applied to solving eigenproblems, a slightly better term is to speak about methods to solve the complete eigenspectrum of a certain matrix. The ultimate goal there is to transform the matrix to diagonal form by similarity transformations, i.e. transformation that do not alter the eigenvalues of the matrix. A prominent representative of this class of methods is the Jacobi method. The basic idea of this method is to transform a matrix $\mathbf{A}$,

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \tag{3.30}$$

into a diagonal matrix $\mathbf{J}^T\mathbf{AJ}$, where the orthogonal matrix $\mathbf{J}$ is given by

$$\mathbf{J} = \begin{pmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi). \end{pmatrix} \tag{3.31}$$

This can be accomplished by selecting the angle ϕ so that the off-diagonal entries of $\mathbf{J}^T\mathbf{AJ}$ will vanish, which is the case when

$$\frac{\cos^2(\varphi) - \sin^2(\varphi}{\sin(\varphi)\cos(\varphi)} = \frac{a_{22} - a_{11}}{a_{12}}. \tag{3.32}$$

Applying this scheme to larger matrices, one manages to eliminate a pair of column and row indices (i, j). Repeating this scheme will lead to a diagonal matrix even though in some steps previously eliminated off-diagonal elements might be refilled. Details to this method can be found in [14] and [15]. To close the gap another important representative of the class of methods for solving the full eigenspectrum is the QR-method. It makes use of the power iteration described below. The QR-method, e.g described in detail in [15] ch. 10, starts with a QR-decomposition, factorizing the matrix $\mathbf{A}$ into a orthogonal matrix $\mathbf{Q}$ times an upper triangular matrix $\mathbf{R}$:

$$\mathbf{A} = \mathbf{Q}\mathbf{R}. \tag{3.33}$$

The QR-factorization is unique, as long as the matrix $\mathbf{A}$ is invertible. The factorization is followed by a QR transformation $\widetilde{\mathbf{A}} = \mathbf{R}\mathbf{Q} = \mathbf{Q}^T\mathbf{A}\mathbf{Q}$. The algorithm then consists of the following iteration

$$\mathbf{A}_1 = \mathbf{A} \tag{3.34a}$$

$$\mathbf{A}_k = \mathbf{Q}_k\mathbf{R}_k \quad k = 1, 2, \ldots \tag{3.34b}$$

$$\mathbf{A}_{k+1} = \mathbf{R}_k\mathbf{Q}_k = \mathbf{Q}^T\mathbf{A}_k\mathbf{Q}_k. \tag{3.34c}$$

For non-symmetric cases this procedure converges to an upper triangular matrix, where the eigenvalue can be taken from the entries on the diagonal . For symmetric matrices, the limit matrix is diagonal.

3.2.2 Subspace Iteration Methods - Selective Spectrum

In practical applications for solving large eigenvalue problems, subspace iteration methods are the most important class of eigenvalue routines. The simplest approach of employing an iterative scheme is the power method. It is based on the idea that successive action of the matrix on a vector will converge to an extreme eigenvector. It yields the dominating eigenvalue λ_1 and its corresponding eigenvector $\mathbf{x}_1$. Consider the eigenvalue problem

$$\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i. \tag{3.35}$$

Starting with an initial starting vector $\mathbf{v}_1$ taken from a random guess, the following iteration prescription can be employed:

$$\widetilde{\mathbf{v}}_2 = \mathbf{A}\mathbf{v}_1, \tag{3.36a}$$

$$\widetilde{\mathbf{v}}_3 = \mathbf{A}\widetilde{\mathbf{v}}_2 = \mathbf{A}^2\mathbf{v}_1, \tag{3.36b}$$

$$\vdots$$

$$\widetilde{\mathbf{v}}_m = \mathbf{A}^{m-1}\mathbf{v}_1. \tag{3.36c}$$

The convergence of this method is characterized by the ratio of the two eigenvalues with largest magnitude $\left|\frac{\lambda_2}{\lambda_1}\right|$. If this ratio is small, as it is the case in many applications, the convergence behaviour is quite poor [16]. Another problem arises when the eigenvalues are degenerate. In this case only one eigenvalue and its corresponding eigenvector is found. For real matrices having complex eigenvalues the method cannot be applied at all, since there are two eigenvalues with largest modulus in these cases, since they appear as complex pairs. One possibility to improve convergence is to apply the power method to a shifted matrix $\mathbf{B}$ with

$$\mathbf{B}(\lambda_0) = \mathbf{A} + \sigma\mathbf{I}, \tag{3.37}$$

where σ represents a user defined target value. The shifted matrix has the same eigenvectors with the corresponding shifted eigenvalues. If one wants the iteration to converge to the lowest eigenvalues instead, the power iteration should be applied to the inverse, thus having the iteration

$$\mathbf{v}_m = \mathbf{A}^{-1}\mathbf{v}_{m-1}. \tag{3.38}$$

Of course the inverse need not be computed directly, since only its action on a vector is of interest. This can be easily achieved by a pre computation of its LU-factorization, followed by solving an upper and lower triangular system at each step. The same techniques will be used when transforming generalized eigenvalue problems to standard form, outlined later in this chapter. The inverse iteration will converge to eigenvalues with lowest modulus, since the eigenvalues of $\mathbf{A}$ and $\mathbf{A}^{-1}$ are inversely related to each other, yet having the same eigenvector. The last two extensions of the power method can be combined to the inverse power iteration with a shift. Here the iteration is applied to the matrix $\mathbf{C}$ with

$$\mathbf{C} = (\mathbf{A} - \sigma\mathbf{I})^{-1}, \tag{3.39}$$

with σ again a user-selected shift. The algorithm will converge to the eigenvalue closest to the shift, since the eigenvalue of $\mathbf{C}$ with largest modulus is $\frac{1}{\lambda_1-\sigma}$. The convergence of this routine is characterized by the ratio

$$\left|\frac{\lambda_1 - \sigma}{\lambda_2 - \sigma}\right|, \tag{3.40}$$

with λ_1 and λ_2 being the closest and next closest eigenvalue to the shift σ.

For Hermitian matrices, an efficient possibility for a shift is to take the Rayleigh quotient of the latest approximated eigenvector $\mathbf{v}_k$. The algorithm for this Rayleigh quotient iteration is to start with a normalized initial vector $\mathbf{v}_0$ and applying the inverse power iteration, where in each step the shift σ_k and the following iterated vector $\mathbf{v}_k$ are given by

$$\sigma_k = \mathbf{v}_{k-1}^H \mathbf{A} \mathbf{v}_{k-1}, \tag{3.41}$$

$$\mathbf{v}_k = \frac{1}{\|\mathbf{v}_k\|} \left(\mathbf{A} - \sigma_k \mathbf{I}\right)^{-1} \mathbf{v}_{k-1}. \tag{3.42}$$

Although this process is globally convergent for Hermitian matrices, it is not widely used in practice, since it requires an expensive LU-factorization in each iteration step [16].

In the following the ideas of the power iteration is extended in such a way that the iterated vectors are used to span a subspace, called *Krylov* subspace, in which the eigenpairs are sought. The principal idea behind these iterative subspace methods is to first project the matrices into the lower-dimensional Krylov subspace. Given a certain matrix $\mathbf{A}$ and an initial vector $\mathbf{v}_0$, the associated Krylov sub-space is spanned by the vectors

$$\mathcal{K}_m = \{\mathbf{v}_0, \, \mathbf{A}\mathbf{v}_0, \, \mathbf{A}^2\mathbf{v}_0, \, \ldots, \mathbf{A}^{m-1}\mathbf{v}_0\}. \tag{3.43}$$

Depending on the algorithm, this sequence will lead to a matrix structure, being either of upper Hessenberg or of tridiagonal form which will then be diagonalized by direct methods. An upper Hessenberg matrix is a matrix whose entries are zero for any pair of indices i, j with $i > j + 1$. Lower Hessenberg matrices are defined accordingly. A tridiagonal matrix can be viewed as both an upper and a lower Hessenberg matrix, since it only has nonzero elements in the main diagonal, and the first diagonal below and above the main diagonal.

3.2.2.1 Arnoldi

The Arnoldi algorithm can be applied to general non-Hermitian matrices. The orthogonal projection onto the *Krylov* subspace $\mathcal{K}_m$ will lead to an upper Hessenberg form. The basic procedure of the algorithm is to build an orthogonal basis of $\mathcal{K}_m$ by first applying the matrix onto a starting vector, followed by an orthogonalization of the resulting vector to a randomly chosen starting vector. This process is repeated where in each step the new vector is orthogonalized to all previous ones. The applied orthogonalization scheme is the modified Gram-Schmidt routine. The ordinary Gram-Schmidt orthogonalization scheme might lead to sever cancellations [16]. The iteration will stop when the norm of the corrected vector falls below a certain threshold. The method is described in Algorithm 1 below. The quantities in the lines 5 and 8 define the Hessenberg matrix $\mathbf{H}_m$ which can then be diagonalized by direct routines.

The fact, that the eigenvalues of $\mathbf{H}_m$ are identical to those the original matrix $\mathbf{A}$, can be illustrated by looking at the following relation [16]:

Algorithm 1 Arnoldi method

```
$\mathbf{v}_1 = \mathbf{v}_1/||\mathbf{v}_1||$
for $j = 1, 2, \ldots, m$ do
    $\mathbf{w} = \mathbf{A}\mathbf{v}_j$
    for $i = 1, 2, \ldots, j$ do
        $\mathbf{H}_{ij} = \mathbf{w}^H \cdot \mathbf{v}_i$
        $\mathbf{w} = \mathbf{w} - \mathbf{H}_{ij}\mathbf{v}_i$
    end for
    $\mathbf{H}_{j+1,j} = ||\mathbf{w}||$
    $\mathbf{v}_{j+1} = \mathbf{w}/\mathbf{H}_{j+1,j}$
    test for convergence
end for
```

$$\mathbf{A}\mathbf{V}_m = \mathbf{V}_m\mathbf{H}_m + h_{m+1,m}\mathbf{v}_{m+1}\mathbf{e}_m^H, \tag{3.44}$$

$$\mathbf{V}_m^H\mathbf{A}\mathbf{V}_m = \mathbf{H}_m. \tag{3.45}$$

Here the columns of the matrix $\mathbf{V}_m$ are the iterated vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ and $\mathbf{e}_m$ denotes the m-th canonical basis vector of the Eucledian space.

It is noted that it may happen in certain cases, that the new iterated vector is already orthogonal to the previous vectors. In this case the algorithm stops and the found eigensolutions are exact.

3.2.2.2 Lanczos

When having Hermitian matrices at hand, the Arnoldi method can be simplified to get the Lanczos algorithm. The basic principle is similar, but in the Lanczos case one only has to store three rather than all vectors. The upper Hessenberg matrix of the Arnoldi method is a tridiagonal and symmetric matrix in the Lanczos case. The method is outlined in Algorithm 2

The eigenvalues of the tridiagonal matrices $\mathbf{T}_m$ will be equal to the lowest eigenvalues of the original matrix $\mathbf{A}$, since the matrices $\mathbf{T}_m$ of the Lanczos iterations are the corresponding projections

$$\mathbf{T}_m = \mathbf{V}_m^H\mathbf{A}\mathbf{V}_m. \tag{3.46}$$

Algorithm 2 Lanczos method

```
$\mathbf{v}_1 = \mathbf{v}_1/||\mathbf{v}_1||, \beta_1 = 0, \mathbf{v}_0 = 0$
for $j = 1, 2, \ldots, m$ do
    $\mathbf{w} = \mathbf{A}\mathbf{v}_j - \beta_j \mathbf{v}_{j-1}$
    $\mathbf{T}_{j,j} = \mathbf{w}^H \cdot \mathbf{v}_j$
    $\mathbf{w} = \mathbf{w} - \mathbf{T}_{j,j}\mathbf{v}_j$
    $\mathbf{T}_{j,j+1} = \mathbf{T}_{j+1,j} = ||\mathbf{w}||$
    $\mathbf{v}_{j+1} = \mathbf{w}/\beta_{j+1}$
    test for convergence
end for
```

3.2.2.3 Generalized Eigenvalue Problems

In this subsection attention is paid to the practically relevant case, where an iterative computation of eigenpairs to the generalized eigenvalue problem is sought. The problem is then given as

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{B}\mathbf{v}_i, \tag{3.47}$$

where $\mathbf{A}$ and $\mathbf{B}$ are general $n \times n$- matrices.

Following chapter X of [15] a few methods to treat such generalized eigenproblems are outlined. A common strategy for solving the large-scaled generalized eigenvalue problem is to find a reduction to standard form, i.e. a standard eigenvalue problem like in (3.35), and then apply the routines from the previous sections, like Arnoldi or Lanczos. In order to get to standard-form, a linear system involving one or both matrices has to be solved in each iteration. The mentioned reduction to a standard eigenvalue problem can be carried out using the following routines. An obvious possibility is to compute the inverse of the matrix on the right-hand side, $\mathbf{B}$, presuming that $\mathbf{B}$ is nonsingular and well conditioned. This procedure is then equivalent to

$$\left(\mathbf{B}^{-1}\mathbf{A}\right)\mathbf{v}_i = \lambda_i \mathbf{v}_i. \tag{3.48}$$

In iterative schemes, however, there is no need to explicitly compute the matrix product $\mathbf{B}^{-1}\mathbf{A}$, since only its action on a vector is required. This is important for sparse matrices, where the above product would generally be dense. Thus, all that is required, is the computation of matrix-vector products, like

$$\mathbf{p} = \left(\mathbf{B}^{-1}\mathbf{A}\right)\mathbf{q}, \tag{3.49}$$

for a given vector $\mathbf{q}$. This computation is most efficiently carried out via a single LU-factorization of the matrix $\mathbf{B}$, i.e. $\mathbf{B} = \mathbf{L}\mathbf{U}$. When evaluating the matrix-vector product of (3.49), one solves

$\mathbf{p}$ from $\mathbf{Bp} = \mathbf{Aq}$ applying forward backward substitutions with the lower and upper triangular matrices $\mathbf{L}$ and $\mathbf{U}$.

The rather expensive LU-factorization can be pre-calculated and the resulting triangular matrices are stored. Attention has to be devoted to the tempting strategy of considering an approximate solution of systems like $\mathbf{Bp} = \mathbf{q}$, instead of the costly LU-factorization. In this case the obtained approximate solution might only satisfy some nearby system, with a matrix $\mathbf{B}'$ which might significantly differ from iteration to iteration in forming the Krylov subspace. Thus, this iterative solution technique can only be used, if it leads efficiently to high accuracy, comparable to a stable direct solution method [16].

If the matrix on the right-hand side is Hermitian, i.e. $\mathbf{B}^H = \mathbf{B}$, one can compute a sparse Cholesky decomposition

$$\mathbf{B} = \mathbf{L}\mathbf{L}^H, \tag{3.50}$$

where $\mathbf{L}$ represents a lower triangular matrix. As a consequence, the equivalent standard eigenvalue problem reads

$$\left(\mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{H-1}\right)\mathbf{L}^H\mathbf{v}_i = \lambda_i\mathbf{L}^H\mathbf{v}_i. \tag{3.51}$$

Of course, like in the previous case, the matrix product in the brackets should not be computed explicitly, since only its action on a given vector is of interest:

$$\mathbf{p} = \left(\mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{H-1}\right)\mathbf{q}. \tag{3.52}$$

The suggested reduction methods have the disadvantage that they can only be applied when the matrix $\mathbf{B}$ is well conditioned. If this is not the case and when one further wishes to look for eigenvalues close to a selected target, a so-called *shift and invert spectral transformation* can be carried out. For a user-specified shift σ the generalized eigenvalue problem can be transformed to

$$\mathbf{C}^{-1}\mathbf{B}\mathbf{v}_i = \mu_i\mathbf{v}_i, \tag{3.53}$$

where $\mathbf{C} = \mathbf{A} - \sigma\mathbf{B}$ and $\mu_i = \frac{1}{\lambda_i - \sigma}$. Note that the eigenvalues closest to the specified target will be transformed to the eigenvalues with largest magnitude and thence most routines will converge first to those eigenvalues. Furthermore, a proper shift not only amplifies the desired eigenvalues, but also might lead to a well-conditioned matrix $\mathbf{C}$.

Applying the *shift and invert spectral transformation* requires in each iteration the evaluation of matrix-vector products of the following form

$$\mathbf{p} = \mathbf{C}^{-1}\mathbf{B}\mathbf{q} = \left((\mathbf{A} - \sigma\mathbf{B})^{-1}\mathbf{B}\right)\mathbf{q}, \tag{3.54}$$

with $\mathbf{q}$ a given vector. Like in the previous cases, efficiency can be improved by the use of a pre- LU- factorization of the matrix $\mathbf{C} = \mathcal{LU}$. As a consequence, the matrix-vector product $\mathbf{p} = \mathbf{C}^{-1}\mathbf{Bq}$ can then be computed by first forming $\mathbf{v} = \mathbf{Bq}$, followed by solving $\mathcal{L}\mathbf{w} = \mathbf{v}$ for $\mathbf{w}$ and finally solving $\mathcal{U}\mathbf{p} = \mathbf{w}$ for $\mathbf{p}$. Again the expensive LU-factorization needs to be carried out once only.

Finally a short discussion is outlined how to handle quadratic eigenproblems like

$$\left(\lambda_i^2\mathbf{A} + \lambda_i\mathbf{B} + \mathbf{C}\right)\mathbf{v}_i = 0. \tag{3.55}$$

The most common way of solving such problems is to convert them to generalized eigenvalue problems which are treated by the routines presented above. Thereby one has to pay the price of increasing the dimensionality of the system. One possible way [16] is to define the vector $\mathbf{u}_i$

$$\mathbf{u}_i = \begin{pmatrix} \lambda_i\mathbf{v}_i \\ \mathbf{v}_i \end{pmatrix} \tag{3.56}$$

and then rewrite (3.55) to

$$\begin{pmatrix} -\mathbf{B} & \mathbf{C} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}\mathbf{u}_i = \lambda_i \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}\mathbf{u}_i. \tag{3.57}$$

It is noted that this is only one out of a large number of different possibilities of rewriting (3.55).

3.2.2.4 Jacobi-Davidson

Finally another algorithm, quite related to the previous two Krylov subspace methods, is the *Jacobi-Davidson* algorithm (JDA). Like *Arnoldi* or *Lanczos*, the JDA extracts eigenpair approximations residing in a search space, which is iteratively extended. Unlike Krylov subspace methods, this extension is performed by correction vectors which have to be solved approximately only, since there is no special structure to be taken care of. The JDA can be decomposed in several steps as illustrated in Algorithm 3 and outlined in detail in [17] and [18]. The aim is to extract eigenpairs $(\mathbf{\Lambda}, \mathbf{Q})$ of the matrix pencil $(\mathbf{A}, \mathbf{B})$. The matrix $\mathbf{\Lambda}$ is a diagonal matrix with the eigenvalues λ_i as diagonal entries and the matrix $\mathbf{Q}$ is composed of the eigenvectors $\mathbf{q}_i$ as columns.

The algorithm is broken down into several steps, which will be explained separately. The first step, line 4, is the one of extracting eigenvalues out of an already found subspace $\mathbf{V}$, e.g. by the use of a Ritz-Galerkin extraction. Thereby one tries to find an eigenpair candidate $(\theta_i, \mathbf{Vz}_i)$, whose residual $\mathbf{r}(\theta_i, \mathbf{Vz}_i) = (\mathbf{A} - \theta_i\mathbf{B})\mathbf{Vz}_i$ should be orthogonal to the search space $\mathbf{V}$. Thus

Algorithm 3 Jacobi-Davidson Method

```
1: JacobiDavidson $(\mathbf{A}, \mathbf{B}, \sigma, \mathbf{V}, \varepsilon, \mathbf{v}_0)\{$
2:     $\mathbf{V} = \mathbf{v}_0$   (Initialize the search space with the starting vector $\mathbf{v}_0$)
3:     do
4:       $[\theta, \mathbf{Z}] :=$ Extract$(\mathbf{A}, \mathbf{B}, \mathbf{V}, \sigma)$   (Find eigenpairs from the subspace $\mathbf{V}$)
5:       $[\mathbf{\Lambda}, \mathbf{Q}, \mathbf{V}, \theta, \mathbf{Z}, \mathbf{r}] :=$ Convergence/Deflation$(\theta, \mathbf{Z}, \varepsilon)$
           (Check for convergence of the suggested eigenpairs)
6:       Solve$(\mathbf{A} - \sigma\mathbf{B})\mathbf{c} \approx -\mathbf{r}$   (Expand the search space $\mathbf{V}$ by a new vector $\mathbf{c}$)
7:       $[\mathbf{V}, \theta, \mathbf{Z}] :=$ Restart$(\mathbf{V}, \theta, \mathbf{Z}, s_{min}, s_{max})$ (Control the size of the subspace)
8:     end
9: }
```

one has to solve a tiny EVP, involving the matrix $\mathbf{G}$, defined as $\mathbf{G} := \mathbf{V}^T\mathbf{A}\mathbf{V}$, since

$$\mathbf{V}^T\mathbf{r} = \mathbf{G}\mathbf{z}_i - \theta_i\mathbf{z}_i = \mathbf{0}. \tag{3.58}$$

The eigenpairs $(\theta_i, \mathbf{V}\mathbf{z}_i)$, termed Ritz pairs, have to undergo a test in line 5, where the residual norm is compared to a required tolerance ε. If

$$\|\mathbf{r}(\theta_i, \mathbf{V}\mathbf{z}_i)\|_2 = \|\mathbf{A}\mathbf{V}\mathbf{z}_i - \theta_i\mathbf{B}\mathbf{V}\mathbf{z}_i\|_2 \leq \varepsilon\|\mathbf{V}\mathbf{z}_i\|_2, \tag{3.59}$$

the tested eigenpair is considered to be accurate enough and added to the space of already converged eigenpairs. Thus θ_i is added to $\mathbf{\Lambda}$, and $\mathbf{V}\mathbf{z}_i$ is added to $\mathbf{Q}$. In order to not search twice in the same directions, all the components pointing into directions of the found eigenvector are purged, an action termed deflation. Alternatively, if the residual norm is above the threshold ε, the algorithm continues by expanding the search space $\mathbf{V}$, i.e. finding a new vector $\mathbf{c}$ which is added to the search space. This is done by solving the correction equation, see [19] for details,

$$\mathbf{P}^T(\mathbf{A} - \sigma\mathbf{B})\mathbf{P}\mathbf{c} = -\mathbf{P}^T\mathbf{r}(\theta, \mathbf{q}) = -\mathbf{r}(\theta, \mathbf{q}), \quad \hat{\mathbf{Q}}^T\mathbf{B}\mathbf{c} = 0, \tag{3.60}$$

where $\mathbf{q}$ is the most recent eigenvector and $\hat{\mathbf{Q}} = (\mathbf{Q}, \mathbf{q})$ is composed of the already accepted eigenvectors $\mathbf{Q}$ and $\mathbf{q}$. The matrix $\mathbf{P}$ is projecting into the $\mathbf{B}$-orthogonal complement of $\hat{\mathbf{Q}}$, ensuring that the new search direction $\mathbf{c}$ is $\mathbf{B}$-orthogonal to the already found eigenvectors. $\mathbf{B}$-orthogonality means that for two vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ the following relation holds:

$$\mathbf{v}_i^T\mathbf{B}\mathbf{v}_j = 0 \quad \text{for} \quad i \neq j. \tag{3.61}$$

The last step, line 6, is needed in order to limit storage requirements. There the search space $\mathbf{V}$ is reduced as soon it exceed an upper limit s_{max}. This is achieved by keeping the $s_{min} - 1$

best Ritz vectors. The whole process is iterated until the required number of eigenpairs has been found.

