DIPLOMARBEIT

Numerical Computation of Topological Properties of Photonic Crystals

ausgeführt am

Institut für
Analysis und Scientific Computing
TU Wien

unter der Anleitung von

Univ.Prof. Dipl.-Ing. Dr.techn. Joachim Schöberl

durch

Amanda Huber
Matrikelnummer: 00271473

Wien, am 20. August 2023

Abstract

In recent years huge advances were made in the field of topological photonics. One area of interest are photonic crystals with broken time reversal symmetry resulting in gaps in their band structure that prevent light propagation within specific frequency ranges. These photonic crystals hold promising applications such as topological photonic insulators. In this context, Chern numbers play an important role in characterizing the optical properties of such components. Numerically calculating the Chern number for an energy band requires solving a certain number of resonance problems. The amount depends on the experimental setup and the chosen computation method.

We apply the finite element method, in combination with a reduced basis approach, to efficiently obtain band structures of 2D photonic crystals. Furthermore, our approach allows us to consider problems with nonlinear frequency-dependent permittivities and permeabilities. Employing this method to compute resonance frequencies, we compare two ways of computing Chern numbers: the first principal calculation and the Wilson loop approach. All implementations are conducted using the high-performance multiphysics finite element software Netgen/NGSolve.

We demonstrate that, even with significantly reduced dimensions of the system, accurate Chern numbers can be obtained. Additionally, we are able to calculate Chern numbers for photonic crystals with highly frequency dependent material parameters.

Acknowledgement

First and foremost I want to thank my supervisor Prof. Joachim Schöberl for all his support during the past three years. He gave me the opportunity to get a glimpse at the scientific work done in his department and I learned a lot from him and his colleagues. Furthermore he was very supportive and accommodating throughout my pregnancy and later after my son was born.

Furthermore I want to thank the entire workgroup on Computational Mathematics in Engineering for making me feel welcome on the third floor and always having an open ear for my questions even at lunch time. My special thanks goes to Michael Leumüller who proofread most of my thesis and gave me very detailed and helpful feedback.

I also want to thank Prof. Florian Libisch for taking the time to discuss some questions I had about the physical assumptions that are commonly made in context of photonic crystals. Great thanks also to my two best friends from school, Hanna and Max, who went on to study physics and chemistry and helped a lot in understanding the general physical background.

The time at TU Vienna studying mathematics will be always in my mind as one of the most fun and interesting ones in my life. This is not at last due to the incredible people I was lucky to meet and become friends with: Theresa, Thomas, Thomas, Hubert, Martin, Lorenz and many more.

Additionally I want to thank my parents for their love and support through everything I did in my life.

Last but not least I want to thank the love of my life who has been a wonderful partner for over 10 years and recently did more that his fair share in caring for our son so that I have the time to finish my studies.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Diplomarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 20. August 2023

Amanda Huber

Contents

1 Introduction

Chern numbers were first described in the context of quantum physics. In 1982 the authors Thouless, Kohmoto, Nightingale and Nijs managed to show, that the Hall conductance for one energy band, arising from an electron in a periodic electron potential under a strong perpendicular magnetic field, is a topological invariant that is always given by an integer multiple of e^2/h, where h denotes Planck constant [18]. Meanwhile, Berry discovered, that if an eigenstate of a quantal system is slowly transported around a closed circuit in a parameter space, it will accumulate a phase factor [4]. This factor would later be known as the Berry phase. A year later Simon managed to relate Berry's phase and the TKNN numbers by expressing both in terms of differential geometry. He was able to show that the Berry phase is the holonomy in a Hermitian line bundle and the TKNN numbers are precisely the integral invariants, called Chern numbers, of such a bundle [4]. Finally, effects analogous to the quantum hall effect in condensed matter physics were discovered in photonic crystals with broken time reversal symmetry [7]. Today Chern numbers play an important role in the field of topological photonics. For more information the work of Lu, Joannopoulos and Soljačić [13] is highly recommended.

In **Chapter 2** we discuss what photonic crystals are and what kind of photonic crystals we consider throughout this thesis. Furthermore we give a derivation of the governing equation for transversal magnetically polarized electromagnetic waves propagating through this kind of photonic crystals. It turns out that the light propagation mathematically comes down to an eigenvalue problem of shape

$$H_{\boldsymbol{k}} u = \omega^2 u, \tag{1.1}$$

where $H_{\boldsymbol{k}}$ is a Hermitian operator depending on the so called wave vector $\boldsymbol{k}$ and ω turns out to be the frequency of the electromagnetic wave. Additionally we formally introduce Chern numbers in an abstract setting.

The overall goal of **Chapter 3** is to describe the numerical methods employed to obtain an approximation for the eigenpairs of (1.1). At first we derive a weak formulation and show that solutions of the kind we are looking for exist. Especially, that we can expect real valued frequencies ω. Then we apply a Galerkin discretization and justify that the solutions of the discrete problem converge to the continuous eigenpairs. If the operator H does not depend on ω, the discretized version of (1.1) constituted a general eigenvalue problem for a fixed wave vector $\boldsymbol{k}$. If however the operator does depend on the frequency, we can instead fix ω and solve a quadratic eigenvalue problem with an eigenvalue related to $\boldsymbol{k}$ and eigenfunction u. This happens, if the permittivity or the permeability of a material contained in the photonic crystal is frequency dependent. Subsequently we apply the reduced basis method as a model order reduction to both problems.

In **Chapter 4** we describe the first principal calculation [24] and the Wilson loop approach [22] to calculate Chern numbers numerically.

The results presented in **Chapter 5** show, that a reduced basis space of relatively small dimension is required to obtain accurate Chern numbers. This holds true for problems with frequency dependent and frequency independent material parameters. Applying the first principal calculation for problems with frequency independent material parameters additionally yields a fast method to compute and plot Berry curvatures. Meanwhile, the Wilson loop approach is not only generally faster, but also better compatible with our way to deal with frequency dependent material parameters. The combination of fixing ω and solving (1.1) for a pair $(u, \boldsymbol{k})$ and employing the Wilson loop approach allows a fast calculation of Chern numbers for highly frequency dependent material parameters. Another advantage of the Wilson loop is the possibility to visually check the plausibility our results. As a side effect we can efficiently compute band structures of photonic crystals, not only for frequency independent material parameters, but also for permittivities and permeabilities that depend non linearly on ω.

2 Photonic Crystals and Chern Numbers

In this chapter we cover the basic physical and mathematical background required to understand what it is we calculate numerically.

2.1 Photonic Crystals

A photonic crystal (PC) is a periodic arrangement of macroscopic media with different dielectric properties. The pattern in which the material is repeated is called a *lattice* [8]. We will consider PCs that have a discrete translational symmetry in two dimensions (the (x, y)-plane) and are homogeneous in the third dimension (along the $z-$axis). This arrangement is called a two-dimensional PC. The crystal can be regarded as the periodic repetition of some *unit cell* Ω. In our case the unit cell consists of a rod of some gyromagnetic material in air. By *gyromagnetic* we mean that the material's dielectric properties are altered by the presence of a magnetic fields. One very prominent example is Yittrium-Iron-Garnet (YIG) [14]. Figure 2.1 shows a schematic visualization.

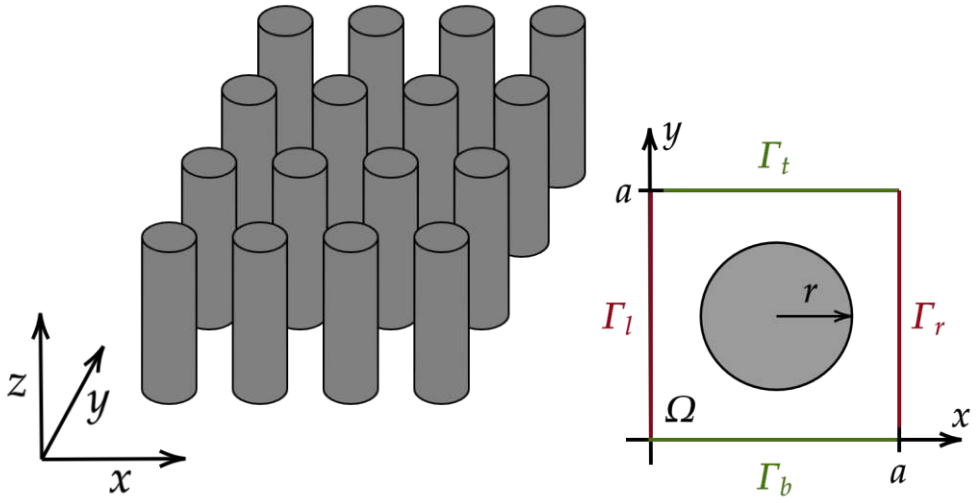


Figure 2.1: 2D PC consisting of gyromagnetic rods in air with unit cell Ω.

2.1.1 Derivation of Governing Equations for 2D Photonic Crystal Modes

To derive Equation (2.8) we will follow the steps given in [8], except that in our case some simplifications cannot be made.

The propagation of electromagnetic waves, and therefore light, is governed by the four macroscopic Maxwell equations

$$\begin{aligned} \operatorname{div} \boldsymbol{B} &= 0, & \operatorname{curl} \boldsymbol{E} + \partial_t \boldsymbol{B} &= \boldsymbol{0}, \\ \operatorname{div} \boldsymbol{D} &= \rho, & \operatorname{curl} \boldsymbol{H} - \partial_t \boldsymbol{D} &= \boldsymbol{J}, \end{aligned} \tag{2.1}$$

where

- $\boldsymbol{H}$ is the magnetic field,
- $\boldsymbol{E}$ is the electric field,
- $\boldsymbol{B}$ is the magnetic induction (or displacement) field,
- $\boldsymbol{D}$ is the electric induction (or displacement) field,
- $\boldsymbol{J}$ is the current density and
- ρ is the free charge.

In our case there are no free charges or currents, so $\rho = 0$ and $\boldsymbol{J} = \boldsymbol{0}$. The Hermitean *permittivity* tensor $\varepsilon_0 \boldsymbol{\varepsilon}$ relates the electric field and the electric induction by $\boldsymbol{D} = \varepsilon_0 \boldsymbol{\varepsilon} \boldsymbol{E}$, where $\varepsilon_0 \approx 8.854 \times 10^{-12}$ Farad/m is the vacuum permittivity. Analogously the Hermitean *permeability* tensor $\mu_0 \boldsymbol{\mu}$ relates the magnetic field and the magnetic induction by $\boldsymbol{B} = \mu_0 \boldsymbol{\mu} \boldsymbol{H}$, where $\mu_0 = 4\pi \times 10^{-7}$ Henry/m is the vacuum permeability. We are interested in gyromagnetic photonic crystals. This means that according to [14] the ferrite permeability tensor is given by

$$\boldsymbol{\mu} = \begin{pmatrix} \mu & i\kappa & 0 \\ -i\kappa & \mu & 0 \\ 0 & 0 & \mu_{3,3} \end{pmatrix}, \quad \mu = 1 + \frac{\omega_m \omega_0}{\omega_0^2 - \omega^2}, \quad \kappa = \frac{\omega_m \omega}{\omega_0^2 - \omega^2} \tag{2.2}$$

with $\omega_0 = \gamma H_0$, $\omega_m = \gamma 4\pi M_s$ and $i = \sqrt{-1}$. The parameters are

- the gyromagnetic ratio γ [C/kg],
- the material dependent magnetic saturation M_s [T] and
- the magnetic field strength H_0 [T].

For our example $\mu_{3,3}$ is neglectable because it will never be used, as we will see later. From now on we assume that the inverse

$$\boldsymbol{\mu}^{-1} = \frac{1}{\mu^2 - \kappa^2} \begin{pmatrix} \mu & -i\kappa & 0 \\ i\kappa & \mu & 0 \\ 0 & 0 & \frac{\mu^2 - \kappa^2}{\mu_{3,3}} \end{pmatrix} \tag{2.3}$$

exits, is piecewise constant and bounded. Furthermore we impose that

$$\frac{\mu}{\mu^2 - \kappa^2} > 0. \tag{2.4}$$

The reason for that will be explained in Chapter 3. The permittivity ε is a non-negative and piece wise constant function of $\boldsymbol{r}$ and bounded as a function of ω.

Altogether the Maxwell Equations (2.1) become

$$\begin{aligned} \operatorname{div}(\mu_0 \boldsymbol{\mu} \boldsymbol{H}) &= 0, & \operatorname{curl} \boldsymbol{E} + \partial_t(\mu_0 \boldsymbol{\mu} \boldsymbol{H}) &= \boldsymbol{0}, \\ \operatorname{div}(\varepsilon_0 \boldsymbol{\varepsilon} \boldsymbol{E}) &= 0, & \operatorname{curl} \boldsymbol{H} - \partial_t(\varepsilon_0 \boldsymbol{\varepsilon} \boldsymbol{E}) &= \boldsymbol{0}. \end{aligned} \tag{2.5}$$

In general $\boldsymbol{H}$ and $\boldsymbol{E}$ are dependent on time and space. We assume that $\boldsymbol{\varepsilon}$ and $\boldsymbol{\mu}$ are time independent. In order to separate the space component $\boldsymbol{r}$ from the time component t, we expand the fields into a set of harmonic modes. According to Fourier's theorem any solution can be build as a combination of these modes (though possibly infinitely many of them). For a given frequency ω we assume time harmonic modes

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{r}, t) &= \boldsymbol{E}(\boldsymbol{r}) e^{-i\omega t}, \\ \boldsymbol{H}(\boldsymbol{r}, t) &= \boldsymbol{H}(\boldsymbol{r}) e^{-i\omega t}. \end{aligned}$$

After inserting above equations into (2.5) we can use the curl components

$$\operatorname{curl} \boldsymbol{E}(\boldsymbol{r}) = i\omega\mu_0 \boldsymbol{\mu} \boldsymbol{H}(\boldsymbol{r}) \tag{2.6}$$

$$\operatorname{curl} \boldsymbol{H}(\boldsymbol{r}) = -i\omega\, \varepsilon_0\, \boldsymbol{\varepsilon} \boldsymbol{E}(\boldsymbol{r}). \tag{2.7}$$

to relate $\boldsymbol{H}$ and $\boldsymbol{E}$. Multiplying (2.6) with $\boldsymbol{\mu}^{-1}$, applying curl to both sides and subsequently using (2.7) yields

$$\operatorname{curl}(\boldsymbol{\mu}^{-1} \operatorname{curl} \boldsymbol{E}(\boldsymbol{r})) = \omega^2\, \varepsilon_0\, \mu_0 \boldsymbol{\varepsilon} \boldsymbol{E}(\boldsymbol{r}).$$

Now we use that the vacuum speed of light c is connected to ε_0 and μ_0 via the formula $1/c^2 = \varepsilon_0\, \mu_0$. Finally, the equation describing the modes of a PC is given by

$$\operatorname{curl}\left(\boldsymbol{\mu}^{-1} \operatorname{curl} \boldsymbol{E}(\boldsymbol{r})\right) = \left(\frac{\omega}{c}\right)^2 \boldsymbol{\varepsilon} \boldsymbol{E}(\boldsymbol{r}). \tag{2.8}$$

We have already established that our PC has a discrete translational symmetry in two dimensions (the (x, y)-plane) and is homogeneous in the third dimension (along the $z-$axis). We are interested in the propagation of light in a 2D plane throughout the PC. Therefore we can assume that $\boldsymbol{r} = \begin{pmatrix} x & y & 0 \end{pmatrix}^T$ and $\boldsymbol{k} = \begin{pmatrix} k_x & k_y & 0 \end{pmatrix}^T$. Provided the lattice grid is spanned by two primitive lattice vectors $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$, meaning they are the smallest vectors pointing from one lattice point to another, any lattice vector can be written as

$$\boldsymbol{R} = \boldsymbol{a}_1 n_1 + \boldsymbol{a}_2 n_1$$

where $\boldsymbol{a}_i \in \mathbb{R}^2$ and $n_i \in \mathbb{N}$. For some arbitrary but fixed lattice point $\boldsymbol{p}$ we define the unit cell Ω as the rectangle with corners $\boldsymbol{p}$, $\boldsymbol{p} + \boldsymbol{a}_1$, $\boldsymbol{p} + \boldsymbol{a}_1 + \boldsymbol{a}_2$, $\boldsymbol{p} + \boldsymbol{a}_2$. The boundary $\Gamma = \partial\Omega$ can be split into $\Gamma = \Gamma^b \cup \Gamma^r \cup \Gamma^t \cup \Gamma^l$ illustrated in Figure 2.1.

Definition 2.1. A function u on Ω is called *periodic on* Γ if $u\big|_{\Gamma^b} = u\big|_{\Gamma^t}$ and $u\big|_{\Gamma^l} = u\big|_{\Gamma^r}$.

Now, the goal is to use the symmetries of the crystal to characterise its electromagnetic modes. The Bloch Theorem states, that we can write these modes as a planar wave $e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$ modulated by a function $\boldsymbol{u}(\boldsymbol{r})$ with the same periodicity as the PC. In this chapter we assume that $\boldsymbol{u} = \begin{pmatrix} u_1 & u_2 & u_3 \end{pmatrix}^T$ with

$$u_j \in C_p^2(\Omega) := \{u \in C^2(\Omega) \mid u \text{ is periodic on } \Gamma\ \}.$$

For a wave vector $\boldsymbol{k}$ the Bloch Theorem can be written as

$$\boldsymbol{E_k}(\boldsymbol{r}) = e^{i\boldsymbol{k}\cdot\boldsymbol{r}}\boldsymbol{u}(\boldsymbol{r}) = e^{i\boldsymbol{k}\cdot\boldsymbol{r}}\boldsymbol{u}(\boldsymbol{r}+\boldsymbol{R}).$$

We are only interested in the modes propagating in the $(x,y)-$plane, therefore we set $\boldsymbol{k} = \begin{pmatrix} k_x & k_y & 0 \end{pmatrix}^T$ in the equation above. This leads to a 2D wave propagating in the $(x,y)-$plane.

We know that $\boldsymbol{E_k}$ is lattice periodic hence

$$\boldsymbol{E_k}(\boldsymbol{r}+\boldsymbol{R}) = e^{i\boldsymbol{k}\cdot(\boldsymbol{r}+\boldsymbol{R})}\boldsymbol{u}(\boldsymbol{r}) = e^{i\boldsymbol{k}\cdot\boldsymbol{R}}\boldsymbol{E_k}(\boldsymbol{r}). \tag{2.9}$$

We call a function that fulfills (2.9) *k-Bloch periodic*. The expression $e^{i\boldsymbol{k}\cdot\boldsymbol{R}}$ is periodic as a function of $\boldsymbol{k}$ with value 1 if $\boldsymbol{k}\cdot\boldsymbol{R} = 2\pi l$ for an integer l. The space containing all the wave vectors is called the *reciprocal lattice space*. Any lattice vector $\boldsymbol{G}$ in the reciprocal space can be written as

$$\boldsymbol{G} = \boldsymbol{b}_1 m_1 + \boldsymbol{b}_2 m_2$$

where $\boldsymbol{b}_i \in \mathbb{R}^2$ and $m_i \in \mathbb{N}$. We want to choose suitable basis vectors $\boldsymbol{b}_i$ that fulfil the condition

$$\boldsymbol{G}\cdot\boldsymbol{R} = l2\pi$$

for some $l \in \mathbb{N}$. In matrix notation this reads as

$$\boldsymbol{G}\cdot\boldsymbol{R} = \begin{pmatrix} n_1 & n_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{a_1}\cdot\boldsymbol{b_1} & \boldsymbol{a_1}\cdot\boldsymbol{b_2} \\ \boldsymbol{a_2}\cdot\boldsymbol{b_1} & \boldsymbol{a_2}\cdot\boldsymbol{b_2} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = l2\pi.$$

For convenience we also want to impose $\boldsymbol{a}_i\cdot\boldsymbol{b}_i = \delta_{ij}2\pi$. This yields the basis vectors

$$\boldsymbol{b_1} = \frac{2\pi}{\boldsymbol{a_1}\times\boldsymbol{a_2}}\begin{pmatrix} a_{22} \\ -a_{21} \end{pmatrix}, \qquad \boldsymbol{b_2} = \frac{2\pi}{\boldsymbol{a_1}\times\boldsymbol{a_2}}\begin{pmatrix} -a_{12} \\ a_{11} \end{pmatrix}.$$

Now, the *first Brillouin zone (BZ)* can be defined as the smallest volume entirely enclosed by planes, that are perpendicular bisectors of the basis reciprocal lattice vectors drawn from the origin.

From now on we will assume a quadratic grid of cell length a, meaning

$$\boldsymbol{a}_1 = a\begin{pmatrix} 1 \\ 0 \end{pmatrix},\ \boldsymbol{a}_2 = a\begin{pmatrix} 0 \\ 1 \end{pmatrix},\ \boldsymbol{b}_1 = \frac{2\pi}{a}\begin{pmatrix} 1 \\ 0 \end{pmatrix},\ \boldsymbol{b}_2 = \frac{2\pi}{a}\begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{2.10}$$

The corresponding first BZ is

$$\left(-\frac{\pi}{a},\frac{\pi}{a}\right)\times\left(-\frac{\pi}{a},\frac{\pi}{a}\right). \tag{2.11}$$

In this case the so called irreducibly Brillouin zone is a triangle with corners

$$\Gamma=\begin{pmatrix}0\\0\end{pmatrix},\quad X=\begin{pmatrix}\frac{\pi}{a}\\0\end{pmatrix},\quad M=\begin{pmatrix}\frac{\pi}{a}\\\frac{\pi}{a},\end{pmatrix}$$

as illustrated in Figure 2.2. It is the smallest area such that the rest of the BZ can be obtained by rotation, mirror-reflection or inversion of the irreducible BZ [8].

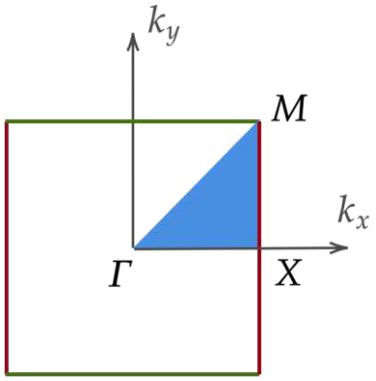


Figure 2.2: First BZ.

Furthermore we only consider $\boldsymbol{k}=\begin{pmatrix}k_x & k_y & 0\end{pmatrix}^T\in \mathrm{BZ}\times\{0\}$, a so called *Bloch wave vector*. The associated mode

$$\boldsymbol{E_k}(\boldsymbol{r})=e^{i\boldsymbol{k}\cdot\boldsymbol{r}}\boldsymbol{u}(\boldsymbol{r}) \tag{2.12}$$

is called the *Bloch mode*.

In 2D PC Bloch modes can be split into two distinct polarizations, the transversal-magnetic (TM) and transversal-electric (TE) modes. TM modes are characterised by a z-polarized electric fields, while TE modes have a z-polarized magnetic field. From now on we will only consider TM modes. All calculations can be done analogously for TE modes.

We make the ansatz

$$\boldsymbol{E_k}=\begin{pmatrix}0\\0\\e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u(\boldsymbol{r})\end{pmatrix}=:\begin{pmatrix}0\\0\\E_{\boldsymbol{k}}\end{pmatrix} \tag{2.13}$$

for a lattice periodic function $u\in C^2_P(\Omega)$. Putting (2.13) into (2.8) and defining $\varepsilon=\varepsilon_{3,3}$ yields

$$\begin{aligned}
\operatorname{curl}(\boldsymbol{\mu}^{-1}\operatorname{curl}\boldsymbol{E_k}) &= \operatorname{curl}\left(\boldsymbol{\mu}^{-1}\begin{pmatrix}\partial_y E_{\boldsymbol{k}}\\-\partial_x E_{\boldsymbol{k}}\\0\end{pmatrix}\right)\\
&= \operatorname{curl}\begin{pmatrix}\mu^{-1}_{1,1}\partial_y E_{\boldsymbol{k}}-\mu^{-1}_{1,2}\partial_x E_{\boldsymbol{k}}\\ \mu^{-1}_{2,1}\partial_y E_{\boldsymbol{k}}-\mu^{-1}_{2,2}\partial_x E_{\boldsymbol{k}}\\0\end{pmatrix}\\
&= \begin{pmatrix}0\\0\\ \partial_x\left(\mu^{-1}_{2,1}\partial_y E_{\boldsymbol{k}}-\mu^{-1}_{2,2}\partial_x E_{\boldsymbol{k}}\right)-\partial_y\left(\mu^{-1}_{1,1}\partial_y E_{\boldsymbol{k}}-\mu^{-1}_{1,2}\partial_x E_{\boldsymbol{k}}\right)\end{pmatrix}\\
&\overset{(2.8)}{=} \left(\frac{\omega}{c}\right)^2\begin{pmatrix}0\\0\\\varepsilon\, E_{\boldsymbol{k}}\end{pmatrix}
\end{aligned}$$

The $x-$ and $y-$coordinates are all trivial and we only have to consider the equation in the $z-$coordinate. Together with the Hermitian nature of $\boldsymbol{\mu}^{-1}$ as defined in (2.3) we get

$$\begin{aligned}\left(\frac{\omega}{c}\right)^2 \varepsilon E_{\boldsymbol{k}} &= \operatorname{div}\left(\mu_{2,1}^{-1}\partial_y E_{\boldsymbol{k}} - \mu_{2,2}^{-1}\partial_x E_{\boldsymbol{k}} - \mu_{1,1}^{-1}\partial_y E_{\boldsymbol{k}} + \mu_{1,2}^{-1}\partial_x E_{\boldsymbol{k}}\right) \\ &= \operatorname{div}\begin{pmatrix} -\mu_{2,2}^{-1} & \mu_{2,1}^{-1} \\ \mu_{1,2}^{-1} & -\mu_{1,1}^{-1}\end{pmatrix}\begin{pmatrix}\partial_x E_{\boldsymbol{k}} \\ \partial_y E_{\boldsymbol{k}}\end{pmatrix} \\ &= \operatorname{div}\left(-\boldsymbol{\mu}^{-1}\nabla E_{\boldsymbol{k}}\right).\end{aligned}$$

From now on we consider $\boldsymbol{k}$ and $\boldsymbol{r}$ to be two-dimensional vectors and $\boldsymbol{\mu}^{-1}$ a two dimensional matrix. Altogether we arrive at the following formulation.

Problem 2.2 (TM modes)**.** Let $\boldsymbol{k} = \begin{pmatrix} k_x & k_y\end{pmatrix}^T \in \mathrm{BZ}$ be an arbitrary but fixed wave vector. Find $u \in C_p^2(\Omega)$ and $\omega \in \mathbb{R}$ such that

$$-\operatorname{div}(\boldsymbol{\mu}^{-1}\nabla(e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u(\boldsymbol{r}))) = \left(\frac{\omega}{c}\right)^2 \varepsilon(\boldsymbol{r})e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u(\boldsymbol{r}). \tag{2.14}$$

2.2 An Abstract Approach to Chern Numbers

This section is built on [19]. Assuming that we have a complex vector space $\mathcal{V}$ with the scalar product $\langle .,. \rangle$, which is linear in the first and semi-linear in the second argument and a parameter space $\mathcal{P}$. The vectors (or *states*) $\boldsymbol{u}(\boldsymbol{k}) \in \mathcal{V}$ depend on the parameters $\boldsymbol{k} \in \mathcal{P}$. Note that there is no function $\boldsymbol{k} \mapsto \boldsymbol{u}(\boldsymbol{k})$. However we impose that for one $\boldsymbol{k}$ all possible values $\boldsymbol{u}(\boldsymbol{k})$ only differ in phase and magnitude (meaning by a factor $z = |z|e^{i\varphi}$). Let

$$\Gamma : [0,1] \to \mathcal{P}, \quad t \mapsto \Gamma(t)$$

be a closed path in the parameter space. Now we want to parallel transport one state $\boldsymbol{u}(\boldsymbol{k})$ around that loop. Therefore we need to define what parallel means in this context.

Definition 2.3. Two states $\boldsymbol{u}^1$ and $\boldsymbol{u}^2$ are called *parallel* if $\langle \boldsymbol{u}^1, \boldsymbol{u}^2\rangle$ is real valued and positive.

For every $t \in [0,1)$, h small with $t + h \in [0,1)$ we impose the condition

$$\operatorname{Im}\ \ln\langle \boldsymbol{u}(\boldsymbol{k}(t)), \boldsymbol{u}(\boldsymbol{k}(t+h))\rangle = 0$$

on our choices of $\boldsymbol{u}$ around the loop Γ. To get a well defined value for the complex logarithm we need to choose a branch. From now on we restrict $\ln z$ to the interval $(-\pi, \pi]$.