This rather qualitative description should outline the basic ideas of the algorithm. Details of how to extract eigenpairs of the subspace and how to solve the correction equation can be found in [18]. The method becomes interesting for practical applications when the dimensionality of the problem increases.

Part II

Applications

4 Eigenvalue Analysis of Electromagnetic Waveguides

This chapter is devoted to an eigenvalue analysis of electromagnetic waveguides. It is the first applications of the fundamentals outlined in the previous chapters. The content of this chapter has been presented by me at the COMPUMAG 2009 in Florianopolis, Brazil [20].

4.1 State of Research

Various numerical methods and different formulations have been presented in the past for solving dielectric waveguide problems. Typical methods are the method of moments [21], spectral-domain methods [22], finite difference methods [23], and finite element methods [24]. Among them, the latter has proved to be a very general and efficient tool. A serious problem for the description of electromagnetic field quantities involving finite elements is the occurrence of spurious modes. In order to overcome these difficulties mainly two approaches have been taken. One is to impose the divergence-free condition, mostly in the case of nodal finite elements approaches. The other is to use $H(curl)$-tangential elements that are capable of correctly representing the properties of the curl-curl operator [25].

In [26] a waveguiding structure showing lossy material properties is considered using triangular hybrid elements. A study covering the anisotropic case is presented in [27]. The work [28] suggests a method to overcome unreliabilities for low frequencies by presenting an algorithm for employing a tree-cotree splitting in order to accomplish an inexact Helmholtz decomposition. There, hierarchical elements are used.

The authors of [29] present a very general framework for the investigation of inhomogeneous waveguiding structures. In their work the above mentioned Helmholtz decomposition is also employed which gives rise to different possible gauges. The authors have presented three gauges and have shown a superior performance of one of the gauges, termed axial gauge, over a field description. The latter differs only by a scaling factor from a gauge where the scalar potential is set to zero. They also use hierarchical elements on triangles.

4.2 Formulation

Let us consider an $\mathbf{A} - V$-formulation for a source free region and isotropic, lossy material properties. Using ε_c as a complex quantity describing both permitivity (ε) and conductivity (σ),

$$\varepsilon_c = \varepsilon - j\frac{\sigma}{\omega}, \tag{4.1}$$

and sinusoidal time dependencies the Maxwell equations in the frequency domain can be written as

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H}, \tag{4.2a}$$

$$\nabla \times \mathbf{H} = j\omega\varepsilon_c\mathbf{E}, \tag{4.2b}$$

$$\nabla \cdot \mathbf{D} = 0, \tag{4.2c}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{4.2d}$$

Introducing a magnetic vector potential $\mathbf{A}$ and a scalar potential ϕ, the electromagnetic field quantities can be rewritten as

$$\mathbf{B} = \nabla \times \mathbf{A}, \tag{4.3a}$$

$$\mathbf{E} = -j\omega\mathbf{A} - c\nabla\phi, \tag{4.3b}$$

$$0 = \nabla \times \mu_r^{-1}\nabla \times \mathbf{A} - k_0\varepsilon_r\left(k_0\mathbf{A} - j\nabla\phi\right), \tag{4.3c}$$

$$0 = \nabla \cdot \varepsilon_r\left(k_0\mathbf{A} - j\nabla\phi\right), \tag{4.3d}$$

where c is the velocity of light and k_0 the free space wave number.

The material properties for a waveguiding structure being uniform in the z-direction and with the splitting of the nabla operator $\nabla \rightarrow \nabla_\tau - \gamma\hat{\mathbf{e}}_z$ the fields and the potentials can be described as

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}(x, y)e^{-\gamma z}e^{j\omega t}, \tag{4.4a}$$

$$\mathbf{H}(\mathbf{r}, t) = \mathbf{H}(x, y)e^{-\gamma z}e^{j\omega t}, \tag{4.4b}$$

$$\mathbf{A} = e^{-\gamma z}\left(\mathbf{A}_\tau(x, y) + A_z(x, y)\hat{\mathbf{e}}_z\right), \tag{4.4c}$$

$$\phi = e^{-\gamma z}V(x, y), \tag{4.4d}$$

with $\mathbf{r}$ being a space vector in the waveguide with its cross-section lying in the x, y-plane. $\gamma = \alpha + j\beta$ is the propagation constant built up by the attenuation constant α and the phase constant β.

As it is pointed out and explained in [28] and [29] it is very important to additionally split the transverse vector potential $\mathbf{A}_t$ into a solenoidal part A_τ^c and a transverse gradient $\nabla_\tau \psi$ in order to separate the null solution resulting from the curl-curl operator:

$$\mathbf{A}_\tau(x, y) = \mathbf{A}_\tau^c + \nabla_\tau \psi. \tag{4.5}$$

Taking the same approach and additionally consider a gauging with the gradient part of the transverse vector potential set to zero leads to the following system of partial differential equations

$$\begin{aligned}
0 &= \nabla_\tau \times \mu_r^{-1} \nabla_\tau \times \mathbf{A}_\tau^c \\
&\quad + \gamma^2 \mathbf{e}_z \times \mu_r^{-1} \mathbf{e}_z \times \mathbf{A}_\tau^c - k_0^2 \varepsilon_r \mathbf{A}_\tau^c \\
&\quad + \gamma \mathbf{e}_z \times \mu_r^{-1} \mathbf{e}_z \times \nabla_\tau A_z + k_0 \varepsilon_r \nabla_\tau (jV), && \text{(4.6a)} \\
0 &= -\gamma \mathbf{e}_z \cdot \left(\nabla_\tau \times \mu_r^{-1} \mathbf{e}_z \times \mathbf{A}_\tau^c \right) \\
&\quad - \mathbf{e}_z \cdot \left(\nabla_\tau \times \mu_r^{-1} \mathbf{e}_z \times \nabla_\tau A_z \right) \\
&\quad - k_0^2 \varepsilon_r A_z - \gamma k_0 \varepsilon_r (jV), && \text{(4.6b)} \\
0 &= \nabla_\tau \cdot \varepsilon_r \left(k_0 \mathbf{A}_\tau^c - \nabla_\tau (jV) \right) \\
&\quad - \gamma \varepsilon_r (jV) - k_0 A_z. && \text{(4.6c)}
\end{aligned}$$

In order to ensure a unique solution of the boundary value problem, the following boundary conditions are imposed for the simplest case of a perfect electric conductor (PEC):

$$\begin{cases} \mathbf{n} \times \mathbf{A}_\tau^c = 0 \\ \quad A_z = 0 \qquad \text{on PEC,} \\ \quad V = 0 \end{cases} \tag{4.7}$$

with $\mathbf{n}$ being the normal vector to the boundary.

4.3 Finite Element Representation

We represent A_z and V by H^1 basis functions (N_i) and $\mathbf{A}_\tau^c$ by $H(curl)$ basis functions $(\mathbf{N}_i)$ according to the function space definitions (3.24), presented in section 3.1.3. The trial functions

$\mathbf{a}_\tau$ and a_z can be expressed as

$$\mathbf{A}^c_\tau = \sum_{i=0}^{n_\tau} c_{\tau i}\mathbf{N}_i, \quad A_z = \sum_{i=0}^{n_z} c_{A_z i} N_i, \quad jV = \sum_{i=0}^{n_z} c_{(jV)i} N_i. \tag{4.8}$$

Applying Galerkin's method to (4.6a) - (4.6c) and considering the boundary conditions, the following polynomial eigenvalue equation is obtained:

$$\begin{aligned} \Bigg[&\begin{pmatrix} A_{AA} - k_0^2 B_{AA} & 0 & k_0 C_{AV} \\ 0 & k_0^2 H_{VV} - E_{VV} & 0 \\ k_0 C_{AV}^T & 0 & -D_{VV} \end{pmatrix} \\ &\qquad - \gamma \begin{pmatrix} 0 & F_{AV} & 0 \\ F_{AV}^T & 0 & -k_0 H_{VV} \\ 0 & -k_0 H_{VV} & 0 \end{pmatrix} \\ &\qquad - \gamma^2 \begin{pmatrix} G_{AA} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -H_{VV} \end{pmatrix} \Bigg] \begin{bmatrix} \mathbf{c}_{\mathbf{A}_\tau} \\ \mathbf{c}_{A_z} \\ \mathbf{c}_{jV} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \end{aligned} \tag{4.9}$$

where $\mathbf{c}_{\mathbf{A}_\tau}$ is the coefficient vector for the edge basis functions and $\mathbf{c}_{A_z}$, $\mathbf{c}_{jV}$ are the respective vectors for nodal basis functions. The matrices are given by

$$[\mathbf{A}_{AA}]_{ij} = \int_{\Omega_e} \nabla_t \times \mathbf{N}_i \cdot \mu_r^{-1} \nabla_t \times \mathbf{N}_j \, d\Omega, \tag{4.10a}$$

$$[\mathbf{B}_{AA}]_{ij} = \int_{\Omega_e} \mathbf{N}_i \cdot \varepsilon_r \mathbf{N}_j \, d\Omega, \tag{4.10b}$$

$$[\mathbf{C}_{AV}]_{ij} = \int_{\Omega_e} \mathbf{N}_i \cdot \varepsilon_r \nabla_t N_j \, d\Omega, \tag{4.10c}$$

$$[\mathbf{D}_{VV}]_{ij} = \int_{\Omega_e} \nabla_t N_i \cdot \varepsilon_r \nabla_t N_j \, d\Omega, \tag{4.10d}$$

$$[\mathbf{E}_{VV}]_{ij} = \int_{\Omega_e} (\hat{\mathbf{e}}_z \times \nabla_t N_i) \left(\mu_r^{-1} \hat{\mathbf{e}}_z \times \nabla_t N_j\right) d\Omega, \tag{4.10e}$$

$$[\mathbf{F}_{AV}]_{ij} = \int_{\Omega_e} (\hat{\mathbf{e}}_z \times \mathbf{N}_i) \cdot \left(\mu_r^{-1} \hat{\mathbf{e}}_z \times \nabla_t N_j\right) d\Omega, \tag{4.10f}$$

$$[\mathbf{G}_{AA}]_{ij} = \int_{\Omega_e} (\hat{\mathbf{e}}_z \times \mathbf{N}_i) \cdot \left(\mu_r^{-1} \hat{\mathbf{e}}_z \times \mathbf{N}_j\right) d\Omega, \tag{4.10g}$$

$$[\mathbf{H}_{VV}]_{ij} = \int_{\Omega_e} N_i \cdot \varepsilon_r N_j \, d\Omega. \tag{4.10h}$$

With the variable transformations $A_z =: \gamma g_1$ and $g_2 := -jV$ the quadratic eigenvalue problem can be linearized resulting in a generalized eigenvalue problem for γ^2:

$$\left[\begin{pmatrix} \mathbf{A}_{AA} - k_0^2\mathbf{B}_{AA} & 0 & -k_0\mathbf{C}_{AV} \\ 0 & 0 & 0 \\ -k_0\mathbf{C}_{AV}^T & 0 & -\mathbf{D}_{VV} \end{pmatrix} -\gamma^2 \begin{pmatrix} \mathbf{G}_{AA} & \mathbf{F}_{AV} & 0 \\ \mathbf{F}_{AV}^T & k_0^2\mathbf{H}_{VV} - \mathbf{E}_{VV} & k_0\mathbf{H}_{VV} \\ 0 & k_0\mathbf{H}_{VV} & -\mathbf{H}_{VV} \end{pmatrix}\right] \times \begin{bmatrix} \mathbf{c}_{A_\tau} \\ \mathbf{c}_{g_1} \\ \mathbf{c}_{g_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{4.11}$$

4.4 Basis Functions

The H^1-nodal basis functions are built from a linear combination of quadratic polynomials in the x-and y-directions. The $H(curl)$-edge basis functions can be expressed as a linear combination of the nodal basis functions $(f_i(\xi,\eta))$ times the gradient of the respective local coordinate. For the reference element they are obtained by

$$\mathbf{N}_i(\xi,\eta) = f_i(\xi,\eta)\cdot\nabla\xi \quad \text{for} \quad i = 1,5, \tag{4.12a}$$

$$\mathbf{N}_i(\xi,\eta) = f_i(\xi,\eta)\cdot\nabla\eta \quad \text{for} \quad i = 6,10, \tag{4.12b}$$

where ξ and η are the coordinates of the reference cell ranging from -1 to 1. The geometry of the finite elements used is depicted in Fig. 4.1.

4.5 Tree co-tree algorithm

The decomposition of the transverse vector potential into a rotational part and a gradient part is accomplished by a tree co-tree splitting of our finite element mesh [30]. Thereby, the set of tree-edges is removed from the finite element graph. The remaining ones, the co-tree edges, are then used to approximate the rotational part of the transverse vector potential. The nodes of the tree-edges would be used for describing the scalar potential if we had not gauged the transverse gradient part to zero in the first place. The algorithm of this splitting is described below.

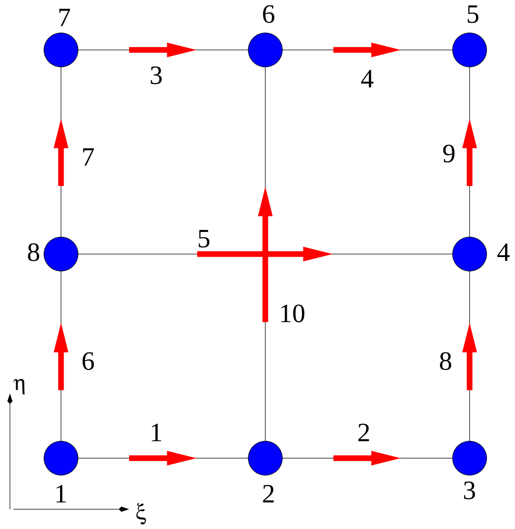


Figure 4.1: Geometry of an element, edges referring to the transverse vector potential are depicted in red, knots referring to the scalar potential in blue

Algorithm 4 Tree co-tree splitting

```
for every edge $\{i,j\}$, reset Mark($\{i,j\}$) = CO-TREE
for every node $V_i$, reset Mark($i$) =NOT DONE
add one inner node to set $Nodes$
**while** $Nodes \neq$ EMPTY
    Set $N_i$ last entry of $Nodes$ and remove it from $Nodes$
    for every inner node $V_j \in N_{neighbour}(i)$ and Mark($j$) = NOT DONE do
        Mark ($\{i,j\}$) = TREE
        Mark ($j$) = DONE
        Add $N_j$ to $Nodes$
**end while**
remove DOFs associated to tree-edges
```

4.6 Eigenvalue Solver

In order to solve the generally complex-symmetric eigenvalue problem we use a Krylov subspace method, namely the Arnoldi method described in section 3.2.2.1 implemented in the public domain software ARPACK. Since we are only interested in a small part of the spectrum, i.e. the dominant modes, and Krylov subspace methods tend to converge to extremal eigenvalues first, we spectrally transform our EVP with a shift-and-invert preconditioner as described below:

$$\mathbf{S}\mathbf{v} = \lambda \mathbf{T}\mathbf{v}, \tag{4.13a}$$

$$\varepsilon := \lambda - \lambda_G, \quad |\varepsilon| \ll \lambda, \tag{4.13b}$$

$$(\mathbf{S} - \lambda_G \mathbf{T})^{-1} T\mathbf{v} = \frac{1}{\varepsilon}\mathbf{v}, \tag{4.13c}$$

where λ_G has to be guessed. Since the action of the inverse has to be computed in every subspace iteration step, this step is done implicitly. The preconditioning matrix is thereby LU-factorized once in order to quickly solve the resulting linear systems by backward and forward substitutions. It is however mentioned that the LU-factorization is limited by memory restraints for very large problems. In such cases a Jacobi-Davidson method is preferable, as it is described in section 3.2.2.4

Special care has to be devoted to the orthogonalization against non-physical solutions as presented in detail in [29]. Although the non-physical solutions in this formulation all yield zero eigenvalues, this still constitutes a challenge at their respective cut-off frequencies. There the propagation constants tend to zero and the physical solution may mix with the approximated null-space eigenvalues. Therefore we employ an orthogonalization of the initial and iterated vectors of the following form

$$\mathbf{v}_i^T \mathbf{T} \mathbf{v}_k = 0 \quad \text{for } i \neq k. \tag{4.14}$$

For the generalized eigenvalue problem resulting from the ψ-gauge formulation, i.e. setting the solenoidal part of the vector potential to zero, the above condition (4.14) can be fulfilled through the following prescription for the initial and iterated vectors during the Arnoldi scheme [29]:

$$\mathbf{c}_{g_1} = (\mathbf{G}_{VV} - k_0^2 \mathbf{H}_{VV})^{-1} \cdot \left(\mathbf{F}_{AV}^T \mathbf{c}_{\mathbf{A}_\tau} + k_0 \mathbf{H}_{VV} \mathbf{c}_{g_2}\right). \tag{4.15}$$

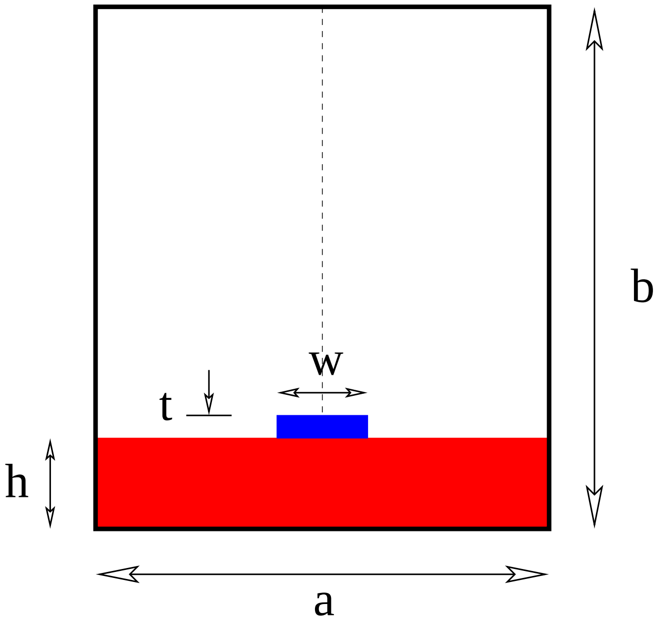


Figure 4.2: Geometry of the waveguide of the numerical example. The dimensions are $a = b = 2.5\,\mathrm{mm}$, $h = w = 0.25\,\mathrm{mm}$ and $t = 0.05\,\mathrm{mm}$. The electrical properties of the substrate are $\varepsilon_r = 9.0$, $\mu_r = 1$ and $\sigma = 0.05\,\mathrm{S/m}$

4.7 Numerical Application

As numerical example we consider the case of a microstrip in a dielectric medium. The geometry and the material properties of the substrate are chosen as described in [26] and sketched in Fig. 4.2. The width and the height of the cross-section is 2.5 mm, the height of the substrate 0.25 mm is equal to the width of the microstrip, and the thickness of the microstrip is taken to be 0.05 mm. The material properties of the substrate used in the simulation are $\varepsilon_r = 9.0$ and $\sigma = 0.05\,\mathrm{S/m}$. We are interested in the determination of the dominant propagation modes and therefore are only considering one half of the geometry due to symmetry, as done in [26] too. For the five dominant modes, the variation of the attenuation and the phase constant with the frequency is depicted in Fig. 4.3 and 4.4. Since the propagation constant is in general a complex quantity for lossy waveguides, there is an ambiguity for judging which mode is more dominant than the other. Here, this problem is resolved by simply comparing the values of the phase constant for different modes.

Due to the non-vanishing conductivity of the substrate the attenuation constant is different form zero even for propagating modes. One cannot clearly see this effect from Fig. 4.3 and 4.4. Therefore Fig. 4.5 is presented, where in addition a comparison is given to a calculation from [26]. The phase and attenuation constant of the dominant mode is depicted depend-

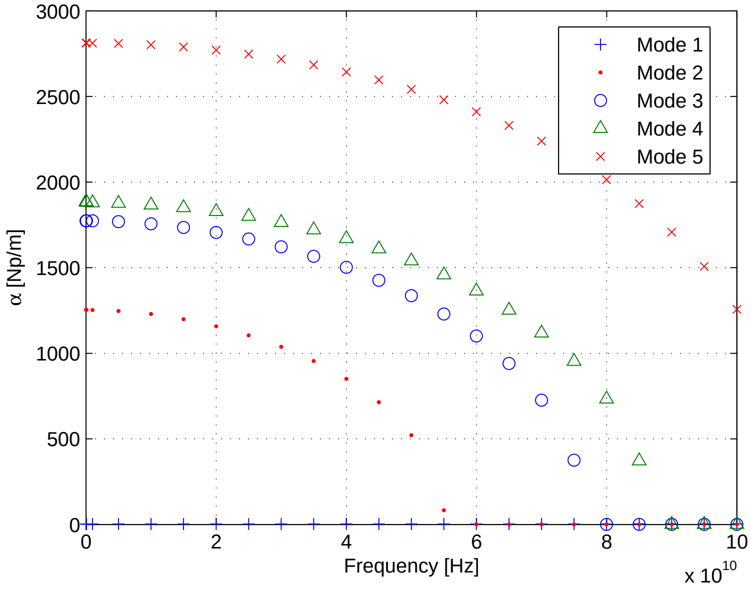


Figure 4.3: The dispersion of the attenuation constant for the five dominant modes

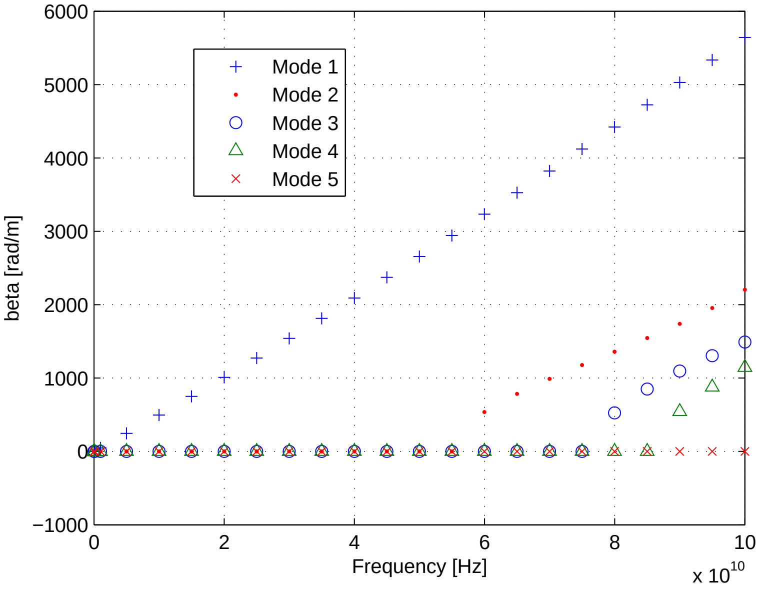
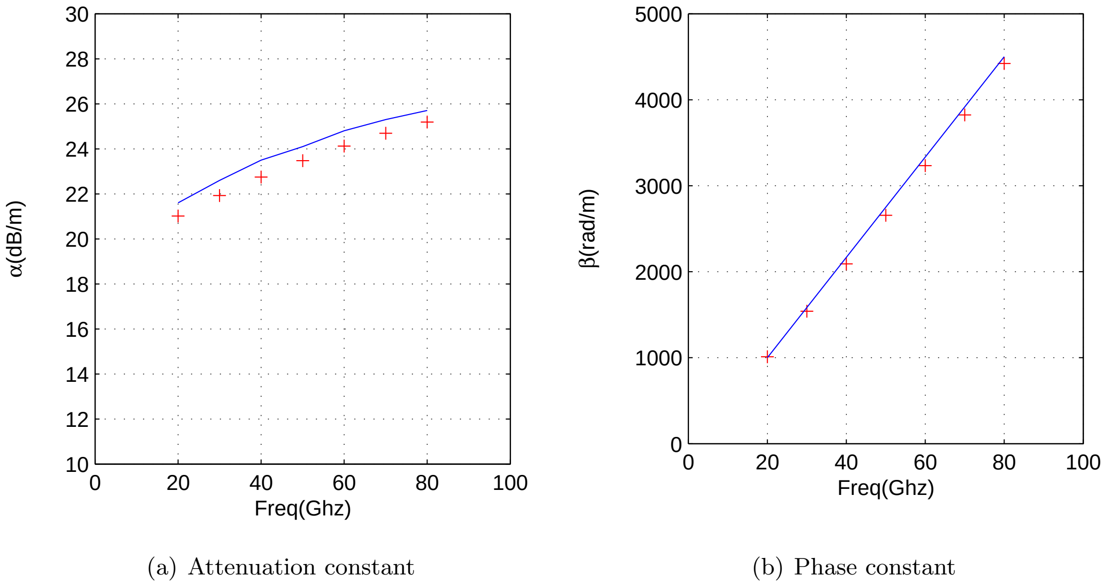


Figure 4.4: The dispersion of the phase constant for the five dominant modes



(a) Attenuation constant

(b) Phase constant

Figure 4.5: The dispersion of the attenuation and the phase constant of the microstrip line. The crosses represent our numerical solution compared to the results of [26].

ing on the frequency. In our calculations we are using a finite element mesh consisting of 1100 elements. This corresponds to a matrix size of 6501 after the enforcement of the homogeneous Dirichlet boundary conditions and the removal of the degrees of freedom related to tree edges. Considering the target value in the shift-and-invert preconditioner, we have chosen the strategy to take the square of the TE_{01} propagation constant of a fictitious rectangular waveguide homogeneously filled with a material having the highest permittivity of our problem. It is worth pointing out that even near the static limit and near cut-off frequencies of certain modes the results are correct.

4.8 New scientific results

In this chapter, a finite-element description for waveguiding structures has been presented. The formulation is capable of yielding solutions free from spurious modes. The work is motivated by the analysis [29], where the authors stressed the importance of an additional splitting of the transverse vector potential. As it is also pointed out there, this splitting gives rise to three possible gauging strategies, involving either the scalar potential, the axial component of the vector potential or the gradient part of the transverse vector potential. The novelty in the present work has been the investigation of the performance of the so called ψ gauge, where the gradient part of the transverse vector potential is set to zero. The formulation has been applied to a waveguiding structure consisting of materials allowing for losses. In addition, a tree-cotree algorithm has been used to accomplish the decomposition of the transverse vector potential into a rotational and a gradient part, and rectangular second-order finite-element basis functions are used. As demonstrated above, the formulation leads to satisfying results, even in the static limit. For future studies, it is certainly of great interest to additionally consider open waveguide problems, including losses due to radiation.

5 Photonic Crystals

Photonic crystals - the optical materials for the future? In recent years numerous scientists have developed novel materials in which light waves show a very similar behaviour to that of electronic signals in semiconductor crystals. This motivates the above question as a possible shift from the electronic to photonic era. Indeed, if one could manipulate light in an as powerful way as one is able to with electronic devices, one could open the door to many fascinating phenomena and promising applications [31].

After summarizing and outlining the basic concepts of photonic crystals and their mathematical description, an overview of photonic band structure calculations is given. The application of the finite-element method is the one chosen in this work and together with an incorporation scheme of periodic boundary conditions the details are described for both the 2D and 3D case. The computation of band structure diagrams requires the repeated calculation of numerically expensive eigenvalue problems. To improve the performance of these computations different model-order reduction schemes are presented and in detail discussed in section 5.4. These methods have in common that the full and expensive eigenvalue problem is only solved for selected parameter points, whereas for the remaining evaluation points a considerably reduced model is set up. This has a strong and positive impact on the run-times whereas the error levels remain comparable to the underlying FE calculation. Finally, the obtained results are systematically discussed in section 5.5.

5.1 Introduction and Fundamental Concepts

Closely following [31], a short introduction is given and the physical background of photonic crystals is demonstrated. In an *electronic* crystal of a semiconductor, electrons are interacting with a potential created by the periodic arrangement of atoms. The allowed energies of electrons in the crystal are given by so called bands which could be separated from each other. The concept of these possible energy gaps is fundamental for describing the fascinating physics of semiconductors. In 1987, Yablonovitch [32] carried over this concept to what would happen if one considered certain crystalline materials showing this behaviour for light waves. Such a wave, with a frequency within the bandgap of this material, could not propagate therein, independent of the direction of propagation. This is the concept of a photonic crystal opposed to the electronic counterpart of solid state physics.

Historically, in 1991 the authors of [33] have successfully constructed an artificial material showing a photonic band gap. They have been drilling holes in each geometrical dimension of a ceramic block. The honeycomb structure of the holes has resulted in a photonic band gap of centimeter wave length. A band gap for infrared light has been designed 10 years later by [34]. The authors have constructed artificial opal stones out of silicate bowls which are oriented in a face-centered cubic lattice structure. This material shows a bandgap at 1500nm. Some more revolutionary methods for the construction of photonic crystals can be found in [35],[36] and [37].

Photonic crystals are significant in various applications. An improvement of radiation characteristics of planar antennas is presented in [38] and [39]. The works of [40] and [41] show how photonic crystals can be used in laser technology. The way two dimensional photonic crystals can be built out of silicon or gallium arsenide is demonstrated in [42]. If a line defect is added to such a two dimensional crystal, the material can be used to assemble a perfect waveguide [43]. Waves with a frequency in the bandgap of the crystal can propagate along the defect. In this spirit it is even possible to bend waves in a crystal like it is shown in [44].

In nature, photonic crystals can be encountered for example in certain opal stones, whose iridescent colours are due to such materials. The same is true of various butterflies. Also diatoms as a part of plankton show these shining colors originated by photonic crystals.

5.2 State of Research

As stated in the introductory paragraphs above, photonic crystals are artificial materials built by periodic arrangements of identical unit cells. In analogy to solid-state physics, photonic crystals possess well-defined band structures and possibly bandgaps where electromagnetic waves cannot propagate. In order to engineer the microwave or optical properties [6] of such materials, the accurate knowledge of their band structures, i.e. their $k-\beta$ dispersion diagrams, is of utmost importance.

Many numerical methods for computing such band structures have been adapted from solid-state physics. They include, but are not restricted to, plane-wave expansions [45], the Korringa-Kohn-Rostoker method [46], and shell methodologies [47]. Recent developments include specialized versions of the finite-element (FE) method [48] and the flexible local approximation method [49].

The finite-element method, which stands out for its flexibility in modelling materials and complicated geometry, is the approach chosen in this work. The general procedure is to reduce the computational domain to one unit-cell with the help of periodic boundary conditions (PBCs), specify the value of the phase-coefficient β, and solve an eigenvalue problem for the dominant wavenumbers k_p. However, for the calculation of band structures, i.e., broadband k-β diagrams, the EVP has to be solved a great number of time for different values of β. In consequence, computational costs tend to become very high.

5.3 The photonic bandgap problem

This section gives a short introduction to the concept of the photonic bandgap focussing on computational aspects. This is a review of reported facts and follows mainly the respective chapters in the books [6] and [9].

The main equation for the analysis of these compounds is given by the vector Helmholtz equation (2.7)

$$\nabla \times \left(\frac{1}{p(\mathbf{r})} \nabla \times \mathbf{F}(\mathbf{r}) \right) = \omega^2 q(\mathbf{r}) \mathbf{F}, \tag{5.1}$$

where $\mathbf{F}$ represents either the electric field $\mathbf{E}$ or the magnetic field $\mathbf{H}$. In the first case p is the magnetic permeability and q the electric permittivity, whereas it is opposite in the latter case. ω and $\mathbf{r}$ stand for the angular frequency of the wave and the space coordinate vector, respectively.

Together with the divergence conditions

$$\nabla \cdot [\mu(\mathbf{r}) \mathbf{H}(\mathbf{r})] = 0, \tag{5.2a}$$

$$\nabla \cdot [\varepsilon(\mathbf{r}) \mathbf{E}(\mathbf{r})] = 0, \tag{5.2b}$$

(5.1) gives all information for the electromagnetic field quantities $\mathbf{E}$ and $\mathbf{H}$. The procedure is to solve the main equation for a given structure $\varepsilon(\mathbf{r})$ subject to the transversality requirement, find the modes ($\mathbf{F}(\mathbf{r})$) and the corresponding frequencies.

Due to the translational invariance of the crystal, Bloch's theorem can be applied and the whole system can be reduced to one unit cell which is infinitely many times repeated with the help of periodic boundary conditions. According to Bloch's theorem a solution to this translationally invariant problem is given by a periodic function multiplied with an exponential term, referred as the *Floquet* or *Bloch* term,

$$\mathbf{F_k}(\mathbf{r}) = \exp(\jmath \mathbf{k} \cdot \mathbf{r}) \cdot \mathbf{u_k}(\mathbf{r}), \tag{5.3}$$

$\mathbf{k}$ is the Bloch wave vector lying in the Brillouin zone, which itself only depends on the crystal structure. The function $\mathbf{u_k}$ is periodic for all lattice vectors $\mathbf{R}$, as illustrated in Fig. 2.2, so that

$$\mathbf{u_k}(\mathbf{r}) = \mathbf{u_k}(\mathbf{r} + \mathbf{R}) \tag{5.4}$$

is satisfied. For each value of the wave vector $\mathbf{k}$ the allowed frequencies ω are given by solving an eigenvalue problem. The obtained dispersion relation $\omega(\mathbf{k})$ is then the sought band structure of the photonic crystal. Note that one can solve the problem also in a reverse order, meaning that one specifies the frequency and solves for the Bloch vector. The chosen approach, however, is to solve for the frequencies. The reason for this will be apparent when the model-order reduction method is discussed.

5.3.1 Plane-Wave Expansion

Among a few methods, one way of computing band structures of photonic crystals is the use of a spectral method employing a planewave basis. In the following a short description of this method is presented. Details can be found in [45].

The idea is to describe the quantities of the governing equation as Fourier series. The periodic part of the electromagnetic quantity, $\mathbf{u}$, can then be written down as

$$\mathbf{u} = \sum_{\mathbf{m} \in \mathbb{Z}^2} \widetilde{\mathbf{u}}(\mathbf{k}_m) \exp(\jmath \mathbf{k_m} \cdot \mathbf{r}), \tag{5.5}$$

with $k_\mathbf{m} = \frac{2\pi}{a}\mathbf{m}$ and a being the crystal dimension. The Fourier coefficients, $\widetilde{\mathbf{u}}$, are yet to be determined. In the $\mathbf{E}$-formulation the full electrical field is given by

$$\mathbf{E} = \mathbf{u} \exp(\jmath \mathbf{K} \cdot \mathbf{r}) = \sum_{\mathbf{m} \in \mathbb{Z}^3} = \widetilde{\mathbf{u}}(\mathbf{m}) \exp(\jmath (\mathbf{k_m} - \mathbf{K}) \cdot \mathbf{r}). \tag{5.6}$$

Since the dielectric permittivity ε is also periodic it can be expanded into a Fourier series as well. To avoid coordinate-dependent coefficients on the right side of the equation, it is better to work with the inverse of the dielectric permittivity $\gamma = \varepsilon^{-1}$.

$$\gamma = \sum_{\mathbf{m} \in \mathbb{Z}^3} \widetilde{\gamma}(\mathbf{m}) \exp(\jmath \mathbf{k_m} \mathbf{r}). \tag{5.7}$$

Having the quantities at hand, a Fourier transform of (5.1) leads to the following equation:

$$\sum_{\mathbf{s}\in\mathbb{Z}^3} |\mathbf{k_m} - \mathbf{K}|^2 \widetilde{\gamma}(\mathbf{m}-\mathbf{s})\widetilde{\mathbf{u}}(\mathbf{s}) = \omega^2 \mu \widetilde{\mathbf{u}}(\mathbf{m}). \tag{5.8}$$

This last equation can be interpreted as an infinite eigenvalue problem which has to be truncated to a finite size in practice. The number of equations which are taken into account is a parameter for the whole computation. Speaking about accuracy it is worth mentioning that the method has to deal with Gibbs phenomenon stating that Fourier transformed quantities show oscillating behaviour at sharp material edges.

5.3.2 FEM for photonic band structure calculations

As already stated above, the finite-element method is the choice in this thesis. Due to its flexibility it is well suited for analyzing complicated structures including sharp material edges. The FE scheme can be applied only to the spatially periodic function $\mathbf{u}$ or directly to the full electromagnetic field quantity. In the latter case one has to explicitly enforce periodic boundary conditions in order to fulfill the Bloch boundary requirement. This will be illustrated in the following sections.

As it is explained above, the FE method considers the weak form of the governing differential equations and uses appropriate functional spaces for the approximation of the quantities in question. The right choice for these spaces depend on the special character of the differential operator. In a 2D analysis, the system is described by the scalar Helmholtz equation which can be approximated by ordinary scalar elements. In a full wave 3D consideration, however, vectorial edge elements, $H(curl)$ elements, have to be used, in order to correctly represent the properties of the vectorial Helmholtz equation. Additionally a further complication results in the rather large null space of the $curl - curl$-operator. This will be discussed in section 5.3.2.2 where the full wave analysis is presented.

The following subsections closely follow published work that I have authored during my PhD time. The first one considers a 2D-treatment of periodic structures, followed by a description of the full-wave case. The emphasis will be on a model-order reduction scheme allowing for an efficient computation of the quantities of interest, i.e. the band structures. The different order-reducing models are presented separately in the sections below.

5.3.2.1 Photonic bandgaps in 2D

We consider the two-dimensional case in the H plane, with lossless and isotropic material properties, $\partial_z \equiv 0$, and $\mathbf{E} = E_z\hat{\mathbf{e}}_z$ for the electrical field. For a rectangular unit cell Ω of size $D_x \times D_y$, the governing BVP reads

$$-\nabla \cdot \mu_r^{-1}\nabla E_z(x,y) - k_p^2\varepsilon_r E_z(x,y) = 0, \tag{5.9a}$$

$$\left.\begin{array}{rcl} E_z(D_x,y) &=& c_x E_z(0,y) \\ E_z(x,D_y) &=& c_y E_z(x,0) \end{array}\right\} \tag{5.9b}$$

$$\left.\begin{array}{rcl} \mu_r^{-1}\partial_x E_z(D_x,y) &=& c_x\mu_r^{-1}\partial_x E_z(0,y) \\ \mu_r^{-1}\partial_y E_z(x,D_y) &=& c_y\mu_r^{-1}\partial_y E_z(x,0) \end{array}\right\}. \tag{5.9c}$$

Here, $k_p = \omega_p/c$ stands for the wavenumber, and ω_p and c denote the angular frequency and the speed of light, respectively. The Floquet coefficients c_x and c_y, linking the fields on opposite boundaries, take the form

$$c_x = e^{-j\beta_x D_x}, \tag{5.10a}$$

$$c_y = e^{-j\beta_y D_y}, \tag{5.10b}$$

wherein β_x and β_y are the components of β in the coordinate directions, satisfying

$$\beta_x^2 + \beta_y^2 = \beta^2. \tag{5.11}$$

By enforcing (5.9b) as an essential interface condition on E_z and choosing test functions W with

$$W(D_x,y) = \frac{1}{c_x}W(0,y), \tag{5.12a}$$

$$W(x,D_y) = \frac{1}{c_y}W(x,0), \tag{5.12b}$$

the weak form [50] of the BVP (5.9) simplifies to

$$\int_\Omega \left(\nabla W \cdot \mu_r^{-1}\nabla E_z - k_p^2 W\varepsilon_r E_z\right) d\Omega = 0. \tag{5.13}$$

Note that (5.10), (5.12), and (5.13) provide a generalization of the formulation for periodically loaded waveguides of [51].

By restricting E_z and W to a space of FE basis functions N_i, the weak form (5.13) leads to a generalized algebraic EVP of the type

$$\mathbf{S}(c_x,c_y)\mathbf{e}_p = -k_p^2\mathbf{T}(c_x,c_y)\mathbf{e}_p, \tag{5.14}$$

wherein $\mathbf{S}(c_x, c_y)$ and $\mathbf{T}(c_x, c_y)$ are parameter-dependent stiffness and mass matrices, respectively, and $\mathbf{e}$ denotes the FE solution vector. Provided that the FE mesh is conforming in the sense that there exist matching pairs of nodes and edges on each of the two surfaces of a periodic boundary, the constraints (5.9b) and (5.12) are easy to enforce, by eliminating one half of the periodic boundaries. Let the subscripts (i, W, E, S, N, c) denote the FE degrees of freedom in the interior (i) of the domain, on the western (W), eastern (E), southern (S) and northern (N) boundaries, and in the four corners, respectively, see Fig. 5.1. Then, the PBC (5.9b) leads to

$$\mathbf{e} = \mathbf{P}\bar{\mathbf{e}}, \tag{5.15}$$

with

$$\mathbf{e} = \begin{bmatrix} \mathbf{e}_i \\ \mathbf{e}_W \\ \mathbf{e}_S \\ \mathbf{e}_E \\ \mathbf{e}_N \\ e_{c_1} \\ e_{c_2} \\ e_{c_3} \\ e_{c_4} \end{bmatrix}, \quad \bar{\mathbf{e}} = \begin{bmatrix} \mathbf{e}_i \\ \mathbf{e}_W \\ \mathbf{e}_S \\ e_{c_1} \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} \mathbf{I} & 0 & 0 & 0 \\ 0 & \mathbf{I} & 0 & 0 \\ 0 & 0 & \mathbf{I} & 0 \\ 0 & c_x\mathbf{I} & 0 & 0 \\ 0 & 0 & c_y\mathbf{I} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & c_x \\ 0 & 0 & 0 & c_y \\ 0 & 0 & 0 & c_x c_y \end{bmatrix}. \tag{5.16}$$

In (5.16), $\mathbf{I}$ stands for the unity matrix. Similarly, the constraints on the test functions (5.12) result in a restriction matrix $\mathbf{R}(c_x, c_y)$ defined by

$$\mathbf{R} = \begin{bmatrix} \mathbf{I} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{I} & 0 & \frac{1}{c_x}\mathbf{I} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{I} & 0 & \frac{1}{c_y}\mathbf{I} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{c_x} & \frac{1}{c_y} & \frac{1}{c_x c_y} \end{bmatrix}. \tag{5.17}$$

By (5.10), c_x and c_y lie on the unit circle. Hence we have

$$1/c_x = c_x^*, \qquad 1/c_y = c_y^*, \tag{5.18}$$

$$\mathbf{R}(c_x, c_y) = \mathbf{P}^H(c_x, c_y). \tag{5.19}$$

Here, superscript * and ^{H} denote complex conjugate and conjugate transpose, respectively. By substituting (5.16) and (5.17) for the solution vector and test functions in (5.14), we arrive at the final EVP

$$\bar{\mathbf{S}}(c_x, c_y)\bar{\mathbf{e}}_p = k_p^2 \bar{\mathbf{T}}(c_x, c_y)\bar{\mathbf{e}}_p. \tag{5.20}$$

In view of (5.19), the system matrices in (5.20) are given by

$$\bar{\mathbf{S}}(c_x, c_y) = \mathbf{P}^H(c_x, c_y)\mathbf{S}^{FE}\mathbf{P}(c_x, c_y), \tag{5.21a}$$

$$\bar{\mathbf{T}}(c_x, c_y) = \mathbf{P}^H(c_x, c_y)\mathbf{T}^{FE}\mathbf{P}(c_x, c_y), \tag{5.21b}$$

with

$$\mathbf{S}_{ij}^{FE} = \int_\Omega \nabla N_i \cdot \mu_r^{-1} \nabla N_j d\Omega, \tag{5.22a}$$

$$\mathbf{T}_{ij}^{FE} = \int_\Omega N_i \varepsilon_r N_j d\Omega. \tag{5.22b}$$

Since the FE matrices of (5.22) are real-valued and symmetric, $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ become Hermitian. They may also be represented as

$$\begin{aligned}\bar{\mathbf{S}}(c_x, c_y) = \mathbf{S}_0 &+ \left(c_x \mathbf{S}_1 + c_x^* \mathbf{S}_1^T\right) + \left(c_y \mathbf{S}_2 + c_y^* \mathbf{S}_2^T\right) \\ &+ \left(c_x c_y \mathbf{S}_3 + c_x^* c_y^* \mathbf{S}_3^T\right) + \left(c_x c_y^* \mathbf{S}_4 + c_y c_x^* \mathbf{S}_4^T\right),\end{aligned} \tag{5.23a}$$

$$\begin{aligned}\bar{\mathbf{T}}(c_x, c_y) = \mathbf{T}_0 &+ \left(c_x \mathbf{T}_1 + c_x^* \mathbf{T}_1^T\right) + \left(c_y \mathbf{T}_2 + c_y^* \mathbf{T}_2^T\right) \\ &+ \left(c_x c_y \mathbf{T}_3 + c_x^* c_y^* \mathbf{T}_3^T\right) + \left(c_x c_y^* \mathbf{T}_4 + c_y c_x^* \mathbf{T}_4\right).\end{aligned} \tag{5.23b}$$

Here, the matrices $\mathbf{S}_0, \ldots, \mathbf{S}_4$ and $\mathbf{T}_0 \ldots \mathbf{T}_4$ are obtained by partitioning the FE matrices of (5.22), performing the pre- and post-multiplications in (5.21), and collecting terms of equal parameter-dependence. Eq. (5.23) is particularly useful for band structure computations, where the EVP (5.20) has to be solved a great number of times, for different values of c_x and c_y, according to the parameters β_x and β_y, respectively. Since the matrices $\mathbf{S}_i$ and $\mathbf{T}_i$ in (5.23) are parameter-independent, most of the assembly process needs to be carried out only once. The PBC formulation of [52] is reported to deliver similar results regarding accuracy and computational cost.