At $\Gamma(0) = \Gamma(1)$ the original vector $\boldsymbol{u}(\boldsymbol{k}(0))$ differs from the parallel transported vector $\boldsymbol{u}(\boldsymbol{k}(1))$ by a complex factor $z \in \mathbb{C}$. We are only interested in the phase difference

$$\phi = \operatorname{Im}\ \ln\langle \boldsymbol{u}(\boldsymbol{k}(0)), \boldsymbol{u}(\boldsymbol{k}(1))\rangle, \tag{2.15}$$

the so called *Berry phase*. In a nutshell the Berry phase is a phase angle accumulated by a vector that is parallel transported around a closed loop in parameter space $\mathcal{P}$.

Now we ask ourselves how the Berry phase ϕ can be calculated. Let $\boldsymbol{k}^1, \dots, \boldsymbol{k}^N$ be a discretization of Γ as visualized in Figure 2.3. Choose some states $\boldsymbol{u}^1(\boldsymbol{k}^1), \dots, \boldsymbol{u}^N(\boldsymbol{k}^N)$ with $\boldsymbol{u}^1 = \boldsymbol{u}^N$. For better readability we omit the explicit dependence on $\boldsymbol{k}$ for now. The goal is to find $\tilde{\boldsymbol{u}}^2, \dots, \tilde{\boldsymbol{u}}^N$ such that $\tilde{\boldsymbol{u}}^N$ is the parallel transported version of $\boldsymbol{u}^1$ around Γ. Assume that $\tilde{\boldsymbol{u}}^j$ is a parallel transported version of $\boldsymbol{u}^1$ and $j+1 \leq N$. We have

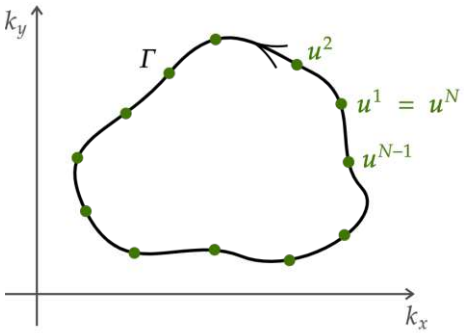


Figure 2.3: Discretized path Γ.

$$\langle \tilde{\boldsymbol{u}}^j, \boldsymbol{u}^{j+1} \rangle = |z| e^{i\varphi},$$

so by setting $\tilde{\boldsymbol{u}}^{j+1} := \langle \tilde{\boldsymbol{u}}^j, \boldsymbol{u}^{j+1} \rangle \boldsymbol{u}^{j+1}$ we get

$$\langle \boldsymbol{u}^j, \tilde{\boldsymbol{u}}^{j+1} \rangle = \langle \boldsymbol{u}^j, \boldsymbol{u}^{j+1} \rangle |z|^2 e^{i\varphi} e^{-i\varphi} = |z|^2 e^{i0},$$

which is exactly the parallel transport condition we have imposed. The last vector is

$$\tilde{\boldsymbol{u}}^N = \langle \boldsymbol{u}^1, \boldsymbol{u}^2 \rangle \langle \boldsymbol{u}^2, \boldsymbol{u}^3 \rangle \dots \langle \boldsymbol{u}^{N-1}, \boldsymbol{u}^N \rangle \boldsymbol{u}^N.$$

Using the fact that $\boldsymbol{u}^1 = \boldsymbol{u}^N$ and therefore $\langle \boldsymbol{u}^N, \boldsymbol{u}^1 \rangle > 0$, the Berry phase can be written as

$$\begin{aligned} \phi &= \operatorname{Im} \ln \langle \boldsymbol{u}^1, \tilde{\boldsymbol{u}}^N \rangle \\ &= -\operatorname{Im} \ln \left(\langle \boldsymbol{u}^1, \boldsymbol{u}^2 \rangle \langle \boldsymbol{u}^2, \boldsymbol{u}^3 \rangle \dots \langle \boldsymbol{u}^{N-1}, \boldsymbol{u}^N \rangle \right) \\ &= \sum_{j=1}^{N-1} -\operatorname{Im} \ln \langle \boldsymbol{u}^j, \boldsymbol{u}^{j+1} \rangle \mod 2\pi. \end{aligned} \tag{2.16}$$

Note that the Berry phase is often defined as $-\phi$ in (2.15). However our scalar products are conjugated in the second argument, so for consistency with (2.16) we define it like that.

We can see that the result in (2.16) does not depend on concrete choices for the phase of the states $\boldsymbol{u}^1, \dots, \boldsymbol{u}^N$. We can apply an arbitrary *gauge transformation* (meaning a multiplication with $e^{i\beta_j}$) to all states. Each of the factors $e^{i\beta_j}$ appear exactly once on the right side of the scalar product and once on the left side. Hence all the factors cancel out.

In a next step we want to get rid of the discretization of Γ and calculate the Berry phase in a continuous way. In a first step we assume that we have a one dimensional parameter space with $\Gamma = [0, 1]$. We choose states $\boldsymbol{u}(t)$ on Γ such that $\boldsymbol{u}(0) = \boldsymbol{u}(1)$ and $t \mapsto \boldsymbol{u}(t)$ is a well defined differentiable function. Note that it does not have to be continuously differentiable. Using Taylor expansion we can write

$$\begin{aligned} \ln \langle \boldsymbol{u}(t), \boldsymbol{u}(t + \Delta t) \rangle &= \ln \langle \boldsymbol{u}(t), \boldsymbol{u}(t) + \Delta t \partial_t \boldsymbol{u}(t) + \dots \rangle \\ &= \ln \left(1 + \Delta t \langle \boldsymbol{u}(t), \partial_t \boldsymbol{u}(t) \rangle + \dots \right) \\ &= \Delta t \langle \boldsymbol{u}(t), \partial_t \boldsymbol{u}(t) \rangle + \dots, \end{aligned}$$

where the last equation holds because $\ln(1+z) = \sum_j (-1)^{j+1} z^j / j$.
Plugging that expression into (2.16) and considering the limit $\Delta t \to 0$ yields yet another expression for the Berry phase

$$\phi = \oint_\Gamma - \operatorname{Im} \langle \boldsymbol{u}(t), \partial_t \boldsymbol{u}(t) \rangle \, \mathrm{d}t \,. \tag{2.17}$$

The integrand

$$A(t) = - \operatorname{Im} \langle \boldsymbol{u}(t), \partial_t \boldsymbol{u}(t) \rangle \tag{2.18}$$

is called the *Berry connection* or *Berry potential*.

Lemma 2.4. *The Berry potential* (2.18) *is not gauge invariant. The Berry phase as defined in* (2.17) *is gauge invariant up to an integer multiple of* 2π.

Proof. Define $\tilde{\boldsymbol{u}}(t) = e^{i\beta(t)} \boldsymbol{u}(t)$ then

$$\begin{aligned} \tilde{A}(t) &= - \operatorname{Im} \langle \tilde{\boldsymbol{u}}(t), \partial_t \tilde{\boldsymbol{u}}(t) \rangle \\ &= - \operatorname{Im} \langle e^{i\beta(t)} \boldsymbol{u}(t), -i\beta'(t) e^{i\beta(t)} \boldsymbol{u}(t) + e^{i\beta(t)} \partial_t \boldsymbol{u}(t) \rangle \\ &= \beta'(t) + A(t). \end{aligned}$$

For the second statement of the lemma we calculate (2.17) for $\tilde{A}$. This yields

$$\oint_\Gamma \tilde{A}(t) \, \mathrm{d}t = \beta(1) - \beta(0) + \oint_\Gamma A(t) \, \mathrm{d}t \,.$$

The condition $\tilde{\boldsymbol{u}}(0) = \tilde{\boldsymbol{u}}(1)$ together with $\tilde{\boldsymbol{u}}(t) = \boldsymbol{u}(t) e^{i\beta(t)}$ implies that $\beta(1) - \beta(0) = 2\pi m$ with $m \in \mathbb{N}$. □

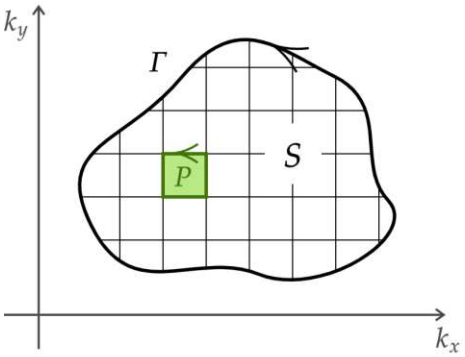


Figure 2.4: Surface S discretized by plaquettes P.

Now let Γ be a closed path in a two–dimensional parameter space. The states depend on $\boldsymbol{k} = (k_x, k_y) \in \Gamma$. The Berry potential (2.18) takes the form $\boldsymbol{A} = (A_x, A_y)$ with

$$A_j = - \operatorname{Im} \langle \boldsymbol{u}(\boldsymbol{k}), \partial_{k_j} \boldsymbol{u}(\boldsymbol{k}) \rangle \tag{2.19}$$

for $j \in \{x, y\}$. The Berry phase (2.17) can be written as

$$\phi = \oint_\Gamma \boldsymbol{A}(\boldsymbol{k}) \cdot \mathrm{d}\mathbf{k}. \tag{2.20}$$

With the same argument as before this expression is well defined up to an integer multiple of 2π.

We set Γ to be a closed loop in a two-dimensional space. Hence it encloses as surface S. We can divide S into plaquettes P as visualized in Figure 2.4. Each P should be tiny enough such that the Berry phase (2.20) calculated around the path ∂P is so small that no integer

multiple is added. In other words the Berry phases around the paths ∂P are unambiguous.

Now we consider the sum

$$\Phi_S = \sum_P \oint_{\partial P} \boldsymbol{A}(\boldsymbol{k}) \cdot \mathrm{d}\mathbf{k}. \tag{2.21}$$

Every path that is not in $(\bigcup_P \partial P) \cap \Gamma$ cancels out because it is integrated over exactly once in each direction. Hence there is an $m \in \mathbb{N}$ and $\phi \in [0, 2\pi)$ such that

$$\oint_\Gamma \boldsymbol{A}(\boldsymbol{k}) \cdot \mathrm{d}\mathbf{k} = \phi + 2\pi m = \Phi_S. \tag{2.22}$$

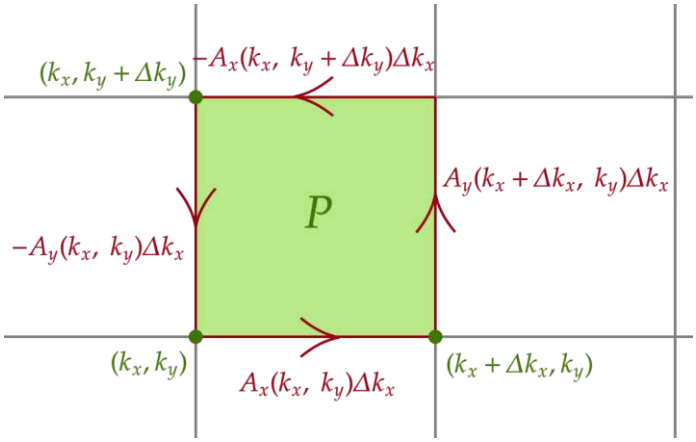


Figure 2.5: Approximate Berry phase per plaquette.

Assume that all plaquettes are axis aligned rectangles. Now let's have a closer look at one of the plaquettes with the lower left corner (k_x, k_y) and the upper right corner $(k_x + \Delta k_x, k_y + \Delta k_y)$ as visualized in Figure 2.5. The Berry phase per plaquette can be approximated by

$$\begin{aligned}\frac{\phi_P}{|P|} &\approx \frac{A_x(k_x,k_y)\Delta k_x + A_y(k_x+\Delta k_x,k_y)\Delta k_y - A_x(k_x,k_y+\Delta k_y)\Delta k_x - A_y(k_x,k_y)\Delta k_y}{\Delta k_x \Delta k_y} \\ &= \frac{A_y(k_x+\Delta k_x,k_y) - A_y(k_x,k_y)}{\Delta k_x} - \frac{A_x(k_x,k_y+\Delta k_y) - A_x(k_x,k_y)}{\Delta k_y}.\end{aligned}$$

Now we can approximate Φ_S from (2.21) by

$$\Phi_S \approx \sum_P \left(\frac{A_y(k_x+\Delta k_x,k_y) - A_y(k_x,k_y)}{\Delta k_x} - \frac{A_x(k_x,k_y+\Delta k_y) - A_x(k_x,k_y)}{\Delta k_y} \right) |P|.$$

Considering the limit $\Delta k_x, \Delta k_y \to 0$ leads to the *Berry flux*

$$\Phi_S = \int_S \Omega(\boldsymbol{k}) \mathrm{d}\mathbf{k}, \tag{2.23}$$

where the integrand $\Omega(\boldsymbol{k}) := \partial_{k_x} A_y(\boldsymbol{k}) - \partial_{k_y} A_x(\boldsymbol{k}) = \operatorname{curl} \boldsymbol{A}(\boldsymbol{k})$ is called the *Berry curvature*. We can calculate

$$\begin{aligned}\Omega(\boldsymbol{k}) = &- \operatorname{Im} \Big(\langle \partial_{k_x} \boldsymbol{u}, \partial_{k_y} \boldsymbol{u} \rangle + \langle \boldsymbol{u}, \partial_{k_x} \partial_{k_y} \boldsymbol{u} \rangle \\ &+ \overline{\langle \partial_{k_x} \boldsymbol{u}, \partial_{k_y} \boldsymbol{u} \rangle} - \langle \boldsymbol{u}, \partial_{k_x} \partial_{k_y} \boldsymbol{u} \rangle \Big) \\ = &- 2 \operatorname{Im} \langle \partial_{k_x} \boldsymbol{u}, \partial_{k_y} \boldsymbol{u} \rangle.\end{aligned}$$

Lemma 2.5. *The Berry curvature is gauge invariant.*

Proof. As in proof of Lemma 2.4 we can determine the two-dimensional Berry potential under gauge transformation $\tilde{\boldsymbol{A}}(\boldsymbol{k}) = \boldsymbol{A}(\boldsymbol{k}) + \nabla \beta(\boldsymbol{k})$. Due to the fact that $\operatorname{curl} \nabla(.) = 0$ we immediately get that $\Omega(\boldsymbol{k}) = \operatorname{curl} \tilde{\boldsymbol{A}}(\boldsymbol{k}) = \operatorname{curl} \boldsymbol{A}(\boldsymbol{k})$. □

In a next step we will investigate where the possible difference of $2\pi m$ in (2.22) comes from. First we remember Stokes' Theorem.

Theorem 2.6 (Stokes)**.** *Let $G \subset \mathbb{R}^2$ be open and bounded and $\boldsymbol{F} = (F_x, F_y) : \overline{G} \to \mathbb{R}^2$ a continuously differentiable vector field. Then*

$$\int_G \operatorname{curl} \boldsymbol{F} = \oint_{\partial G} \boldsymbol{F}.$$

That means we only have $m = 0$ in (2.22) if the Berry connection $\boldsymbol{A}$ is continuously differentiable in $\overline{S}$. Remember that $\boldsymbol{A}$ is not gauge invariant, meaning it depends on the chosen gauge of the states $\boldsymbol{u}(\boldsymbol{k})$. So the real question is if we can find a smooth enough representation of $\boldsymbol{k} \mapsto \boldsymbol{u}(\boldsymbol{k})$ that is valid everywhere on $\overline{S}$. In general this does not have to be the case. Actually we will later see that the interesting problems do not have a globally smooth gauge. However locally smooth representations for $\boldsymbol{u}$ must exist everywhere on S in order to get a well defined expression for the Berry curvature $\Omega(k)$.

Now we have everything we need to formulate and proof the Chern Theorem.

Theorem 2.7 (Chern)**.** *Let S be a closed two–dimensional manifold. Then the Berry flux Φ_S is quantized to be 2π times an integer,*

$$\Phi_S = \int_S \Omega(\boldsymbol{k})\, d\boldsymbol{k} = 2\pi C$$

for some well defined $C \in \mathbb{Z}$. This C is called the Chern number*.*

Proof. The uniqueness of Φ_S follows directly from the gauge invariance of $\Omega(\boldsymbol{k})$. In the beginning of the chapter we required $\boldsymbol{u}(\boldsymbol{k})$ to be unique up to a complex factor z. The Berry connection $\boldsymbol{A}(\boldsymbol{k})$ and hence the Berry curvature $\Omega(\boldsymbol{k})$ are independent of the magnitude of $\boldsymbol{u}(\boldsymbol{k})$. Hence we can assume $z = e^{i\varphi(\boldsymbol{k})}$, which is exactly a gauge transformation.

It remains to proof that $C \in \mathbb{N}$. We know that $\boldsymbol{A}$ is locally continuously differentiable. Hence we find an atlas of patches P_j with $S \subset \bigcup_j P_j$ such that $\boldsymbol{A}_j = \boldsymbol{A}\big|_{P_j}$ is a continuously

differentiable Berry connection on $\overline{P_j}$ and $(\overline{P_j} \cap \overline{P_l}) \subset (\partial P_j \cup \partial P_l)$. Now we can apply Stokes' Theorem 2.6 and get a unique value

$$\Phi_j = \int_{P_j} \Omega = \phi_j + m_j 2\pi$$

for each patch. The manifold has no boundary. This means for every patch the boundary is traced twice, once in each direction. The difference between these integrals must be an integer multiple of 2π. This is exactly the Chern number C. □

Remark 2.8. *Non trivial Chern numbers can only arise if no gauge can be found such that* $\boldsymbol{A}$ *is continuously differentiable everywhere on* S*. In the context of PCs that happens if time reversal symmetry is broken. Time reversal symmetry means that if* $\boldsymbol{u}$ *satisfies* (2.14) *the complex conjugate* $\overline{\boldsymbol{u}}$ *is a solution as well for fixed* ω *and* $\boldsymbol{k}$ *[8]. For PCs containing gyromagnetic materials, the breaking of time reversal symmetry can be observed in the presence of strong magnetic fields [7].*

3 Numerical Methods for Calculating Photonic Crystal Modes

Throughout this Chapter we derive a weak formulation for Problem 2.2 and discuss the existence and properties of its solutions. Subsequently we use a finite element method to get a discretized version of the problem at hand. The latter can either be solved as a general eigenvalue problem (GEP) or a quadratic eigenvalue problem (QEP). Hence we describe methods to arrive at solutions for both. Last we give a quick overview over the model order reduction method employed in this thesis. The idea of solving (2.14) as a weak general eigenvalue problem, employing the FEM together with a model order reduction is not new. This was already done for example in [15].

In this chapter we will use the notation $\simeq, \lesssim, \gtrsim$ for "$=, \leq, \geq$ up to a constant". Matrices and vectors are written in bold letters. Entries are accessed as it is done in the programming language Python or by subscript. For a matrix $\boldsymbol{M}$ the Hermitian is denoted as $\boldsymbol{M}^*$. If not stated otherwise $\langle\ .\ ,\ .\ \rangle$ represents the Euclidean scalar product where the second entry is conjugated.

All implementations are conducted using the high performance multiphysics finite element software Netgen/NGSolve [16], [17].

3.1 Finite Element Method

Before a finite dimensional approximation of Problem 2.2 can be discussed we first have to take a closer look at the properties of the problem itself. We derive a weak formulation and use Galerkin discretization to arrive at a matrix form of the original problem.

3.1.1 Weak Formulation

In a first step we formulate a weak version of Problem 2.2. For that purpose let $\boldsymbol{k}$ be fixed but arbitrary. Let Ω be the unit cell of our lattice grid with the boundary $\Gamma = \partial\Omega$. Remember the Sobolev space

$$\mathcal{H}^1(\Omega) := \left\{u \in \mathcal{L}^2(\Omega) \mid \nabla u \in \mathcal{L}^2(\Omega)\right\}$$

with the scalar product

$$\langle u, v\rangle_{\mathcal{H}^1} = \int_\Omega u\overline{v} + \nabla u \cdot \overline{\nabla v}.$$

We know that a solution of (2.14) must be a Bloch Bloch periodic function. Hence define the space

$$\mathcal{H}^1_{\boldsymbol{k}}(\Omega) := \left\{ u \in \mathcal{H}^1(\Omega) \mid u \text{ is } \boldsymbol{k} - \text{Bloch periodic} \right\}, \tag{3.1}$$

with the same scalar product as for $\mathcal{H}^1$. In Definition 2.1 point wise evaluation of a function u is required. This is in general not possible for functions in $\mathcal{H}^1$. Hence we impose the periodicity of u in a weak sense on the trace of u. The Bloch Theorem states that all functions $\varphi \in \mathcal{H}^1_k(\Omega)$ are of the form

$$\varphi(\boldsymbol{r}) = e^{i\boldsymbol{k}\cdot\boldsymbol{r}} v(\boldsymbol{r})$$

for some v that is periodic on Γ. Therefore it makes sense to define

$$\mathcal{H}^1_p(\Omega) := \left\{ v \in \mathcal{H}^1(\Omega) \mid v \text{ is periodic on } \Gamma \right\}.$$

Now we have everything we need to state a candidate for a weak formulation of Problem 2.2.

Problem 3.1 (weak formulation)**.** Let $\boldsymbol{k} = \begin{pmatrix} k_x & k_y \end{pmatrix} \neq 0$ be an arbitrary but fixed wave vector in the BZ. Find $u \in \mathcal{H}^1_p(\Omega)$ and a real $\lambda > 0$ such that

$$a_{\boldsymbol{k}}(u, v) = \lambda b(u, v) \quad \forall v \in \mathcal{H}^1_p(\Omega) \tag{3.2}$$

with

$$\begin{aligned} a_{\boldsymbol{k}}(u, v) &= \int_\Omega \boldsymbol{\mu}^{-1} \left(\nabla u(\boldsymbol{r}) + i\boldsymbol{k}u(\boldsymbol{r})\right) \cdot \overline{\left(\nabla v(\boldsymbol{r}) + i\boldsymbol{k}v(\boldsymbol{r})\right)} \mathrm{d}\mathbf{r}, \\ b(u, v) &= \int_\Omega \varepsilon \; u(\boldsymbol{r})\overline{v(\boldsymbol{r})} \mathrm{d}\mathbf{r}, \\ \lambda &= \left(\frac{\omega}{c}\right)^2. \end{aligned} \tag{3.3}$$

Remark 3.2. *From the definition $\lambda = (\omega/c)^2$ for a real valued ω we immediately get that λ must be a non-negative real number. As we will see later, this is not a restriction if $\varepsilon > 0$ and (2.4) holds. In this case we deal with a compact self adjoint positive operator.*

We have to show that Problem 3.1 really is a well posed weak formulation of Problem 2.2. Similar problems are discussed in [2].

Theorem 3.3. *Let Ω be open and bounded and $\partial\Omega = \Gamma \in C^1$. Then the following two statements hold.*

(i) If (u, λ) is a solution for Problem 2.2 then it also solves Problem 3.1.

(ii) If (u, λ) is a solution for Problem 3.1 and additionally $u \in C^1(\overline{\Omega}) \cap C^2(\Omega) \cap \mathcal{H}^1_p(\Omega)$ then it also solves Problem 2.2.

Proof. Ad (i): Let $v \in \mathcal{H}^1_p(\Omega)$ be arbitrary. Multiplying (2.14) with $e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}v(r)$ on both sides and integrating by parts yields

$$\int_\Omega \boldsymbol{\mu}^{-1}\nabla\left(e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u(\boldsymbol{r})\right)\cdot\nabla\left(e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}\overline{v(\boldsymbol{r})}\right) - \int_\Gamma \boldsymbol{\mu}^{-1}\nabla\left(e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u(\boldsymbol{r})\right)e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}\overline{v(\boldsymbol{r})}\cdot\boldsymbol{n}$$
$$= \left(\frac{\omega}{c}\right)^2\int_\Omega \varepsilon\, u(\boldsymbol{r})\overline{v(\boldsymbol{r})}.$$

The boundary term vanishes because of the periodicity of the integrand. Hence u fulfills

$$\int_\Omega \boldsymbol{\mu}^{-1}\nabla\left(e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u(\boldsymbol{r})\right)\cdot\nabla\left(e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}\overline{v(\boldsymbol{r})}\right) = \left(\frac{\omega}{c}\right)^2\int_\Omega \varepsilon\, u(\boldsymbol{r})\overline{v(\boldsymbol{r})}.$$

Applying the gradient on the left hand side yields

$$\int_\Omega \boldsymbol{\mu}^{-1}\left(\nabla u(\boldsymbol{r}) + i\boldsymbol{k}u(\boldsymbol{r})\right)\cdot\left(\nabla\overline{v(\boldsymbol{r})} - i\boldsymbol{k}\overline{v(\boldsymbol{r})}\right) = \left(\frac{\omega}{c}\right)^2\int_\Omega \varepsilon\, u(\boldsymbol{r})\overline{v(\boldsymbol{r})}.$$

The test function v was arbitrary, hence we have proven (i).

Ad (ii): First we choose an arbitrary real valued test function

$$v \in C_0^\infty = \{v \in C^\infty \mid v|_\Gamma = 0\} \subset \mathcal{H}^1_p,$$

multiply the integrands in (3.3) by $1 = e^{i\boldsymbol{k}\cdot\boldsymbol{r}}e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}$ and apply partial integration. Due to our choice of test functions the boundary term vanishes and we get

$$0 = a_{\boldsymbol{k}}(u,v) - \lambda b(u,v)$$
$$= \int_\Omega \left(-\operatorname{div}\left(\boldsymbol{\mu}^{-1}\nabla\left(e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u(\boldsymbol{r})\right)\right) - \lambda\,\varepsilon\, e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u(\boldsymbol{r})\right)e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}\overline{v(\boldsymbol{r})}\mathrm{d}\mathbf{r}.$$

The fundamental lemma of the calculus of variations and the smoothness of u yield

$$0 = \left(-\operatorname{div}\left(\boldsymbol{\mu}^{-1}\nabla\left(e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u(\boldsymbol{r})\right)\right) - \lambda\,\varepsilon\, e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u(\boldsymbol{r})\right)e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}.$$

If the real or imaginary parts of $e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}$ vanish, so do the ones of the complex conjugate $e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$. Hence

$$0 = -\operatorname{div}\left(\boldsymbol{\mu}^{-1}\nabla\left(e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u(\boldsymbol{r})\right)\right) - \lambda\,\varepsilon\, e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u(\boldsymbol{r})$$

everywhere in Ω. The requirement $u \in \mathcal{H}^1_p(\Omega)$ together with the smoothness of u guarantees the periodicity on Γ. □

Remark 3.4. *Theorem 3.3 requires smooth boundaries. Remember that the boundary is the union of the edges $\Gamma = \Gamma^b \cup \Gamma^r \cup \Gamma^t \cup \Gamma^l$ as is illustrated in Figure2.1. This was also used in the proof of Theorem 3.3, so we need to justify why this is okay to do. Without loss of generality we can assume that the unit cell Ω is placed such that the origin is in the center. The length of one edge is assumed to be r_0. Now the boundary can be described the following way in polar coordinates*

$$\Gamma = r(\varphi)\begin{pmatrix}\cos(\varphi)\\ \sin(\varphi)\end{pmatrix} \tag{3.4}$$

where

$$r(\varphi) = \begin{cases} r_{l,r}(\varphi) \;\; \text{if } \varphi \in [0, \frac{\pi}{4}) \cup [\frac{3\pi}{4}, \frac{5\pi}{4}) \cup [\frac{7\pi}{4}, 2\pi) \\ r_{t,b}(\varphi) \;\; \text{if } \varphi \in [\frac{\pi}{4}, \frac{3\pi}{4}) \cup [\frac{5\pi}{4}, \frac{7\pi}{4}) \end{cases}$$

with

$$\begin{aligned} r_{l,r}(\varphi) &= \frac{r_0}{2}\sqrt{1+\sin(\varphi)^2}, \\ r_{t,b}(\varphi) &= \frac{r_0}{2}\sqrt{1+\cos(\varphi)^2}. \end{aligned}$$

This parametrization is not smooth at the corners $\varphi_c \in \{\pi/4, 3\pi/4, 5\pi/4, 7\pi/4\}$. *We choose an arbitrary small* $\epsilon > 0$. *Observe that* $r(\varphi_c + \epsilon) = r(\varphi_c - \epsilon) =: r_\epsilon$, $r'(\varphi_c - \epsilon) =: s_\epsilon$ *and* $r'(\varphi_c + \epsilon) =: -s_\epsilon$. *Our goal is to find a function* $\tilde{r}(\varphi)$ *such that*

$$\begin{aligned} \tilde{r}(\varphi_c - \epsilon) &= r_\epsilon \\ \tilde{r}(\varphi_c + \epsilon) &= r_\epsilon \\ \tilde{r}'(\varphi_c - \epsilon) &= s_\epsilon \\ \tilde{r}'(\varphi_c + \epsilon) &= -s_\epsilon. \end{aligned} \tag{3.5}$$

Making the ansatz

$$\tilde{r}(\varphi) = a\varphi^2 + b\varphi + c$$

and using (3.5) *yields*

$$a = -\frac{s_\epsilon}{2\epsilon}, \quad b = \frac{s\varphi_c}{\epsilon}, \quad c = -\frac{s_\epsilon \varphi_c^2}{2\epsilon} + \frac{s_\epsilon \epsilon}{2} + r_\epsilon.$$

Replacing $r(\varphi)$ *with* $\tilde{r}(\varphi)$ *in* (3.4) *for* $\varphi \in [\varphi_c - \epsilon, \varphi_c + \epsilon]$ *yields a* C^1 *approximation* Γ_ϵ *of* Γ. *It remains to adapt Definition 2.1 for* Γ_ϵ. *A function* $u \in \mathcal{H}^1$ *is called periodic on* Γ_ϵ *if we identify*

$$r(\varphi)\begin{pmatrix}\cos(\varphi)\\ \sin(\varphi)\end{pmatrix} \;\; \textit{with} \;\; r(\varphi)\begin{pmatrix}-\cos(\varphi)\\ \sin(\varphi)\end{pmatrix}$$

for $\varphi \in [0, \frac{\pi}{4}) \cup [\frac{3\pi}{4}, \frac{5\pi}{4}) \cup [\frac{7\pi}{4}, 2\pi)$ *and*

$$r(\varphi)\begin{pmatrix}\cos(\varphi)\\ \sin(\varphi)\end{pmatrix} \;\; \textit{with} \;\; r(\varphi)\begin{pmatrix}\cos(\varphi)\\ -\sin(\varphi)\end{pmatrix}$$

for $\varphi \in [\frac{\pi}{4}, \frac{3\pi}{4}) \cup [\frac{5\pi}{4}, \frac{7\pi}{4})$.
Let Ω_ϵ *be the open set enclosed by* Γ_ϵ. *Theorem 3.3 holds for every* $\epsilon > 0$ *and* $\bigcup_\epsilon \Omega_{\epsilon>0} = \Omega$. *Hence according to [5] the statement also holds for* Ω. *The last part of the proof was conducted only for the limes* Ω. *However it can be done in the same way for* Ω_ϵ.