Our FE code is based on hierarchical basis functions of third order and uses the ARPACK implementation of the Arnoldi method with shift-and-invert preconditioning to solve for the dominating eigenvalues of the EVP (5.20).

To track a specific mode $(k_p^2, \bar{\mathbf{e}}_p)$ over a wide range of phase coefficients, even when cross-over occurs as in Fig. 5.9, we exploit the modal orthogonality equation

$$\bar{\mathbf{e}}_i^H \bar{\mathbf{T}} \bar{\mathbf{e}}_j = \delta_{ij}, \tag{5.24}$$

which holds for any Hermitian EVP. Provided that consecutive evaluation points $(c_x, c_y)_n$ are sufficiently close, the dominant eigenvectors do not vary substantially, and the approximations

$$\bar{\mathbf{e}}_i^H((c_x, c_y)_{n-1})\,\bar{\mathbf{T}}((c_x, c_y)_n)\,\bar{\mathbf{e}}_i((c_x, c_y)_n) \approx 1, \tag{5.25a}$$

$$\bar{\mathbf{e}}_i^H((c_x, c_y)_{n-1})\,\bar{\mathbf{T}}((c_x, c_y)_n)\,\bar{\mathbf{e}}_j((c_x, c_y)_n) \approx 0, \tag{5.25b}$$

provide a good indicator for tracking a mode pattern from one evaluation point to the next.

5.3.2.2 Photonic Bandgaps in 3D

In contrast to the formulation in the H-plane, in the 3d-case the governing equation is the vectorial Helmholtz equation

$$\nabla \times \left(\tfrac{1}{\mu_r}\nabla \times \mathbf{E}\right) - \varepsilon_r k_0^2 \mathbf{E} = 0. \tag{5.26}$$

Again we consider one unit cell of the photonic crystal and link the the cells with the help of periodic boundary conditions. These constraints are then given by

$$\mathbf{E}_t(D_x, y, z) = c_x \mathbf{E}_t(0, y, z), \tag{5.27a}$$

$$\mathbf{E}_t(x, D_y, z) = c_y \mathbf{E}_t(x, 0, z), \tag{5.27b}$$

$$\mathbf{E}_t(x, y, D_z) = c_z \mathbf{E}_t(x, y, 0), \tag{5.27c}$$

$$\left(\tfrac{1}{\mu_r}\nabla \times \mathbf{E}\right)_t (D_x, y, z) = c_x \left(\tfrac{1}{\mu_r}\nabla \times \mathbf{E}\right)_t (0, y, z), \tag{5.27d}$$

$$\left(\tfrac{1}{\mu_r}\nabla \times \mathbf{E}\right)_t (x, D_y, z) = c_y \left(\tfrac{1}{\mu_r}\nabla \times \mathbf{E}\right)_t (x, 0, z), \tag{5.27e}$$

$$\left(\tfrac{1}{\mu_r}\nabla \times \mathbf{E}\right)_t (x, y, D_z) = c_z \left(\tfrac{1}{\mu_r}\nabla \times \mathbf{E}\right)_t (x, y, 0). \tag{5.27f}$$

Here, $c_x = \exp(-j\beta_x D_x)$ represents the Floquet coefficient.

In a similar way, these periodic conditions can be enforced by choosing test functions satisfying

$$\mathbf{W}_t(D_x, y, z) = \tfrac{1}{c_x}\mathbf{W}_t(0, y, z), \tag{5.28}$$

$$\mathbf{W}_t(x, D_y, z) = \tfrac{1}{c_y}\mathbf{W}_t(x, 0, z), \tag{5.29}$$

$$\mathbf{W}_t(x, y, D_z) = \tfrac{1}{c_z}\mathbf{W}_t(x, y, 0). \tag{5.30}$$

The weak form then simplifies to

$$\int_\Omega \left(\nabla \times \mathbf{W} \tfrac{1}{\mu_r} \nabla \times \mathbf{E} - k_0^2 \mathbf{W} \cdot \varepsilon_r \mathbf{E}\right) d\Omega = 0. \tag{5.31}$$

That the last relation holds can be easily verified [48]. The weak form of (5.26) is obtained by applying the vector form of Green's theorem to the residual

$$R = \int_\Omega \Big(\nabla \times \mathbf{W} \cdot \tfrac{1}{\mu_r} \nabla \times \mathbf{E} - k_0^2 \mathbf{W} \cdot \varepsilon_r \mathbf{E} \Big)\, d\Omega. \tag{5.32}$$

This gives

$$R = \int_\Omega \mathbf{W} \cdot \Big(\nabla \times \tfrac{1}{\mu_r} \nabla \times \mathbf{E} - k_0^2 \varepsilon_r \mathbf{E} \Big)\, d\Omega - \oint_{\partial\Omega} \Big(\mathbf{W} \times \tfrac{1}{\mu_r} \nabla \times \mathbf{E} \Big) \cdot \mathbf{n}\, dS. \tag{5.33}$$

Setting the last equation to zero requires that the curl-curl equation (5.26) is fulfilled and that the surface integral above vanishes. The latter is the case when the test functions are chosen periodically according to (5.28), since the parts of the surface integral on opposite boundaries cancel. For the x-direction and with the normal vector $\mathbf{n}$ on one boundary being anti-parallel to the counterpart on the other boundary, the boundary integral reads

$$\int_{x=0} \mathbf{n} \times \mathbf{W}_t(0,y,z) \cdot \left[\Big(\tfrac{1}{\mu_r} \nabla \times \mathbf{E} \Big)_t (0,y,z) - \frac{1}{c_x} \Big(\tfrac{1}{\mu_r} \nabla \times \mathbf{E} \Big)_t (D_x,y,z) \right] dS = 0. \tag{5.34}$$

Since this relation remains true for all test functions, the natural boundary condition (5.27d) follows.

Restricting the trial functions $\mathbf{E}$ and the test functions $\mathbf{W}$ to the space of vectorial curl-conforming FE basis functions $\mathbf{N}_i$ results again in a generalized eigenvalue problem of the form

$$\mathbf{S}(c_x, c_y, c_z)\mathbf{e}_p = k_p^2 \mathbf{T}(c_x, c_y, c_z)\mathbf{e}_p, \tag{5.35}$$

with the stiffness and mass matrix defined as

$$S_{ij} = \int_\Omega \nabla \times \mathbf{N}_i \cdot \tfrac{1}{\mu_r} \nabla \times \mathbf{N}_j d\Omega, \tag{5.36}$$

$$T_{ij} = \int_\Omega \mathbf{N}_i \cdot \varepsilon_r \mathbf{N}_j d\Omega. \tag{5.37}$$

The implementation of the periodic boundary conditions follows in complete analogy to the 2d case. The usual principle is to reorder the degrees of freedom into blocks of the same type of boundary followed by an elimination of the slave variables. Their contribution is then linked to the master variables with the help of the Floquet coefficients. The complexity, however, slightly increases for the 3d case given the existence of 27 different types of boundaries. To close the picture let us recall the procedure for singly and doubly periodic structures. In the first case, there is only one pair of periodic boundaries, i.e. three different types. Thus with the notation of section 5.3.2.1 the vector of variables and the expansion matrices have the form

$$\mathbf{e} = \begin{bmatrix} \mathbf{e}_i \\ \mathbf{e}_- \\ \mathbf{e}_+ \end{bmatrix}, \bar{\mathbf{e}} = \begin{bmatrix} \mathbf{e}_i \\ \mathbf{e}_- \end{bmatrix}, \mathbf{P} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{I} \\ 0 & c_x\mathbf{I} \end{bmatrix}, \mathbf{R} = \begin{bmatrix} \mathbf{I} & 0 & 0 \\ 0 & \mathbf{I} & \frac{1}{c_x}\mathbf{I} \end{bmatrix}, \tag{5.38}$$

where the subscript $-$ denotes the master, $+$ the slave boundary and i stands for the inner degrees of freedom.

For the model-order reduction purposes of the following chapters the matrices are needed in a Floquet-parameter independent form. This is obtained by substituting (5.38) for the solution vector. The generalized eigenvalue problem then reads

$$\left(\mathbf{S}_0 + c_x\mathbf{S}_1 + \tfrac{1}{c_x}\mathbf{S}_1^T\right)\bar{\mathbf{e}}_p = k_p^2\left(\mathbf{T}_0 + c_x\mathbf{T}_1 + \tfrac{1}{c_x}\mathbf{T}_1^T\right)\bar{\mathbf{e}}_p, \tag{5.39}$$

with the reduced matrices $\mathbf{S}_0$, $\mathbf{S}_1$ given as

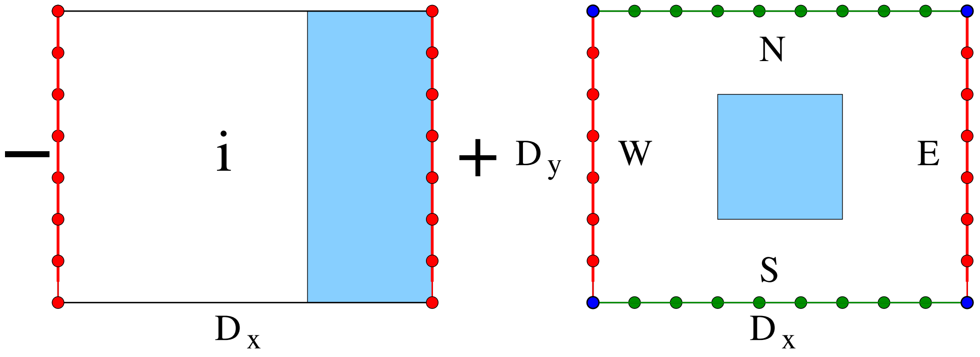
$$\mathbf{S}_0 = \begin{bmatrix} \mathbf{S}_{ii} & \mathbf{S}_{i-} \\ \mathbf{S}_{-i} & \mathbf{S}_{--} + \mathbf{S}_{++}. \end{bmatrix}, \quad \mathbf{S}_1 = \begin{bmatrix} 0 & \mathbf{S}_{i+} \\ 0 & \mathbf{S}_{-+} \end{bmatrix}. \tag{5.40}$$

The expressions for the mass submatrices are calculated similarly.

In the case of two pairs of periodic boundaries, there are nine different types. Variables on the corner of the unit cell have to be treated separately since they are subject to both periodic counterparts. With the expansion matrices of section 5.3.2.1 the parameter independent EVP is now built out of $3^2 = 9$ different terms

$$\begin{aligned} &\left(\mathbf{S}_0 + c_x\mathbf{S}_1 + \tfrac{1}{c_x}\mathbf{S}_1^T + c_y\mathbf{S}_2 + \tfrac{1}{c_y}\mathbf{S}_2^T + c_xc_y\mathbf{S}_3 + \tfrac{1}{c_xc_y}\mathbf{S}_3^T + \tfrac{c_x}{c_y}\mathbf{S}_4 + \tfrac{c_y}{c_x}\mathbf{S}_4^T\right)\bar{\mathbf{e}}_p = \\ k_p^2 &\left(\mathbf{T}_0 + c_x\mathbf{T}_1 + \tfrac{1}{c_x}\mathbf{T}_1^T + c_y\mathbf{T}_2 + \tfrac{1}{c_y}\mathbf{T}_2^T + c_xc_y\mathbf{T}_3 + \tfrac{1}{c_xc_y}\mathbf{T}_3^T + \tfrac{c_x}{c_y}\mathbf{T}_4 + \tfrac{c_y}{c_x}\mathbf{T}_4^T\right)\bar{\mathbf{e}}_p. \end{aligned} \tag{5.41}$$

With the notation of Fig. 5.1 the parameter independent matrices $\mathbf{S}_0, \ldots, \mathbf{S}_4$ are related to



(a) Singly periodic

(b) Doubly periodic

Figure 5.1: Unit cell for the singly and doubly periodic unit cell

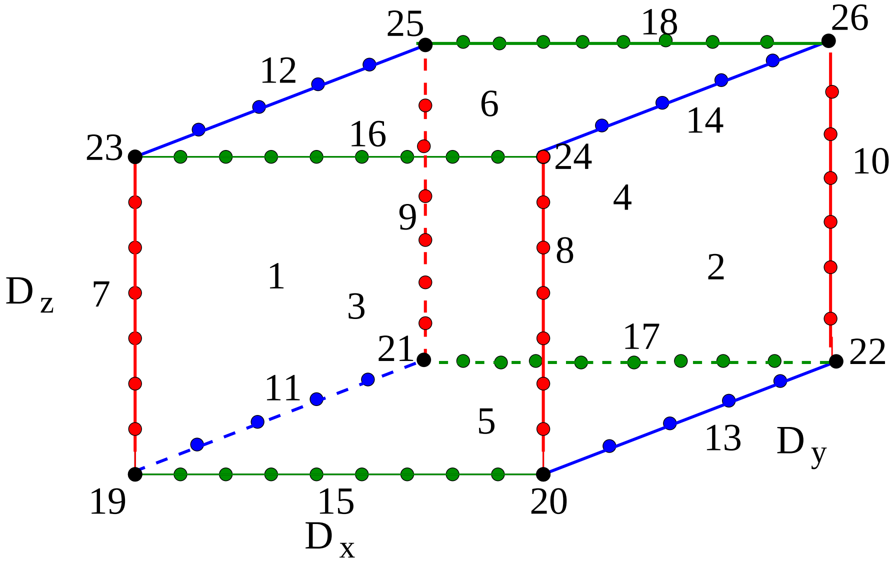


Figure 5.2: Unit cell for the triply periodic unit cell

the original FE matrix $\mathbf{S}$ by

$$\mathbf{S}_0 = \begin{bmatrix} \mathbf{S_{ii}} & \mathbf{S_{iW}} & \mathbf{S_{iS}} & \mathbf{S_{ic_1}} \\ \mathbf{S_{Wi}} & \mathbf{S_{WW}} + \mathbf{S_{EE}} & \mathbf{S_{WS}} & \mathbf{S_{Wc_1}} + \mathbf{S_{Ec_2}} \\ \mathbf{S_{Si}} & \mathbf{S_{SW}} & \mathbf{S_{SS}} + \mathbf{S_{NN}} & \mathbf{S_{Sc_1}} + \mathbf{S_{Nc_3}} \\ \mathbf{S_{c_1i}} & \mathbf{S_{c_1W}} + \mathbf{S_{c_1E}} & \mathbf{S_{c_1S}} + \mathbf{S_{c_1N}} & \sum_i S_{c_ic_i} \end{bmatrix}, \tag{5.42a}$$

$$\mathbf{S}_1 = \begin{bmatrix} \mathbf{0} & \mathbf{S_{iE}} & \mathbf{0} & \mathbf{S_{ic_2}} \\ \mathbf{0} & \mathbf{S_{WE}} & \mathbf{0} & \mathbf{S_{Wc_2}} \\ \mathbf{0} & \mathbf{S_{SE}} & \mathbf{0} & \mathbf{S_{Sc_2}} + \mathbf{S_{Nc_4}} \\ \mathbf{0} & \mathbf{S_{c_1E}} & \mathbf{0} & S_{c_1c_2} + S_{c_3c_4} \end{bmatrix}, \tag{5.42b}$$

$$\mathbf{S}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{S_{iN}} & \mathbf{S_{ic_3}} \\ \mathbf{0} & \mathbf{0} & \mathbf{S_{WN}} & \mathbf{S_{Wc_3}} + \mathbf{S_{Ec_4}} \\ \mathbf{0} & \mathbf{0} & \mathbf{S_{SN}} & \mathbf{S_{Sc_3}} \\ \mathbf{0} & \mathbf{0} & \mathbf{S_{c_1N}} & S_{c_1c_3} + S_{c_2c_4} \end{bmatrix}, \tag{5.42c}$$

$$\mathbf{S}_3 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{S_{ic_4}} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{S_{Wc_4}} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{S_{Sc_4}} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & S_{c_1c_4} \end{bmatrix}, \tag{5.42d}$$

$$\mathbf{S}_4 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S_{NE}} & \mathbf{0} & \mathbf{S_{Nc_2}} \\ \mathbf{0} & \mathbf{S_{c_3E}} & \mathbf{0} & S_{c_3c_2} \end{bmatrix}. \tag{5.42e}$$

In a completely similar way one calculates the expressions for the mass matrices $\mathbf{T}_0, \cdots, \mathbf{T}_4$.

If we now consider the most general case of a triply periodic unit cell, i.e. with three pairs of periodic boundaries, we have 27 types of boundaries. This is illustrated in Fig. 5.2. In this case it is no longer useful to directly evaluate the expansion matrices $\mathbf{P}$ and $\mathbf{R}$, but to separate the matrices directly in the assembly process. The parameter independent form is then composed of $3^3 = 27$ entries

$$\left(\sum_{k,l,m=\{-1,0,1\}} c_x^k c_y^l c_z^m \mathbf{S}_{klm} \right) \bar{\mathbf{e}}_p = k_p^2 \left(\sum_{k,l,m=\{-1,0,1\}} c_x^k c_y^l c_z^m \mathbf{T}_{klm} \right) \bar{\mathbf{e}}_p. \tag{5.43}$$

5.3.2.3 Solving the eigenvalue problem

Finally we are again confronted with the problem of finding a few selected eigen-functions of a sparse generalized eigenvalue problem. Given the specific null space of the curl-curl operator a orthogonalization scheme is required that filters out all solution with corresponding zero eigenvalue so that the Arnoldi algorithm will converge to the desired eigen-solutions.

Thus each suggested new vector in the Arnoldi iteration ($\bar{\mathbf{e}}$) has to be projected into the orthogonal complement of the null space, i.e.

$$\mathbf{e} = \bar{\mathbf{e}} - \mathbf{G}\mathbf{g}, \tag{5.44}$$

so that the resulting vector $\mathbf{e}$ fulfills

$$\mathbf{G}^T\mathbf{T}\mathbf{e} = 0. \tag{5.45}$$

Here $\mathbf{G}$ denotes the gradient matrix connecting nodes and edges in the finite-element mesh. The vector $\mathbf{g}$ remains to be calculated.

Inserting (5.44) into (5.45) leads to

$$\begin{aligned} \mathbf{G}^T\mathbf{T}\left(\bar{\mathbf{e}} - \mathbf{G}\mathbf{g}\right) &= 0 \\ \underbrace{\mathbf{G}^T\mathbf{T}\mathbf{G}}_{\mathbf{E}_{VV}}\mathbf{g} &= \underbrace{\mathbf{G}^T\mathbf{T}}_{\mathbf{C}_{AV}^T}\bar{\mathbf{e}}. \end{aligned} \tag{5.46}$$

Here the notation of section 4.3 is chosen for the matrices $\mathbf{E}_{VV}$ and $\mathbf{C}_{AV}$. Thus the required correction vector g can be calculated as

$$\mathbf{g} = \mathbf{E}_{VV}^{-1}\mathbf{C}_{AV}^T\bar{\mathbf{e}}. \tag{5.47}$$

In each Arnoldi iteration, such a correction has to be done. Of course, the matrix inversion need not be computed directly, since only its action on a vector is needed. Hence the matrix S_{VV} is factorized only once resulting in a very efficient solving routine. The matrices $\mathbf{E}_{VV}$ and $\mathbf{C}_{AV}$ can be calculated directly using the matrix $\mathbf{G}$.

There is, however, one complication to this scheme stemming from the fact that some variables are of slave nature. As a consequence there are some master edges having only one corresponding node. Since the eigenvalue problem is solved after eliminating the slave variables, we have to take care of the contribution of the edges pointing into slave nodes. Thus the gradient matrix $\mathbf{G}$ cannot simply be reduced to its master coordinates. We want to accomplish that the product $\mathbf{g}_E = \mathbf{G}\mathbf{g}_N$ results in a vector of active edge variables given a vector of active node variables. Before the truncation of the slave variables, the gradient matrix cast in blocks, reads:

$$\begin{bmatrix} \mathbf{g}_E^a \\ \mathbf{g}_E^s \end{bmatrix} = \begin{bmatrix} \mathbf{G}^{aa} & \mathbf{G}^{as} \\ \mathbf{G}^{sa} & \mathbf{G}^{ss} \end{bmatrix} \cdot \begin{bmatrix} g_N^a \\ \mathbf{g}_N^s \end{bmatrix}, \tag{5.48}$$

where the superscripts a and s denote active and slave variables, respectively. In the end, the vector $\mathbf{g}_E^a$ is sought given the input $\mathbf{g}_V^a$.