It remains to justify that we can expect real eigenvalues $\lambda > 0$ for Problem 3.1. Our goal is to use the Spectral Theorem for self adjoint compact operators. Therefore we need an operator setting of our problem.

Definition 3.5. The *solution operator* $T: \mathcal{H}_p^1 \to \mathcal{H}_p^1$ is defined by

$$a_{\boldsymbol{k}}(Tf, v) = b(f, v) \quad \forall f, v \in \mathcal{H}_p^1. \tag{3.6}$$

From the definition we see that for an eigenpair (u, λ) of Problem 3.1, $(u, 1/\lambda)$ is an eigenpair of the solution operator. Hence studying T makes sense. In the course *Numerics of partial differential equations: instationary problems (101.507)* we performed this analysis for a solution operator of that kind in a real Hilbert space. The notation and line of argumentation is very similar to what we did in that course but all statements can also be found in [5].

Lemma 3.6. *The sesquilinear form $a_{\boldsymbol{k}}$ defines a scalar product on $\mathcal{H}_p^1$, meaning it is*

(i) sesquilinear: $a_{\boldsymbol{k}}(u + \alpha v, w) = a_{\boldsymbol{k}}(u, w) + \alpha a_{\boldsymbol{k}}(v, w)$ and $a_{\boldsymbol{k}}(u, v + \alpha w) = a_{\boldsymbol{k}}(u, v) + \overline{\alpha} a_{\boldsymbol{k}}(u, w)$

(ii) hermitian: $a_{\boldsymbol{k}}(u, v) = \overline{a_{\boldsymbol{k}}(v, u)}$

(iii) positive definite: $a_{\boldsymbol{k}}(u, u) \geq 0$ and $a_{\boldsymbol{k}}(u, u) = 0$ iff $u = 0$.

Furthermore the induced norm $\sqrt{a_{\boldsymbol{k}}(u, u)}$ is equivalent to $\|u\|_{\mathcal{H}^1}$.

Proof. Ad (i), (ii): Follows directly from the definition and the linearity of the integral.
Ad (iii): Follows from the norm equivalence.
Ad norm equivalence: From the hermitian property we get $a_{\boldsymbol{k}}(u, u) \in \mathbb{R}$ by calculating

$$\begin{aligned} a_{\boldsymbol{k}}(u, u) &= \frac{1}{2}\left(a_{\boldsymbol{k}}(u, u) + a_{\boldsymbol{k}}(u, u)\right) \\ &= \frac{1}{2}\left(a_{\boldsymbol{k}}(u, u) + \overline{a_{\boldsymbol{k}}(u, u)}\right) \\ &= \operatorname{Re} a_{\boldsymbol{k}}(u, u). \end{aligned}$$

Using the Cauchy Schwarz inequality (CS) we show continuity by estimating

$$\begin{aligned} |a_{\boldsymbol{k}}(u, v)| &= \left| \int_\Omega \boldsymbol{\mu}^{-1} \left(\nabla u(\boldsymbol{r}) + i\boldsymbol{k}u(\boldsymbol{r})\right) \cdot \overline{\left(\nabla v(\boldsymbol{r}) + i\boldsymbol{k}v(\boldsymbol{r})\right)} \mathrm{d}\mathbf{r} \right| \\ &\overset{\boldsymbol{\mu}^{-1} \text{ is bounded}}{\lesssim} \left| \int_\Omega \left(\nabla u(\boldsymbol{r}) + i\boldsymbol{k}u(\boldsymbol{r})\right) \cdot \overline{\left(\nabla v(\boldsymbol{r}) + i\boldsymbol{k}v(\boldsymbol{r})\right)} \mathrm{d}\mathbf{r} \right| \\ &\overset{\text{CS}}{\leq} \|\nabla u + i\boldsymbol{k}u\|_{\mathcal{L}^2} \|\nabla v + i\boldsymbol{k}v\|_{\mathcal{L}^2} \\ &\overset{\text{triangle inequality}}{\lesssim} \left(\|u\|_{\mathcal{L}^2} + \|\nabla u\|_{\mathcal{L}^2}\right)\left(\|v\|_{\mathcal{L}^2} + \|\nabla v\|_{\mathcal{L}^2}\right) \\ &\approx \|u\|_{\mathcal{H}^1} \|v\|_{\mathcal{H}^1}. \end{aligned}$$

It remains to show ellipticity. It suffices to consider functions in $C^\infty(\Omega) \cap \mathcal{H}_p^1(\Omega)$ because $C^\infty(\Omega)$ is dense in $\mathcal{H}^1(\Omega)$. The domain Ω is a torus so $u \in C^\infty(\Omega) \cap \mathcal{H}_p^1(\Omega)$ can be written as

$$u = \sum_{(m_x,m_y)\in\mathbb{Z}^2} c_{(m_x,m_y)} e^{i2\pi/a(m_x x + m_y y)}.$$

For better readability we choose an arbitrary basis function

$$u = e^{i2\pi/a(m_x x + m_y y)} =: e^{i2\pi/a(\boldsymbol{m}\cdot\boldsymbol{r})} \tag{3.7}$$

and calculate

$$\boldsymbol{z} := \nabla u + i\boldsymbol{k}u = ie^{i2\pi/a(\boldsymbol{mr})} \begin{pmatrix} 2\pi/am_x + k_x \\ 2\pi/am_y + k_y \end{pmatrix}.$$

Remember that

$$\boldsymbol{\mu}^{-1} = \frac{1}{\mu^2 - \kappa^2} \begin{pmatrix} \mu & -i\kappa \\ i\kappa & \mu \end{pmatrix}.$$

We consider the following expression

$$\begin{aligned} \boldsymbol{\mu}^{-1}\boldsymbol{z}\cdot\overline{\boldsymbol{z}} &= \frac{1}{\mu^2-\kappa^2} \begin{pmatrix} \mu & -i\kappa \\ i\kappa & \mu \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \cdot \begin{pmatrix} \overline{z_1} \\ \overline{z_2} \end{pmatrix} \\ &= \frac{1}{\mu^2-\kappa^2} \begin{pmatrix} \mu z_1 - i\kappa z_2 \\ i\kappa z_1 + \mu z_2 \end{pmatrix} \cdot \begin{pmatrix} \overline{z_1} \\ \overline{z_2} \end{pmatrix} \\ &= \frac{1}{\mu^2-\kappa^2} \left(\mu z_1\overline{z_1} - i\kappa z_2\overline{z_1} + i\kappa z_1\overline{z_2} + \mu z_2\overline{z_2}\right) \\ &= \frac{\mu}{\mu^2-\kappa^2}\boldsymbol{z}\cdot\overline{\boldsymbol{z}}. \end{aligned} \tag{3.8}$$

The last equality follows immediately by plugging in our definition of $\boldsymbol{z}$. We can observe

that assumption (2.4) is necessary for a to be elliptic. We carry on by estimating

$$\begin{aligned}
a_{\boldsymbol{k}}(u,u) &= \int_\Omega \boldsymbol{\mu}^{-1}\left(\nabla u(\boldsymbol{r}) + i\boldsymbol{k}u(\boldsymbol{r})\right)\cdot\overline{\left(\nabla u(\boldsymbol{r}) + i\boldsymbol{k}u(\boldsymbol{r})\right)}\,\mathrm{d}\boldsymbol{r}\\
&\overset{(3.8)}{\approx} \langle \nabla u + i\boldsymbol{k}u, \nabla u + i\boldsymbol{k}u\rangle_{\mathcal{L}^2}\\
&= \langle\nabla u,\nabla u\rangle_{\mathcal{L}^2} + i\langle \boldsymbol{k}u,\nabla u\rangle_{\mathcal{L}^2} - i\langle\nabla u,\boldsymbol{k}u\rangle_{\mathcal{L}^2} + \langle\boldsymbol{k}u,\boldsymbol{k}u\rangle_{\mathcal{L}^2}\\
&= \langle\nabla u,\nabla u\rangle_{\mathcal{L}^2} - 2\,\mathrm{Im}\,\langle\boldsymbol{k}u,\nabla u\rangle_{\mathcal{L}^2} + \langle\boldsymbol{k}u,\boldsymbol{k}u\rangle_{\mathcal{L}^2}\\
&\overset{(3.7)}{=} \frac{4\pi^2}{a^2}|\boldsymbol{m}|^2\|u\|^2_{\mathcal{L}^2} + |\boldsymbol{k}|^2\|u\|^2_{\mathcal{L}^2} + \frac{4\pi}{a}\boldsymbol{k}\cdot\boldsymbol{m}\|u\|^2_{\mathcal{L}^2}\\
&= \|u\|^2_{\mathcal{L}^2}\left(\frac{2\pi}{a}m_x\left(\frac{2\pi}{a}m_x + 2k_x\right) + \frac{2\pi}{a}m_y\left(\frac{2\pi}{a}m_y + 2k_y\right) + k_x^2 + k_y^2\right)\\
&\geq \|u\|^2_{\mathcal{L}^2}\left(\frac{2\pi}{a}|m_x|\left(\frac{2\pi}{a}|m_x| - 2|k_x||m_x|\right) + \frac{2\pi}{a}|m_y|\left(\frac{2\pi}{a}|m_y| - 2|k_y||m_y|\right) + k_x^2 + k_y^2\right)\\
&\overset{(*)}{\geq} \|u\|^2_{\mathcal{L}^2}\left(\frac{2\pi}{a}|m_x|^2\left(\frac{2\pi}{a} - \frac{2\pi}{a} + \varepsilon\right) + \frac{2\pi}{a}|m_y|^2\left(\frac{2\pi}{a} - \frac{2\pi}{a} + \varepsilon\right) + k_x^2 + k_y^2\right)\\
&= \|u\|^2_{\mathcal{L}^2}\left(\frac{4\pi^2}{a^2}|m_x|^2\left(\frac{a}{2\pi}\varepsilon\right) + \frac{4\pi^2}{a^2}|m_y|^2\left(\frac{a}{2\pi}\varepsilon\right) + k_x^2 + k_y^2\right)\\
&= \frac{a}{2\pi}\varepsilon\,\|\nabla u\|^2_{\mathcal{L}^2} + |\boldsymbol{k}|^2\|u\|^2_{\mathcal{L}^2}\\
&\simeq \|u\|^2_{\mathcal{H}^1}.
\end{aligned}$$

Inequality $(*)$ holds because $\boldsymbol{k}$ is in the first BZ (2.11). Hence there is an $\varepsilon > 0$ such that $|k_x|, |k_y| \leq \pi/a - \varepsilon$. The constant in above estimate gets worse if $\boldsymbol{k}$ is near the boundary of the Brillouin zone. Finally we can conclude

$$\|u\|^2_{\mathcal{H}^1} \lesssim a_{\boldsymbol{k}}(u,u) \lesssim \|u\|^2_{\mathcal{H}^1}.$$

□

Lemma 3.7. *The solution operator T as introduced in Definition 3.5 is well defined and fulfills*

(i) T is compact

(ii) T is self adjoint

(iii) T is positive on $\mathcal{H}^1_p$ meaning $a(Tu,u) > 0\ \forall u \in \mathcal{H}^1_p\backslash\{0\}$.

Proof. Ad well definedness: From Lemma 3.6 we know that $a_{\boldsymbol{k}}$ is a scalar product on $\mathcal{H}^1_p$. Furthermore b is a bounded sesquilinear form on $\mathcal{H}^1_p$. Hence Lax-Milgram yields that there is a unique continuous linear operator $T : \mathcal{H}^1_p \to \mathcal{H}^1_p$ such that

$$b(f,v) = a_{\boldsymbol{k}}(Tf, v) \quad \forall f, v \in \mathcal{H}^1_p.$$

Ad (i): From Rellich Compactness Theorem we know that $\mathcal{H}^1$ and therefore $\mathcal{H}^1_p$ is compactly embedded in $\mathcal{L}^2$. Hence

$$T : \mathcal{H}^1_p \hookrightarrow\hookrightarrow \mathcal{L}^2 \to \mathcal{H}^1_p$$

is compact as a composition of a continuous and a compact operator.
Ad(ii): For arbitrary $u, v \in \mathcal{H}_p^1$ there holds

$$a_{\boldsymbol{k}}(Tu, v) = b(u, v) = \overline{b(v, u)} = \overline{a_{\boldsymbol{k}}(Tv, u)} = a_{\boldsymbol{k}}(u, Tv).$$

Ad (iii): The statement follows directly from

$$a_{\boldsymbol{k}}(Tu, u) = b(u, u) = \int_\Omega \varepsilon\, u\overline{u} \simeq \|u\|^2_{\mathcal{L}^2}.$$

□

Theorem 3.8. *All eigenvalues of Problem 3.1 are real valued and non-negative.*

Proof. According to Lemma 3.7 the solution operator T is a compact, self adjoint, positive operator. Hence the Spectral Theorem yields the statement. □

Theorem 3.9. *Let* $\boldsymbol{k} = \begin{pmatrix} k_x & k_y \end{pmatrix}^T \neq 0$ *be an arbitrary but fixed wave vector in the BZ. There exists a series* $(u_n, \lambda_n)_{n\in\mathbb{N}} \subset \mathcal{H}_p^1 \times \mathbb{R}$ *that fulfills:*

(i) The pairs (u_n, λ_n) *solve (3.2).*

(ii) The eigenvalues are real and form an unbounded sequence $0 < \lambda_1 \leq \lambda_2 \leq \ldots$.

(iii) The eigenvectors $(u_n)_{n\in\mathbb{N}}$ *are an orthonormal basis (ONB) of* $\mathcal{H}_p^1$.

Proof. According to Lemma 3.7 the solution operator T is compact, self adjoint and positive. Hence the Spectral Theorem yields that the eigenvalues $\tilde{\lambda}_n$ of T are real valued, positive and have 0 as their only accumulation point. Due to the fact that $\lambda_n = 1/\tilde{\lambda}_n$ statement (ii) follows immediately. Let u_n be the corresponding eigenvectors of T, hence (u_n, λ_n) solves (3.2). Statement (iii) is another immediate consequence of the Spectral Theorem. □

3.1.2 Galerkin Discretization

Up until now all problems are posed in an infinite dimensional space. For numerical computations this setting is not suitable. Therefore we want to find a discrete problem for which the solutions approximate the eigenpairs of Problem 3.1 reasonably well.

Problem 3.10 (Galerkin discretization)**.** Let $\mathcal{V}_h \subset \mathcal{H}_p^1$ be a finite dimensional Hilbert subspace. Then the *Galerkin discretization* of Problem 3.1 reads as follows. Find $\lambda_h > 0$ and $u_h \in \mathcal{V}_h \backslash \{0\}$ such that

$$a_{\boldsymbol{k}}(u_h, v_h) = \lambda_h b(u_h, v_h) \quad \forall v_h \in \mathcal{V}_h. \tag{3.9}$$

The associated *discrete solution operator* $T_h : \mathcal{V}_h \to \mathcal{V}$ is defined as

$$a_{\boldsymbol{k}}(T_h f_h, v_h) = b(f, v) \quad \forall f_h, v_h \in \mathcal{V}_h. \tag{3.10}$$

Next we want to talk about solutions of the weak Problem 3.1 and their relationship to solutions of the discrete Problem 3.10. This section is based on [5]. The setting there is described in Problem 3.11.

Problem 3.11. Let V_1, V_2 be complex Hilbert spaces and let $a, b : V_1 \times V_2 \to \mathbb{C}$ be sesquilinear forms. They are assumed to have the following properties.

(i) a is continuous, meaning

$$|a(v_1, v_2)| \lesssim \|v_1\|_{V_1}\|v_2\|_{V_2} \quad \forall v_1 \in V_1, \forall v_2 \in V_2.$$

(ii) b is continuous with respect to a compact norm, meaning there exists a norm $\|\cdot\|_{H_1}$ such that any bounded sequence in V_1 has a Cauchy subsequence with respect to $\|\cdot\|_{H_1}$ and

$$|b(v_1, v_2)| \lesssim \|v_1\|_{H_1}\|v_2\|_{V_2} \quad \forall v_1 \in V_1, v_2 \in V_2.$$

(iii) a fulfills the *inf-sup condition*, meaning there exists a $\gamma > 0$ such that

$$\inf_{v_1 \in V_1} \sup_{v_2 \in V_2} \frac{|a(v_1, v_2)|}{\|v_1\|_{V_1}\|v_2\|_{V_2}} \geq \gamma,$$
$$\sup_{v_1 \in V_1} |a(v_1, v_2)| > 0 \quad \forall v_2 \in V_2 \backslash \{0\}.$$

Find $\lambda \in \mathbb{C}$ and $u \in V_1$ with $u \neq 0$ such that

$$a(u, v) = \lambda b(u, v) \quad \forall v \in V_2.$$

Lemma 3.12. *Problem 3.1 fulfills the assumptions of Problem 3.11 with* $V_1 = V_2 = \mathcal{H}^1_p$, $\|u\|_{H_1} = \|u\|_\varepsilon := \sqrt{\int_\Omega \varepsilon\, u\overline{u} d\boldsymbol{r}}$, $a = a_{\boldsymbol{k}}$ *and* $b = b$.

Proof. Ad (i): We have already shown continuity in the proof of Lemma 3.6.
Ad (ii): From Rellich's compactness theorem we already know that $\mathcal{H}^1$ is compactly embedded in $\mathcal{L}^2$. The boundedness of b follows with Cauchy Schwarz

$$\begin{aligned} |b(u, v)| &= \left| \int_\Omega \varepsilon\; u(\boldsymbol{r})\overline{v(\boldsymbol{r})} \mathrm{d}\mathbf{r} \right| \\ &\leq \|\sqrt{\varepsilon} u\|_{\mathcal{L}^2}\|v\|_{\mathcal{L}^2} \\ &\leq \|u\|_\varepsilon \|v\|_{\mathcal{H}^1}. \end{aligned}$$

Ad (iii): In Lemma 3.6 we have already show that $|a(u, u)| = a(u, u)$ is bounded from below by $\gamma\|u\|^2_{\mathcal{H}^1}$ for some positive γ. Using that result yields that the inf-sup condition

$$\begin{aligned} \inf_{u \in \mathcal{H}^1_p} \sup_{v \in \mathcal{H}^1_p} \frac{|a_{\boldsymbol{k}}(u, v)|}{\|u\|_{\mathcal{H}^1}\|v\|_{\mathcal{H}^1}} &\geq \inf_{u \in \mathcal{H}^1_p} \frac{|a_{\boldsymbol{k}}(u, u)|}{\|u\|^2_{\mathcal{H}^1}} \\ &\geq \inf_{u \in \mathcal{H}^1_p} \frac{\gamma\|u\|^2_{\mathcal{H}^1}}{\|u\|^2_{\mathcal{H}^1}} \\ &= \gamma > 0 \end{aligned}$$

is fulfilled. □

Theorem 3.13. *Assume that we choose a sequence of Hilbert subspaces $\mathcal{V}_h \subset \mathcal{H}_p^1$ such that the discrete solution operator T_h defined in (3.10) converges to the solution operator T defined in (3.6) as $h \to 0$. Then for a converging series of discrete eigenpairs (u_h, λ_h) of (3.10) exist eigenpairs (u, λ) of (3.1) such that*

$$|\lambda - \lambda_h| \overset{h\to 0}{\longrightarrow} 0 \quad \textit{and} \quad \|u - u_h\|_{\mathcal{H}_p^1} \overset{h\to 0}{\longrightarrow} 0.$$

Proof. Problem 3.1 fulfills the setting of Problem 3.11. Hence the convergence follows from Babuška-Osborn theory [2] which is discussed in Chapter 9 of [5]. □

Assume that the Hilbert subspace $\mathcal{V}_h$ is spanned by basis functions $\{\phi_1, ..., \phi_{N_h}\}$. An eigenvector u_h that fulfills (3.9) can be written as

$$u_h = \sum_{j=1}^{N_h} u_j \phi_j.$$

Problem 3.10 can then be written as find $\boldsymbol{u} \in \mathbb{C}$ and $\lambda > 0$ such that

$$\boldsymbol{A_k u} = \lambda \boldsymbol{B u}, \quad u_j \in \mathbb{C} \tag{3.11}$$

with

$$\boldsymbol{A_k}_{j,l} = a_{\boldsymbol{k}}(\phi_j, \phi_l) \quad \text{and} \quad \boldsymbol{B}_{j,l} = b(\phi_j, \phi_l). \tag{3.12}$$

The matrix $\boldsymbol{A_k}$ can be split into components independent of $\boldsymbol{k} = \begin{pmatrix} k_x & k_y \end{pmatrix}^T$. Expanding the left-hand side of (3.2) yields

$$\left((ik_x)^2 + (ik_y)^2\right) a_{k_x^2,k_y^2}(u,v) + ik_x a_{k_x}(u,v) + ik_y a_{k_y}(u,v) + a_1(u,v) = \lambda b(u,v)$$

with

$$\begin{aligned} a_{k_x^2,k_y^2}(u,v) &= \int_\Omega -\mu_{1,1}^{-1} u\overline{v}, \\ a_{k_x}(u,v) &= \int_\Omega u\boldsymbol{\mu}_{:,1}^{-1} \cdot \overline{\nabla v} - \nabla u \cdot \overline{v}\boldsymbol{\mu}_{1,:}^{-1}, \\ a_{k_y}(u,v) &= \int_\Omega u\boldsymbol{\mu}_{:,2}^{-1} \cdot \overline{\nabla v} - \nabla u \cdot \overline{v}\boldsymbol{\mu}_{2,:}^{-1}, \\ a_1(u,v) &= \int_\Omega \mu^{-1} \nabla u \cdot \overline{\nabla v}. \end{aligned}$$

Consequently the matrix $\boldsymbol{A_k}$ can be written as

$$\boldsymbol{A_k} = \left((ik_x)^2 + (ik_y)^2\right) \boldsymbol{A}^{k_x^2,k_y^2} + ik_x \boldsymbol{A}^{k_x} + ik_y \boldsymbol{A}^{k_y} + \boldsymbol{A}^1 \tag{3.13}$$

with

$$\boldsymbol{A}^\alpha_{j,l} = a_\alpha(\phi_j, \phi_l). \tag{3.14}$$

Remark 3.14. *Lemma 3.7 holds true for the discrete solution operator T_h defined in (3.10). Hence the same statement as in Theorem 3.9 can be made for the discrete Problem 3.10, with the only difference that $u \in \mathcal{V}_h$. Assume that $\mathcal{V}_h$ has dimension N_h. Due to the fact that all norms are equivalent in a finite dimensional Hilbert space, there are eigenvectors of (3.11) that constitute an ONB of $\mathbb{C}^{N_h}$.*

Before we go on, we want to introduce the notion of *frequency bands*. Let $\boldsymbol{k}$ be a wave vector in the first BZ and $0 < \lambda_{\min} < \lambda_{\max}$. Assume that there are eigenvalues $\lambda_1, \dots, \lambda_m \in (\lambda_{\min}, \lambda_{\max})$ of (3.11) with simple multiplicity for some $m \in \mathbb{N}$ and corresponding eigenvectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m$. These are members of m different bands. According to remark 3.14 and Theorem 3.9 the eigenvectors are orthogonal. The matrix $\boldsymbol{A_k}$ in (3.11) depends continuously on the wave vector $\boldsymbol{k}$. Hence solving (3.11) for a $\tilde{\boldsymbol{k}}$ with $\|\tilde{\boldsymbol{k}} - \boldsymbol{k}\|$ small yields eigenvalues $\tilde{\lambda_1}, \dots, \tilde{\lambda_m} \in (\lambda_{\min}, \lambda_{\max})$ and corresponding eigenvectors $\tilde{\boldsymbol{u}_1}, \dots, \tilde{\boldsymbol{u}_m}$. Assume that all eigenvectors are normalised, then we have that

$$\langle \boldsymbol{u}_j, \tilde{\boldsymbol{u}_j} \rangle \approx 1 \quad \forall j \in \{1, \dots, m\}.$$

We define λ_j and $\tilde{\lambda_j}$ as belonging to the same band, if there are $j, l \in \{1, \dots, m\}$ with $j \neq l$ such that

$$\langle \tilde{\boldsymbol{u}_j}, \tilde{\boldsymbol{u}_l} \rangle \approx 1$$

we say that bands j and l *intersect* each other. If there is no intersection of bands for all eligible wave vectors $\boldsymbol{k}$ we say that the bands are *separate*.

In the following chapter we will need the definitions of three kinds of matrix eigenvalue problems.

Definition 3.15. For $\boldsymbol{H} \in \mathbb{C}^{n \times n}$ a solution $(\lambda, \boldsymbol{u})$ of the *linear eigenvalue problem* fulfills

$$\boldsymbol{H}\boldsymbol{u} = \lambda \boldsymbol{u}. \tag{3.15}$$

For $\boldsymbol{C}, \boldsymbol{G} \in \mathbb{C}^{n \times n}$ a solution $(\lambda, \boldsymbol{u})$ of the *general eigenvalue problem (GEP)* fulfills

$$\boldsymbol{C}\boldsymbol{u} = \lambda \boldsymbol{G}\boldsymbol{u}. \tag{3.16}$$

For $\boldsymbol{M}, \boldsymbol{D}, \boldsymbol{K} \in \mathbb{C}^{n \times n}$ a solution $(\lambda, \boldsymbol{u})$ of the *quadratic eigenvalue problem (QEP)* fulfills

$$\left(\lambda^2 \boldsymbol{M} + \lambda \boldsymbol{D} + \boldsymbol{K}\right) \boldsymbol{u} = \boldsymbol{0}. \tag{3.17}$$

3.2 Interpretation as General Eigenvalue Problem

For a fixed $\boldsymbol{k}$ the Galerkin discretization of Problem 3.1 results in a GEP.

Problem 3.16. Let $\boldsymbol{k} = \begin{pmatrix} k_x & k_y \end{pmatrix} \neq \boldsymbol{0}$ be an arbitrary but fixed wave vector in the BZ. Find $u \in \mathbb{C}^{N_h}$ and a real $\lambda > 0$ that solve the GEP (3.16) with

$$\begin{aligned} \boldsymbol{C} &= -(k_x^2 + k_y^2)\boldsymbol{A}^{k_x^2, k_y^2} + ik_x \boldsymbol{A}^{k_x^2} + ik_y \boldsymbol{A}^{k_y^2} + \boldsymbol{A}^1, \\ \boldsymbol{G} &= \boldsymbol{B}, \\ \lambda &= \left(\frac{\omega}{c}\right)^2, \end{aligned}$$

where the matrices are defined in (3.12) and (3.14).

3.2.1 LOBPCG

Problem 3.16 can be solved by using a locally optimal block preconditioned conjugate gradient (LOBPCG) solver already available in Netgen/NGSolve. The LOBPCG method in use was introduced in [10]. An earlier version is discussed in [9]. A more extensive description can be found in [11]. The matrices $\boldsymbol{C}$ and $\boldsymbol{G}$ must be hermitian and positive definite. In our case that is true as we have shown in the previous section. We will now describe the idea behind LOBPCG in a nutshell. More detailed information can be found in the papers referenced above.

Definition 3.17. For matrices $\boldsymbol{C}, \boldsymbol{G} \in \mathbb{C}^{N_h \times N_h}$ defined in Problem 3.16 and $\boldsymbol{u} \in \mathbb{C}^{N_h}$, the *Rayleigh quotient* ρ is defined as

$$\rho(\boldsymbol{u}) = \frac{\langle \boldsymbol{C}\boldsymbol{u}, \boldsymbol{u} \rangle}{\langle \boldsymbol{G}\boldsymbol{u}, \boldsymbol{u} \rangle}. \tag{3.18}$$

Theorem 3.18. *Let ρ be the Rayleigh quotient from Definition 3.17. Let $\lambda_1 \leq \lambda_2 \leq \dots$ be the eigenvalues of Problem 3.16 with corresponding eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots$, then $\lambda_j = \rho(\boldsymbol{u}_j)$. Define*

$$E_m := \operatorname{span}\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_m\}$$

and

$$E_m^{\perp} := \{\boldsymbol{v} \in \mathbb{C}^{N_h} \mid \langle \boldsymbol{v}, \boldsymbol{w} \rangle = 0 \quad \forall \boldsymbol{w} \in E_m\}.$$

It holds that

$$\lambda_1 = \min_{\boldsymbol{u} \in \mathbb{C}^{N_h}} \rho(\boldsymbol{u}), \quad \boldsymbol{u} = \arg\min_{\boldsymbol{u} \in \mathbb{C}^{N_h}} \rho(\boldsymbol{u})$$

and for $m > 1$ it holds that

$$\lambda_m = \min_{\boldsymbol{u} \in V_{m-1}^{\perp}} \rho(\boldsymbol{u}), \quad \boldsymbol{u} = \arg\min_{\boldsymbol{u} \in V_{m-1}^{\perp}} \rho(\boldsymbol{u})$$

Proof. This is proven in [5]. The idea is to use Theorem 3.9 and the fact that $\langle \boldsymbol{C}\cdot, \cdot \rangle \simeq \langle \cdot, \cdot \rangle$. We write $\boldsymbol{v} \in \mathbb{C}^{N_h}$ as

$$\boldsymbol{v} = \sum_{j=1}^{N_h} \langle \boldsymbol{C}\boldsymbol{v}, \boldsymbol{u}_j \rangle \boldsymbol{u}_j.$$

Due to Parceval's Theorem the Rayleigh quotient takes the form

$$\rho(\boldsymbol{v}) = \frac{\sum_{j=1}^{N_h} \lambda_j |\langle \boldsymbol{C}\boldsymbol{v}, \boldsymbol{u}_j \rangle|^2}{\sum_{j=1}^{N_h} |\langle \boldsymbol{C}\boldsymbol{v}, \boldsymbol{u}_j \rangle|^2}.$$

□

In a first step we are only interested in the smallest eigenvalue λ_1. From Theorem 3.18 we know that

$$\lambda_1 = \min_{\boldsymbol{u}\in\mathbb{C}^{N_h}} \rho(\boldsymbol{u}).$$

Hence minimization of the Rayleigh quotient seams to be a good route to determine the first eigenvalue and an associated eigenvector. As already suggested by the name of LOBPCG the method is a form of a conjugate gradient algorithm. The idea of this kind of iterative procedures is the following. Let $\boldsymbol{u}^{(j)}$ be an approximation to $\hat{\boldsymbol{u}} := \arg\min_{\boldsymbol{u}\in\mathbb{C}^n} \rho(\boldsymbol{u})$ after j iteration steps. The basic idea is to get the next iteration is to set

$$\boldsymbol{u}^{(j+1)} = \boldsymbol{u}^{(j)} + \alpha \boldsymbol{d}^{(j+1)} \tag{3.19}$$

with the condition

$$\rho(\boldsymbol{u}^{(j+1)}) = \min_{\alpha} \rho(\boldsymbol{u}^{(j)} + \alpha \boldsymbol{d}^{(j+1)}).$$

In a steepest descent algorithm the search direction $\boldsymbol{d}^{(j+1)}$ would be simply chosen as $-\nabla\rho(\boldsymbol{u}^{(j)})$. For a *conjugate gradient* algorithm we additionally impose the orthogonality

$$\langle \boldsymbol{C}\boldsymbol{d}^{(j)}, \boldsymbol{d}^{(j+1)} \rangle = 0,$$

hence

$$\boldsymbol{d}^{(j+1)} = -\nabla\rho(\boldsymbol{u}^{(j)}) + \beta \boldsymbol{d}^{(j)}.$$

Next we calculate the gradient

$$\begin{aligned}
\nabla\rho(\boldsymbol{u}) &= \frac{\nabla\langle \boldsymbol{C}\boldsymbol{u}, \boldsymbol{u}\rangle \langle \boldsymbol{G}\boldsymbol{u}, \boldsymbol{u}\rangle - \langle \boldsymbol{C}\boldsymbol{u}, \boldsymbol{u}\rangle \nabla\langle \boldsymbol{G}\boldsymbol{u}, \boldsymbol{u}\rangle}{\langle \boldsymbol{G}\boldsymbol{u}, \boldsymbol{u}\rangle^2} \\
&= \frac{2\boldsymbol{C}\boldsymbol{u}\langle \boldsymbol{G}\boldsymbol{u}, \boldsymbol{u}\rangle - \langle \boldsymbol{C}\boldsymbol{u}, \boldsymbol{u}\rangle 2\boldsymbol{G}\boldsymbol{u}}{\langle \boldsymbol{G}\boldsymbol{u}, \boldsymbol{u}\rangle^2} \\
&= \frac{2}{\langle \boldsymbol{G}\boldsymbol{u}, \boldsymbol{u}\rangle} \left(\boldsymbol{C}\boldsymbol{u} - \rho(\boldsymbol{u})\boldsymbol{G}\boldsymbol{u}\right) \\
&\simeq \boldsymbol{r},
\end{aligned}$$

where $\boldsymbol{r}$ denotes the residual $\boldsymbol{C}\boldsymbol{u} - \rho(\boldsymbol{u})\boldsymbol{G}\boldsymbol{u}$. Plugging the resulting search direction into (3.19) yields

$$\boldsymbol{u}^{(j+1)} = \boldsymbol{u}^{(j)} + \alpha \left(-\boldsymbol{r}^{(j)} + \beta \boldsymbol{d}^{(j)}\right)$$

The *locally optimal* in LOBPCG means that the parameters α and β are optimized at once, because

$$\min_{\alpha,\beta} \rho\left(\boldsymbol{u}^{(j)} + \alpha(-\boldsymbol{r}^{(j)} + \beta \boldsymbol{d}^{(j)})\right) \leq \min_{\alpha} \rho\left(\boldsymbol{u}^{(j)} + \alpha(-\boldsymbol{r}^{(j)} + \beta \boldsymbol{d}^{(j)})\right).$$

In general the pencil $\boldsymbol{C} - \lambda\boldsymbol{G}$ can be ill conditioned. To tackle that problem a *preconditioner* $\boldsymbol{P}$ is introduced. For a linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ a good preconditioner is one that approximates $\boldsymbol{A}$. For an eigenvalue problem, where λ varies it is not obvious what the

optimal target should be. In [11] it is argued why a preconditioner $\boldsymbol{P}$ that approximates $\boldsymbol{C}^{-1}$ is a reasonable choice.

Defining $\boldsymbol{w}^{(j)} := \boldsymbol{P}\boldsymbol{r}^{(j)}$ the next value of the preconditioned iterative algorithm reads as

$$\boldsymbol{u}^{(j+1)} = \boldsymbol{u}^{(j)} + \alpha\left(-\boldsymbol{w}^{(j)} + \beta\boldsymbol{d}^{(j)}\right).$$

The minimum for $\rho(\boldsymbol{u}^{(j+1)})$ is exactly the minimal eigenvalue of the 3×3 eigenvalue problem

$$\left(\begin{pmatrix}\boldsymbol{u}^{(j)*}\\ -\boldsymbol{w}^{(j)*}\\ \boldsymbol{d}^{(j)*}\end{pmatrix}\boldsymbol{C}\begin{pmatrix}\boldsymbol{u}^{(j)} & -\boldsymbol{w}^{(j)} & \boldsymbol{d}^{(j)}\end{pmatrix}\right)\begin{pmatrix}\tau\\ \eta\\ \gamma\end{pmatrix} = \lambda\left(\begin{pmatrix}\boldsymbol{u}^{(j)*}\\ -\boldsymbol{w}^{(j)*}\\ \boldsymbol{d}^{(j)*}\end{pmatrix}\boldsymbol{G}\begin{pmatrix}\boldsymbol{u}^{(j)} & -\boldsymbol{w}^{(j)} & \boldsymbol{d}^{(j)}\end{pmatrix}\right)\begin{pmatrix}\tau\\ \eta\\ \gamma\end{pmatrix}$$

with $\alpha = \eta/\tau$ and $\beta = \gamma/\eta$.

The LOPCG method yields the smallest eigenvalue. The *block* version LOBPCG discussed in [11] will return the m smallest eigenvalues. This can be made plausible by the use of Theorem 3.18 and the fact that the vectors are orthogonalized.

3.3 Interpretation as Quadratic Eigenvalue Problem

The approach described in Section 3.2 does not work if the permeability $\boldsymbol{\mu}$ or the permittivity ε and therefore the matrices in Problem 3.16 depend on ω. To be able to solve this case we choose ω arbitrary but fixed. Additionally we impose a linear relationship between k_x and k_y by setting

$$\boldsymbol{k} = \begin{pmatrix}k_x\\ k_y\end{pmatrix} = \begin{pmatrix}p_x\\ p_y\end{pmatrix} + \tilde{\lambda}\begin{pmatrix}s_x\\ s_y\end{pmatrix} = \boldsymbol{p} + \tilde{\lambda}\boldsymbol{s}, \tag{3.20}$$

for $\boldsymbol{s}, \boldsymbol{p} \in \mathbb{R}^2$.

Problem 3.19. Let $\omega \in \mathbb{R}$ be arbitrary but fixed. Find $u \in \mathbb{C}^{N_h}$ and $\lambda \in \mathbb{C}$ that solve the QEP (3.17) with

$$\begin{aligned}\boldsymbol{M} &= -(p_x^2 + p_y^2)\boldsymbol{A}^{k_x^2,k_y^2} + ip_x\boldsymbol{A}^{k_x^2} + ip_y\boldsymbol{A}^{k_y^2} + \boldsymbol{A}^1 - \left(\frac{\omega}{c}\right)^2\boldsymbol{B}\\ \boldsymbol{D} &= i2(s_xp_x + s_yp_y)\boldsymbol{A}^{k_x^2,k_y^2} + s_x\boldsymbol{A}^{k_x^2} + s_y\boldsymbol{A}^{k_y^2}\\ \boldsymbol{K} &= (s_x^2 + s_y^2)\boldsymbol{A}^{k_x^2,k_y^2}.\end{aligned}$$

where the matrices are defined in (3.12) and (3.14).

The algorithms we will use approximate the eigenvalues with the biggest absolute values the best. Therefore it makes sense to solve for eigenvalue

$$\lambda := -i/\tilde{\lambda}. \tag{3.21}$$

in Problem 3.19 instead of $\tilde{\lambda}$ as defined in (3.20).

The methods described in this section were previously discussed and implemented in Netgen/NGSolve by the author as part of the course *AKNUM Seminar Computational Mathematics (101.845)*.

Let the eigenpair $(\lambda, \boldsymbol{u})$ be a solution of (3.17). Instead of solving a QEP of size $n \times n$ we consider an equivalent GEP (3.16) of size $2n \times 2n$ by defining

$$\boldsymbol{C} = \begin{pmatrix} -\boldsymbol{D} & -\boldsymbol{K} \\ \boldsymbol{I} & \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{G} = \begin{pmatrix} \boldsymbol{M} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}, \quad \tilde{\boldsymbol{u}} = \begin{pmatrix} \lambda \boldsymbol{u} \\ \boldsymbol{u} \end{pmatrix}.$$

For $\boldsymbol{G}$ invertible the problem can further be transformed into a linear eigenvalue problem (3.15) of size $2n \times 2n$ by defining

$$\boldsymbol{H} := \boldsymbol{G}^{-1}\boldsymbol{C} = \begin{pmatrix} -\boldsymbol{M}^{-1}\boldsymbol{D} & -\boldsymbol{M}^{-1}\boldsymbol{K} \\ \boldsymbol{I} & \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{I} & \boldsymbol{0} \end{pmatrix} \tag{3.22}$$

with $\boldsymbol{A} := -\boldsymbol{M}^{-1}\boldsymbol{D}$ and $\boldsymbol{B} := -\boldsymbol{M}^{-1}\boldsymbol{K}$. In order to do that we need $\boldsymbol{M}$ to be invertible, which is the case for $\boldsymbol{M}$ defined in Problem 3.19. However linearization comes with a few drawbacks. First, the problem size is doubled. Second, the structural properties of (3.17) are lost.

3.3.1 Rayleigh-Ritz Method

To avoid linearization of Problem (3.17) the second-order Krylov subspace is introduced, see [3].

Definition 3.20. For $\boldsymbol{H} \in \mathbb{C}^{n\times n}$ and $\boldsymbol{r} \in \mathbb{C}^n$ the *N-dimensional Krylov subspace* is defined as

$$\mathcal{K}_N(\boldsymbol{H}, \boldsymbol{r}) = \text{span}\{\boldsymbol{r}, \boldsymbol{H}\boldsymbol{r}, .., \boldsymbol{H}^{N-1}\boldsymbol{r}\}.$$

For $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$ and $\boldsymbol{r}_{-1}, \boldsymbol{r}_0 \in \mathbb{C}^n$ the *N-dimensional second-order Krylov subspace* is defined as

$$\mathcal{G}_N(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{r}_{-1}, \boldsymbol{r}_0) = \text{span}\{\boldsymbol{r}_{-1}, \boldsymbol{r}_0, .., \boldsymbol{r}_{N-1}\} \tag{3.23}$$

with

$$\boldsymbol{r_j} = \boldsymbol{A}\boldsymbol{r}_{j-1} + \boldsymbol{B}\boldsymbol{r}_{j-2} \quad \text{for } j \geq 1.$$

From now on we will use the starting values $\boldsymbol{r}_{-1} = \boldsymbol{0}$ and $\boldsymbol{r}_0 = \boldsymbol{u}$ for $\boldsymbol{u} \in \mathbb{C}^n$ and denote the corresponding second-order Krylov subspace as

$$\mathcal{G}_N(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{u}) := \mathcal{G}_N(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{0}, \boldsymbol{u}).$$

For the choice $\boldsymbol{A} = -\boldsymbol{M}^{-1}\boldsymbol{D}$, $\boldsymbol{B} = -\boldsymbol{M}^{-1}\boldsymbol{K}$ and the starting vector $\boldsymbol{x} = [\boldsymbol{u}^T, \boldsymbol{0}]^T$ the Krylov spaces $\mathcal{G}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{u})$ and $\mathcal{K}(\boldsymbol{H}, \boldsymbol{x})$ are connected. We see that

$$\begin{aligned}
\boldsymbol{H}\boldsymbol{x} &= \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{I} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}\boldsymbol{u} \\ \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_0 \end{pmatrix} \\
\boldsymbol{H}^2\boldsymbol{x} &= \begin{pmatrix} \boldsymbol{A}^2\boldsymbol{u} + \boldsymbol{B} \\ \boldsymbol{A}\boldsymbol{u} \end{pmatrix} = \begin{pmatrix} \boldsymbol{r}_2 \\ \boldsymbol{r}_1 \end{pmatrix} \\
&\vdots \\
\boldsymbol{H}^j\boldsymbol{x} &= \begin{pmatrix} \boldsymbol{r}_j \\ \boldsymbol{r}_{j-1} \end{pmatrix}.
\end{aligned}$$

Therefore we have

$$\mathcal{K}_N(\boldsymbol{H}, \boldsymbol{x}) = \operatorname{span}\left\{ \begin{pmatrix} \boldsymbol{r}_0 \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_0 \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{r}_{N-1} \\ \boldsymbol{r}_{N-2} \end{pmatrix} \right\}.$$

This means, that we can build $\mathcal{K}_N(\boldsymbol{H}, \boldsymbol{x})$ entirely with vectors contained in $\mathcal{G}_N(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{u})$. So we can expect an equally good convergence of eigenvalues, if we project (3.17) on $\mathcal{G}_N(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{u})$, as if the problem is linearised first and then projected on $\mathcal{K}_N(\boldsymbol{H}, \boldsymbol{x})$.
To describe this connection in a more general way, let $\boldsymbol{V}_N$ be a matrix with basis vectors $\{\boldsymbol{v}_0, ..., \boldsymbol{v}_{N-1}\}$ of $\mathcal{K}_N(\boldsymbol{H}, [\boldsymbol{r}_0, \boldsymbol{r}_{-1}]^T)$. Also denote the matrix containing basis vectors $\{\boldsymbol{q}_0, ..., \boldsymbol{q}_{N-1}\}$ of $\mathcal{G}_N(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{r}_{-1}, \boldsymbol{r}_0)$ as $\boldsymbol{Q}_N$. Then we have

$$\operatorname{span}\{\boldsymbol{V}_N[0 : n, :]\} = \operatorname{span}\{\boldsymbol{q}_0, \ldots, \boldsymbol{q}_{N-1}\}, \tag{3.24}$$
$$\operatorname{span}\{\boldsymbol{V}_N[n : 2n, :]\} = \operatorname{span}\{\boldsymbol{q}_{-1}, \ldots, \boldsymbol{q}_{N-2}\} \tag{3.25}$$

holds. Thus $\boldsymbol{Q}_N$ can be written as

$$\boldsymbol{Q}_N = \operatorname{span}\{\boldsymbol{V}_N[: n, :], \boldsymbol{V}_N[n : 2n, :]\}. \tag{3.26}$$

Equation (3.26) indicates that a basis of $\boldsymbol{Q}_N$ could be constructed using $\boldsymbol{V}_N$. However, this would be computationally expensive, because it would require working with a matrix of size $2n \times N$ to get a matrix of size $n \times N$.

Before we discuss the construction of a good basis, we need to define two special cases.

Definition 3.21. Let $\mathcal{G}_j(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{r}_{-1}, \boldsymbol{r}_0)$ be the second-order Krylov subspace spanned by $\{\boldsymbol{r}_{-1}, ..., \boldsymbol{r}_{j-1}\}$. The case

$$\mathcal{G}_j(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{r_{-1}}, \boldsymbol{r_0}) = \mathcal{G}_{j+1}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{r_{-1}}, \boldsymbol{r_0})$$

is called *deflation*. If

$$\mathcal{G}_j(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{r}_{-1}, \boldsymbol{r}_0) = \mathcal{G}_i(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{r}_{-1}, \boldsymbol{r}_0) \quad \forall i \geq j$$

a *breakdown* occurs.

The definition of a breakdown can be applied to the linear Krylov subspace as well.

Now a modified Rayleigh-Ritz procedure, described in Algorithm 1, can be introduced. A huge advantage of that method is, that the structure of the original eigenvalue problem is preserved.

Algorithm 1 Rayleigh-Ritz for QEP

Input: $\boldsymbol{M}, \boldsymbol{D}, \boldsymbol{K} \in \mathbb{C}^{n\times n}$, $N \in \mathbb{N}$ with $N < n$
Output: $\boldsymbol{u}^j \in \mathbb{C}^n$, $\lambda^j \in \mathbb{C}$ for $j \in \{1, ..., N\}$

1: define $\boldsymbol{A} = -\boldsymbol{M}^{-1}\boldsymbol{D}$ and $\boldsymbol{B} = -\boldsymbol{M}^{-1}\boldsymbol{K}$
2: compute an orthonormal basis $\boldsymbol{Q}_N$ of N-dimensional subspace $\mathcal{G}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{u})$
3: compute N eigenpairs $\{(\boldsymbol{g}^1, \lambda^1), ..., (\boldsymbol{g}^N, \lambda^N)\}$ of the QEP $(\lambda^2\boldsymbol{Q}_N^*\boldsymbol{M}\boldsymbol{Q}_N + \lambda\boldsymbol{Q}_N^*\boldsymbol{D}\boldsymbol{Q}_N + \boldsymbol{Q}_N^T\boldsymbol{K}\boldsymbol{Q})\boldsymbol{g} = \boldsymbol{0}$
4: get approximate eigenpairs $(\boldsymbol{u}^j, \lambda^j) = (\boldsymbol{Q}_N\boldsymbol{g}^j, \lambda^j)$

Last but not least, we have to choose some kind of error estimate to evaluate the quality of the eigenpairs we have calculated. As in [3] we use the relative residual norms to get a backwards error estimate

$$\varepsilon = \frac{\|(\lambda^2\boldsymbol{M} + \lambda\boldsymbol{D} + \boldsymbol{K})\boldsymbol{z}\|}{|\lambda|^2\|\boldsymbol{M}\| + |\lambda|\|\boldsymbol{D}\| + \|\boldsymbol{K}\|}. \tag{3.27}$$

The length of vectors is measured by the Euklidean norm. Hence, the naturally induced norm for a matrix $\boldsymbol{M} \in \mathbb{C}^{n\times n}$ would be the spectral norm

$$\|\boldsymbol{M}\| := \sqrt{\max\{|\mu| \mid \mu \text{ is eigenvalue of } \boldsymbol{M}^*\boldsymbol{M}\}}.$$

However, the spectral norm is quite expensive to compute. Therefore, the Frobenius norm

$$\|\boldsymbol{M}\| := \sqrt{\sum_{i,j=0}^{n-1} |m_{i,j}|^2}$$

is used instead.

3.3.2 Arnoldi Based Procedures to Construct a Krylov Subspace Basis

The second-order Arnoldi (SOAR) and two-level orthogonal Arnoldi (TOAR) procedures introduced in [3] and [12] for constructing an ONB of $\mathcal{G}_N(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{u})$ are both based on the Arnoldi procedure. Hence, as a first step we are going to discuss the Arnoldi procedure which results in an ONB of $\mathcal{K}_N(\boldsymbol{H}, \boldsymbol{x})$. We use the notation $\boldsymbol{V}_N := (\boldsymbol{v}_0, \ldots, \boldsymbol{v}_{N-1}) \in \mathbb{C}^{2n\times N}$ where $\boldsymbol{v}_i$ are basis vectors of $\mathcal{K}_N(\boldsymbol{H}, \boldsymbol{x})$ and $\boldsymbol{Q}_N := (\boldsymbol{q}_0, \ldots, \boldsymbol{q}_{N-1}) \in \mathbb{C}^{n\times N}$ is a basis matrix of $\mathcal{G}_N(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{u})$.

The Arnoldi procedure can be expressed as

$$\boldsymbol{H}\boldsymbol{V}_m = \boldsymbol{V}_m\boldsymbol{T}[:m,:m] + \boldsymbol{v}_m\boldsymbol{e}_{m-1}^T t_{m,m-1} = \boldsymbol{V}_{m+1}\boldsymbol{T}[:m+1,:m] \tag{3.28}$$

where $\boldsymbol{T}$ is an upper Hessenberg matrix and $\boldsymbol{e}_j$ is the j-th Euklidean unit vector (starting at 0) [3]. Using that all $\boldsymbol{v}_j$ are pairwise orthonormal we can calculate

$$\boldsymbol{V}_m^*\boldsymbol{H}\boldsymbol{V}_m = \underbrace{\boldsymbol{V}_m^*\boldsymbol{V}_m}_{\boldsymbol{I}}\boldsymbol{T}[:m,:m] + \underbrace{\boldsymbol{V}_m^*\boldsymbol{v}_m}_{\boldsymbol{0}}\boldsymbol{e}_{m-1}^T t_{m,m-1} = \boldsymbol{T}[:m,:m].$$

Therefore, the eigenvalues of $\boldsymbol{H}$ can be approximated by the ones of $\boldsymbol{T}$, which is a Hessenberg matrix and allows more efficient eigenvalue computation using QR decomposition.

Algorithm 2 Arnoldi

Input: $\boldsymbol{H} \in \mathbb{C}^{n\times n}$, $\boldsymbol{x} \in \mathbb{C}$, $N \in \mathbb{N}$
Output: ONB $\boldsymbol{V}_N \in \mathbb{C}^{n\times N}$, Hessenberg matrix $\boldsymbol{T} \in \mathbb{C}^{N\times N}$

$t_{i,j} = 0$ for $0 \le i,j \le N-1$
$\boldsymbol{v_0} = \boldsymbol{x}/\|\boldsymbol{x}\|$
for $j = 0,\dots,N-2$ **do**
  $\boldsymbol{w} = \boldsymbol{H}\boldsymbol{v_j}$
  for $i = 0,..,j$ **do**
    $t_{ij} = \langle \boldsymbol{w}, \boldsymbol{v_i}\rangle$
    $\boldsymbol{w} = \boldsymbol{w} - t_{ij}\boldsymbol{v_i}$
  $t_{j+1,j} = \|\boldsymbol{w}\|$
  if $t_{j+1,j} == 0$ **then**
    stop ▷ *breakdown*
  $\boldsymbol{v_{j+1}} = \boldsymbol{w}/t_{j+1,j}$

The SOAR procedure, as introduced in [3], generates an orthonormal basis (ONB) of $\mathcal{G}_N(\boldsymbol{A},\boldsymbol{B},\boldsymbol{u})$.