First the slave component of the the vector corresponding to the nodal variables has to be calculated. This vector is linked to the active variables by the respective Floquet coefficients, properly collected into the matrix $\mathbf{Z}^{sa}$:

$$\mathbf{g}_N^a = \mathbf{Z}^{sa} \cdot \mathbf{g}_V^a. \tag{5.49}$$

Once $\mathbf{Z}^{sa}$ is built, the sought vector $\mathbf{g}_E^a$ is then computed by the following relation

$$\mathbf{g}_E^a = \begin{bmatrix} \mathbf{G}^{aa} & \mathbf{G}^{as} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{Z}^{sa} \end{bmatrix} \mathbf{g}_V^a. \tag{5.50}$$

5.4 Model-Order Reduction

This chapter is devoted to the problem of an efficient evaluation of the eigenvalue problems of the previous sections for a set of varying parameters. This is needed to obtain a band structure diagram of a photonic crystal, but the goal here is to avoid a repetitive diagonalization of the usually big matrices for each point of the curve. The proposed outcome is to construct a specific reduced model that covers the necessary information but is much easier to solve. The field quantities of the original problem can then be recovered from the solutions of the reduced model.

In the following section, two different approaches are presented. The first uses solutions at selected so called expansion points to construct the model. The original problem is then projected onto the space spanned by the model and all the eigenvalue computations are performed with matrices of heavily reduced size. Of course, error levels should remain comparable to the underlying finite-element solutions. By the way of construction, it is obvious that this is the case in the expansion points, but the results below will demonstrate that for all evaluation points the error is reasonably low. The section is divided into subsections discussing the 2d and 3d case. As it is outlined above, the first case is captured by the scalar Helmholtz equation whereas the latter is described by its vectorial counterpart. This has a consequence for the reduced model, since the null space of the curl-curl operator will cause non-physical solutions in the parameter sweep even though the solutions at the expansion points are correctly computed. A possible way to deal with this fact will be presented below.

On the other hand, it is possible to construct a model-order reduction scheme out of only one full solution of the original problem. Further information for the model stems form a series expansion of the respective quantities followed by a mode matching procedure. This creates a system of equations which has to be solved only once. The solution of the latter, along with the solution of the original eigenvalue problem at the expansion point creates the reduced model. This serves as basis for the parameter sweep.

5.4.1 Multi-Point Approach

In the following the main ideas of multi-point approaches are presented. The formulations and results of the two-dimensional case have been presented at the CEFC-conference in Chicago 2010. This chapter follows quite closely the outlines of our published work [53]. After that the discussion is extended to the three dimensional full-wave case. The novelty of this work has been presented at the FEM workshop in Meredith 2010 [54].

5.4.1.1 Two dimensional case

Motivated by [55], the key idea of the multi-point approach of this work is to restrict the trial and test functions in (5.20) to the subspace spanned by the dominant P eigenvectors at a small number N of expansion points $(c_x, c_y)_n$. Hence the original EVP (5.20) needs to be solved N times only. From its solutions, we construct a unitary matrix $\mathbf{Q} \in \mathbb{C}^{M \times PN}$ with

$$\operatorname{colsp} \mathbf{Q} = \{\bar{\mathbf{e}}_{11}, \ldots, \bar{\mathbf{e}}_{PN}\}, \tag{5.51}$$

where colsp denotes the column space, restrict the solution to

$$\bar{\mathbf{e}}_p = \mathbf{Q}\widetilde{\mathbf{e}}_p, \tag{5.52}$$

and replace $\mathbf{S}_i$ and $\mathbf{T}_i$ in (5.23) by the Galerkin projections

$$\widetilde{\mathbf{S}}_i = \mathbf{Q}^H \mathbf{S}_i \mathbf{Q}, \tag{5.53}$$

$$\widetilde{\mathbf{T}}_i = \mathbf{Q}^H \mathbf{T}_i \mathbf{Q}, \tag{5.54}$$

with $i = 0, 1, 2, 3, 4$.

Thus, the ROM takes the form

$$(\widetilde{\mathbf{S}}_0 + c_x\widetilde{\mathbf{S}}_1 + c_x^*\widetilde{\mathbf{S}}_1^T + \cdots + c_y c_x^*\widetilde{\mathbf{S}}_4^T)\widetilde{\mathbf{e}}_p \;=\; k_p^2(\widetilde{\mathbf{T}}_0 + c_x\widetilde{\mathbf{T}}_1 + c_x^*\widetilde{\mathbf{T}}_1^T + \cdots + c_y c_x^*\widetilde{\mathbf{T}}_4^T)\widetilde{\mathbf{e}}_p. \tag{5.55}$$

Note that (5.55) preserves the parameterization and Hermitian structure of the original EVP. Since $PN \ll M$, the reduced EVP (5.55) is very fast to solve. The high speed of the proposed method offers the possibility to compute band structures not only along lines between symmetry points on the boundary of the reduced Brillouin zone [6], which are usually sufficient for bandgap calculations, but at any point in the (c_x, c_y) or (β_x, β_y) spaces, respectively.

5.4.1.2 Adaptivity issues

A priori, it is neither known how many expansion points are needed to produce an accurate ROM, nor where they should be placed. To fill this gap, an adaptive method is proposed that subdivides the parameter domain successively by placing a new point in the middle of the subregion that exhibits the worst error indicator for the eigenvalue k_p. For brevity, Algorithm 5 just presents the one-parameter case, with $\beta_y = 0$. First, a strictly increasing sequence of evaluation phase coefficients $\mathcal{B}$ is defined, then the FE problem is solved at the end points of the interval, followed by a modified Gram-Schmidt (MGS) step to generate the projection matrix $\mathbf{Q}_2$. Lines 5–8 compute the initial ROM, which is then solved at each evaluation point.

At the end of the initialization phase, the whole interval is marked as the region of worst error. The main loop starts with computing the location of the expansion point, splitting the associated sub-interval, and solving the FE system. In Lines 18–22, the ROM is updated. The following loop is over all evaluation points: it computes the solution of the ROM as well as a simple incremental error indicator Δk_p^q for mode p at iteration step q, defined by

$$\Delta k_p^q(c_x, c_y) = k_p^q(c_x, c_y) - k_p^{q-1}(c_x, c_y). \tag{5.56}$$

The algorithm tracks the three error measures

$$E_1(q) = \frac{1}{N_b \cdot N_e} \sum_{p=1}^{N_b} \sum_{i=1}^{N_e} |\Delta k_p^q((c_x, c_y)_i)|, \tag{5.57a}$$

$$E_2(q) = \left(\frac{1}{N_b \cdot N_e} \sum_{p=1}^{N_b} \sum_{i=1}^{N_e} |\Delta k_p^q((c_x, c_y)_i)|^2 \right)^{1/2}, \tag{5.57b}$$

$$E_\infty(q) = \max_{i,p} |\Delta k_p^q((c_x, c_y)_i)|, \tag{5.57c}$$

where N_b and N_e stand for the number of bands under consideration and the number of evaluation points, respectively. The measure E_∞ is taken as the error indicator, which is checked against a user-defined tolerance $E_{\infty,tol}$ in Line 28 to signal convergence.

5.4.1.3 Three dimensional full wave case

This section summarizes my work presented at the FEM 2010 workshop in Meredith, New Hampshire [54]. It can be regarded as a natural extension of the previous chapter to a full wave analysis. The novelty of this work is the application of a model-order reduction scheme to photonic band structure cases. The governing equation of the three dimensional formulation is the vectorial Helmholtz equation as described in section 5.3.2.2. This requires the use of $H(curl)$ - elements which have to be treated with some care when constructing the reduced order multi-point model. Although a proper orthogonalization scheme guarantees the solutions at the expansion points to be free of non-physical solutions, this is no longer the case for the evaluation points in between when performing the parameter sweep.

The strategy remains the same as in the 2d case: The full finite-element problem is solved at selected expansion points. The obtained eigenvectors are collected into a matrix $\mathbf{Q}$. Due to stability reasons it is orthogonalized and serves as a projection basis.

$$\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_n], \tag{5.58}$$

Algorithm 5 Adaptive multi-point MOR

```
$\mathcal{B} = \{\beta_1, \ldots, \beta_{N_e}\}$;
Solve $\bar{\mathbf{S}}(c_x(\beta_1), 1)\mathbf{X}_1 = k_p^2 \bar{\mathbf{T}}(c_x(\beta_1), 1)\mathbf{X}_1$
Solve $\bar{\mathbf{S}}(c_x(\beta_2), 1)\mathbf{X}_2 = k_p^2 \bar{\mathbf{T}}(c_x(\beta_2), 1)\mathbf{X}_1$
$\mathbf{Q}_2 = \mathrm{MGS}([\mathbf{X}_1, \mathbf{X}_2])$ {Calculate projection matrix}
for $i = 0$ to 4 do
    $\widetilde{\mathbf{S}}_i = \mathbf{Q}_2^H \mathbf{S}_i \mathbf{Q}_2$
    $\widetilde{\mathbf{T}}_i = \mathbf{Q}_2^H \mathbf{S}_i \mathbf{Q}_2$
end for
for all $\beta_i \in \mathcal{B}$ do
    Compute eigenvalues $k_p^2(c_x(\beta_i), 1)$
end for
$\mathcal{B}_1 = [\beta_1, \beta_{N_e}]$ {First sub-interval}
$\hat{n} = 1$ {Index of sub-interval with highest error indicator}
for $q = 3$ to $q_{max}$ do
    $\hat{\beta} = \frac{1}{2}(\min(\mathcal{B}_{\hat{n}}) + \max(\mathcal{B}_{\hat{n}}))$
    $\mathcal{B}_{\hat{n}} \leftarrow [\min(\mathcal{B}_{\hat{n}}), \hat{\beta}]$ and $\mathcal{B}_{q-1} = [\hat{\beta}, \max(\mathcal{B}_{\hat{n}})]$
    Solve $\bar{\mathbf{S}}(c_x(\hat{\beta}), 1)\mathbf{X} = k_p^2 \bar{\mathbf{T}}(c_x(\hat{\beta}), 1)\mathbf{X}$
    $\mathbf{Q}_q = [\mathbf{Q}_{q-1}, \mathbf{Q}_*] = \mathrm{MGS}([\mathbf{Q}_{q-1}, \mathbf{X}])$
    for $i = 0$ to 4 do
        $\widetilde{\mathbf{S}}_i = \begin{bmatrix} \widetilde{\mathbf{S}}_i & \mathbf{Q}_{q-1}^H \mathbf{S}_i \mathbf{Q}_* \\ \mathbf{Q}_*^H \mathbf{S}_i \mathbf{Q}_{q-1} & \mathbf{Q}_*^H \mathbf{S}_i \mathbf{Q}_* \end{bmatrix}$,
        $\widetilde{\mathbf{T}}_i = \begin{bmatrix} \widetilde{\mathbf{T}}_i & \mathbf{Q}_{q-1}^H \mathbf{T}_i \mathbf{Q}_* \\ \mathbf{Q}_*^H \mathbf{T}_i \mathbf{Q}_{q-1} & \mathbf{Q}_*^H \mathbf{T}_i \mathbf{Q}_* \end{bmatrix}$
    end for
    for all $\beta_i \in \mathcal{B}$ do
        Compute eigenvalues $k_p^2(c_x(\beta_i), 1)$
        $\Delta k_p^q(c_x, 1) = k_p^q(c_x, 1) - k_p^{q-1}(c_x, 1)$
    end for
    Update $E_1(q), E_2(q), E_\infty(q)$
    if $E_\infty(q) < E_{\infty,tol}$ then
        return(converged)
    end if
    $\hat{n} \leftarrow$ index of sub-interval with highest error indicator
end for
```

with the range of

$$\operatorname{ran}\mathbf{Q} = \operatorname{span}(\bar{\mathbf{e}}_{11}, \ldots, \bar{\mathbf{e}}_{PN}). \tag{5.59}$$

In complete analogy to the 2d-case, the system matrices $(\mathbf{S}_i,\, \mathbf{T}_i)$ are replaced by the projected ones $(\widetilde{\mathbf{S}}_i,\, \widetilde{\mathbf{T}}_i)$

$$\widetilde{\mathbf{S}}_i = \mathbf{Q}^H \mathbf{S}_i \mathbf{Q}, \tag{5.60a}$$

$$\widetilde{\mathbf{T}}_i = \mathbf{Q}^H \mathbf{T}_i \mathbf{Q}, \tag{5.60b}$$

and the resulting eigenvalue problem of much smaller dimension is solved for the set of parameters. In order to choose the correct modes, the same mode tracking strategy is applied as in the 2d-case.

Unfortunately it turns out, that although the gradients are successfully filtered out when solving the full problem, the reduced model produces non-physical solutions. Since these are difficult to distinguish from the correct modes, a scheme is required to get rid of them. In the most ideal case, an algorithm should be used capable of intrinsically filtering out those modes. Since such a procedure is very complicated we employ a different solution. Instead of avoiding those modes in the first place, they will be detected and thence disregarded.

For all physical solutions the mode-orthogonality relation

$$\mathbf{G}^T \mathbf{T} \mathbf{e} = 0 \tag{5.61}$$

holds. This correspond to the divergence condition in its weak form.

The idea is now to simply test this condition for the modes resulting from the reduced model. In other words, the test condition t_c,

$$t_c = \left\| \mathbf{G}^T \mathbf{T} \hat{\mathbf{e}} \right\|_2 , \tag{5.62}$$

should be very small for physical and relatively big for non-physical modes. Here, $\hat{\mathbf{e}}$ stands for the eigenvector of the full problem being calculated from the ROM-modes $\bar{\mathbf{e}}$

$$\hat{\mathbf{e}} = \mathbf{Q}\bar{\mathbf{e}}. \tag{5.63}$$

5.4.2 Single-Point Approach

In the next section, a topic summarized that I have been presenting at the *IGTE Symposium 2010* [56]. It is devoted to a model-order reduction scheme, now created out of one finite-element solution only.

We consider the singly-periodic case along the x-direction. Following the arguments above, a FE-discretizations with periodic boundary conditions results in a generalized wave-number dependent eigenvalue problem:

$$\underbrace{\left(\mathbf{S}_0 + c_x\mathbf{S}_1 + \tfrac{1}{c_x}\mathbf{S}_2\right)}_{\mathbf{S}}\mathbf{v}(c_x) = k^2(c_x)\underbrace{\left(\mathbf{T}_0 + c_x\mathbf{T}_1 + \tfrac{1}{c_x}\mathbf{T}_2\right)}_{\mathbf{T}}\mathbf{v}(c_x), \tag{5.64}$$

where $\mathbf{S}$ and $\mathbf{T}$ stand for the global finite-element matrices and $c_x = e^{-j\beta_x D_x}$ for the Floquet coefficient linking the periodic boundaries with respect to the x-direction. The relations $S_2 = S_1^T$ and $T_2 = T_1^T$ hold, but the notation will be easier to read if the matrices are defined separately. The structure of the matrices $\mathbf{S}$ and $\mathbf{T}$, formulated in the dimensionless quantity $\nu := \beta_x D_x$, is given by:

$$\mathbf{S}(\beta) = \mathbf{S}_0 + e^{-j\nu}\mathbf{S}_1 + e^{j\nu}\mathbf{S}_2, \tag{5.65}$$

$$\mathbf{T}(\beta) = \mathbf{T}_0 + e^{-j\nu}\mathbf{T}_1 + e^{j\nu}\mathbf{T}_2, \tag{5.66}$$

with $\mathbf{S}_2 = \mathbf{S}_1^T$ and $\mathbf{T}_2 = \mathbf{T}_1^T$.

In contrary to the previous section where we solved the full-model at different parameter points, our goal is now to use the solution at one expansion point (ν) only. The obtained solution will be used to construct a proper model for the parameter sweep, similar to the analysis of waveguides described in [57]. In contrast to those cases, the matrix structure in our problem is not polynomial, but exponential in the parameter β_x. Let us take on a similar idea and expand the quantities in (5.64) by Taylor series around the central parameter ν_0. This yields

$$k^2(\nu) = \sum_{i=0}^{N} k_i^2(\nu - \nu_0)^i, \tag{5.67}$$

$$\mathbf{v}(\nu) = \sum_{i=0}^{N} \mathbf{v}_i(\nu - \nu_0)^i, \tag{5.68}$$

$$\mathbf{S}(\nu) = \mathbf{S}_0 + e^{-j\nu_0}\mathbf{S}_1 + e^{j\nu_0}\mathbf{S}_2 + \sum_{i=1}^{N} \frac{(j(\nu - \nu_0))^i}{i!} \left[(-1)^i e^{-j\nu}\mathbf{S}_1 + e^{j\nu}\mathbf{S}_2\right], \tag{5.69}$$

$$\mathbf{T}(\nu) = \mathbf{T}_0 + e^{-j\nu_0}\mathbf{T}_1 + e^{j\nu_0}\mathbf{T}_2 + \sum_{i=1}^{N} \frac{(j(\nu - \nu_0))^i}{i!} \left[(-1)^i e^{-j\nu}\mathbf{T}_1 + e^{j\nu}\mathbf{T}_2\right]. \tag{5.70}$$

The following equations for the derivative of $\mathbf{v}(\nu)$ and $k^2(\nu)$ with respect to ν can be obtained by collecting equal powers of β_x:

$$\begin{aligned}
\nu^0 &: \left[\mathbf{S}(\nu_0) - k_0^2\mathbf{T}(\nu_0)\right]\mathbf{v}_0 = 0,\\
\nu^1 &: \left[\mathbf{S}(\nu_0) - k_0^2\mathbf{T}(\nu_0)\right]\mathbf{v}_1 = k_1^2\mathbf{T}(\nu_0)\mathbf{v}_0 - j\left[-\mathbf{S}_1 + \mathbf{S}_2 - k_0^2\left(-\mathbf{T}_1 + \mathbf{T}_2\right)\right]\mathbf{v}_0,\\
\nu^2 &: \left[\mathbf{S}(\nu_0) - k_0^2\mathbf{T}(\nu_0)\right]\mathbf{v}_2 = k_2^2\mathbf{T}(\nu_0)\mathbf{v}_0 - j\left[-\mathbf{S}_1 + \mathbf{S}_2 - k_0^2\left(-\mathbf{T}_1 + \mathbf{T}_2\right)\right]\mathbf{v}_1\\
&\quad - \frac{1}{2}\left[-\mathbf{S}_1 - \mathbf{S}_2 - k_0^2\left(-\mathbf{T}_1 - \mathbf{T}_2\right)\right]\mathbf{v}_0,\\
&\quad + \frac{1}{2}k_1^2\left[\mathbf{T}(\nu_0)\mathbf{v}_1 + j\left(-\mathbf{T}_1 + \mathbf{T}_2\right)\mathbf{v}_0\right]\\
&\quad\ \vdots\\
\nu^P &: \left[\mathbf{S}(\nu_0) - k_0^2\mathbf{T}(\nu_0)\right]\mathbf{v}_P = k_P^2\mathbf{T}(\nu_0)\mathbf{v}_0\\
&\quad - \sum_{i=1}^{P} \frac{j^i}{i!}\left[(-1)^i\mathbf{S}_1 + \mathbf{S}_2 - k_0^2\left((-1)^i\mathbf{T}_1 + \mathbf{T}_2\right)\right]\mathbf{v}_{P-i}\\
&\quad + \sum_{i=1}^{P-1} k_i^2 \cdot \left[\mathbf{T}(\nu_0)\mathbf{v}_{P-1} + \sum_{m=1}^{P-i}\frac{(j)^m}{m!}\left((-1)^m\mathbf{T}_1 + \mathbf{T}_2\right)\mathbf{v}_{P-i-m}\right].
\end{aligned} \tag{5.71}$$

This system of equations is solved recursively. First, the full eigenvalue problem is solved for ν_0 to obtain $\mathbf{v}_0$ and k_0^2. Then the procedure is to pre-multiply the equations in (5.71) with $\mathbf{v}_0^H$, to calculate k_i^2 and in the following $\mathbf{v}_i$. The eigenvalues are obtained from a ROM that is created as the projection of the matrices in (5.64) by a matrix $\mathbf{V}$ with

$$\text{range}(\mathbf{V}) = \text{span}\{\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_P\}. \tag{5.72}$$

For stability reasons, the model matrix $\mathbf{V}$ is again orthogonalized using a Gram-Schmidt scheme. The resulting eigenvalue problem is of course of much smaller dimension and hence faster to solve.

5.4.2.1 Singularity of the linear system

One problem arises from the fact that the left-hand side of the above system of equation is singular for an exact eigenvalue k_0^2. Since the eigenvalue is only approximately computed, one can factorize the matrix and the method will yield useful results. In a mathematically proper way one would have to solve iteratively the projected linear system

$$\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{b}, \tag{5.73}$$

with $\mathbf{A} = \mathbf{S}(\nu_0) - k_0^2\mathbf{T}(\nu_0)$ and $\mathbf{P}$ being the orthogonal projector

$$\mathbf{P} = \mathbf{I} - \mathbf{v}_0\mathbf{v}_o^T. \tag{5.74}$$

This comes at the expense of two drawbacks. First, one can no longer factorize the matrix $\mathbf{A}$ once and then simply calculate the forward/backward substitutions. The other and major disadvantage is the fact that an iterative solver has to be implemented. A standard CG-method would not function in a very efficient way, especially for electrically large field regions, even when an incomplete Cholesky-decomposition is used as a pre-conditioner.

A less rigorous outcome, however, is employed in [57]. The authors suggest to simply remove one row and its corresponding column from the matrix making it non-singular. When the corresponding component is also removed form the right-hand side, the condition number can be improved. More importantly, the eigenvalues, resulting from a model with the input vectors stemming from the changed linear system remain unchanged. In this thesis I stick to the mathematically less rigorous but practically working case. In practice, one first selects the component of the eigenvector $\mathbf{v}_0$ having the largest magnitude and removes the corresponding rows and columns of the matrix $\mathbf{A}$ to get $\mathbf{A}'$. The linear equation system $\mathbf{A}'$ will be factorized once, so that the frequency derivatives $\mathbf{v}_i (i = 1, 2, \ldots, P)$ can be computed efficiently. Note that the corresponding entry of the right hand side has to be modified accordingly.

5.5 Results

Finally, this chapter presents the results obtained from the formulations presented in the previous sections. It is organized as follows. As a first step singly-periodic structures are investigated serving as a benchmark. The calculated band-structures stemming from direct FE-calculations are compared to reported results and the model-order reduction performance is analyzed thereafter. These structures, along with doubly-periodic crystals, are calculated using both 2d and 3d formulations. The results are presented separately and then compared. Of course, for crystals periodic in each spatial dimension, only a three dimensional full wave simulation can serve as a proper model. The multi-point model-order reduction scheme is applied to all three cases, whereas the single-point expansion technique is analyzed solely on cases with periodicity in one dimension.