To describe the algorithm, we first assume that neither a breakdown nor deflation occurs. These cases are discussed later. Similar to the Arnoldi procedure the SOAR Algortihm 3 produces an upper Hessenberg matrix $\boldsymbol{T} \in \mathbb{C}^{N\times N}$. Let $\boldsymbol{P}_m$ be the matrix consisting of the helping vectors $\{\boldsymbol{p}_0,\dots,\boldsymbol{p}_m\}$ as columns. The m-th Euklidean unit vector is denoted by $\boldsymbol{e}_m$ (starting at 0). The following equations

$$\boldsymbol{A}\boldsymbol{Q}_m + \boldsymbol{B}\boldsymbol{P}_m = \boldsymbol{Q}_m\boldsymbol{T}[:m,:m] + \boldsymbol{q}_m\boldsymbol{e}_{m-1}^T t_{m,m-1}, \tag{3.29}$$

$$\boldsymbol{Q}_m = \boldsymbol{P}_m\boldsymbol{T}[:m,:m] + \boldsymbol{p}_m\boldsymbol{e}_{m-1}^T t_{m,m-1} \tag{3.30}$$

hold. By definition $\boldsymbol{p}_0 = \boldsymbol{0}$ and therefore (3.30) can be written as

Algorithm 3 SOAR – simple version

Input: $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$, $\boldsymbol{u} \in \mathbb{C}^n$
Output: ONB $\boldsymbol{Q} \in \mathbb{C}^{n\times N}$, Hessenberg matrix $\boldsymbol{T} \in \mathbb{C}^{N\times N}$

$t_{i,j} = 0$ for $0 \leq i, j \leq N-1$
$\boldsymbol{q}_0 = \boldsymbol{u}/\|\boldsymbol{u}\|$
$\boldsymbol{p}_0 = \boldsymbol{0}$
for $j = 0, .., N-2$ **do**
    $\boldsymbol{r} = \boldsymbol{A}\boldsymbol{q}_j + \boldsymbol{B}\boldsymbol{p}_j$
    $\boldsymbol{s} = \boldsymbol{q}_j$
    for $i = 0, .., j$ **do**
        $t_{ij} = \langle \boldsymbol{r}, \boldsymbol{q}_i \rangle$
        $\boldsymbol{r} = \boldsymbol{r} - t_{ij}\boldsymbol{q}_i$
        $\boldsymbol{s} = \boldsymbol{s} - t_{ij}\boldsymbol{p}_i$
    $t_{j+1,j} = \|\boldsymbol{r}\|$
    if $t_{j+1,j} == 0$ **then**
        if $\boldsymbol{s} \in \text{span}\{\boldsymbol{p}_i \mid i : \boldsymbol{q}_i = \boldsymbol{0},\ 0 \leq i \leq j\}$ **then**
            stop ▷ breakdown
        else
            reset $t_{j+1,j} = 1$ ▷ deflation
            $\boldsymbol{q}_{j+1} = \boldsymbol{0}$
            $\boldsymbol{p}_{j+1} = \boldsymbol{s}$
    $\boldsymbol{q}_{j+1} = \boldsymbol{r}/t_{j+1,j}$
    $\boldsymbol{p}_{j+1} = \boldsymbol{s}/t_{j+1,j}$

$$\begin{aligned}\boldsymbol{Q}_m &= \boldsymbol{P}_m[:, 1:]\boldsymbol{T}[1:m, :m] + \boldsymbol{p}_m \boldsymbol{e}_{m-1}^T t_{m,m-1} \\ &= \boldsymbol{P}_{m+1}[:, 1:]\boldsymbol{T}[1:m+1, :m].\end{aligned}$$

Using that relation the helping vectors $\boldsymbol{p}_j$ in (3.29) can be eliminated entirely by

$$\boldsymbol{P}_m = \begin{pmatrix}\boldsymbol{0} & \boldsymbol{Q}_{m-1}\boldsymbol{T}[1:m, :m-1]^{-1}\end{pmatrix} = \boldsymbol{Q}_m \begin{pmatrix}\boldsymbol{0} & \boldsymbol{T}[1:m, :m-1]^{-1} \\ 0 & \boldsymbol{0}^T\end{pmatrix} =: \boldsymbol{Q}_m \boldsymbol{S}_m.$$

Plugging that into (3.29) leads to

$$\boldsymbol{A}\boldsymbol{Q}_m + \boldsymbol{B}\boldsymbol{Q}_m\boldsymbol{S}_m = \boldsymbol{Q}_m\boldsymbol{T}[:m, :m] + \boldsymbol{q}_m \boldsymbol{e}_{m-1}^T t_{m,m-1}. \tag{3.31}$$

Now we can extract $\boldsymbol{q}_m$ from (3.31) by calculating

$$\begin{pmatrix}\boldsymbol{0} & \dots & \boldsymbol{0} & \boldsymbol{q}_m\end{pmatrix} t_{m,m-1} = \boldsymbol{A}\boldsymbol{Q}_m + \boldsymbol{B}\boldsymbol{Q}_{m-1}\begin{pmatrix}\boldsymbol{0} & \boldsymbol{T}[1:m, :m-1]^{-1}\end{pmatrix} - \boldsymbol{Q}_m\boldsymbol{T}_m[:m, :m]$$

hence

$$\boldsymbol{q}_m = \frac{1}{t_{m,m-1}}\left(\boldsymbol{A}\boldsymbol{q}_{m-1} + \boldsymbol{B}\underbrace{\boldsymbol{Q}_{m-1}\boldsymbol{T}[1:m, :m-1]^{-1}[:, m-2]}_{=:\boldsymbol{f}} - \underbrace{\boldsymbol{Q}_m\boldsymbol{T}[:, m-1]}_{\sum_{i=0}^{m-1}\boldsymbol{q}_i t_{j,m-1}}\right).$$

This results in a more efficient version of SOAR described in Algorithm 4.

Note that $\boldsymbol{T}[1:m, :m-1]$ is an upper triangular matrix and can as such be inverted in $\mathcal{O}(m^2)$ using backward substitution. However, the matrix might be ill conditioned. Subsequently, the algorithm is not always numerically stable. The TOAR procedure described in [12] later resolves that issue.

According to [3] Algorithm 3 breaks down at a step j if and only if Algorithm 2 breaks down at the same step j. We know that this is the case if

$$\boldsymbol{v}_j \in \operatorname{span}\left\{\begin{pmatrix}\boldsymbol{q}_i \\ \boldsymbol{p}_i\end{pmatrix} \mid 0 \le i < j\right\}.$$

This requires, that

$$\boldsymbol{0} = \boldsymbol{r} - \sum_{i=0}^{j-1}\langle \boldsymbol{r}, \boldsymbol{v}_i\rangle \boldsymbol{v}_i$$

with $\boldsymbol{r}$ defined as in line 5 in Algorithm 3 or line 5 in Algorithm 4. This results in $\boldsymbol{q}_j = \boldsymbol{0}$. It remains to check if

$$\boldsymbol{v}_j[n:] \in \operatorname{span}\{\boldsymbol{p}_i \mid i \ : \ \boldsymbol{q}_i = \boldsymbol{0},\ 0 \le i < j\}.$$

Algorithm 4 SOAR – efficient version

Input: $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$, $\boldsymbol{u} \in \mathbb{C}^n$
Output: ONB $\boldsymbol{V} \in \mathbb{C}^{n\times N}$, Hessenberg matrix $\boldsymbol{T} \in \mathbb{C}^{N\times N}$

1: $t_{i,j} = 0$ for $0 \leq i, j \leq N-1$
2: $\boldsymbol{q}_0 = \boldsymbol{u}/\|\boldsymbol{u}\|$
3: $\boldsymbol{f} = \boldsymbol{0}$
4: **for** $j = 0, .., N-2$ **do**
5: $\quad \boldsymbol{r} = \boldsymbol{A}\boldsymbol{q}_j + \boldsymbol{B}\boldsymbol{f}$
6: $\quad$ **for** $i = 0, .., j$ **do**
7: $\quad\quad t_{ij} = \langle \boldsymbol{r}, \boldsymbol{q}_i \rangle$
8: $\quad\quad \boldsymbol{r} = \boldsymbol{r} - t_{ij}\boldsymbol{q}_i$
9: $\quad t_{j+1,j} = \|\boldsymbol{r}\|$
10: $\quad$ **if** $t_{j+1,j} == 0$ **then**
11: $\quad\quad$ reset $t_{j+1,j} = 1$
12: $\quad\quad \boldsymbol{q}_{j+1} = \boldsymbol{0}$
13: $\quad\quad \boldsymbol{f} = \boldsymbol{Q}_j \boldsymbol{T}[1 : j+1, : j]^{-1}[:, j-1]$
14: $\quad\quad$ save $\boldsymbol{f}$ and check deflation and breakdown (see lines 14-16 in Algorithm 3)
15: $\quad$ **else**
16: $\quad\quad \boldsymbol{q}_{j+1} = \boldsymbol{r}/t_{j+1,j}$
17: $\quad\quad \boldsymbol{f} = \boldsymbol{Q}_j \boldsymbol{T}[1 : j+1, : j]^{-1}[:, j-1]$

As mentioned before, the SOAR algorithm can become numerically unstable due to inverting a potentially ill conditioned upper triangular matrix in Equation (3.31). To cure that problem the two-level orthogonal Arnoldi procedure is introduced in [12]. It creates not only an orthonormal basis $\boldsymbol{Q}_m$ of $\mathcal{G}_m(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{u})$, but it also ensures the orthonormality of the column vectors of the basis matrix $\boldsymbol{V}_m$ of the corresponding linear Krylov space $\mathcal{K}_m(\boldsymbol{H}, \boldsymbol{x})$.

Equations (3.24) and (3.25) justify the representation

$$\boldsymbol{V}_m = \begin{pmatrix} \boldsymbol{V}_m[0:n,:] \\ \boldsymbol{V}_m[n:2n,:] \end{pmatrix} = \begin{pmatrix} \boldsymbol{Q}_m \boldsymbol{U}_{m,0} \\ \boldsymbol{Q}_m \boldsymbol{U}_{m,1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Q}_m & \\ & \boldsymbol{Q}_m \end{pmatrix} \begin{pmatrix} \boldsymbol{U}_{m,0} \\ \boldsymbol{U}_{m,1} \end{pmatrix} =: \boldsymbol{Q}_{[m]} \boldsymbol{U}_m \tag{3.32}$$

where $\boldsymbol{U}_{m,0}$ and $\boldsymbol{U}_{m,1}$ are upper triangle matrices. The goal is that $\boldsymbol{Q}_m$ as well as $\boldsymbol{V}_m$ have orthonormal columns. Therefore, $\boldsymbol{U}_m$ needs to be orthonormal as well. In contrast to [12] we assume $\boldsymbol{r}_{-1} = \boldsymbol{0}$ and $\boldsymbol{r}_0$ from the beginning. As before we think about deflation and breakdown later and first assume that none of the two occur.

For $m = 1$ we have

$$\boldsymbol{V}_1 = \frac{1}{\|\boldsymbol{u}\|} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|} & \boldsymbol{0} \\ \boldsymbol{0} & \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{Q}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Q}_1 \end{pmatrix} \begin{pmatrix} \boldsymbol{U}_{1,0} \\ \boldsymbol{U}_{1,1} \end{pmatrix} = \boldsymbol{Q}_{[1]} \boldsymbol{U}_1.$$

Assume, that $\boldsymbol{Q}_j$, $\boldsymbol{U}_j$ and thus $\boldsymbol{V}_j$ are computed for $j \leq m$. In a fist step we want to calculate $\boldsymbol{q}_m$.

Algorithm 5 TOAR

Input: $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$, $\boldsymbol{u} \in \mathbb{C}^n$
Output: ONB $\boldsymbol{V} \in \mathbb{C}^{n\times N}$

$t_{i,j} = 0$ for $0 \leq i, j \leq N-1$
$\boldsymbol{q}_0 = \boldsymbol{u}/\|\boldsymbol{u}\|$
$\boldsymbol{U}_{0,1} = 1$
$\boldsymbol{U}_{0,1} = 0$
$\eta_0 = 1$
for $j = 0, .., N-2$ **do**
  $\boldsymbol{r} = \boldsymbol{A}\left(\boldsymbol{Q}_j \boldsymbol{U}_{j,0}[:, j]\right) + \boldsymbol{B}(\boldsymbol{Q}_j \boldsymbol{U}_{j,1}[:, j])$
  for $i = 0, \ldots, \eta_j - 1$ **do**
    $s_i = \langle \boldsymbol{r}, \boldsymbol{q}_i \rangle$
    $\boldsymbol{r} = \boldsymbol{r} - s_i \boldsymbol{q}_i$
  $\alpha = \|\boldsymbol{r}\|$
  $\boldsymbol{s} = (s_0, \ldots, s_{\eta_j - 1})$
  $\boldsymbol{w} = \boldsymbol{U}_{j,0}[:, j]$
  for $i = 0, \ldots, j$ **do**
    $t_{i,j} = \langle \boldsymbol{s}, \boldsymbol{U}_{j,0}[:, i] \rangle + \langle \boldsymbol{w}, \boldsymbol{U}_{j,1}[:, i] \rangle$
    $\boldsymbol{s} = \boldsymbol{s} - t_{i,j} \boldsymbol{U}_{j,0}[:, i]$
    $\boldsymbol{w} = \boldsymbol{w} - t_{i,j} \boldsymbol{U}_{j,1}[:, i]$
  $t_{j+1,j} = (\alpha^2 + \|\boldsymbol{s}\|^2 + \|\boldsymbol{w}\|^2)^{\frac{1}{2}}$
  if $t_{j+1,j} == 0$ **then**
    stop ▷ breakdown
  if $\alpha == 0$ **then**
    $\eta_{j+1} = \eta_j$ ▷ deflation
    $\boldsymbol{Q}_{j+1} = \boldsymbol{Q}_j$
    $\boldsymbol{U}_{j+1,0} = \begin{pmatrix} \boldsymbol{U}_{j,0} & \boldsymbol{s}/t_{j+1,j} \end{pmatrix}$
    $\boldsymbol{U}_{j+1,1} = \begin{pmatrix} \boldsymbol{U}_{j,1} & \boldsymbol{w}/t_{j+1,j} \end{pmatrix}$
  else
    $\eta_{j+1} = \eta_j$
    $\boldsymbol{Q}_{j+1} = \begin{pmatrix} \boldsymbol{Q}_j & \boldsymbol{r}/\alpha \end{pmatrix}$
    $\boldsymbol{U}_{j+1,0} = \begin{pmatrix} \boldsymbol{U}_{j,0} & \boldsymbol{s}/t_{j+1,j} \\ 0 & \alpha/t_{j+1,j} \end{pmatrix}$
    $\boldsymbol{U}_{j+1,1} = \begin{pmatrix} \boldsymbol{U}_{j,1} & \boldsymbol{w}/t_{j+1,j} \\ 0 & 0 \end{pmatrix}$

From Lemma 3.1 in [12] we know, that

$$\operatorname{span}\{\boldsymbol{q}_0, \ldots, \boldsymbol{q}_m\} = \operatorname{span}\{\boldsymbol{q}_0, \ldots, \boldsymbol{q}_{m-1}, \boldsymbol{r}\}$$

with

$$\boldsymbol{r} = \boldsymbol{A}\boldsymbol{Q}_m \boldsymbol{U}_{m,0}[:, m-1] + \boldsymbol{B}\boldsymbol{Q}_m \boldsymbol{U}_{m,1}[:, m-1] \tag{3.33}$$

Thus, $\boldsymbol{q}_m$ can be computed by orthogonalizing $\boldsymbol{r}$ against $\boldsymbol{Q}_m$ and subsequent normalization. That yields

$$\boldsymbol{q}_m = (\boldsymbol{I} - \boldsymbol{Q}_m \boldsymbol{Q}_m^*)\boldsymbol{r}/\alpha \tag{3.34}$$

with $\alpha \in \mathbb{R}$, such that $\|\boldsymbol{q}_m\| = 1$. This step is done in lines 7 – 12 and 31 in Algorithm 5. It remains to calculate $\boldsymbol{U}_{m+1,0}$ and $\boldsymbol{U}_{m+1,1}$, such that $\boldsymbol{U}_{m+1}$ is orthonormal.
From (3.24) and (3.25) we know that

$$\boldsymbol{v}_m = \begin{pmatrix} \boldsymbol{v}_m[:n] \\ \boldsymbol{v}_m[n:2n] \end{pmatrix} \in \begin{pmatrix} \operatorname{span}\{\boldsymbol{q}_0, \dots, \boldsymbol{q}_m\} \\ \operatorname{span}\{\boldsymbol{q}_0, \dots, \boldsymbol{q}_{m-1}\} \end{pmatrix}.$$

Hence, we can write

$$\begin{aligned}
\boldsymbol{V}_{m+1} &= \begin{pmatrix} \boldsymbol{V}_m & \boldsymbol{v}_m \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{Q}_m \boldsymbol{U}_{m,0} & \boldsymbol{Q}_m \boldsymbol{s} + \beta \boldsymbol{q}_m \\ \boldsymbol{Q}_m \boldsymbol{U}_{m,1} & \boldsymbol{Q}_m \boldsymbol{w} \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{Q}_{m+1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Q}_{m+1} \end{pmatrix} \begin{pmatrix} \boldsymbol{U}_{m,0} & \boldsymbol{s} \\ \boldsymbol{0}^T & \beta \\ \boldsymbol{U}_{m,1} & \boldsymbol{w} \\ \boldsymbol{0}^T & 0 \end{pmatrix} \begin{matrix} \} = \boldsymbol{U}_{m+1,0} \\ \\ \} = \boldsymbol{U}_{m+1,1} \\ \\ \end{matrix} \\
&= \boldsymbol{Q}_{[m+1]} \boldsymbol{U}_{m+1}
\end{aligned}$$

with $\beta \neq 0$, if no deflation occurs and $\beta = 0$ otherwise.

The calculation of $\boldsymbol{s}, \boldsymbol{w}$ and β is now based on the Arnoldi decomposition (3.28). Defining $\underline{\boldsymbol{T}}_{m+1} = \boldsymbol{T}[:m+1, :m]$ and plugging (3.32) into (3.28) yields

$$\boldsymbol{H}\boldsymbol{Q}_{[m]} \boldsymbol{U}_m = \boldsymbol{Q}_{[m+1]} \boldsymbol{U}_{m+1} \underline{\boldsymbol{T}}_{m+1}. \tag{3.35}$$

Expanding the left-hand side and only considering column $m-1$ we get

$$\begin{aligned}
\boldsymbol{H}\boldsymbol{Q}_{[m]} \boldsymbol{U}_m[:, m-1] &\overset{(3.22)}{=} \begin{pmatrix} (\boldsymbol{A}\boldsymbol{Q}_m \boldsymbol{U}_{m,0} + \boldsymbol{B}\boldsymbol{Q}_m \boldsymbol{U}_{m,1})[:, m-1] \\ \boldsymbol{Q}_m \boldsymbol{U}_{m,0}[:, m-1] \end{pmatrix} \\
&\overset{(3.33)}{=} \begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{Q}_m \boldsymbol{U}_{m,0}[:, m-1] \end{pmatrix}.
\end{aligned}$$

The last column of the right-hand and side of (3.35) can be written as

$$\begin{aligned}
\boldsymbol{Q}_{[m+1]} \boldsymbol{U}_{m+1} \underline{\boldsymbol{T}}_{\boldsymbol{m+1}}[:, m-1] &= \boldsymbol{Q}_{[m+1]} \begin{pmatrix} \boldsymbol{U}_{m,0} & \boldsymbol{s} \\ \boldsymbol{0}^T & \beta \\ \boldsymbol{U}_{m,1} & \boldsymbol{w} \\ \boldsymbol{0}^T & 0 \end{pmatrix} \underline{\boldsymbol{T}}_{m+1}[:, m-1] \\
&= \boldsymbol{Q}_{[m+1]} \left(\begin{pmatrix} \boldsymbol{U}_{m,0} \\ \boldsymbol{0}^T \\ \boldsymbol{U}_{m,1} \\ \boldsymbol{0}^T \end{pmatrix} \boldsymbol{T}[:m, m-1] + \begin{pmatrix} \boldsymbol{s} \\ \beta \\ \boldsymbol{w} \\ 0 \end{pmatrix} t_{m,m-1} \right).
\end{aligned}$$

Using the fact that $\boldsymbol{Q}_{[m+1]}$ is an orthogonal matrix, we can calculate

$$\begin{pmatrix} \boldsymbol{s} \\ \beta \\ \boldsymbol{w} \\ 0 \end{pmatrix} t_{m,m-1} = \begin{pmatrix} \boldsymbol{Q}_{m+1}^* \boldsymbol{r} \\ \boldsymbol{Q}_{m+1}^* \boldsymbol{Q}_m \boldsymbol{U}_{m,0}[:, m-1] \end{pmatrix} - \begin{pmatrix} \boldsymbol{U}_{m,0} \\ \boldsymbol{0}^T \\ \boldsymbol{U}_{m,1} \\ \boldsymbol{0}^T \end{pmatrix} \boldsymbol{T}[:m, m-1]$$

where $t_{m,m-1}$ is chosen such that $\| \begin{pmatrix} \boldsymbol{s}^T & \beta & \boldsymbol{w}^T & 0 \end{pmatrix}^T \| = 1$. Furthermore the vector $\begin{pmatrix} \boldsymbol{s}^T & \beta & \boldsymbol{w}^T & 0 \end{pmatrix}^T$ needs to be orthogonalized against $\boldsymbol{V}_m$. This is done in lines 15-18 of Algorithm 5.
In contrast to the SOAR procedure, TOAR as described in Algorithm 5 generates an ONB for $\mathcal{K}(\boldsymbol{H}, \boldsymbol{x})$ as well. Therefore, a breakdown in step j occurs if $\boldsymbol{v}_j = \boldsymbol{0}$. This is checked in line 21.
As in SOAR, a deflation is characterised by $\boldsymbol{q}_i = \boldsymbol{0}$. In Algorithm 5 this case corresponds to $\alpha = 0$ and is dealt with in lines 25-28.

3.3.3 Selection of Eligible Values

In our scenario the matrices in Problem 3.19 depend on ω. Hence Algorithm 1 returns N values for every single value of ω. However we are only interested in eigenpairs where λ fulfills certain criterions.

The first thing we know is that the components of $\boldsymbol{k}$ must be real valued. From Equation (3.20) together with (3.21) follows that we can ignore all eigenpairs $(\boldsymbol{u}^j, \lambda^j)$ with $|\operatorname{Re} \lambda^j|$ bigger than a certain threshold.

To enforce any criterion that restrict $\boldsymbol{k}$ to a certain area, we only have to consider either k_x or k_y due to the linear connection of the two components. Without loss of generality we can assume that $\lambda \mapsto k_x$ is not constant. Otherwise just do the exact same calculations for k_y.
Assume that we want $\boldsymbol{k}$ to be in the first BZ. Then there exist values $a, b \in \mathbb{R}$ such that $k_x \in [a, b]$. Again using Equations (3.20) and (3.21) we only consider eigenpairs $(\boldsymbol{u}^j, \lambda^j)$ with

$$\operatorname{Im} \lambda^j \in \left[\frac{s_x}{b - p_x}, \frac{s_x}{a - p_x} \right].$$

Of course we can restrict the values of interest even further by defining $a = a_1 < \ldots < a_L = b$ and only considering eigenpairs $(\boldsymbol{u}^j, \lambda^j)$ with

$$\operatorname{Im} \lambda^j \in \bigcup_{l=1}^{L-1} \left[\frac{s_x}{a_{l+1} - p_x}, \frac{s_x}{a_l - p_x} \right].$$

For later application purposes we also want to consider a third selection criterion. Remember that the matrices in Problem (3.19) depend on ω. Assume that we solve it for many different values ω in the frequency range $[\omega_{\min}, \omega_{\max}]$. Allowing all $k_x \in [a, b]$ we can expect to see different frequency bands as discussed in Section 3.1.2. Examles of such band

structures can be seen in Chapter 5.

Algorithm 6 Choose values near k_x

Input:
$\mathcal{P}_\omega = [\omega^1, \ldots, \omega^R]$
$\mathcal{P}_\lambda = [\lambda^1, \ldots, \lambda^L]$
Output: `vecs`, `lams`, `omegas`

`vecs`, `lams`, `omegas` $= [\ \underbrace{[\ ], \ldots, [\ ]}_{L\text{–times}}\ ]$
for ω in $\mathcal{P}_\omega$ **do**
    update $\omega-$dependent matrices in Problem 3.19
    solve Problem 3.19 using Algorithm 1 and get eigenpairs `EP` $:= \left[(\boldsymbol{u}^1, \lambda^1), \ldots, (\boldsymbol{u}^N, \lambda^N)\right]$
    for $(\boldsymbol{u}, \lambda)$ in `EP` **do**
        $\lambda = -i/\lambda$
        if Im $\lambda > \delta_{\text{imag}}$ **or** $\operatorname{Re}\lambda < \min \mathcal{P}_\omega - \delta_{\text{th}}$ **or** $\operatorname{Re}\lambda > \max \mathcal{P}_\omega + \delta_{\text{th}}$ **then**
            continue
        $j = \arg\min |\mathcal{P}_\lambda - \lambda|$
        if $|\mathcal{P}_\lambda[j] - \lambda| < \delta_{\text{th}}$ **then**
            `lams`$[j]$.append(λ)
            `vecs`$[j]$.append$(\boldsymbol{u})$
            `omegas`$[j]$.append(ω)
for $l \in \{1, \ldots L\}$ **do**
    if `lams`$[l]$ == [ ] **then**
        continue
    sort `vecs`$[l]$, `lams`$[l]$, `omegas`$[l]$ simultaneously in ascending order such that $|$`lams`$[l][j] - \mathcal{P}_\lambda[l]| < |$`lams`$[l][j+1] - \mathcal{P}_\lambda[l]|$ for all j with `lams`$[l][j+1]$ exists
    $j = 2$
    while $j \leq$ len(`vecs`$[l]$) **do**
        for c in $\{1, \ldots, j-1\}$ **do**
            if $\langle$`vecs`$[l][j]$, `vecs`$[l][c]\rangle > \delta_\perp$ **then**
                delete `lams`$[l][c]$, `vecs`$[l][c]$, `omegas`$[l][c]$ from respective lists
                break

Algorithm 7 Sort into bands

Input:
eigenpairs with associated frequency EP := $[(\boldsymbol{u}^1, \lambda^1, \omega^1), \ldots, (\boldsymbol{u}^N, \lambda^N, \omega^N)]$
Output: vecs, lams, omegas = [$\underbrace{[\,], \ldots, [\,]}_{\text{amount of bands times}}$]

sort EP along λ
srt_vecs = $[\boldsymbol{u}^1, \ldots \boldsymbol{u}^N]$
srt_lams = $[\lambda^1, \ldots \lambda^N]$
srt_omegas = $[\omega^1, \ldots \omega^N]$
bands = [[0]]
dist = 0
for $i = 1, \ldots,$ len(EP) **do**
    nearest_band = None
    for $j = 1, \ldots,$ len(bands) **do**
        b = bands[j]
        if $|\langle$srt_vecs[$b[-1]$], srt_vecs[i]$\rangle| >$ dist **then**
            nearest_band = b
            dist = $|\langle$srt_vecs[$b[-1]$], srt_vecs[i]$\rangle|$
    if dist $<$ th **then**
        bands.append($[i]$)
    else
        nearest_band.append(i)
vecs, lams, omegas = [ ]
for b in bands **do**
    vecs.append(srt_vecs[b])
    lams.append(srt_lams[b])
    omegas.append(srt_omegas[b])
sort order of bands by mean frequency of the bands

Assume that we have a set of discrete values $\{k_x^1, \ldots, k_x^L\}$. For each point k_x^j in the set and each band in the frequency range we want to find one k_x such that $|k_x^j - k_x|$ is as small as possible.
To this end we need a method to automatically separate the bands independent of how broad the band gap is or if a band gap exists at all. Consider two solutions $(\boldsymbol{u}^1, \lambda^1)$ and $(\boldsymbol{u}^2, \lambda^2)$ of problems with respective frequencies ω^1, ω^2. Further assume that $\lambda^1 \approx \lambda^2$. As already discussed in Section 3.1.2 $(\boldsymbol{u}^1, \omega^1)$ and $(\boldsymbol{u}^2, \omega^2)$ can be interpreted as eigenpairs of (3.11). For $\|u^1\| = \|u^2\| = 1$ and a threshold $\delta_\perp > 0$, we get the simple criterion

$$|\boldsymbol{u}^1 \cdot \boldsymbol{u}^2| \leq \delta_\perp \quad \Rightarrow \quad u^1, u^2 \text{ belong to different bands,}$$
$$|\boldsymbol{u}^1 \cdot \boldsymbol{u}^2| > \delta_\perp \quad \Rightarrow \quad u^1, u^2 \text{ belong to the same band.}$$

Above considerations are summarized in Algorithm 6. A modified version that just sorts existing solutions into a priorly unknown amount of bands is described in Algorithm 7

3.4 Reduced Basis

In this section we will cover the basics of a reduced basis (RB) method as described in [1] and apply it to Problems 3.16 and 3.19. First define the following generic problem.

Problem 3.22. Let $\mathcal{P} = \{\nu_1, \dots, \nu_L\}$ be a set of parameters and $\nu \in \mathcal{P}$. Find $\boldsymbol{u} \in \mathbb{C}^{N_h}$ and $\lambda \in \mathbb{C}$ such that

$$\boldsymbol{A}(\lambda, \nu)\boldsymbol{u} = \boldsymbol{0} \tag{3.36}$$

with

$$\boldsymbol{A}(\lambda, \nu) = \sum_{q=1}^{m_A} \Phi_A^q(\lambda)\Theta_A^q(\nu)\boldsymbol{A}^q \in \mathbb{C}^{N_h \times N_h}. \tag{3.37}$$

Maybe we also want to impose some restrictions that indicate which solutions $(\boldsymbol{u}, \lambda)$ are eligible. Problem 3.22 should be solved for a huge set of parameters $\mathcal{P}$. This can be computationally quite expensive since in general N_h can be very large. This is why it makes sense to instead look for a smaller problem, the so called *reduced problem*

$$\widehat{\boldsymbol{A}}(\lambda, \nu)\widehat{\boldsymbol{u}} = 0 \tag{3.38}$$

with $\widehat{\boldsymbol{A}} \in \mathbb{C}^{\widehat{N} \times \widehat{N}}$ such that the so called *reduced solution* $(\widehat{\boldsymbol{u}}, \lambda) \in \mathbb{C}^{\widehat{N}} \times \mathbb{C}$ is still a good enough approximation to the solution of our original problem and $\widehat{N} \ll N_h$. Note that the matrices $\boldsymbol{A}, \widehat{\boldsymbol{A}}$ and therefore the solutions $\boldsymbol{u}, \widehat{\boldsymbol{u}}$ depend on the parameter ν. We leave out that dependence for better readability, but it is important to keep in mind.