5.5.1 2D - formulation

Keeping consistent to the logic of the thesis an overview of the results stemming from the H-plane discretization is given first. Thereby structures showing singly and doubly periodic repetition of their material properties are distinguished. Triply periodic structures cannot be captured with a 2D-formulation and will be treated later on. In presenting the results, focus will be put on the behaviour of the model-order reduction schemes. The 2d problems are treated with a finite-element code using scalar hierarchical basis functions up to third order. The code is written in MATLAB and the same FE code is used for both the single-point and multi-point model-order reduction schemes.

5.5.1.1 Singly-Periodic Case

Let us consider a simple photonic crystal consisting of a dielectric material and periodic in one dimension. The unit cell, sketched in Fig. 5.3, consists of a dielectric inset repeated in x-direction and having an electric permittivity of $\varepsilon_r = 9$. The band structure obtained by a ROM resulting from a two-dimensional formulation is depicted in Fig. 5.4. The results are in perfect agreement with the first example of [48]. In this example, the structure is obtained by varying the exponent $\beta_x D_x$ of the Floquet coefficient $c_x = \exp(-j\beta_x D_x)$ from 0 to π. The component in y-direction, c_y, is set to one which means that the unit cell is extended to infinity in the y-direction. We note the appearance of bandgaps, i.e. frequency regions in which a wave cannot propagate through this crystal. The gap between the lowest and the second mode is calculated to have the range $[4.24, 8.19]$ rad/m, whereas the second gap between mode two and three is at $[11.47, 13.87]$rad/m.

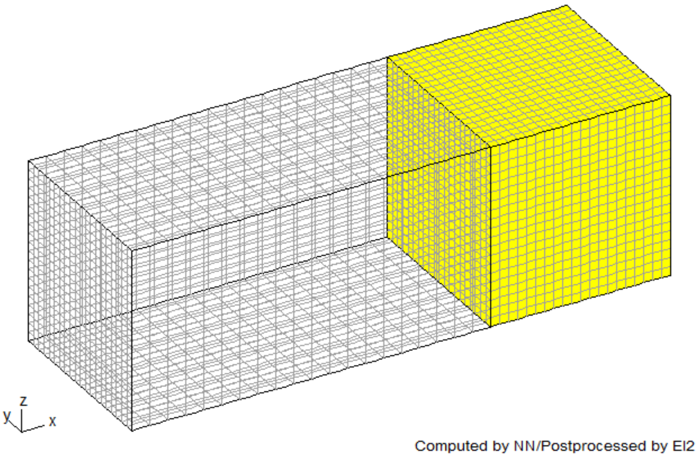


Figure 5.3: Unit cell of a crystal periodic in one dimension.

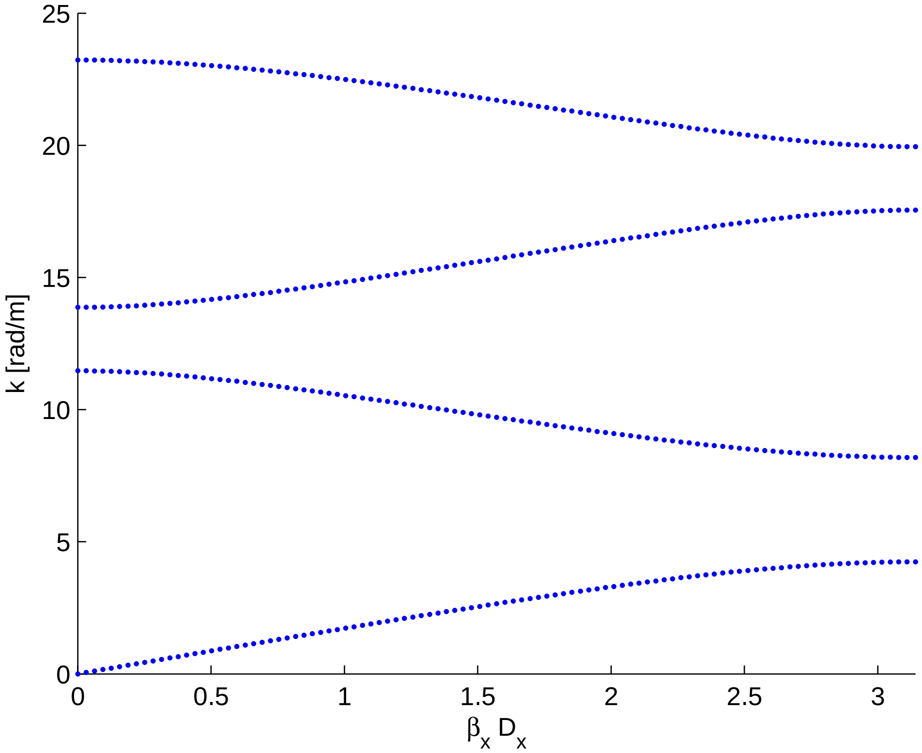


Figure 5.4: Band structure of the single periodic crystal

Fig. 5.5 and Fig. 5.6 demonstrate the accuracy of solutions based on a ROM calculation.

In the first case two expansion points are used for setting up the ROM, one at $\beta_x = 0$ and the second at $\beta_x = \pi/D_x$. The eigenvalues are compared to the results of the underlying FE calculation solved with conventional eigenvalue routines. One can see that the absolute error always remains below 10^{-5} and consequently drops to zero at the expansion points. Since the present crystal is of rather simple structure, the band structure problem is rather smooth, and so the error is just a noise around 10^{-6} when adding a third expansion point at $\beta_x = \pi/(2*D_x)$. In other words a reduced-order model stemming from three solutions is adequate to describe the band structure. Of course, in the general case one does not have the a priori information, how many solutions are needed for a proper model. Therefore, the described adaptive scheme should be applied. Results of this scheme are presented for the doubly-periodic case. Given that the three point solution is already adequate enough here, the adaptive algorithm stops after having compared the eigenvalues of the three point solution with the two-point one.

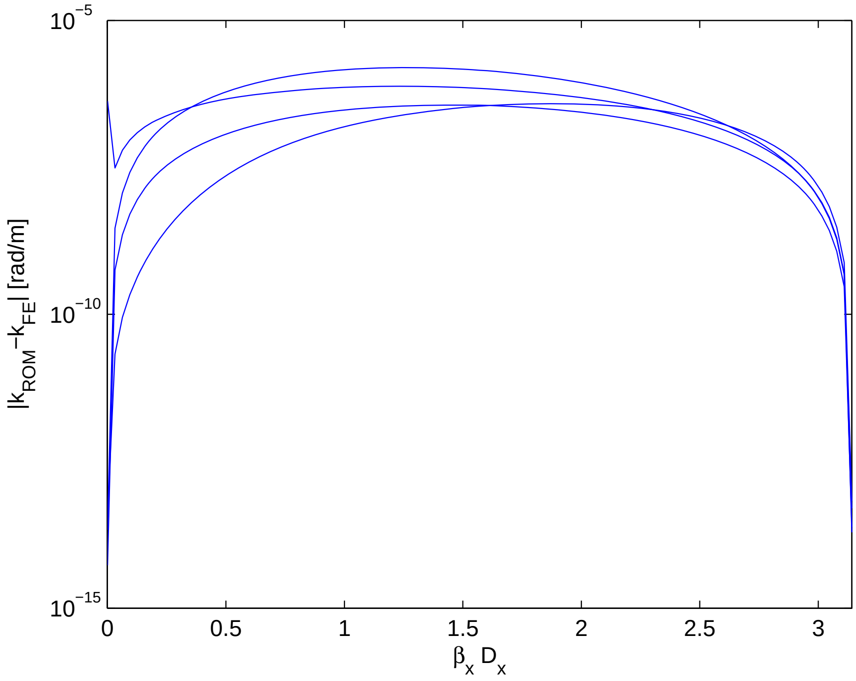


Figure 5.5: Absolute error of eigenvalues obtained by a ROM calculation compared to the underlying FE solution; 2 expansion points for the ROM creation

Although this first case can be solved rather efficiently with traditional methods, a look at the run-times demonstrate the superiority of the reduced-order model case. A comparison can be found in Table. 5.1. The ROM calculation is almost 15 times faster than the full FE case.

Finally a ROM solution obtained from a single-point model order reduction scheme is investigated.

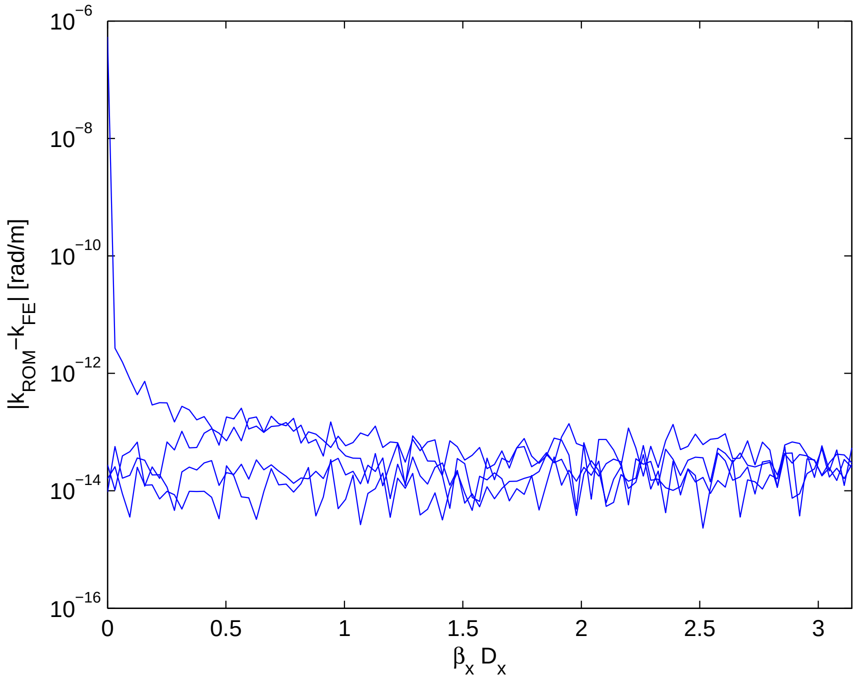
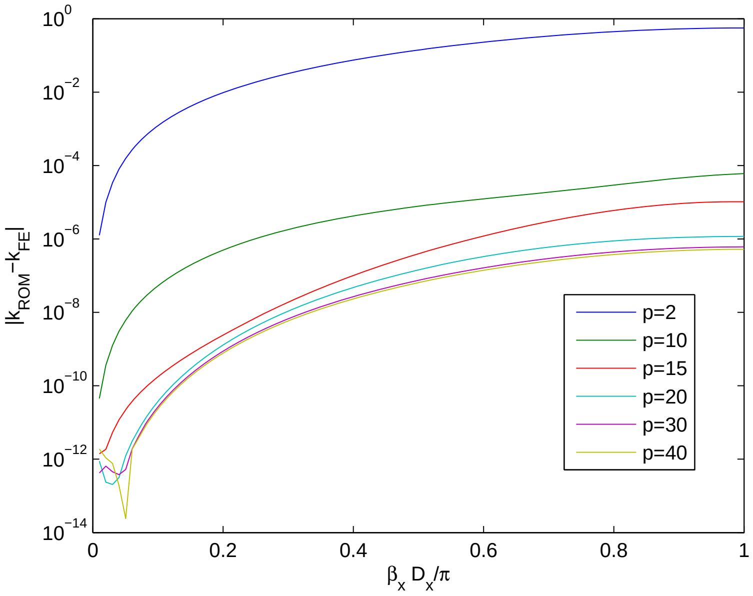


Figure 5.6: Absolute error of eigenvalues obtained by a ROM calculation compared to the underlying FE solution; 3 expansion points for the ROM creation

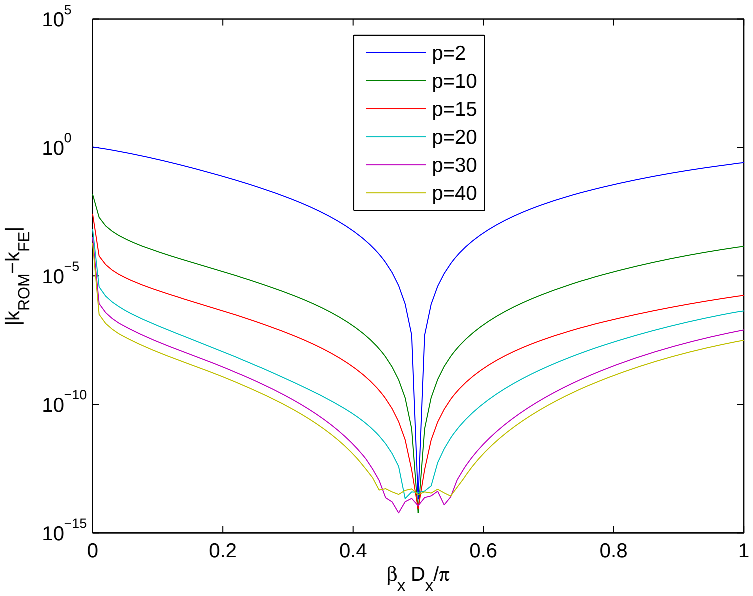
Table 5.1: Computational Data

Parameter	**FE**	**ROM**	**Run-time (s)**	**FE**	**ROM**
			Matrix assembly	0.87	0.89
Model dimension	1897	1897	FE solver	27.94	0.78
Evaluation points	100	100	ROM generation	-	0.09
Expansion points	-	3	ROM evaluation	-	0.17
			TOTAL	**28.81**	**1.93**

The obtained band structure is identical to the one depicted in Fig. 5.4. When looking at the error of the single-point method compared to an FE calculation, Fig. 5.7 illustrates the convergence behaviour with increasing model order P. The expansion points are chosen to be $\beta_x = 0$ and $\beta_x = \pi/(2 * D_x)$. Comparing these error diagrams to the one obtained for a multi-point solution one notes that the error levels are similar for a model-order of around $p = 10$.



(a) Expansion point $\beta_x = 0$



(b) Expansion point $\beta_x = \pi/(2 * D_x)$

Figure 5.7: Error plot for the single-point model-reduction calculation for various model orders p.

5.5.1.2 Doubly-Periodic Case

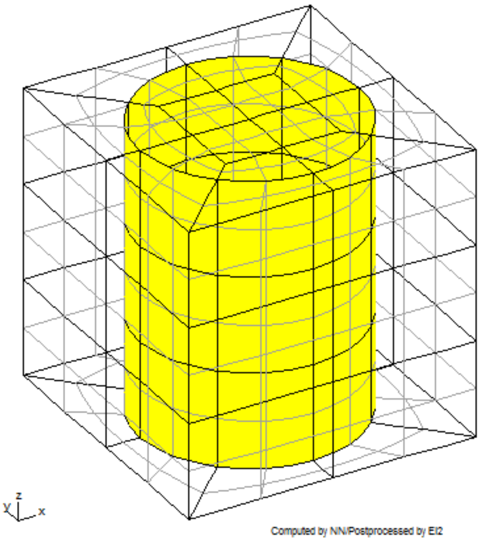


Figure 5.8: Unit cell of a crystal being composed of dielectric rods.

For the doubly-periodic case we consider a photonic crystal consisting of dielectric rods; see the inset of Fig. 5.8. The band structure obtained by the multi-point ROM, which is depicted in Fig. 5.9, is in perfect agreement with [49]: for the first two photonic bandgaps, we have obtained the intervals $[0.2455, 0.2676]$ and $[0.4075, 04519]$, respectively, and, in [49], they were reported to be $[0.2457, 0.2678]$ and $[0.4081, 0.4527]$. We have computed the band structure along the symmetry lines M-Γ-X-M, i.e. for a path around the whole reduced Brillouin zone. For a termination criterion of $E_{\infty,tol} = 10^{-6}$, the adaptive procedure uses 9 expansion points. Fig. 5.10 presents an error plot with respect to full FE calculations. It can be seen that the true error for any of the modes considered is always below the threshold of 10^{-6}. Computational data are given in Table 5.2. The total computer run-time for the MOR process, covering matrix assembly, 9 FE solutions at the expansion points, the ROM generation process, and 450 evaluations, is 82.71s, which is 17 times faster than the corresponding FE solutions. When the overhead of the adaptive procedure is included, run-times rise to 254.71s, still 5.5 times faster than conventional FE calculations. It is emphasized that the adaptive loop, including mode tracking and error estimation, is purely experimental MATLAB code that leaves great room for improvement.

In a final test, we have investigated the asymptotic convergence properties of the proposed adaptive procedure. Fig. 5.11(a) illustrates the behaviour of the error measures E_1, E_2, and E_∞ of (5.57) for the method above, proceeding along the symmetry lines M-Γ-X-M, and Fig. 5.11(b)

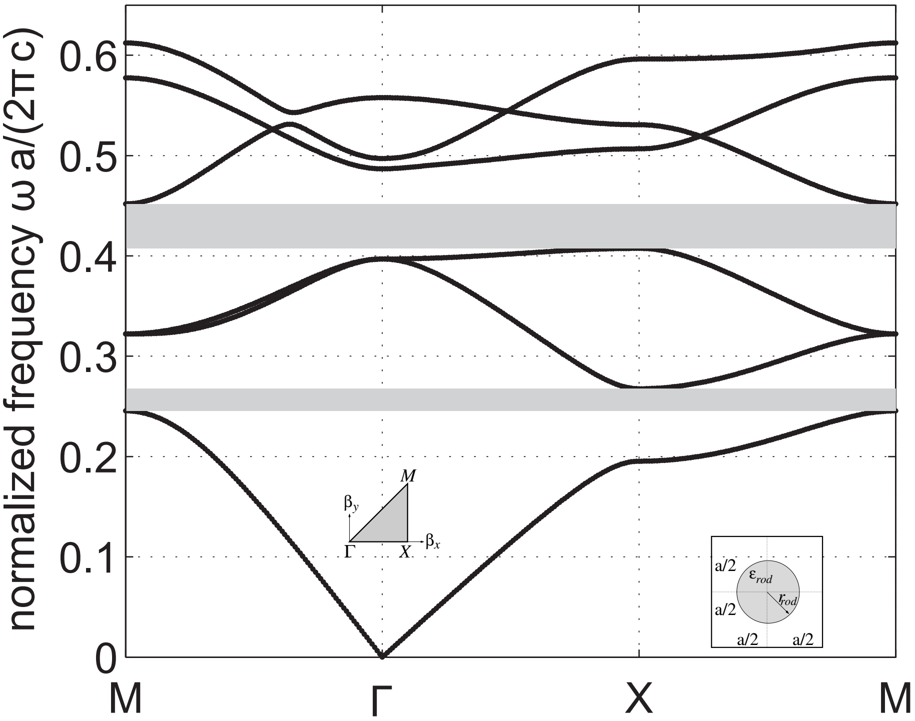


Figure 5.9: Band structure of the dominant six modes of a two-dimensional photonic crystal consisting of dielectric rods. Parameters: $\varepsilon_r = 9$, $r_{rod} = 0.38$ mm, $a = 1$ mm. The right inset shows the geometry of the unit cell. The reduced Brillouin zone with the respective symmetry points is depicted in the left inset. The shaded areas denote photonic bandgaps.

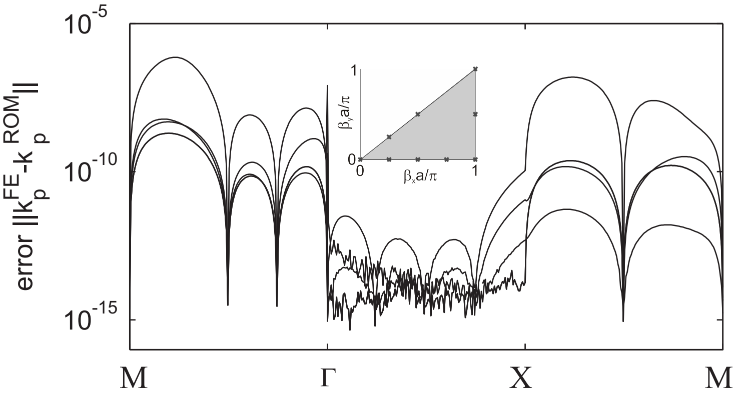
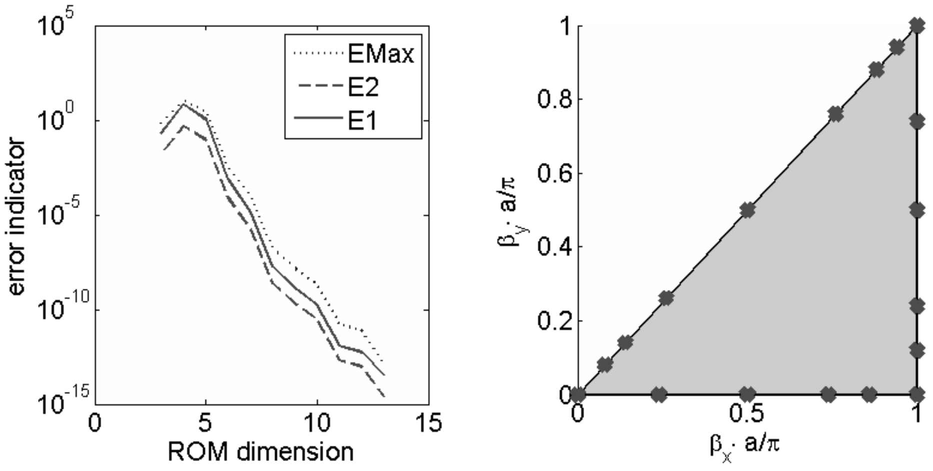


Figure 5.10: Absolute error in ROM wavenumber along the path M-Γ-X-M with respect to FE calculations for a termination criterion of $E_{\infty,tol} = 10^{-6}$. Inset shows locations of adaptively chosen expansion points.