Now we are going to construct such a reduced problem. The idea is to choose so called *snapshot parameters* $\{\nu^1, ..., \nu^{\widehat{N}}\}$ and calculate the associated solutions $\{\boldsymbol{u}^1, ..., \boldsymbol{u}^{\widehat{N}}\}$. They span a subspace $\mathcal{V}_{\widehat{N}} \subset \mathbb{C}^{N_h}$. To get a good basis $\{\boldsymbol{\zeta}^1, ..., \boldsymbol{\zeta}^{\widehat{N}}\}$ of $\mathcal{V}_{\widehat{N}}$ we are going to orthogonalize $\{\boldsymbol{u}^1, ..., \boldsymbol{u}^{\widehat{N}}\}$ by applying the Gram-Schmidt process. Using these ζ_j as columns of a matrix, we get the *transformation matrix* $\boldsymbol{Q} \in \mathbb{C}^{N \times \widehat{N}}$. For a solution $\widehat{\boldsymbol{u}}$ of the reduced problem the vector $\boldsymbol{Q}\widehat{\boldsymbol{u}}$ should be a good approximation for a solution of the original problem 3.22.

What we need is a notion of what a good approximation is. Therefore we define the *residual vector*

$$\boldsymbol{\rho}(\widehat{\boldsymbol{u}}, \lambda) = \boldsymbol{A}(\lambda, \nu)\boldsymbol{Q}_{\widehat{N}}\widehat{\boldsymbol{u}} \tag{3.39}$$

and further the residual as

$$\mathrm{res}(\widehat{\boldsymbol{u}}, \lambda) = \|\boldsymbol{\rho}(\widehat{\boldsymbol{u}}, \lambda)\|^2. \tag{3.40}$$

Note that $\boldsymbol{\rho}$ depends on $\nu, \widehat{\boldsymbol{u}}$ and λ. One way to obtain (3.38) is to force the residual vector to be orthogonal to the subspace $\mathcal{V}_{\widehat{N}}$. Or in other words the orthogonal projection of $\boldsymbol{\rho}$ on $\mathcal{V}_{\widehat{N}}$ has to vanish. In matrix form this criterion can be written as

$$\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}\widehat{\boldsymbol{u}} = \boldsymbol{0}.$$

Hence the reduced matrix in (3.38) is $\hat{\boldsymbol{A}} = \boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}$.

There has been little progress so far: Although solving the reduced Problem (3.38) is much cheaper than solving the big Problem (3.36), the matrix $\widehat{\boldsymbol{A}}$ must be assembled for every parameter ν. This calculation still depends on the large dimension N_h.

Taking a closer look at the matrix $\boldsymbol{A}$ as defined in (3.37) we see that it is *affine parametric dependent*. Meaning $\theta_A^q(\nu)$ and $\Phi_A^q(\lambda)$ are scalar functions and the $\boldsymbol{A}^q$ are independent of the parameter ν. The reduced version is the following.

Problem 3.23. Find $\widehat{\boldsymbol{u}} \in \mathbb{C}^{\widehat{N}}$ and $\lambda \in \mathbb{C}$ such that

$$\widehat{\boldsymbol{A}}(\lambda, \nu)\widehat{\boldsymbol{u}}(\nu) = \boldsymbol{0},$$

with

$$\widehat{\boldsymbol{A}}(\lambda, \nu) = \sum_{q=1}^{m_A} \Phi_A^q(\lambda)\Theta_A^q(\nu)\widehat{\boldsymbol{A}^q}$$

and the matrices

$$\widehat{\boldsymbol{A}^q} = \boldsymbol{Q}_{\widehat{N}}^* \boldsymbol{A}^q \boldsymbol{Q}_{\widehat{N}}.$$

The reduced problem has to be built only once in a possibly expensive *offline* stage and allows a rapid assembly during a cheap *online* stage of the calculation.

Remark 3.24. *In our calculation we consider a reduced basis matrix* $\boldsymbol{Q} \in \mathbb{R}^{N_h \times 2\widehat{N}}$ *instead of* $\boldsymbol{Q} \in \mathbb{C}^{N_h \times \widehat{N}}$.

3.4.1 A Greedy Choice of Snapshots

In this subsection we are going to explore how to construct a projection matrix $\boldsymbol{Q}$. To that end we need a method to choose good snapshots correspondent to parameters $\{\nu, \ldots, \nu^{\widehat{N}}\}$. In the literature one can find different approaches to that problem. One very popular approach to construct a RB space is to use a so called greedy algorithm [1]. In a nutshell the idea of a greedy construction of a RB space is to iteratively add new basis vectors at each step that fulfil some kind of local optimality criterion.

Assume that we already have the RB space $\mathcal{V}_{\widehat{N}} = \operatorname{span}\{\boldsymbol{u}_h^1, \ldots, \boldsymbol{u}^{\widehat{N}}\}$ with orthonormalized basis vectors $\{\boldsymbol{\zeta}^1, \ldots, \boldsymbol{\zeta}^{\widehat{N}}\}$ which constitute the columns of $\boldsymbol{Q}_{\widehat{N}}$.

We ask ourselves which vector $\boldsymbol{u}_h^{\widehat{N}+1}$ should be added to the set of basis vectors that will span $\mathcal{V}_{\widehat{N}+1}$. A simple answer would be to solve the reduced Problem 3.23 for all $\nu \in \mathcal{P}$. This yields a set of solutions $\mathcal{S} = \{(\widehat{\boldsymbol{u}}^1, \lambda^1), \ldots, (\widehat{\boldsymbol{u}}^L, \lambda^L)\}$. Now we choose the ν belonging to

$$(\widehat{\boldsymbol{u}}^{N+1}, \lambda^{N+1}) := \underset{(\widehat{\boldsymbol{u}}, \lambda) \in \mathcal{S}}{\arg\max}\ \operatorname{res}(\widehat{\boldsymbol{u}}, \lambda).$$

For this parameter ν we solve (3.36). It could have more than one eligible solution. Therefore we only choose $\boldsymbol{u}$ as a snapshot solution if the associated λ is in close proximity to λ^{N+1}. Besides we want to avoid adding redundant information to our basis matrix $\boldsymbol{Q}_{\widehat{N}}$.

Assume that $\|\boldsymbol{u}\| = 1$ and let $\boldsymbol{u}_p$ be the projection of $\boldsymbol{u}$ onto the space spanned by the columns of $\boldsymbol{Q}_{\widehat{N}}$. If $\|\boldsymbol{u} - \boldsymbol{u}_p\|$ is smaller than a certain threshold, we also refrain from adding it.

However this version of a greedy algorithm would make the RB construction very expensive for two reasons. First solving (3.23) has to be done for every parameter ν in the potentially very large set $\mathcal{P}$. Additionally the residual has to be calculated for each eligible solution. There can be none, one or even more such solutions for every $\nu \in \mathcal{P}$. Second, the cost of calculating the residual for one value $(\widehat{\boldsymbol{u}}, \lambda)$ is in $\mathcal{O}(N_h^2 \widehat{N})$. How expensive it is to calculate a solution depends on the exact problem. However alone the overall calculation of the residual can be expected to be in $\mathcal{O}(N_h^2 \widehat{N} L)$ for each newly added snapshot.

The second problem can be addressed by choosing a random subset $\mathcal{P}_{\widehat{L}} \subseteq \mathcal{P}$ of size

$$|\mathcal{P}_{\widehat{L}}| = \widehat{L} \quad \ll \quad L = |\mathcal{P}_L|$$

in each iteration. This brings a significant speedup, but can lead to worse results and not deterministic behaviour of the algorithm.
To reduce the computational costs of the residual we write (3.40) as

$$\begin{aligned}
\operatorname{res}(\widehat{\boldsymbol{u}}, \lambda)^2 &= \|\boldsymbol{\rho}(\widehat{\boldsymbol{u}}, \lambda)\|^2 \\
&= \langle \boldsymbol{\rho}(\widehat{\boldsymbol{u}}, \lambda), \boldsymbol{\rho}(\widehat{\boldsymbol{u}}, \lambda) \rangle \\
&\overset{(3.39)}{=} \langle \boldsymbol{A}(\lambda, \nu) \boldsymbol{Q}_{\widehat{N}} \widehat{\boldsymbol{u}}, \boldsymbol{A}(\lambda, \nu) \boldsymbol{Q}_{\widehat{N}} \widehat{\boldsymbol{u}} \rangle \\
&\overset{\text{ONB}}{=} \langle \boldsymbol{A}(\lambda, \nu) \sum_{j=1}^{\widehat{N}} \widehat{u}_j \boldsymbol{\zeta}_j, \boldsymbol{A}(\lambda, \nu) \sum_{k=1}^{\widehat{N}} \widehat{u}_k \boldsymbol{\zeta}_k \rangle \\
&= \sum_{j,k=1}^{\widehat{N}} \widehat{u}_j \overline{\widehat{u}_k} \langle \boldsymbol{A}(\lambda, \nu) \boldsymbol{\zeta}_j, \boldsymbol{A}(\lambda, \nu) \boldsymbol{\zeta}_k \rangle \\
&\overset{(3.37)}{=} \sum_{j,k=1}^{\widehat{N}} \widehat{u}_j \overline{\widehat{u}_k} \langle \sum_{q=1}^{m_A} \Phi_A^q(\lambda) \Theta_A^q(\nu) \boldsymbol{A}^q \boldsymbol{\zeta}_j, \sum_{p=1}^{m_A} \Phi_A^p(\lambda) \Theta_A^p(\nu) \boldsymbol{A}^p \boldsymbol{\zeta}_k \rangle \\
&= \sum_{q,p=1}^{m_A} \Phi_A^q(\lambda) \Theta_A^q(\nu) \overline{\Phi_A^p(\lambda) \Theta_A^p(\nu)} \sum_{j,k=1}^{\widehat{N}} \widehat{u}_j \overline{\widehat{u}_k} \langle \boldsymbol{A}^q \boldsymbol{\zeta}_j, \boldsymbol{A}^p \boldsymbol{\zeta}_k \rangle \\
&= \sum_{q,p=1}^{m_A} \Phi_A^q(\lambda) \Theta_A^q(\nu) \overline{\Phi_A^p(\lambda) \Theta_A^p(\nu)} \widehat{\boldsymbol{u}}^* \boldsymbol{R}^{q,p} \widehat{\boldsymbol{u}}
\end{aligned} \tag{3.41}$$

with

$$R_{j,k}^{q,p} = \sum_{j,k=1}^{\widehat{N}} \langle \boldsymbol{A}^q \boldsymbol{\zeta}_j, \boldsymbol{A}^p \boldsymbol{\zeta}_k \rangle.$$

The matrices $\boldsymbol{R}^{q,p}$ have to be calculated once in the offline stage and can subsequently be used to calculate the residual online for all values $\nu \in \mathcal{P}$. Note that $\boldsymbol{R}^{p,q} = \boldsymbol{R}^{q,p*}$, so only

$(m_A+1)m_A/2$ matrices have to be calculated and stored. We will call the residual in form (3.41) the *cheap residual*.

All together the adapted greedy procedure to build a RB space is summarized in Algorithm 8.

Algorithm 8 greedy RB space

Input:
set of parameters $\mathcal{P}$
amount of randomly chosen parameters $\widehat{L}$
initial RB space base vectors stored in columns of matrix $\boldsymbol{Q}_0$
Output: RB space base vectors stored in columns of matrix $\boldsymbol{Q}$

for $j \in \{0, ..., \text{MAX-ITER}\}$ **do**
  build the reduced Problem 3.23 for $\boldsymbol{Q}_j$
  build the cheap residual (3.41) for $\boldsymbol{Q}_j$
  randomly choose subset $\mathcal{P}_{\widehat{L}} \subseteq \mathcal{P}$ with $|\mathcal{P}_{\widehat{L}}| = \widehat{L}$
  solve Problem 3.23 for all $\nu \in \mathcal{P}_{\widehat{L}}$
  store the solutions in $\widehat{\boldsymbol{S}} := [(\nu^1, \widehat{\boldsymbol{u}}^1, \lambda^1), (\nu^2, \widehat{\boldsymbol{u}}^2, \lambda^2), \ldots]$ such that $\text{res}((\widehat{\boldsymbol{u}}^1, \lambda^1) \geq \text{res}((\widehat{\boldsymbol{u}}^2, \lambda^2) \geq \ldots$ and $\|\widehat{\boldsymbol{u}}\| = 1$ ▷ note that it could be $\nu^l = \nu^{l+1}$
  if $\text{res}((\widehat{\boldsymbol{u}}^1, \lambda^1) < \text{THRESHOLD}$ **then** break
  for $(\nu, \widehat{\boldsymbol{u}}, \lambda) \in \widehat{\boldsymbol{S}}$ **do**
    `snapshot_added` = False
    solve Problem 3.22 for ν
    store the solutions $(\lambda_\nu, \boldsymbol{u}_\nu)$ with $\|\boldsymbol{u}_\nu\| = 1$ in $\boldsymbol{S}_\nu$ ▷ $\boldsymbol{S}_\nu$ could also be empty
    for $(\lambda_\nu, \boldsymbol{u}_\nu) \in \boldsymbol{S}_\nu$ **do**
      if $|\lambda - \lambda_\nu| > \delta_1$ **then** continue
      define $\boldsymbol{u}_\nu^\perp$ as $\boldsymbol{u}_\nu$ orthogonalized against the columns of $\boldsymbol{Q}_j$
      if $\|\boldsymbol{u}_\nu - \boldsymbol{u}_\nu^\perp\| < \delta_2$ **then** continue
      set $\boldsymbol{Q}_{j+1}[:, : j+1] = \boldsymbol{Q}$ and $\boldsymbol{Q}_{j+1}[:, j+1] = \boldsymbol{u}_\nu^\perp / \|\boldsymbol{u}_\nu^\perp\|$
      `snapshot_added` = True
    if `snapshot_added` **then**
      return $\boldsymbol{Q}_{j+1}$

3.4.2 A Reduced Version of the GEP and QEP

Problem 3.16 consists of two matrices $\boldsymbol{C}, \boldsymbol{G} \in \mathbb{C}^{N_h \times N_h}$. The first one is affine parametric dependent on parameter $\boldsymbol{k}$ and the second does not depend on any parameter. Hence it fits the structure of Problem 3.22.

In Problem 3.19 the matrix $\boldsymbol{M}$ depends on the parameter ω. If the material parameters ε and $\boldsymbol{\mu}$ do not depend on ω the matrix $\boldsymbol{M}$, $\boldsymbol{D}$ and $\boldsymbol{K}$ are affine parametric dependent. Otherwise we have a look at (3.3) again. In a first step assume that the permeability $\boldsymbol{\mu}$ is

frequency independent. The permittivity ε only appears in

$$b(u,v) = \int_\Omega \varepsilon(\omega, \boldsymbol{r}) u(\boldsymbol{r}) v(\boldsymbol{r}).$$

We know that Ω consists of two separate materials, the integration area is composed of $\Omega = \Omega_{\text{inner}} \cup \Omega_{\text{outer}}$. The outer material is air, so its relative permittivity is 1. The inner material is homogeneous, hence the permittivity is a function independent of the position $\boldsymbol{r}$. This means we can split up the integral

$$\begin{aligned} b(u,v) &= \int_\omega \varepsilon(\omega, \boldsymbol{r}) u(\boldsymbol{r}) v(\boldsymbol{r}) \\ &= \varepsilon(\omega) \int_{\Omega_{\text{inner}}} u(\boldsymbol{r}) v(\boldsymbol{r}) + \int_{\Omega_{\text{outer}}} u(\boldsymbol{r}) v(\boldsymbol{r}) \\ &= \varepsilon(\omega) \int_\Omega \mathbb{1}_{\Omega_{\text{inner}}} u(\boldsymbol{r}) v(\boldsymbol{r}) + \int_\Omega \mathbb{1}_{\Omega_{\text{outer}}} u(\boldsymbol{r}) v(\boldsymbol{r}) \\ &=: \varepsilon(\omega) b_{\text{inner}}(u,v) + b_{\text{outer}}(u,v). \end{aligned}$$

That is how we arrive at an affine parametric version for the matrix $\boldsymbol{M}$. For a frequency dependent permeability we repeat that process for the sesquilinear forms containing $\boldsymbol{\mu}$.

4 Numerical Methods for Calculating Chern Numbers

In Section 2.2 we stated the Chern Theorem 2.7 in the abstract setting that the states $\boldsymbol{u}$ belong to the complex vector space $\mathcal{V}$ with scalar product $\langle ., . \rangle$. The integration area S was a closed two dimensional manifold.

In our case S is the first BZ which is a torus and hence a closed manifold. As vector space we choose $\mathcal{V}_h \subset \mathcal{H}_p^1$ from Section 3.1.2. The corresponding scalar product is

$$\langle u_{\boldsymbol{k}^1}, u_{\boldsymbol{k}^2} \rangle = \int_B \varepsilon(\omega, \boldsymbol{r}) e^{i(\boldsymbol{k}^1 - \boldsymbol{k}^2)\cdot \boldsymbol{r}} u(\boldsymbol{r}) \overline{v(\boldsymbol{r})} \mathrm{d}\mathbf{r}$$

where ε is the permittivity.
We will only compute Chern numbers of frequency bands that are separated. What that means can be seen in Section 3.1.2. From now all eigenvectors are assumed to belong to the same frequency band.

4.1 First Principal Calculation

The first principal calculation of Chern numbers was proposed in [24]. The idea is to approximate the Berry flux (2.23) by discretizing the BZ into a finite amount of square shaped patches. Each patch P has corners $\boldsymbol{k}_P^1, \boldsymbol{k}_P^2, \boldsymbol{k}_P^3, \boldsymbol{k}_P^4$. The area $|P|$ should be small enough such that the Berry phase around the boundary ∂P is unambiguous. This is the same approach as in (2.21). For better readability we define a *link* as

$$U_{\boldsymbol{k}_P^j \to \boldsymbol{k}_P^l} := \frac{\langle u_{\boldsymbol{k}_P^j}, u_{\boldsymbol{k}_P^l} \rangle}{|\langle u_{\boldsymbol{k}_P^j}, u_{\boldsymbol{k}_P^l} \rangle|}.$$

To approximate the Berry phase around one patch we calculate

$$\phi_P = \mathrm{Im}\ \ln\left(U_{\boldsymbol{k}_P^1 \to \boldsymbol{k}_P^2} U_{\boldsymbol{k}_P^2 \to \boldsymbol{k}_P^3} U_{\boldsymbol{k}_P^3 \to \boldsymbol{k}_P^4} U_{\boldsymbol{k}_P^4 \to \boldsymbol{k}_P^1} \right)$$

only using the states at the corners. The normalisation is not relevant for the overall result. It is only done for numerical stability. Note that the values on the boundary only have to be computed either for Γ^l and Γ^r or Γ^b and Γ^t because $u\big|_{\Gamma^t} = u\big|_{\Gamma^b}$ and $u\big|_{\Gamma^l} = u\big|_{\Gamma^r}$.

The Berry flux (2.23) can now be computed as

$$\Phi_S = \sum_P \phi_P.$$

According to the Chern Theorem 2.7 the Chern number C is defined as the integer

$$C = \frac{\Phi_S}{2\pi}.$$

4.2 Wilson Loop Approach

The Wilson loop approach is detailed described in [22] and implemented in [20]. As already discussed in Section 2.2 the presence of non trivial Chern numbers means that no gauge can be chosen that is continuously differentiable everywhere on S. However for a fixed k_y we can choose a smooth gauge representation that is periodic along the direction k_x. Consequently for the first BZ (2.11) we can calculate the Berry flux (2.23) as

$$\begin{aligned}
\Phi_S &= \int_{-\pi/a}^{\pi/a}\int_{-\pi/a}^{\pi/a} \Omega(\boldsymbol{k})\mathrm{dk_x dk_y} \\
&= \int_{-\pi/a}^{\pi/a}\int_{-\pi/a}^{\pi/a} \partial_{k_x} A_y(\boldsymbol{k}) - \partial_{k_y} A_x(\boldsymbol{k})\mathrm{dk_x dk_y} \\
&= \int_{-\pi/a}^{\pi/a} \underbrace{\int_{-\pi/a}^{\pi/a} \partial_{k_x} A_y(\boldsymbol{k})\mathrm{dk_x}}_{=0 \text{ because of smooth gauge}} - \int_{-\pi/a}^{\pi/a} \partial_{k_y} A_x(\boldsymbol{k})\mathrm{dk_x dk_y} \\
&= \int_{-\pi/a}^{\pi/a}\int_{-\pi/a}^{\pi/a} \partial_{k_y} \operatorname{Im}\, \langle \boldsymbol{u_k}, \partial_{k_x} \boldsymbol{u_k}\rangle \mathrm{dk_x dk_y} \\
&= \int_{-\pi/a}^{\pi/a} \partial_{k_y} \int_{-\pi/a}^{\pi/a} \operatorname{Im}\, \langle \boldsymbol{u_k}, \partial_{k_x} \boldsymbol{u_k}\rangle \mathrm{dk_x dk_y} \\
&= \int_{-\pi/a}^{\pi/a} \partial_{k_y} \phi(k_y)\mathrm{dk_y}.
\end{aligned}$$

When approximating $\partial_{k_y}\phi(k_y)\mathrm{dk_y}$ we have to remember that the Berry phase is only unique up to an integer multiple of 2π. Assume that we restrict ϕ on the interval $[-\pi, \pi)$. Now we consider a small change Δk_y. Instead of approximating the corresponding change in the Berry phase $\Delta\phi$ with $\Delta\phi \approx \phi(k_y + \Delta k_y) - \phi(k_y)$ we define

$$d_m(k_y, k_y + \Delta k_y) = \phi(k_y + \Delta k_y) + 2\pi m - \phi(k_y).$$

For two neighbouring approximation points k_y^1, k_y^2 we choose m such that

$$m = \underset{m\in\{-1,0,1\}}{\arg\min} |d_m(k_y^1, k_y^2)|.$$

For a discretization $-\pi/a = k_y^1 < k_y^2 < \ldots < k_y^l < \pi/a$ we can approximate the Berry flux as

$$\Phi_S = \sum_{j=1}^{l-1} \frac{d_m(k_y^j, k_y^{j+1})}{k^{j+1} - k^j}(k^{j+1} - k^j) = \sum_{j=1}^{l-1} d_m(k_y^j, k_y^{j+1}).$$

Again the Chern Theorem 2.7 yields a Chern number C that is defined as

$$C = \frac{\Phi_S}{2\pi}.$$

This approach allows an interesting topological interpretation: the Chern numbers correspond to the winding number of the Barry phases around the Torus. More information on that topic can be found in [22].

5 Results

In this chapter we first define four model problems for which we calculate band structures and, if possible, Chern numbers. Two of them are already benchmarked in the literature. Hence we use them to verify the correctness of our methods. Furthermore we discuss the speedup that we gain from employing a reduced basis model order reduction. Additionally we investigate what role the degreed of freedom of the FEM space and the the accuracy of the RB approximation play in calculating correct Chern numbers. Last we compare the WLA and FPC for different use cases.

The source code for all numerical experiments conducted for this thesis (and more) can be found in https://github.com/huberamanda/dispersion.git.

5.1 Model Problems

In this section we discuss which parameters we choose for Problem 2.2. The variable parameters are

- the radius r of the rod inside the unit cell Ω as depicted in Figure 2.1,
- the permittivity $\varepsilon(\omega)$ and
- the permeability tensor $\boldsymbol{\mu}(\omega)$.

Having a closer look at the form of $\boldsymbol{\mu}$ as defined in (2.2) we see that it actually depends on the the gyromagnetic ratio γ, the material dependent magnetic saturation M_s, and the magnetic field strength H_0. All of them depend on the material or the magnetic field.According to [14] γ can be chosen as 1.759×10^{11}. However, we will set $\gamma = 1.75784 \times 10^{11}$ to get the same values μ and κ as in the benchmark problem used in [21], [24] and [20]. In these papers a dispersion free version of the permeability tensor is considered, meaning that the dependency of (2.2) on ω is ignored. Instead, a constant frequency $\tilde{\omega}$ of the magnetic field is introduced. The permeability is then given as defined in (2.2) just with $\tilde{\omega}$ instead of ω.

Experiment I consists of Yittrium-Iron-Garnet rods in air. It is examined in [21], [24] and [20]. According model parameters can be found in [14].

Experiment II is benchmarked in [23]. It serves as a control problem to check if our implementation is correct for frequency dependent material parameters.

Experiment III consists of the same YIG rods as I. The only difference is that we now consider the permeability tensor (2.2) for a variable frequency. In our setup we have always

	I	II	III	IV
r [m]	$0.11a$	$0.2a$	$0.11a$	$0.11a$
$\varepsilon(\omega)$ [F/m]	15	$1-\frac{(1914\times 2\pi 10^{12})^2}{\omega^2-i\omega 8.34\times 2\pi 10^{12}}$	15	$15+8\sin(-\frac{\pi}{2}+\frac{2\pi\omega}{c})$
$\tilde{\omega}$ [Hz]	$4.28\times 2\pi 10^9$	0	200ω	$4.28\times 2\pi 10^9$
H_0 [T]	0.16	0	0.16	0.16
$4\pi M_s$ [T]	0.178	–	0.178	0.178

Table 5.1: Parameters for different experiments.

considered a unit cell length of $a = 1$. However the frequency scales with the length of the unit cell. So actually III models a PC with $a = 0.005$ meters.

Experiment IV again considers the same YIG rods as I except for a frequncy dependence of the permittivity ε that is completely made up for demonstration purposes.

We will conduct our calculations with different values for the following parameters

- `ndof`: number of degrees of freedom for the FEM space,
- `nparam`: amount of parameters $\mathcal{P}$ in Algorithm 8,
- `th_res`: threshold for residual in Algorithm 8,
- `nval`: $\hat{L}$ in Algorithm 8.

The `ndof` depend on the maximum height `maxh` of the triangles in the triangulation and the `order` of the FEM space.

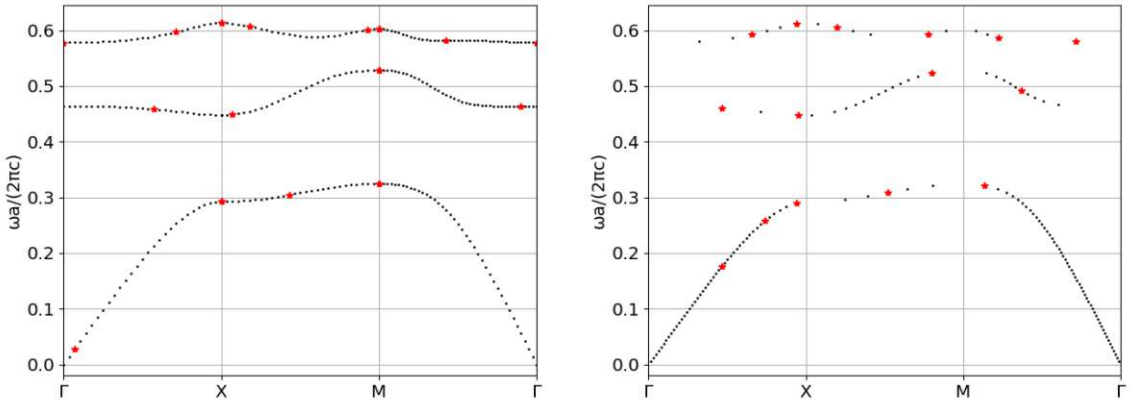
5.2 Band Structures

In the literature the bands for wave vectors along the irreducible BZ, as illustrated in Figure 2.2, are called the *band structure* of a PC [8]. Remember that for a quadratic Brillouin zone the boundary of the irreducible Brillouin zone is the triangle connecting the points

$$\Gamma = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad X = \begin{pmatrix} \frac{\pi}{a} \\ 0 \end{pmatrix}, \quad M = \begin{pmatrix} \frac{\pi}{a} \\ \frac{\pi}{a} \end{pmatrix}.$$

5.2.1 Frequency Independent Material Parameters

We want to compare the construction of the band structure by solving GEPs as described in Section 3.2 to solving QEPs as described in Section 3.3. The only problem in Table 5.1 where both methods can be employed is I. As parameters $\mathcal{P}$ we choose equidistant points along the irreducible Brillouin zone for the GEP and equidistant values between a minimal and maximal scaled frequency for the QEP.



(a) Solutions of GEPs. The greedy algorithm terminated after 3.94 seconds and produced a RB space of dimension 30. The online calculation time amounted to 0.50 seconds.

(b) Solutions of QEPs. The greedy algorithm terminated after 15.77 seconds and produced a RB space of dimension 30. The online calculation time amounted to 4.71 seconds.