Table 5.2: Computational Data

Parameter	FE	ROM	Run-time (s)	FE	ROM
			Matrix assembly	14.84	14.35
Model dimension	9361	9361	FE solver	1391.86	25.99
Evaluation points	450	450	ROM generation	-	3.59
Expansion points	-	9	ROM evaluation	-	38.78
			TOTAL (non-adapt.)	**1405.69**	**82.71**
Expansion points	-	9	Adaptive loop	-	183.46
			TOTAL (adaptive)	-	**254.71**

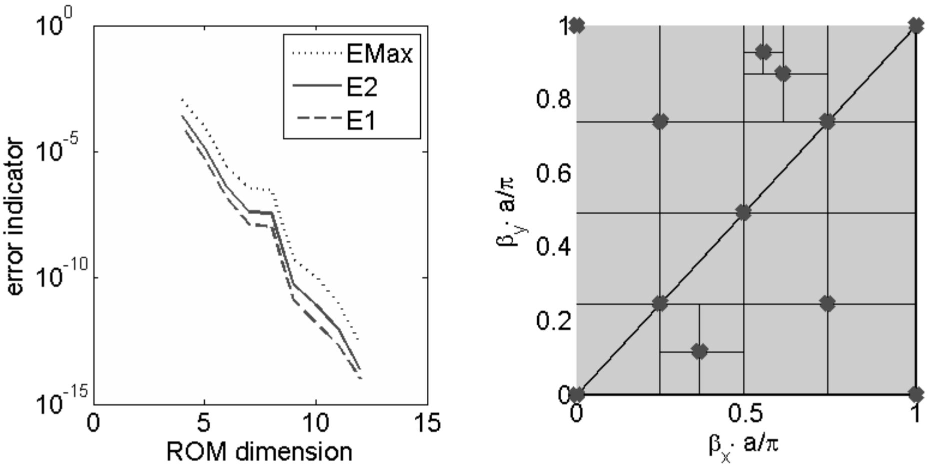
shows the corresponding locations of the expansion points after 13 iterations. Fig. 5.12 presents similar data for a more general two-parameter ROM that covers the entire domain of the Brillouin zone. In both approaches, the behavior of the error measures indicates a very rapid rate of convergence.



(a) Error indicators.

(b) Expansion points.

Figure 5.11: Asymptotic behavior of adaptive single-parameter ROM along the symmetry lines M-Γ-X-M.



(a) Error indicators.

(b) Expansion points.

Figure 5.12: Asymptotic behavior of adaptive two-parameter ROM for the whole Brillouin zone.

5.5.2 3D - formulation

Regarding the full-wave case of the three dimensional formulations, we will investigate structures periodic in one, two and three dimensions. For the first two cases, the models from the 2d formulation are extended to the 3d case in order to compare the results. The need to apply an additional scheme to filter out non-physical modes will be demonstrated. Having demonstrated this on the benchmark cases, attention will be devoted to two triply periodic structures.

For the 3D cases, the MATLAB FE-code based on scalar basis functions cannot be used, since this is based on a two-dimensional H-plane formulation. Instead the in-house software package Elefant3D has been extended to capture periodic boundary conditions. As it is outlined in section there are 27 different types of boundaries in a three dimensional calculation. Therefore a scheme has to be developed to correctly identify these types in the program. This information together with the FE-matrices is then transformed in order to get the splitting into parameter independent matrices. So, for our purposes the assembly process is carried out in Elefant3D, whereas the solutions of the eigenvalue problems and the model-order reduction schemes have been programmed in MATLAB.

5.5.2.1 Singly-Periodic Case

As the first case serves the example of a crystal being periodic in the x-direction only. Applying the full-wave formulation without any further orthogonalization as described in section 5.4.1.3 results in a structure correctly capturing the true modes but also producing additional solutions which are non-physical. As described above, this is due to the fact that the model does not capture the correct operator structure although it is generated out of true solutions since no non-physical modes are present at the expansion points. The band structure including these non-physical modes is depicted in Fig. 5.13. The Floquet coefficient is varied along the x-direction, keeping the y and z values at zero. The expansion points for creating the reduced-order model are at $\beta_x D_x = 0, \pi/2, \pi$. The eigensolutions printed in blue coincide with the true modes obtained from the 2d formulation. The green modes denote the spurious solutions polluting the spectrum, since one does not know a priori how two distinguish between them.

Applying the filtering scheme described in section 5.4.1.3 yields the correct structure as already computed in the two dimensional formulation, see Fig. 5.4. When evaluating the ROM, i.e. performing the parameter sweep in the reduced matrices, each obtained solution is tested for the testing condition, t_c of Eq. (5.62). As sketched in Fig. 5.14, the values for t_c remain below 10^{-10} for the physical modes. These values for the spurious modes, are found

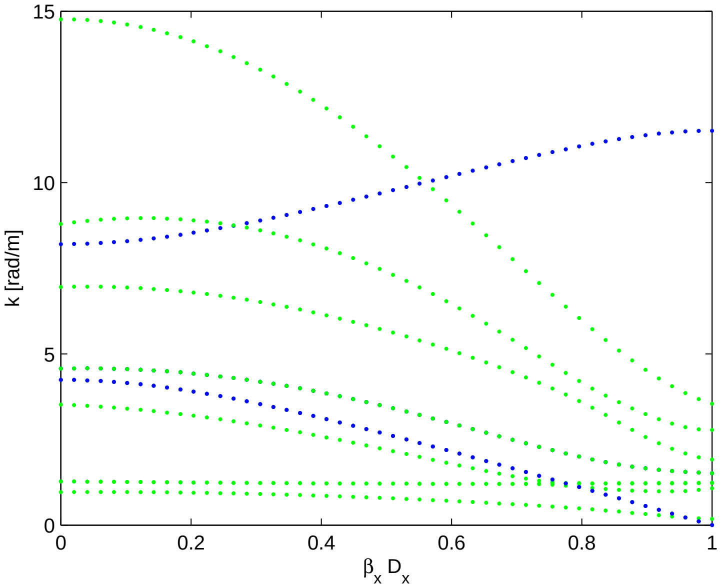


Figure 5.13: Band structure of the single periodic crystal; Blue modes donate correct modes whereas green modes are spurious solutions

to be between 10^{-1} to 10^{0}. In the code, the threshold value for the mode-filtering is set to to 10^{-6}.

Again, the accuracy of the ROM calculation is demonstrated by a comparison with the results of the underlying FE calculation. Fig. 5.20 shows that when using three expansion points the error is already negligible.

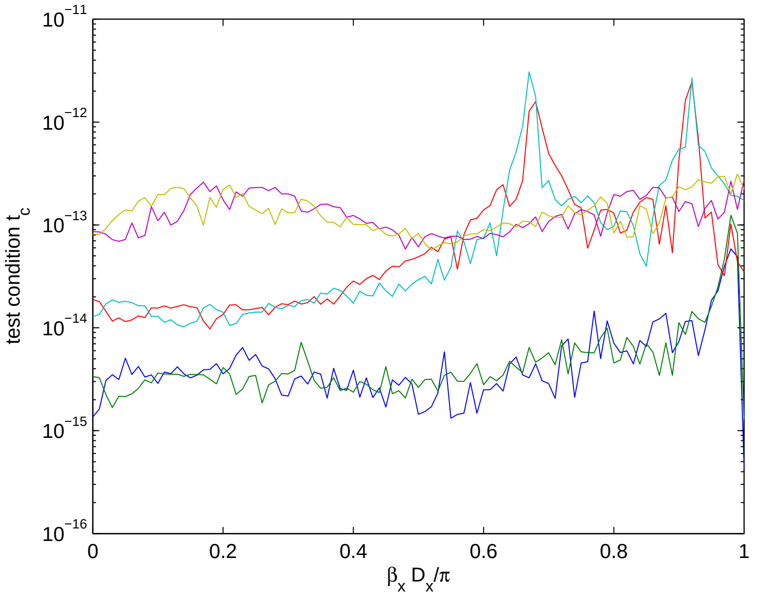


Figure 5.14: Testing condition t_c for the physical modes

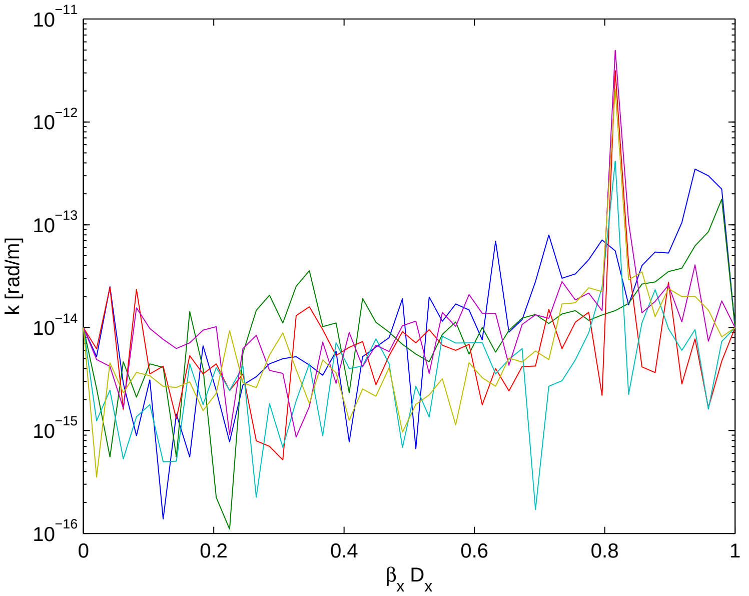


Figure 5.15: Error of a ROM calculation for the singly-periodic crystal

5.5.2.2 Doubly-Periodic Case

For the doubly-periodic case let us reconsider the crystal being composed of dielectric rods, see Fig. 5.8. The results of the full-wave calculation are depicted in Fig. 5.16. There the parameters are varied along the line $\Gamma - X$ in the Brillouin zone, i.e. β_x varied from 0 to π/a, where a denotes the dimension of the unit cell. Again it is remarkable to notice the appearance of non-physical modes depicted in green in Fig. 5.16. Applying the filtering scheme against these non-physical modes, the band structure of diagram Fig. 5.17 is obtained. Of course, the structure should be consistent with the one obtained from a 2d-calculation. Comparing Fig. 5.17 with the Γ-X-part of Fig. 5.9, one notes additional physical modes depicted in red. These modes are physical solutions belonging to an E-plane formulation. These cannot be captured with the H-plane formulation in the two dimensional example.

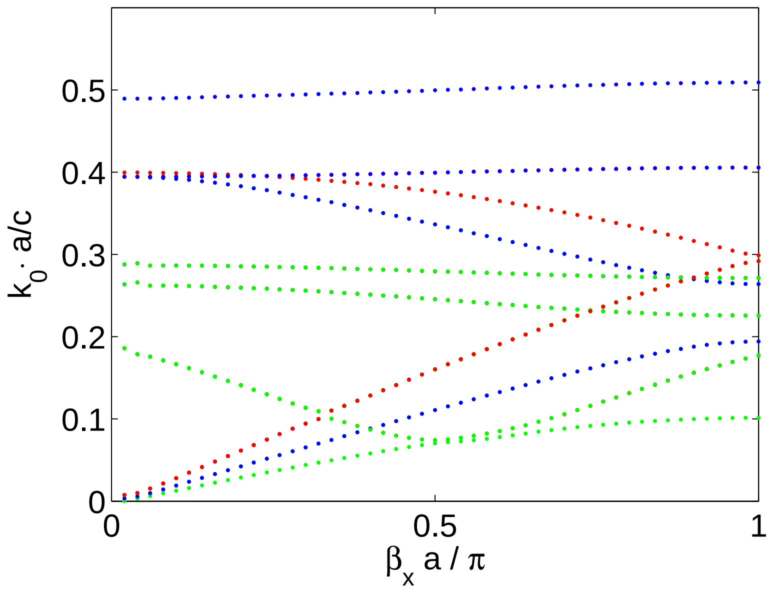


Figure 5.16: Band structure of a crystal consisting of dielectric rods stemming from a 3d-calculation. The blue modes coincide with the 2d-case, red modes donate correct solutions whereas the green solutions are non-physical

The threshold value t_c for accepting a mode to be physical is set to be 10^{-2} in this. The value for all non-physical modes is tested to lie above this value. On the contrary the value for the physical modes, also being captured by the H-plane formulation lie in a region of 10^{-6}, whereas the additional true modes have value of around 10^{-2}. The corresponding values for the spurious modes are not smaller than 10^{-1}.

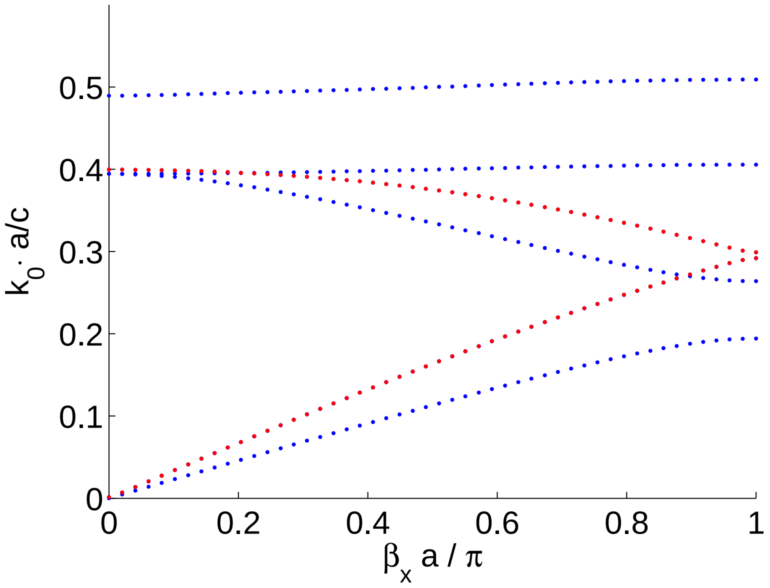


Figure 5.17: Band structure of a crystal consisting of dielectric rods stemming from a 3d-calculation without non-physical solutions

5.5.2.3 Triply-Periodic Case

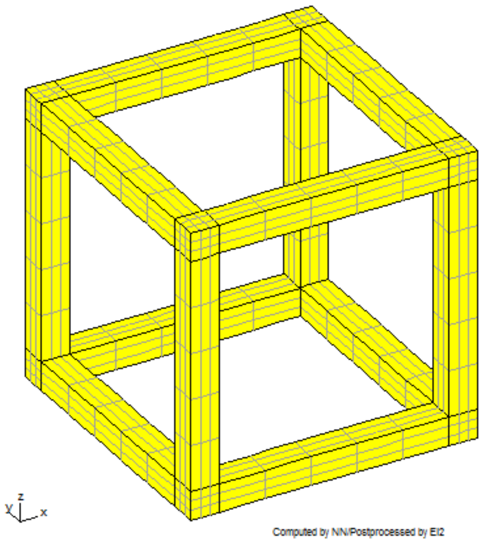


Figure 5.18: Unit cell of a crystal being periodic in three dimensions. $\varepsilon_r = 12.96$ in the yellow area, $\varepsilon_r = 1$ elsewhere;

Finally let us put our focus on crystals being periodic in three dimensions. The first structure we are analyzing is depicted in Fig. 5.18. Continuing the unit cell along the three axes, the crystal looks like a scaffold. The ratio of the thickness of the yellow rods and the dimension of the unit cell is chosen to be 1/10. The relative permittivity of the dielectric medium is $\varepsilon_r = 12.96$. Fig. 5.19 shows a band structure when varying the wave vector along the Brillouin zone's symmetry points $\Gamma - X - M - R$, as it is explained in section 2.2.2. The results are obtained from a multi-point model order reduction scheme, when using 9 expansion points, placed along the chosen symmetry lines. One can see a nice qualitative agreement with [58]. The small discrepancy in absolute values is attributed to the fact, that the exact dimensions of the unit cell are not reported and could only be guessed. The performance of the ROM is tested against the underlying full FE calculation and an error plot can be found in Fig. 5.20. The absolute error of the computed modes always remains below 10^{-3}.

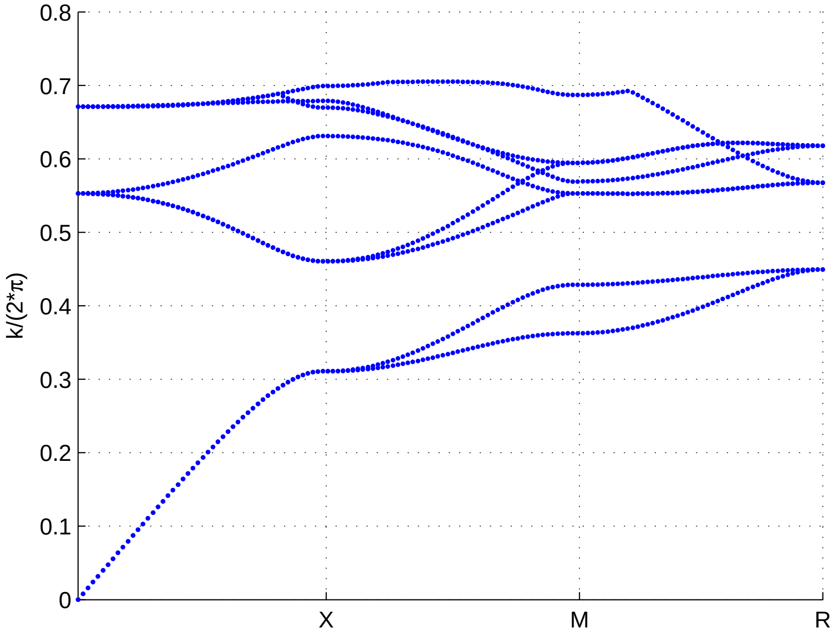


Figure 5.19: Band structure of the triply periodic scaffold structure

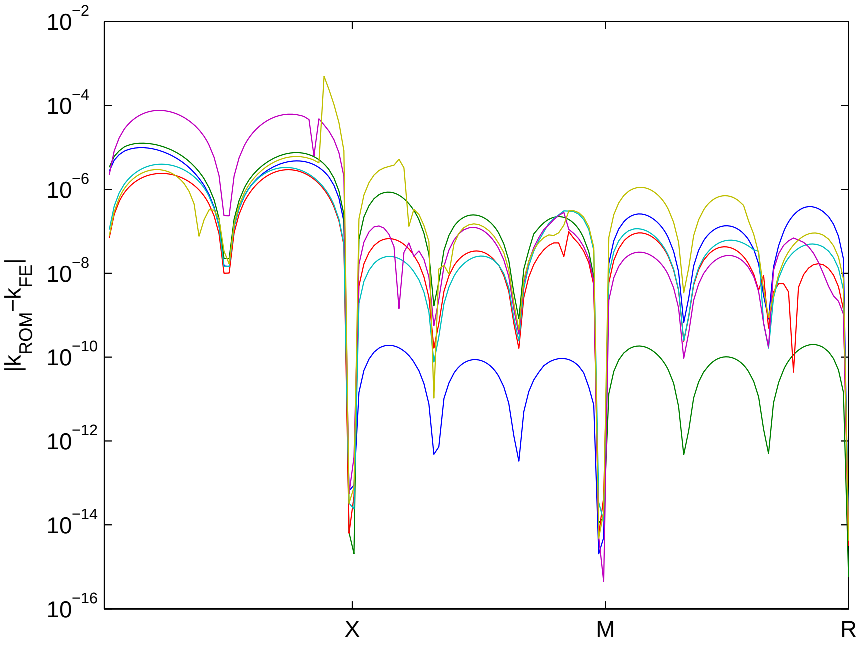
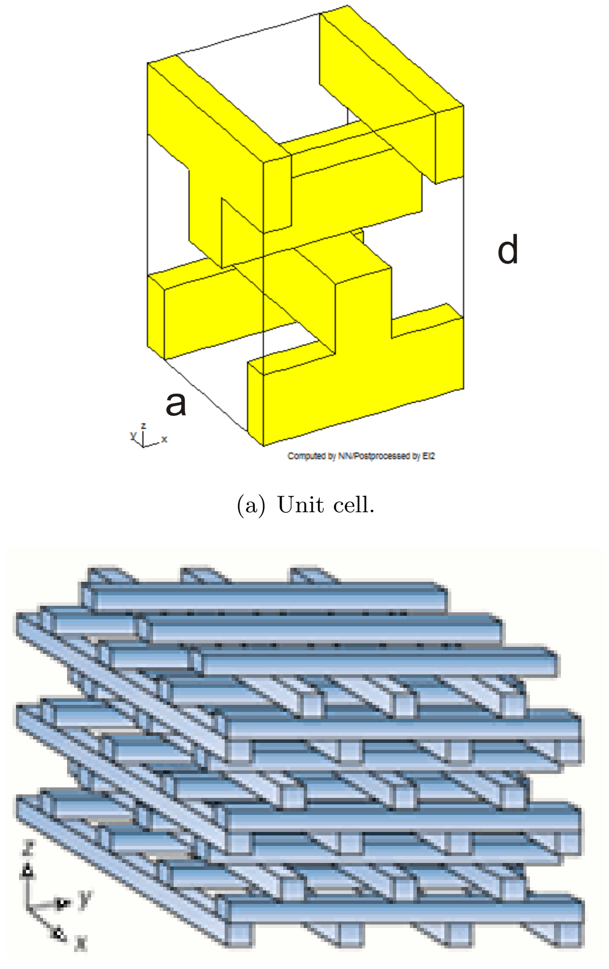


Figure 5.20: ROM error of the triply periodic scaffold crystal

Finally let us put our focus onto a crystal structure having a complete bandgap. The unit cell of the so called woodpile crystal is depicted in Fig. 5.21. The structure is composed of layers of dielectric rods with a stacking sequence repeating itself every four layers with a repeat distance of d. Within the layers, the rods are arranged with their axes parallel and separated by the distance a. The orientations of the axes are rotated by 90 degrees between adjacent layers [59]. Setting $d = \sqrt{2}a$ the lattice can be considered as a face-centered cubic lattice [60]. In contrast to the symmetric scaffold case the unit cell of this woodpile structure is asymmetric in the three dimensions.



(a) Unit cell.

(b) Lattice.

Figure 5.21: Woodpile photonic crystal.

The band structure was obtained on successive runs along the symmetry points of the Brillouin zone. The complete band structure diagram over the whole Brillouin zone is depicted in Fig. 5.22. A full photonic band gap from $0.465c/a$ to $0.568c/a$ is observed, comparing perfectly well with reported results [60]. c denotes the vacuum speed of light. To justify the accuracy of the ROM, Fig. 5.23 and Fig. 5.24 show plots of the absolute error of the ROM solutions compared to the underlying FE calculations. In the latter diagrams the first two symmetry lines are investigated, the lines $\Gamma - X$ and $X - M$. In the first case, three expansion points are used to create the reduced-order model, whereas in the second case a fourth point is added. The diagrams demonstrate a perfect convergence of the ROM solutions.

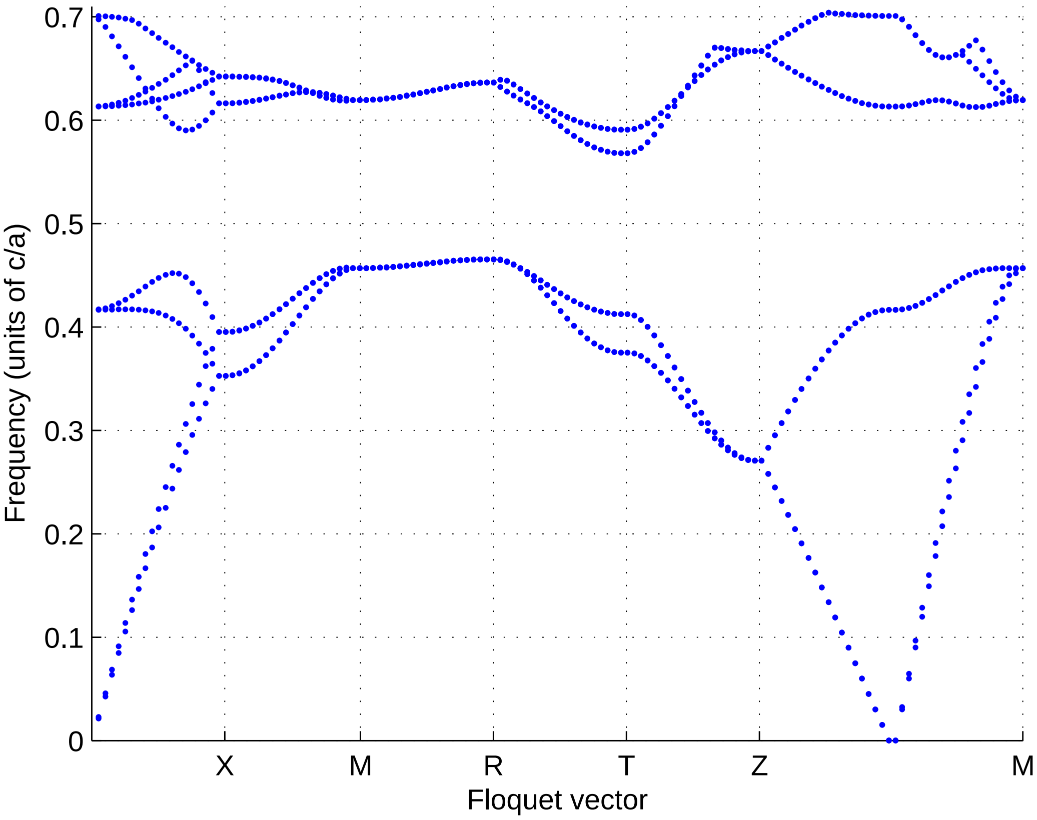


Figure 5.22: Band structure of the woodpile structure; Notation according to section 2.2.2

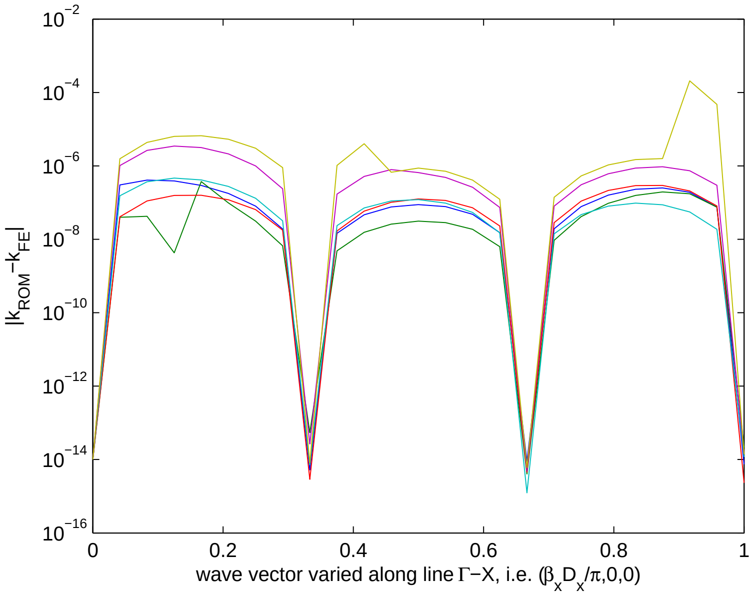


Figure 5.23: Error of the ROM calculation compared to the underlying FE solution; The Floquet coefficient is varied along the line $\Gamma - X$.

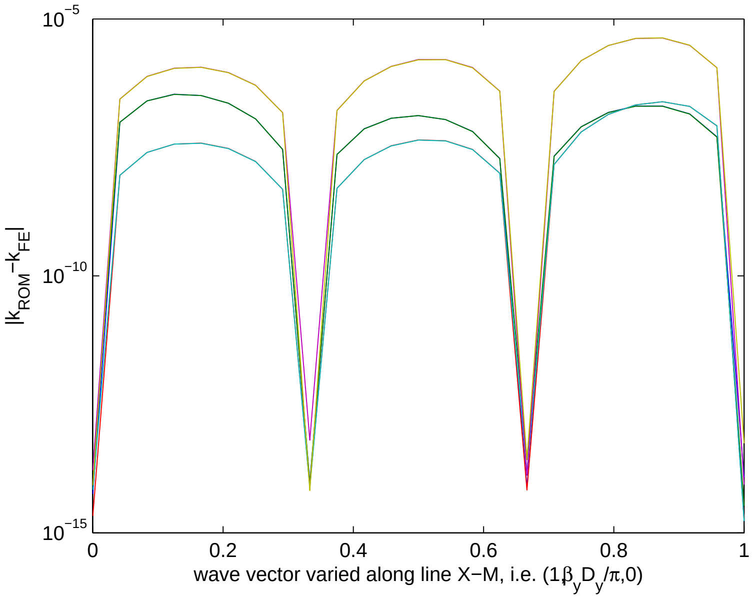


Figure 5.24: Error of the ROM calculation compared to the underlying FE solution; The Floquet coefficient is varied along the line $X - M$.

5.6 New scientific results

This chapter has dealt with methods capable of fast and accurate computations of photonic band structure diagrams. The finite-element method, along with an implementation of periodic boundary conditions, has led to a parameterized generalized eigenvalue problem which has to be solved many times in order to resolve a band structure. The scientific contribution consists of the presentation and application of efficient model-order reduction schemes allowing for accurate computations at considerably reduced computational costs. Multi-point model-order reduction schemes have been presented for two and three dimensional photonic crystal structures. The run-time advantage of these methods over conventional finite-element calculations has been demonstrated, and it has been shown that the approximation error of the reduced-order models is negligibly low. For two dimensional structures, the presentation of an adaptive algorithm has addressed the question of how to choose the expansion points for the model. In addition, a possibility of a direct evaluation of the whole *Brillouin* zone, instead of the surroundings only, has been discussed. A single-point model-order reduction scheme has been applied to a two dimensional structure. In contrast to previous studies, where single-point methods have been applied to polynomial matrix structures, the reduced-order model has been set up for an exponential structure in this thesis. Again, the accuracy and computational efficiency of the suggested scheme has been demonstrated.

6 Conclusion and Outlook

In this thesis eigenvalue problems in the context of electromagnetic field simulations are treated. The finite-element method has served as numerical tool to analyze wave propagation problems. In the opening chapters, the necessary physical and mathematical theory is outlined and reviewed in order to put the applications into an appropriate thematic context. Thereby, the electromagnetic wave equation has been derived and it has been shown how to formulate a feasible discretization using the finite-element method. In this way, boundary value problems are transformed into algebraic systems of equations, or as it is the case in this thesis, eigenvalue problems. Therefore an overview of iterative eigenvalue solvers has been given focusing on routines taking advantage of sparsely occupied matrices.

The first application consists of a consistent description of dispersion relations of waveguiding structures. Based on the studies of [4] a formulation has been tested that successfully describes these quantities without polluting the spectrum with non-physical modes. As it is shown, this is of special importance near the static limit. A focus of ongoing studies is to look at open waveguiding structures, thereby including losses due to radiation. The application of fast frequency sweeps, similar to the methods presented in the second part of this thesis, thereby considerably improves the computation times.

In the second part of the applications, photonic crystal are investigated. Thereby the focus is again on an efficient finite-element formulation, paired with the implementation of periodic boundary conditions. Both multi-point and single-point model-order reduction methods have been presented that allow for a very efficient computation of band structure diagrams. So far these reduction-techniques have been presented for polynomial matrix structures. In this work's context the parameter dependence is of exponential nature. Especially for the single-point scheme, there is no comparable case reported in the literature so far. In this chapter it is briefly mentioned that there exists a problem when solving the equations systems needed for the model creation. This is due to the singularity of the linear system's matrices needed to be solved in each step. Here, a feasible but mathematically rather ambiguous approach is applied, but in the future a nice and rigorous description poses an interesting research topic. In this work the single-point scheme has been applied for singly periodic structures in the physical context of photonic crystals. In the future, an extension to two or multi-dimensional cases is certainly of great interest.

On the other hand, multi-point routines have been suggested and their application to photonic crystals has been demonstrated, for both a 2d and 3d formulation. An adaptive scheme has been outlined that regulates where to put the expansion points for the model creation, since this is a priori not known. In the three dimensional full wave case it has been pointed out that some care has to be devoted to the fact that the model produces non-physical solutions. A scheme is introduced how to handle these modes by successfully filtering out the spurious solution when performing the parameter sweep. As a suggestion for future studies, a routine could be of great interest where the reduced model does not produce non-physical solution in the first place.

Acknowledgements

First of all I would like to thank my advisor Prof. Oszkár Bíró for giving me the opportunity to work in this interesting field within his nice group.

I am grateful to Prof. Romanus Dyczij-Edlinger for the ongoing support and for all the encouraging discussions on and off professional topics. I fully enjoyed my stay in Saarbrücken.

Furthermore I would like to thank all my colleagues at the IGTE institute for making my PhD time as enjoyable as it was. Special thanks go to Andy, not only for listening to my occassional complaints. I would also like to thank my friends, Armin, Peter, Heinz, Markus, Ralf, for the *rare* coffee and lunch breaks and Michael for all the fruitful discussions.

Most of all I am grateful to my mother and my father for their unconditional support through all the years and to Diana for being patient with me.

List of Figures

Bibliography

[1] Y. Zhu and A. Cangellaris, *Multigrid Finite Element Methods for Electromagentic Field Modeling.* New York: IEEE press/John Wiley, 2006. 4, 18

[2] V. Thomé, "From finite differences to finite elements: A short history of numerical analysis of partial differential equations," *J. of Comp. and Applied Math.*, vol. 128, pp. 1–54, 2001. 4

[3] D. S. P. Houston, I. Perugia, "Mixed discontinuous galerkin approximation of the maxwell operator: Non-stabilized formulation," *J. Sci. Comput.*, vol. 22, pp. 315–346, 2005. 4

[4] O. Farle, V. Hill, and R. Dyczij-Edlinger, "Finite-element waveguide solvers revisited," *IEEE Trans. Magn.*, vol. 40, no. 2, pp. 1468–1471, Mar. 2004. 6, 91

[5] J. Jackson, *Classical Electrodynamics*, 3rd ed. New York: Wiley, 1998. 8

[6] J. D. Joannopoulos, S. G. Johnson, J. N. Winn, and R. D. Meade, *Molding the flow of light.* Princeton: University Press, 2008. 9, 15, 49, 50, 64

[7] N. Ashcroft and N. Mermin, *Solid State Physics.* Fort North: Saunders College Publishing, 1976. 13

[8] E. Schachinger. Lecture Notes on Solid State Physics. 16, 94

[9] I. Tsukerman, *Computational methods for nanoscale applications.* New York: Springer, 2008. 18, 19, 50

[10] A. Bondeson, T. Rylander, and P. Ingelström, *Computational Electromagnetics.* New York: Springer Science, 2005. 20, 22

[11] J. Nédélec, "Mixed finite elements in $\mathcal{R}^3$," *Numerische Mathematik*, vol. 35, pp. 315–341, Sep. 1980. 22

[12] O. Farle, "Ordnungsreduktionsverfahren für die finite-elemente-simulation parameterabh¨angiger passiver mikrowellenstrukturen," Ph.D. dissertation, Dissertation, 2007. 23

[13] L. Trefethen and D. Bau, *Numerical Linear Algebra.* Philadelphia: SIAM, 1997. 25

[14] B. Parlett, *The Symmetric Eigenvalue Problem.* Englewood Cliffs: Prentice Hall, 1980. 27

[15] H. van der Vorst, *Computational Methods for Large Eigenvalue Problems.* North Holland: Elsevier, 2000. 27, 31

[16] Y. Saad, *Numerical Methods for Large Eigenvalue Problems.* Manchester: University Press, 1992. 28, 29, 32, 33

[17] P. Arbenz and O. Chinellato, "On solving complex-symmetric eigenvalue problems arising in the design of axisymmetric vcsel devices," *Applied Numerical Mathematics*, vol. 58, no. 4, pp. 381–394, Apr. 2010. 33

[18] O. Chinelatto, "The complex-symmetric jacobi-davidson algorithm and its application to the computation of some resonance frequencies of anisotropic lossy axisymmetric cavities," Ph.D. dissertation, Diss. ETH No. 16243, 2005. 33, 35

[19] G. Sleijpen and H. van der Vorst, "A jacobi-davidson iteration method for linear eigenvalue problems," *SIAM J. Matrix Anal. Appl.*, vol. 17, pp. 401–425, 1996. 34

[20] C. Scheiber and O. Bíró, "Eigenvalue analysis of lossy waveguide structures using hybrid $h(curl)$ second order finite elements," *Proceedings of the 17th Conference on the Computation of Electromagnetic Fields*, pp. 472–473, 2009. 37

[21] C. G. Williams and G. Cambrell, "Numerical solution of surface waveguide modes using transverse field components," *IEEE Trans. Microwave Theory*, vol. MIT-22, pp. 329–320, Mar. 1974. 37

[22] T. Itoh, "Spectral domain inmitance approach for dispersion characteristics ofgeneralized printed transmission lines," *IEEE Trans. Microwave Theory*, vol. MIT-28, pp. 733–736, Jul. 1980. 37

[23] E. Schwig and W. Bridges, "Computer analysis of dielectric waveguides: A finite-difference method," *IEEE Trans. Microwave Theory*, vol. MIT-32, pp. 531–541, May 1984. 37

[24] S. Bardi, O. Bíró, K. Preis, G. Vrisk, and K. Richter, "Nodal and edge element analysis of inhomogeneously loaded 3d cavitiies," *IEEE Trans. Magn.*, vol. 29, pp. 1466–1469, 1993. 37

[25] D. Sun and Z. Cendes, "New vector finite elements for the three-dimensional magnetic fields computations," *J. Appl. Phys.*, vol. 61, no. 8, pp. 3919–3921, Apr. 1987. 37

[26] J.-F. Lee, "Finite element analysis of lossy dielectric waveguides," *IEEE Trans. Microwave Theory*, vol. 42, pp. 1025–1031, Jun. 1994. 37, 44, 46, 94

[27] S. Polstyanko and J.-F. Lee, "$h_1(curl)$ tangential vector finite element method for modelling anisotropic optical fibers," *Journ. of lightwave technology*, vol. 13, no. 11, pp. 2290–2295, Nov. 1995. 37

[28] S.-C. Lee and J.-F. lee, "Hierarchical vector finite elements for analyzing waveguiding structures," *IEEE Trans. Microw. Theory and Techniques*, vol. 51, pp. 1897–1905, Aug. 2003. 37, 39

[29] O. Farle, V. Hill, and R. Dyczij-Edlinger, "Finite-element waveguide solvers revisited," *IEEE Trans. Magn.*, vol. 40, no. 2, pp. 1468–1471, Mar. 2004. 37, 39, 43, 46

[30] R. Albanese and R. Rubinacci, "Solution of three dimensional eddy current problems by integral and differential methods," *IEEE Trans. Magn.*, vol. 24, pp. 98–101, 1988. 41

[31] R. Scharf. (2003, Jul.) http://www.pro-physik.de/. 48

[32] E. Yablonovitch, "Inhibited spontaneous emission in solid-state physics and electronics," *Phys. Rev. Lett.*, vol. 63, pp. 1950–1953, 1987. 48

[33] E. Yablonovitch, T. Gmitter, and K. Leung, "Photonic band structure: The face-centered-cubic case employing nonsperical atoms," *Phys. Rev. Lett.*, vol. 67, pp. 2295–2298, 1991. 49

[34] A. Blanco, E. Chomski, S. Grabtchak, M. Ibisate, S. John, S. Leonard, C. Lopez, F. Meseguer, H. Miguez, and J. Mondia, "Large-scale synthesis of a silicon photonic crystal with a complete three-dimensional bandgap near 1.5 micrometres," *Nature*, vol. 405, pp. 437–440, 2000. 49

[35] Y. Vlasov, X.-Z. Bo, J. Sturm, and D. Norris, "On-chip natural assembly of silicon photonic bandgap crystals," *Nature*, vol. 414, pp. 289–293, 2001. 49

[36] M. Campbell, D. Sharp, M. Harrison, R. Denning, and A. Turberfield, "Fabrication of photonic crystals for the visible spectrum by holographic lithography," *Nature*, vol. 404, pp. 53–56, 2000. 49

[37] Y. Miklyaev, D. Meisel, A.Blanco, G. von Freymann, K. Busch, W. Koch, C. Enkrich, M. Deubel, and M. Wegener, "Three-dimensional face-centered-cubic photonic crystal templates by laser holography: fabrication, optical characterization, and band-structure calculations," *Applied Physical Letters*, vol. 82, no. 8, pp. 1284–1286, Feb. 2003. 49

[38] E. Brown and O. McMahon, "High zenithal directivity from a dipole antenna on a photonic crystal," *Applied Physical Letters*, vol. 68, no. 9, pp. 1300–1302, Feb. 1996. 49

[39] E. Brown, O. McMahon, and C. Parker, "Photonic-crystal antenna substrates," *Lincoln Laboratory Journall 11*, vol. 11, no. 2, pp. 159–173, 1998. 49

[40] A. Koenderik, L. Bechger, H. Schriemer, A. Lagendijk, and W. Vos, "Broadband fivefold reduction of vacuum fluctuations probed by dyes in photonic crystals," *Phys. Rev. Lett.*, vol. 88, no. 14, p. 143903, 2002. 49

[41] A. Koenderik, L. Bechger, A. Lagendijk, and W. Vos, "An experimental study of strongly modified emission in inverse opal photonic crystals," *physica status solidi (a)*, vol. 197, no. 3, pp. 648–661, 2003. 49

[42] R. Wehrsporn and J. Schilling, "A model system for photonic crystals: macroporous silicon," *physica status solidi (a)*, vol. 197, no. 3, pp. 673–687, 2003. 49

[43] T. Krauss, "Planar photonic crystal waveguide devices for integrated optics," *physica status solidi (a)*, vol. 197, no. 3, pp. 688–702, 2003. 49

[44] J. Smajic, C. Hafner, and D. Erni, "Design and optimization of an achromatic photonic crystal bend," *Optics Express*, vol. 11, no. 12, pp. 1378–1384, 2003. 49

[45] R. D. Meade, A. M. Rappe, K. D. Brommer, and J. Joannopoulos, "Accurate theoretical analysis of photonic band-gap materials," *Phys. Rev. B*, vol. 48, no. 11, pp. 8434–8437, 1993. 49, 51

[46] X. Wang, X. G. Zhang, Q. Yu, and B. Harmon, "Multiple-scattering theory for electromagnetic waves," *Phys. Rev. B*, vol. 47, no. 8, pp. 4161–4167, 1993. 49

[47] J. Pendry, "Photonic band structures," *J. Mod. Opt.*, vol. 41, no. 2, pp. 209–229, 1994. 49

[48] A. Tavallaee and J. P. Webb, “Finite element modeling of evanescent modes in the stopband of periodic structures,” *IEEE Trans. Magn.*, vol. 44, no. 6, pp. 1358–1361, Jun. 2008. 49, 57, 71

[49] I. Tsukerman and F. Cajko, “Photonic band structure computation using flame,” *IEEE Trans. Magn.*, vol. 44, no. 6, pp. 1382–1385, 2008. 50, 76

[50] C. Mias, J. Webb, and R. Ferrari, “Finite element modelling of electromagnetic waves in doubly and triply periodic structures,” *IEE Proc. Optoelectron*, vol. 146, no. 2, pp. 111–118, Apr. 1999. 53

[51] P. P. Silvester and R. L. Ferrari, *Finite elements for electrical engineers*, 3rd ed. Cambridge: Cambridge University Press, 1996. 53

[52] S. Coco, A. Laudani, G. Pollicino, and P. Tirro, “Finite element electromagnetic analysis of twt slow-wave structures in grid environment,” *IEEE Trans. Magn.*, vol. 45, no. 3, pp. 1843–1846, Mar. 2009. 55

[53] C. Scheiber, A. Schultschick, O. Bíró, and R. Dyczij-Edlinger, “A model order reduction method for efficient band structure calculations of photonic crystals,” *IEEE Trans. Magn.*, vol. 47, no. 5, pp. 1534–1537, May 2011. 63

[54] C. Scheiber, O. Bíró, and R. Dyczij-Edlinger, “Full-wave band structure calculation of photonic crystals with a model-order reduction scheme,” *Workshop on Finite Elements for Microwave Engineering*, pp. 48–48, 2010. 63, 65

[55] A. Schultschik, O. Farle, and R. Dyczij-Edlinger, “A model order reduction method for the finite-element simulation of inhomogeneous waveguides,” *IEEE Trans. Magn.*, vol. 44, no. 6, pp. 1394–1397, Jun. 2008. 64

[56] C. Scheiber, O. Bíró, and R. Dyczij-Edlinger, “A single point model-order reduction scheme for band structure calculations of photonic crystals,” *Proceedings of the 14th International IGTE Symposium*, pp. 103–103, 2010. 68

[57] S.-H. Lee, T.-Y. Huang, and R.-B. Wu, “Fast waveguide eigenanalysis by wide-band finite-element model-order reduction,” *IEEE Trans. Microw. Theory Techniques*, vol. 53, no. 8, pp. 2552–2558, Aug. 2005. 68, 70

[58] D. Dobson, J. Gopalakrishnan, and J. Pasciak, "An efficient method for band structure calculations in 3d photonic crystals," *Journal of Comp. Physics*, vol. 161, pp. 668–679, 2000. 85

[59] K. Ho, C. Chan, C. Soukoulis, R. Biswas, and M. Sigalas, "Photonic band gaps in three dimensions: New layer-by-layer periodic structures," *Solid State Comm.*, vol. 89, no. 5, pp. 413–416, 1994. 87

[60] S. Lin, J. Fleming, D. Hetherington, B. Smith, R. Biswas, K. M. Ho, M. Sigalas, W. Zubrzycki, S. Kurtz, and J. Bur, "A three-dimensional photonic crystal operating at infrared wavelengths," *Nature*, vol. 394, pp. 251–253, 1998. 87, 88