Figure 5.1: Band structures of the first 3 frequency bands for parameters `ndof` $= 1296$ (`maxh` $= 0.1$, `order` $= 3$), `nparam` $= 100$, `th_res` $= 10^{-3}$ and `nval` $= 50$. The red stars indicate the snapshot parameters.

As can be seen in Figure 5.1 the results in 5.1a are not only better than the ones in 5.1b, but the online and the offline phases are significantly faster. The visual quality of the band structures differ that much because the slope of each band is steeper as a function of ω than as a function of $\boldsymbol{k}$. It is worth mentioning that technically there is no function that represents one band as a function of ω because there can be more values for one ω, but by restricting the range, a local representation can be found. There are several reasons for the longer computation times. To illustrate that we consider the online stage of the computation. A linear connection (3.20) between k_x and k_y is required. Hence, for each side of the triangle `nparam` QEPs have to be solved. In contrast to the LOBPCG algorithm more eigenvalues then the ones we want are computed by the TOAR procedure. This is necessary to get the desired precision. To arrive at a band structure of similar density than in Figure 5.1a we need `nval` to be about 10000, as can be seen in Figure 5.2. Less parameters are necessary if they are not chosen equidistantly but denser in range of flat bands.

Another advantage of solving a GEP is that the amount of desired bands is an input parameter of LOBPCG. If a QEP is solved, we can only specify a range of parameters ω. If bands we are not interested in cross that range, the greedy algorithm will choose a basis that is optimized for them too, which results in larger reduced systems than necessary. This behaviour can be observed when calculating the first 4 bands for I as can be seen in Figure 5.3.

In Figure 5.4 we investigate the computational advantage gained from performing a model order reduction.

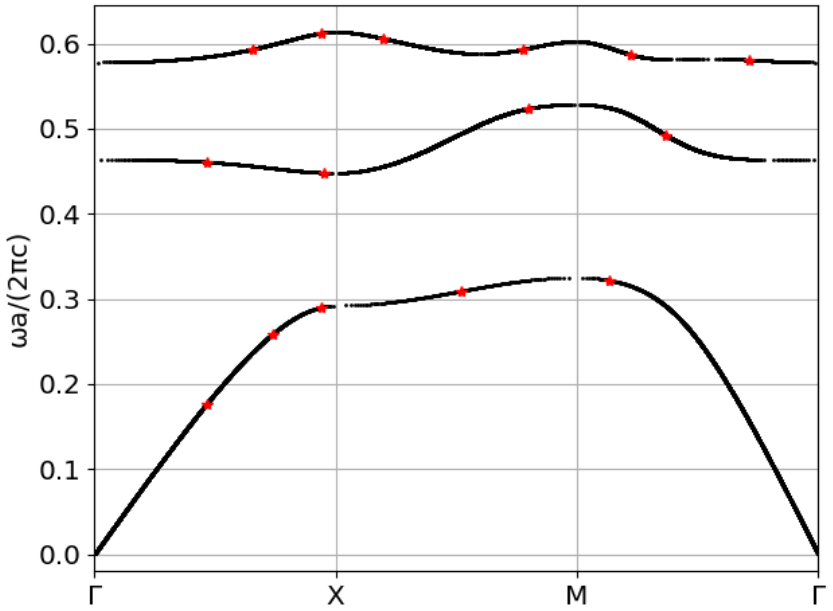
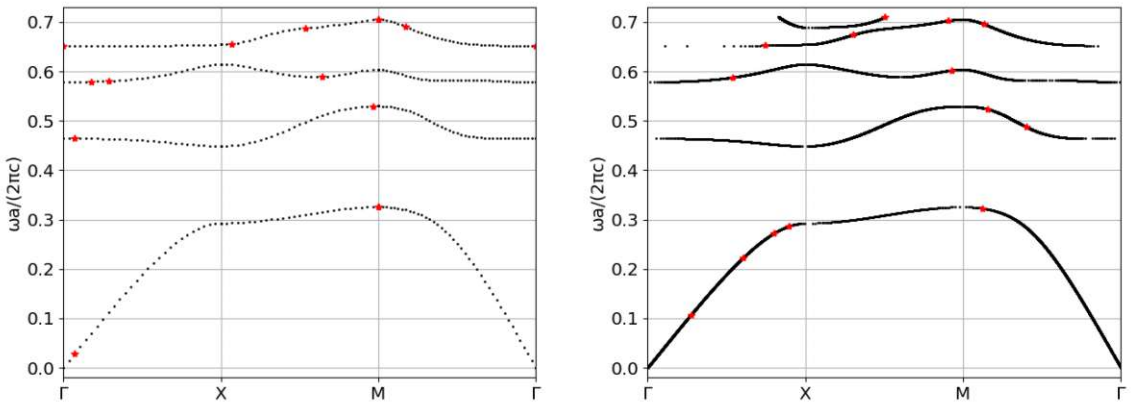


Figure 5.2: Band structure calculated with the same RB space as in Figure 5.1b for parameters `np.concatenate([np.linspace(1e-4, 0.35, 2000), np.linspace(0.42, 0.625, 8000)])`. The online computation time amounted to 446.15 seconds.



(a) Solutions of GEPs. The greedy algorithm terminated after 3.76 seconds and produced a RB space of dimension 23.

(b) Solutions of QEPs. The greedy algorithm terminated after 87.44 seconds and produced a RB space of dimension 26.

Figure 5.3: Band structure for the first 4 frequency bands of I. The RB space is built for parameters `ndof` = 328 (`maxh` = 0.2, `order` = 2), `nparam` = 100, `th_res` = 10^{-3} and `nval` = 50. The red stars indicate the snapshot parameters. The online calculation for the QEP is done for 10000 not equidistant parameters.

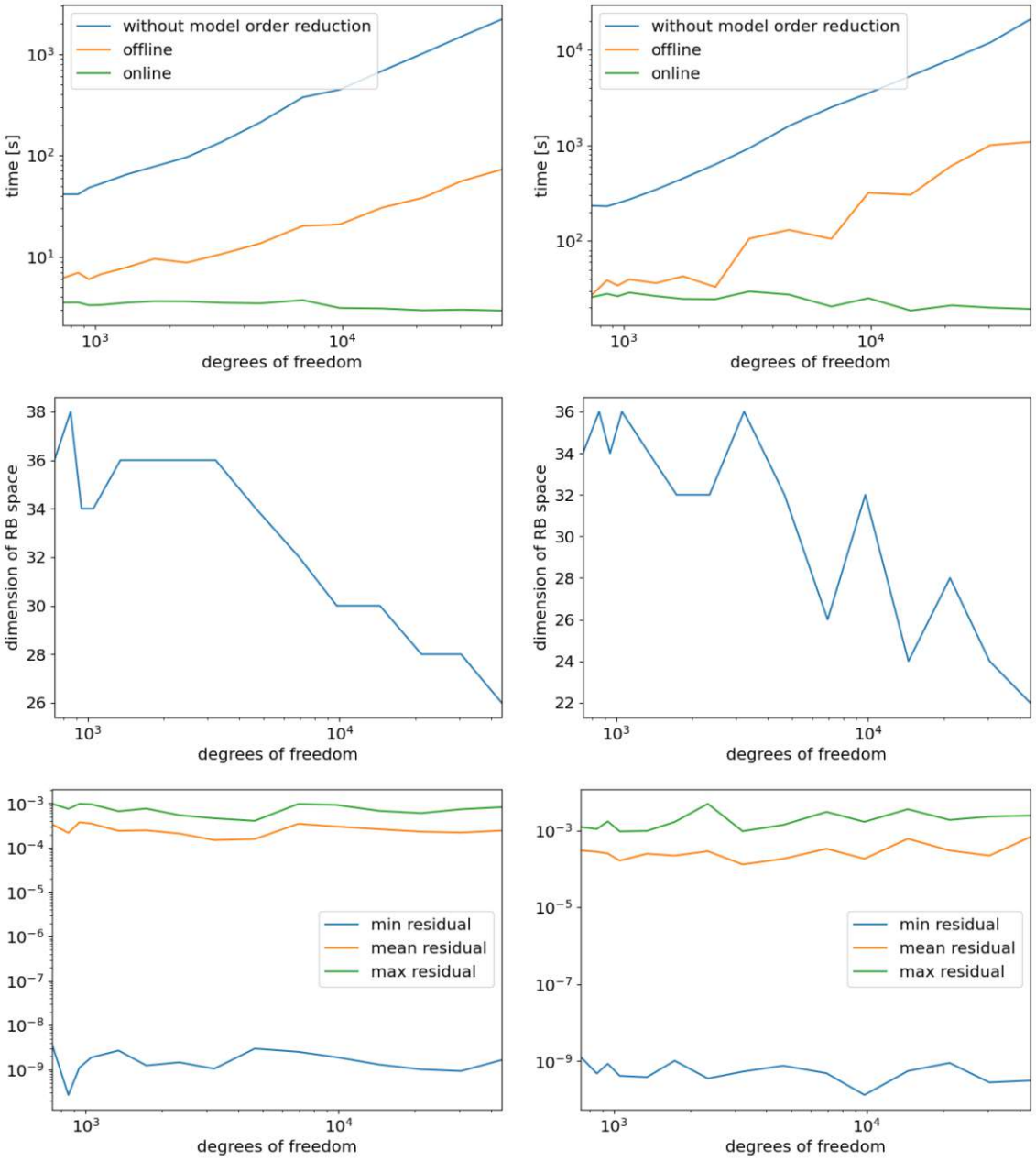


Figure 5.4: Computation times, dimension of RB space and residual for calculation of the same band structure as displayed in Figure 5.3 with `nparam` $= 500$, `nval` $= 50$ and `th_res` $= 1e-3$. The parameters were chosen equidistantly. The figures on the left side show the results for solving a GEP, on the right side the results for solving a QEP are displayed.

The band structure for I concurs with the benchmarks in [21], [24] and [20].

5.2.2 Frequency Dependent Material Parameters

We saw that calculating PC modes by solving GEPs has many benefits over solving QEPs. However for frequency dependent material parameters this can not be done with the methods described in Section 3.2. Hence we resort to employing the methods of Section 3.3.

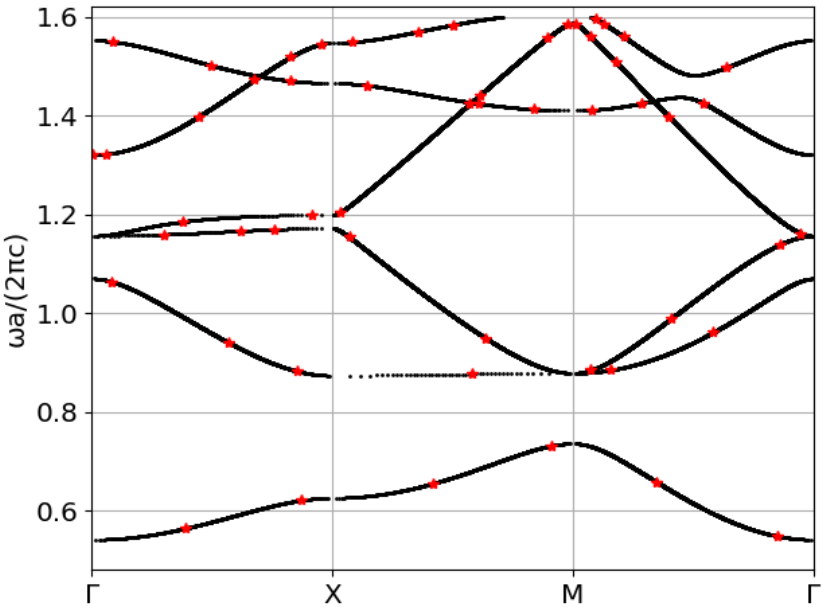


Figure 5.5: Band structure of II. The RB space is built for parameters `ndof` = 909 (`maxh` = 0.1, `order` = 3), `nparam` = 10000, `th_res` = 10^{-3} and `nval` = 100. The red stars indicate the snapshot parameters. The greedy algorithm terminated after 464.816 seconds and produced a RB space of dimension 102. The parameters were chosen as equidistantly between $\omega_{\min} = 0.5$ and $\omega_{\max} = 1.6$. The online computation time amounted to 3202.68 seconds.

To test the correctness of our algorithm for dispersive materials we calculate the band structure of II and compare it to Figure 2a in [23]. Our results are plotted in Figure 5.5 and concur with the reference data. Furthermore we see that every band, except the one belonging to the ground state, intersects at least one other band. Also, we can observe that there are 6 bands in the chosen frequency range, hence a relatively large RB space is necessary to get values of the desired accuracy.

In addition to calculating the band structure of III, which can be seen in Figure 5.7, we investigate how the permeability tensor changes dependent on the scaled frequency. From (2.2) we see that $\boldsymbol{\mu}$ has a singularity at $\omega = \omega_0/200$, which happens at $\omega a/2\pi c \approx 0.075$. In Figure 5.6 we plot the entries of 2D permeability tensor over the range of frequencies we are interested in. Note that in the displayed range criterion (2.4) is still met.

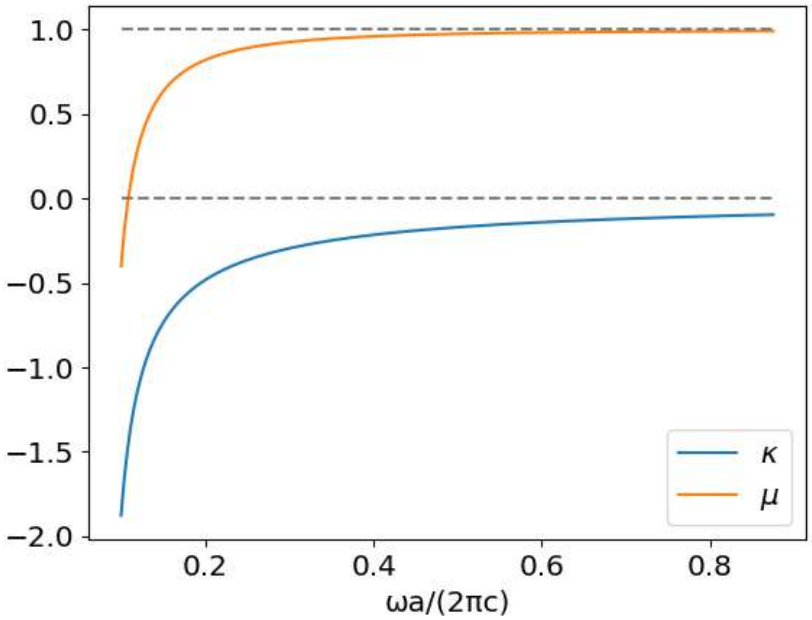


Figure 5.6: Parameters of the permeability tensor (2.2) for III.

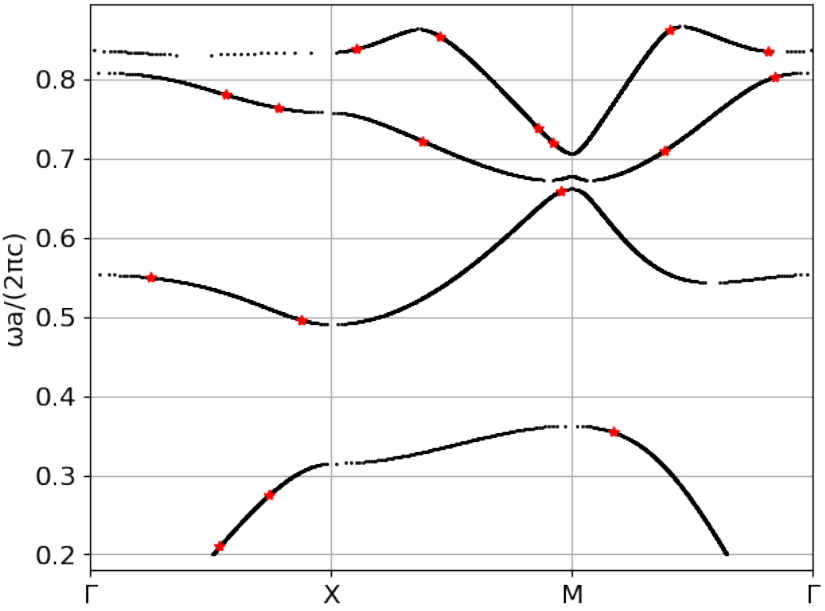


Figure 5.7: Band structure of III. The RB space is built for parameters `ndof` = 1296 (`maxh` = 0.1, `order` = 3), `nparam` = 2000, `th_res` = 10^{-3} and `nval` = 100. The red stars indicate the snapshot parameters. The greedy algorithm terminated after 39.44 seconds and produced a RB space of dimension 34. The parameters were chosen as `np.concatenate([np.linspace(0.2, 0.38, 500), np.linspace(0.49, 0.875, 1500)])`. The online computation time amounted to 157.30 seconds.

Finally we look at the band structure for IV. The results can be seen in Figure 5.8. The very flat bands require a fine resolution of parameters to avoid an incomplete band structure. Also we see that modulating the permittivity yields a visibly different behaviour. As we will discover later the band structures of I and IV are also topologically distinct, meaning they do not have the same Chern numbers.

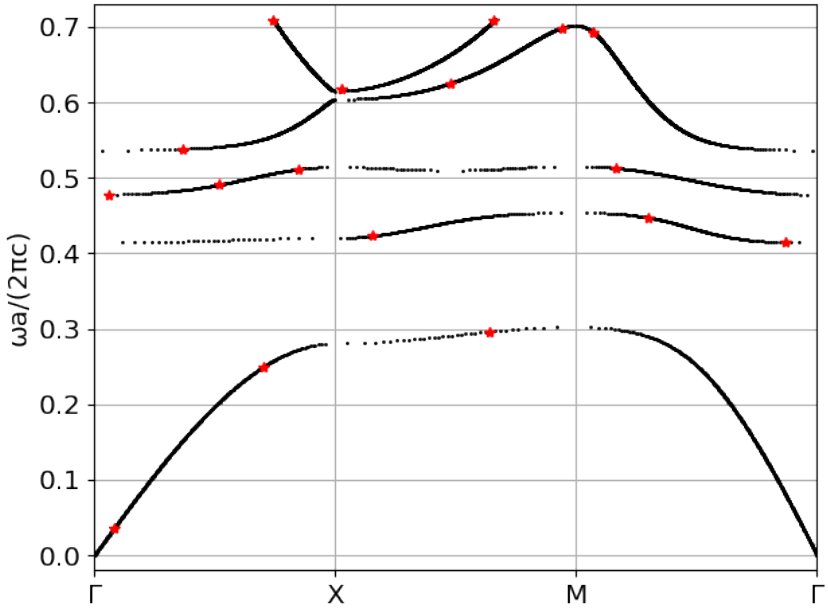


Figure 5.8: Band structure of IV. The RB space is built for parameters `ndof` = 1296 (`maxh` = 0.1, `order` = 3), `nparam` = 2000, `th_res` = 10^{-3} and `nval` = 100. The red stars indicate the snapshot parameters. The greedy algorithm terminated after 49.169 seconds and produced a RB space of dimension 36. The parameters were chosen as `np.concatenate([np.linspace(1e-4, 0.32, 500), np.linspace(0.4, 0.71, 1500)])`. The online computation time amounted to 138.56 seconds.

5.3 Chern Numbers

In this section we investigate the employment of the FPC and the WLA for the calculation of Chern numbers for problems with frequency dependent and frequency independent material parameters. In contrast to the band structure calculation no residuals are computed in the offline phase of the computation of PC modes to ensure time efficiency.

In summary it can be concluded that the RB model order reduction in combination with FEM is very suitable for the calculation of Chern numbers. In general the WLA will be the superior method over the FPC, especially for PCs with frequency dependent material parameters. One exception might be if explicit values for the Berry curvature, as visualized in Figure 5.14, are desired. Furthermore it can be said that the choice of parameters and fine enough discretization for k_y are crucial to get accurate and fast results.

5.3.1 Frequency Independent Material Parameters

From the calculation of band structures we already know that solving a GEP should be preferred over solving a QEP. In case of frequency independent material parameters this is possible. Hence for this section we calculate all modes as described in Section 3.2.

For the calculation of CNs for I we want to compare the first principal calculation (FPC) described in Section 4.1 with the Wilson loop approach (WLA) described in Section 4.2. Therefore we discretize the BZ using a grid of size `ngrid` $\times$ `ngrid`. Note that we only need to calculate data for points either on the right or the left side and on the top or bottom side of the BZ because of the periodic boundary conditions. This is visualized for a 4×4 grid in Figure 5.9. We will calculate the required eigenpairs by solving a GEP. As already established in Section 5.2 this approach needs significantly less computation time than solving a QEP.

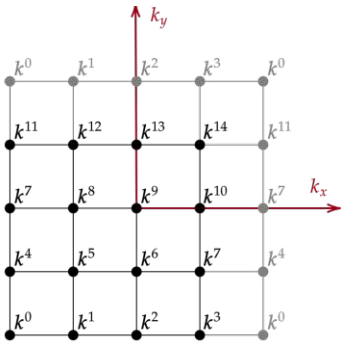


Figure 5.9: Brillouin zone discretized by a 4×4 grid

As can be seen in Table 5.2 the accuracy of the FEM approximation and the residual of the RB method can be chosen quite imprecisely. But even with this small amount of degrees of freedom the model order reduction is beneficial. We ask ourselves if even a smaller RB could be enough to still arrive at accurate Chern numbers. The experiments conducted for Table 5.3 show that this is not the case. Hence we will stick with a threshold residual of 10^{-2} for the construction of the reduced basis from now on.

Furthermore we can observe in Table 5.2 that the WLA takes significantly less time to compute for the same amount of data points. For a grid of size `ngrid` $\times$ `ngrid` the WLA requires the computation of `ngrid` Berry phases, while `ngrid`^{2} Berry phases are needed for the FPC. However a finer discretization of the BZ is necessary to get accurate results for

ngrid	dim RB space	time [s] values	time [s] CNs	method	CN band 1	CN band 2	CN band 3	CN band 4
2	None	0.348	0.037	FPC	0	0	0	0
2	25	1.045	0.039	FPC	0	0	0	0
2	None	0.348	0.011	WLA	0	0	0	0
2	25	1.045	0.012	WLA	0	0	0	0
3	None	0.442	0.087	FPC	0	1	-2	-1
3	27	0.179	0.087	FPC	0	1	-2	-1
3	None	0.442	0.022	WLA	0	1	0	-1
3	27	0.179	0.017	WLA	0	1	0	-1
4	None	0.719	0.150	FPC	0	1	-2	-1
4	29	0.272	0.124	FPC	0	1	-2	-1
4	None	0.719	0.031	WLA	0	1	0	-1
4	29	0.272	0.033	WLA	0	1	0	-1
5	None	1.124	0.178	FPC	0	1	-2	-1
5	31	0.459	0.175	FPC	0	1	-2	-1
5	None	1.124	0.053	WLA	0	1	0	-1
5	31	0.459	0.048	WLA	0	1	0	-1
6	None	1.641	0.253	FPC	0	1	-2	-1
6	31	0.613	0.252	FPC	0	1	-2	-1
6	None	1.641	0.071	WLA	0	1	-1	-1
6	31	0.613	0.077	WLA	0	1	-1	-1
7	None	2.741	0.317	FPC	0	1	-2	-1
7	33	0.837	0.341	FPC	0	1	-2	-1
7	None	2.741	0.109	WLA	0	1	-2	-1
7	33	0.837	0.108	WLA	0	1	-2	-1
8	None	3.730	0.531	FPC	0	1	-2	-1
8	33	0.582	0.151	FPC	0	1	-2	-1
8	None	3.730	0.470	WLA	0	1	-2	-1
8	33	0.582	0.122	WLA	0	1	-2	-1

Table 5.2: Chern numbers for 4 bands calculated by solving a GEP problem. If the dimension of the RB space is `None` the calculation is done without using a model order reduction. The calculation time of the values is the sum of the online and offline phase, meaning it includes building the RB space, if a model order reduction is applied. The desired accuracy `th_res` for Algorithm 8 is 10^{-2}. All parameters are used for construction of the RB space (`nval` = `ngrid` × `ngrid`). The FEM space has 328 degrees of freedom (`maxh` = 0.2, `order` = 2). The coloured rows indicate the optimal results for both methods.

ngrid	dim RB space	time [s] values	time [s] CNs	method	CN band 1	CN band 2	CN band 3	CN band 4
2	13	0.841	0.036	FPC	0	0	0	0
2	13	0.841	0.008	WLA	0	0	0	0
3	13	0.062	0.059	FPC	0	1	-2	0
3	13	0.062	0.014	WLA	0	1	0	0
4	13	0.119	0.101	FPC	0	1	-2	0
4	13	0.119	0.024	WLA	0	1	0	0
5	13	0.160	0.153	FPC	0	1	-2	0
5	13	0.160	0.037	WLA	0	1	0	0
6	13	0.216	0.218	FPC	0	1	-2	0
6	13	0.216	0.053	WLA	0	1	-1	0
7	13	0.277	0.297	FPC	0	1	-2	0
7	13	0.277	0.072	WLA	0	1	-1	0
8	13	0.363	0.384	FPC	0	1	-2	0
8	13	0.363	0.096	WLA	0	1	-2	0

Table 5.3: Chern numbers for 4 bands calculated by solving a GEP problem. The calculation time of the values is the sum of the online and offline phase, meaning it includes building the RB space. The desired accuracy `th_res` for Algorithm 8 is 10^{-1}. All parameters are used for construction of the RB space (`nval` = `ngrid` × `ngrid`). The FEM space has 328 degrees of freedom (`maxh` = 0.2, `order` = 2).

the WLA, so in sum it takes more time to compute. If we look at the derivation in Section 4.2 we can guess that maybe less data points are required to calculate the Berry phase (up to an integer multiple of 2π) of a fixed k_y than we are using when considering `ngrid`×`ngrid` values. We can test this by discretizing the Brillouin zone with a `ngrid_x` × `ngrid_y` grid, where `ngrid_x` ≤ `ngrid_y`. Indeed this yields correct results for `ngrid_x` ≥ 3. As can be seen by comparing the optimal results for the FPC in Table 5.2 to the optimal results in Table 5.4, this makes the WLA faster than the FPC, at least for our example.

If large CNs can be expected, the WLA might be inferior to a FPC. To make that plausible we can observe that the BZ is nothing else than a torus because of its periodic boundary conditions. Now we can regard the CN as a winding number around the torus as it is done in [20]. For large CNs it is nessecary to have a high `ngrid_y` so not to "miss" a winding around the torus. This is even better visible if we just plot the Berry phases for each fixed k_y in the flat BZ as can be seen in Figures 5.10, 5.11, 5.12 and 5.13.

Another advantage of the FPC is that an approximation of the Berry curvature is calculated for each band. The curvature over the first BZ, discretized by a 100 × 100 grid, is plotted in Figure 5.14. Remember that we concluded in Section 2.2 that non trivial chern numbers can only arise if the Berry potential is not continuously differentiable, which would lead to singularities in the Berry curvature. This is exactly what we can observe in Figures 5.14b, 5.14c and 5.14d.

ngrid_x	dim RB space	time [s] values	time [s] CNs	CN band 1	CN band 2	CN band 3	CN band 4
2	None	0.938	0.026	0	0	0	0
2	28	2.266	0.027	0	0	0	0
3	None	1.066	0.051	0	1	-2	-1
3	28	0.149	0.042	0	1	-2	-1
4	None	1.352	0.054	0	1	-2	-1
4	30	0.491	0.052	0	1	-2	-1

Table 5.4: Chern numbers for 4 bands calculated by solving a GEP problem. The parameters are the same as in Table 5.2 except that only the WLA is applied and the Brillouin zone is discretized by a ngrid_x × ngrid_y grid with ngrid_y = 7.

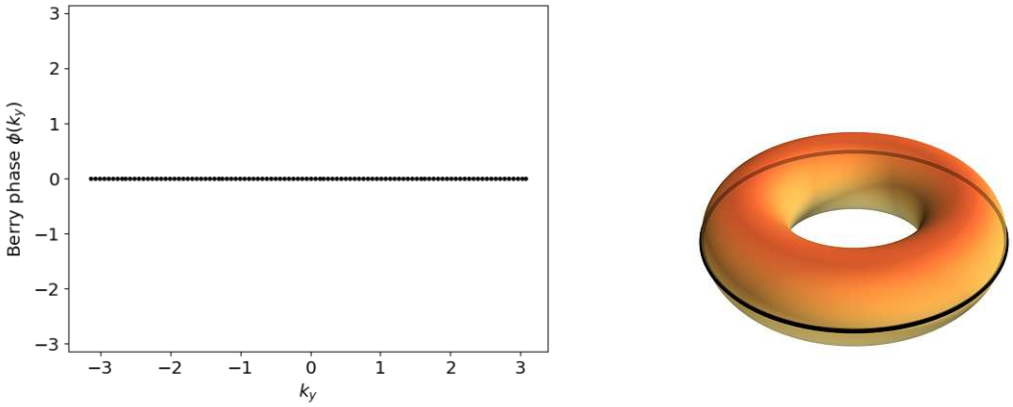


Figure 5.10: Band 1 of I with CN 0.

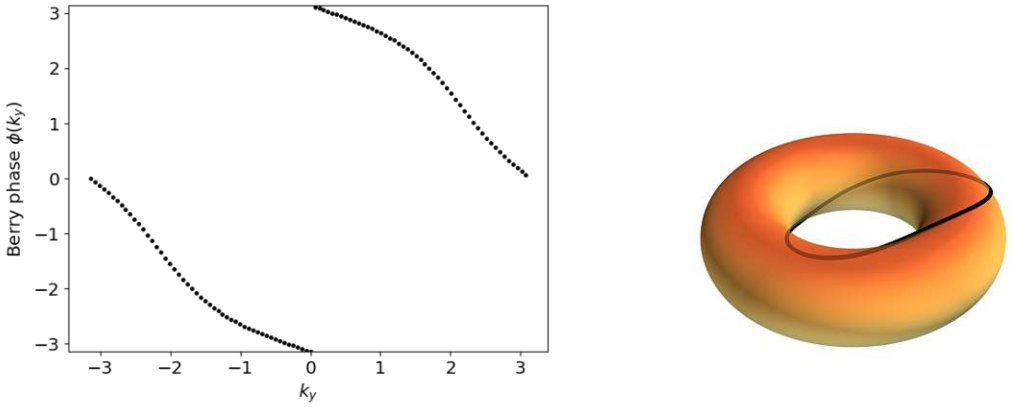


Figure 5.11: Band 2 of I with CN 1.

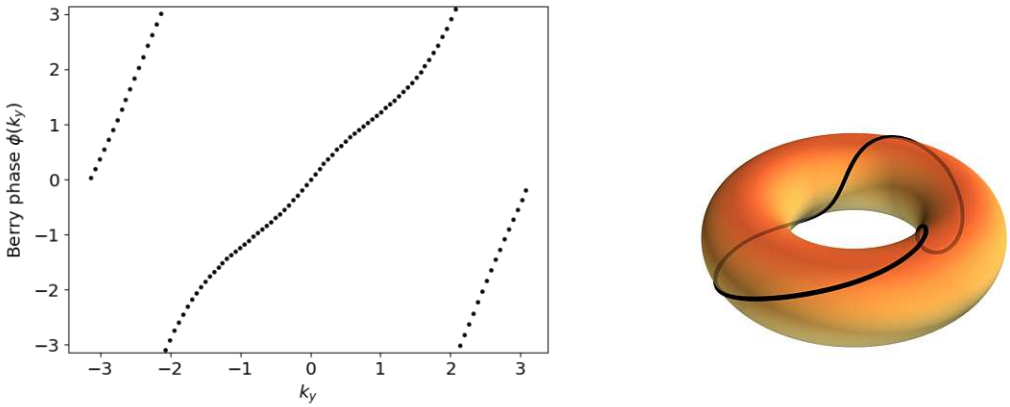


Figure 5.12: Band 3 of I with CN -2.

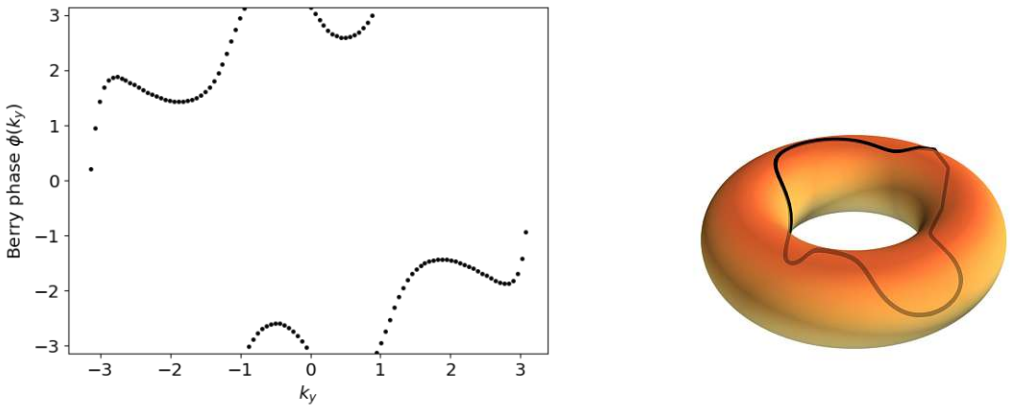
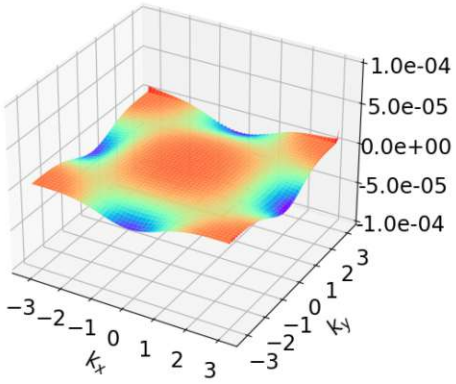
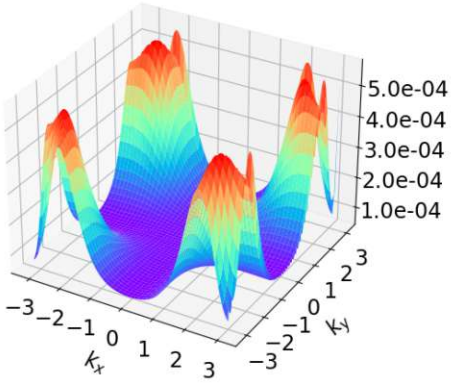


Figure 5.13: Band 4 of I with CN -1.

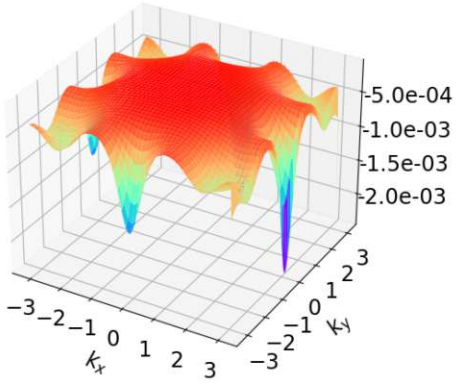
The results for the CNs of the first 3 bands concur with the results in [21], [20] and [24]. The CN for the 4th band is only calculated in [21]. However we get a CN of -1 while the value in [21] is given as 1. In [6] the same benchmark problem is considered. Their results are the same as ours.



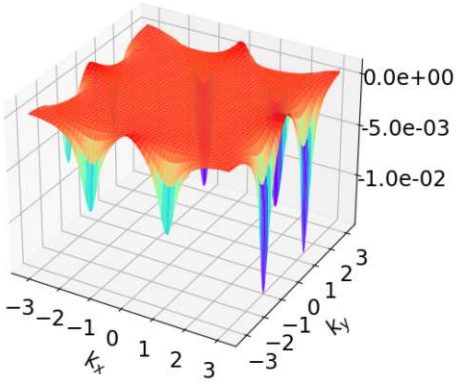
(a) Band 1 of I with CN 0.



(b) Band 2 of I with CN 1.



(c) Band 3 of I with CN -2.



(d) Band 4 of I with CN -1.

Figure 5.14: Approximate Berry curvature of the first 4 bands of I. The Brillouin zone is discretized by a 100×100 grid. The calculation was done solving a GEP.

5.3.2 Frequency Dependent Material Parameters

In contrast to problems with frequency independent material parameters, we are going to solve a QEP in this section. To get comparable results we first solve I as a QEP and subsequently calculate the CNs for III and IV. It does not make sense to consider CNs for II because no bands above the ground state are separated.

Not regarding (2.14) as a GEP brings a few drawbacks. As discussed in Section 5.2 the problem has to be solved for a lot more parameters ω to get a band structure with enough

data points. Another drawback of solving a QEP is that the bands have to be separated after calculating all the eigenpairs. This is done as described in Algorithm 7.

From the experiments documented in Table 5.4 we already know that `ngrid_y` $= 7$ is sufficient to get correct CNs of I for some value `ngrid_x`. Hence for the WLA we consider `ngrid_y` $= 7$. When solving a QEP we want to know how many parameters ω are needed for each path

$$\Gamma_{k_y} : (0,1) \to \mathrm{BZ}, \quad s \mapsto \begin{pmatrix} -\frac{\pi}{a} + s\frac{2\pi}{a} \\ k_y \end{pmatrix} \tag{5.1}$$

with

$$k_y = -\frac{\pi}{a} + m\frac{2\pi}{\texttt{ngrid\_y}\ a} \quad \text{for } m \in \{0, \ldots, \texttt{ngrid\_y} - 1\}.$$

If we prescribe equidistant parameters ω not only the quantity of the parameters matters, but also if the chosen frequencies lie on the relatively flat bands. This can be observed in Table 5.5. Different CNs are calculated correctly for a different amount of equidistant parameters. Fortunately the WLA allows a visual interpretation of the calculated CN as the winding number around the torus. This means we have an easy way of checking the plausibility of our results.

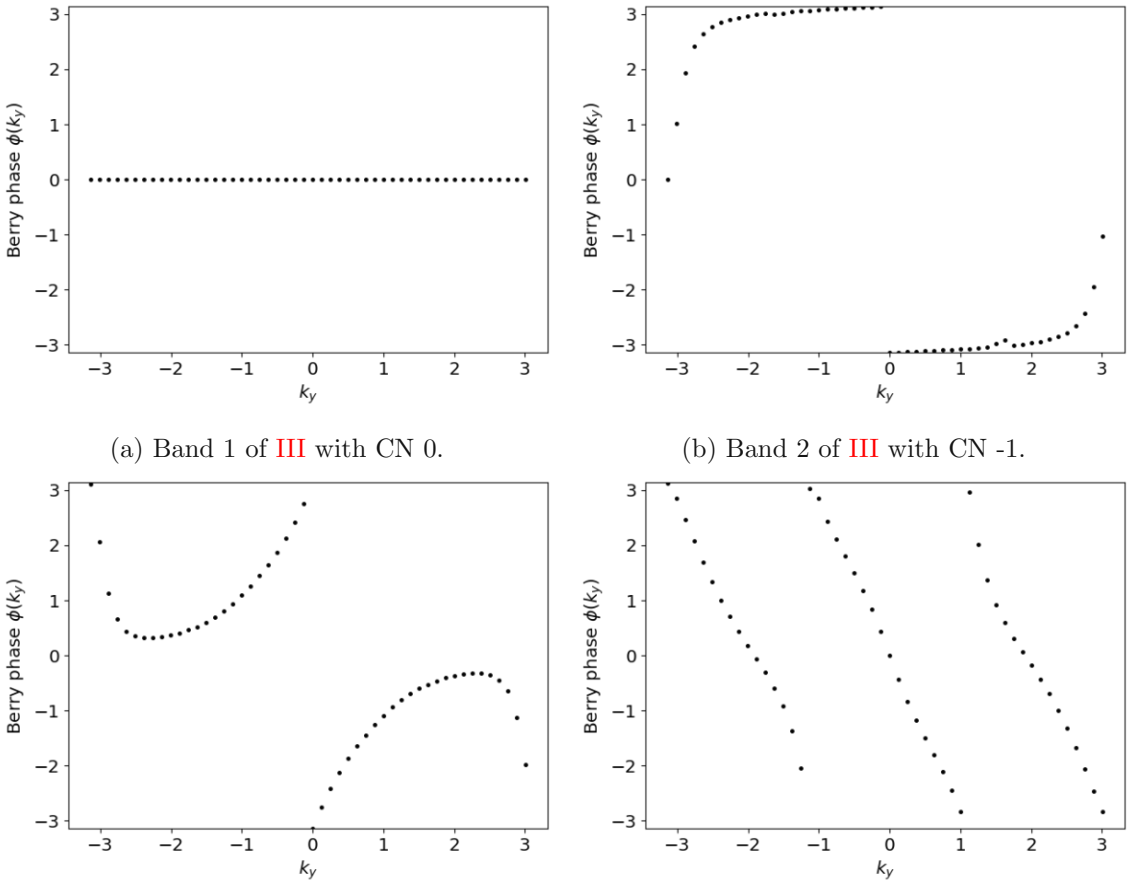
nparam	dim RB space	time [s] values	time [s] CNs	CN band 1	CN band 2	CN band 3	CN band 4
100	34	22.318	0.225	0	0	0	-1
110	34	25.096	0.223	0	1	0	0
120	34	26.722	0.248	0	0	-2	0
130	34	29.759	0.276	0	1	-2	0
140	34	31.369	0.308	0	1	0	-1
150	34	34.287	0.303	0	1	-2	-1
160	34	33.857	0.324	0	1	-2	0
170	34	39.074	0.350	0	1	-2	-1
180	34	41.810	0.747	0	1	-2	-1
190	34	40.268	0.385	0	1	-2	-1
200	36	49.988	0.401	0	1	-2	-1
210	36	48.660	0.413	0	1	-2	-1
220	36	48.329	0.427	0	1	-2	-1
230	36	50.297	0.468	0	1	-2	-1
240	36	57.411	0.480	0	1	-2	-1
250	36	64.245	0.556	0	1	-2	-1

Table 5.5: Chern numbers for 4 bands for I calculated by solving a QEP problem and using the WLA for `ngrid_y` $= 7$. The calculation time of the values is the sum of the online and offline phase and the sorting of values into their bands by application of Algorithm 7. The desired accuracy `th_res` for Algorithm 8 is 10^{-2}. The parameters are chosen equidistantly between $\omega_{\min} = 10^{-4}$ and $\omega_{\max} = 0.71$. All parameters are used for construction of the RB space (`nval` $=$ `nparam`). The FEM space has 328 degrees of freedom (`maxh` $= 0.2$, `order` $= 2$).

This fact is especially beneficial if we have no prior knowledge of the correct CNs. Hence, before investigating their convergence for III and IV we calculate the Berry phases for a lot of paths (3.4). The results can be seen in Figures 5.15 and 5.16. The calculated CNs concur with what we would expect by looking at the plots of the Berry phases around the respective paths Γ_{k_y}.

nparam	dim RB space	time [s] values	time [s] CNs	CN band 1	CN band 2	CN band 3	CN band 4
100	16	14.623	0.301	0	0	-2	1
110	16	13.682	0.193	0	1	-2	0
120	16	13.123	0.258	0	0	-1	0
130	16	13.978	0.238	0	1	-2	0
140	16	14.604	0.242	0	1	-2	1
150	16	15.076	0.251	0	1	-2	-1
160	16	17.101	0.435	0	1	-2	0
170	16	18.604	0.291	0	1	-2	-1
180	16	18.010	0.678	0	1	-2	-1
190	16	22.302	0.463	0	1	-2	0
200	16	23.251	0.399	0	1	-2	0
210	16	23.879	0.440	0	1	-2	-1
220	16	23.471	0.490	0	1	-2	-1
230	16	23.097	0.445	0	1	-2	-1
240	16	26.687	0.626	0	1	-2	-1
250	16	26.103	0.477	0	1	-2	-1

Table 5.6: Chern numbers for 4 bands for I calculated by solving a QEP problem and using the WLA for `ngrid_y` $= 7$. The calculation time of the values is the sum of the online and offline phase and the sorting of values into their bands by application of Algorithm 7. The desired accuracy `th_res` for Algorithm 8 is 10^{-1}. The parameters are chosen equidistantly between $\omega_{\min} = 10^{-4}$ and $\omega_{\max} = 0.71$. All parameters are used for construction of the RB space (`nval` $=$ `nparam`). The FEM space has 328 degrees of freedom (`maxh` $= 0.2$, `order` $= 2$).



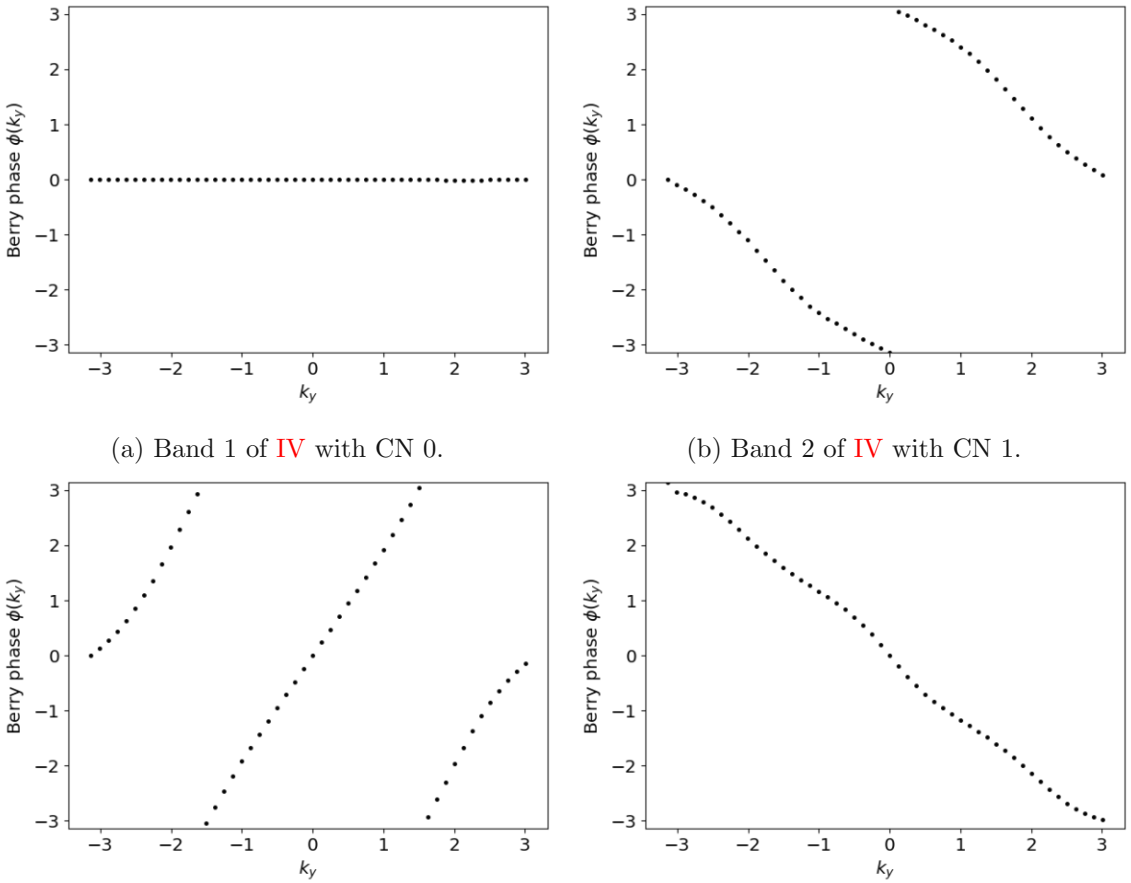
(a) Band 1 of III with CN 0.

(b) Band 2 of III with CN -1.

(c) Band 3 of III with CN 0.

(d) Band 4 of III with CN 3.

Figure 5.15: Chern number of the first 4 bands of III. The calculation was done solving a QEP and using the WLA. 50 equidistant values $\boldsymbol{k}_y$ where chosen.



(a) Band 1 of IV with CN 0.

(b) Band 2 of IV with CN 1.

(c) Band 3 of IV with CN -2.

(d) Band 4 of IV with CN 1.

Figure 5.16: Chern number of the first 4 bands of IV. The calculation was done solving a QEP and using the WLA. 50 equidistant values $\boldsymbol{k}_y$ where chosen.

ngrid	CN band 1	CN band 2	CN band 3	CN band 4
4	0	-1	0	-1
5	0	-1	0	-2
6	0	-1	0	1
7	0	-1	0	1
8	0	-1	0	1
9	0	-1	0	3
10	0	-1	0	3

Table 5.7: Chern numbers for 4 bands of III calculated by solving a QEP problem. The desired accuracy `th_res` for Algorithm 8 is 10^{-2}. The RB space was constructed for `nparam` $= 2000$ values and `nval` $= 400$. The FEM space has 328 degrees of freedom (`maxh` $= 0.2$, `order` $= 2$).

Before we investigate how big we have to choose `nparam`, we first determine the necessary amount of Berry phases to calculate. We have already established that the accuracy of the FEM and the RB space play minor roles so we choose the same residual thereshold and number of degrees of freedom for the FEM as in the frequency independent case. The results are written down in Tables 5.7 and 5.8.

ngrid	CN band 1	CN band 2	CN band 3	CN band 4
4	0	1	0	1
5	0	1	-2	1
6	0	1	-2	1

Table 5.8: Chern numbers for 4 bands of IV calculated by solving a QEP problem. The desired accuracy `th_res` for Algorithm 8 is 10^{-2}. The RB space was constructed for `nparam` $= 2000$ values and `nval` $= 400$. The FEM space has 328 degrees of freedom (`maxh` $= 0.2$, `order` $= 2$).

High values for `nparam` and `nval` are prescribed so it is very unlikely to miss crucial points of some bands. Additionally to that we can visually check the results for plausibility. However, the plots associated with the experiments in Tables 5.7 and 5.8 are not shown here.

nparam	dim RB space	time [s] values	time [s] CNs	CN band 1	CN band 2	CN band 3	CN band 4
100	28	33.210	0.593	0	-1	0	-
110	28	32.518	0.435	0	-1	0	1
120	30	37.576	0.467	0	-1	0	3
130	30	38.653	0.509	0	-1	0	1
140	30	43.134	0.544	0	-1	0	1
150	30	44.275	0.677	0	-1	0	1
160	30	49.003	0.619	0	-1	0	1
170	30	50.290	0.665	0	-1	0	3
180	30	57.883	0.711	0	-1	0	3
190	30	60.884	0.762	0	-1	0	1
200	30	62.520	0.837	0	-1	0	1
210	30	67.069	0.994	0	-1	0	1
220	30	67.446	0.856	0	-1	0	3
230	30	72.661	0.896	0	-1	0	3
240	30	72.754	0.941	0	-1	0	3
250	32	79.964	1.014	0	-1	0	1
260	32	85.736	1.011	0	-1	0	1
270	32	90.257	1.188	0	-1	0	3
280	32	97.101	1.339	0	-1	0	3
290	32	97.880	1.241	0	-1	0	3
300	32	98.531	1.186	0	-1	0	3

Table 5.9: Chern numbers for 4 bands of III calculated by solving a QEP problem and using the WLA for `ngrid_y` $= 9$. The calculation time of the values is the sum of the online and offline phase and the sorting of values into their bands by application of Algorithm 7. The desired accuracy `th_res` for Algorithm 8 is 10^{-2}. The parameters are chosen equidistantly between $\omega_{\min} = 0.2$ and $\omega_{\max} = 0.875$. All parameters are used for construction of the RB space (`nval` $=$ `nparam`). The FEM space has 328 degrees of freedom (`maxh` $= 0.2$, `order` $= 2$). No results for CNs indicate that the respective bands were not found.

Subsequently we calculate the CNs for the minimum amount of k_y and a varying quantity `nparam` of equidistant parameters per path Γ_k. To get deterministic results we choose `nval` $=$ `nparam`. As we can see in Tables 5.9 and 5.10 significantly more parameters are needed for IV than for III in order to get reliable results for the first 3 bands. The reason is that these bands of IV are very flat, which can also be observed in Figure 5.8.

nparam	dim RB space	time [s] values	time [s] CNs	CN band 1	CN band 2	CN band 3	CN band 4
300	48	217.641	0.656	0	1	0	1
330	62	156.188	0.645	0	1	0	1
360	62	139.347	0.713	0	1	0	1
390	62	156.276	0.899	0	1	0	1
420	62	172.221	0.939	0	1	0	1
450	62	182.655	0.945	0	1	0	1
460	44	180.054	0.957	0	1	0	1
470	40	107.647	0.965	0	1	0	1
480	62	195.174	1.025	0	1	-2	1
490	40	111.912	0.958	0	1	-2	1
500	40	114.410	0.974	0	1	-2	1
510	62	208.391	1.027	0	1	-2	1
520	40	121.582	1.078	0	1	-2	1
530	40	132.643	1.083	0	1	-2	1
540	62	224.107	1.150	0	1	-2	1
550	40	129.392	1.118	0	1	-2	1
560	40	130.671	1.088	0	1	-2	1
570	62	235.080	1.206	0	1	-2	1
600	62	252.183	1.209	0	1	-2	1

Table 5.10: Chern numbers for 4 bands of IV calculated by solving a QEP problem and using the WLA for `ngrid_y` $= 5$. The calculation time of the values is the sum of the online and offline phase and the sorting of values into their bands by application of Algorithm 7. The desired accuracy `th_res` for Algorithm 8 is 10^{-2}. The parameters are chosen equidistantly between $\omega_{\min} = 10^{-4}$ and $\omega_{\max} = 0.71$. All parameters are used for construction of the RB space (`nval` $=$ `nparam`). The FEM space has 328 degrees of freedom (`maxh` $= 0.2$, `order` $= 2$).

In remains to investigate the FPC in combination with solving a QEP. To employ that method we need specific wave vectors $\boldsymbol{k}$, hence a very fine sample of frequencies ω is required. The computational effort can be mitigated by estimating at which frequencies ω we can expect to find the desired wave vectors $\boldsymbol{k}$. So for a grid of size `ngrid_x` $\times$ `ngrid_y` it makes sense to first look at the band structure along (5.1).

nparam	dim RB space	time [s] values	time [s] CNs	CN band 1	CN band 2	CN band 3
500	30	-	-	-	-	-
600	30	63.042	0.086	0	1	-2
700	30	65.885	0.093	0	1	-2
800	30	-	-	-	-	-
900	30	83.703	0.087	0	1	-2
1000	30	92.861	0.085	0	1	-2
1100	30	102.098	0.087	0	1	-2
1200	30	111.150	0.084	0	1	-2
1300	30	120.615	0.086	0	1	-2
1400	30	129.595	0.084	0	1	-2
1500	30	138.212	0.087	0	1	-2
1600	30	-	-	-	-	-
1700	30	-	-	-	-	-
1800	30	-	-	-	-	-
1900	30	-	-	-	-	-
2000	30	-	-	-	-	-
2100	30	-	-	-	-	-
2200	30	-	-	-	-	-
2300	30	-	-	-	-	-
2400	30	-	-	-	-	-
2500	30	-	-	-	-	-
2600	30	-	-	-	-	-
2700	30	245.171	0.084	0	1	-2
2800	30	254.713	0.082	0	1	-2
2900	30	263.619	0.082	0	1	-2
3000	30	273.225	0.085	0	1	-2
3100	30	281.824	0.088	0	1	-2
3200	30	291.175	0.092	0	1	-2
3300	30	300.190	0.083	0	1	-2
3400	30	309.017	0.086	0	1	-2
3500	30	318.336	0.087	0	1	-2

Table 5.11: Chern numbers for 3 bands calculated by solving a QEP problem and using the FPC with equidistant parameters ω between $\omega_{\min} = 1e-4$ and $\omega_{\max} = 0.625$. The calculation time of the values is the sum of the online and offline phase and the sorting of values into their bands. The desired accuracy `th_res` for Algorithm 8 is 10^{-2}. All parameters are used for construction of the RB space (`nval` = `nparam`). The FEM space has 328 degrees of freedom (`maxh` = 0.2, `order` = 2). The Brillouin zone is discretized by a 4×4 grid. The accuracy threshold δ_{th} in Algorithm 6 is chosen as $2\pi/(3.5\,$`ngrid`$)$. If there are no entries for the times and CNs when Algorithm 6 was not able to find the required amount of frequencies for each wave vector $\boldsymbol{k}$ on the grid.

In a second step we only have to sample frequencies around certain $\tilde{\omega}$. However this ap-

proach requires some manual effort or prior knowledge about the band structures and is still relatively time consuming. It gets even more messy if there is no interval $[\omega_1, \omega_2]$ for every band such that no other band is in that range, as it is the case in Figure 5.3. For that reason we only calculate Chern numbers of the first 3 bands of I employing the FPC. Table 5.11 shows the results and computation times for different amounts of parameters ω if no initial guess is used. As can be seen in Table 5.12 even with very good initial guesses relatively many parameters ω need to be considered to get the desired result. For deterministic behaviour in both cases all parameters are used when building the RB space.

`nparam` δ	`nparam`	dim RB space	time [s] values	time [s] CNs	CN band 1	CN band 2	CN band 3
15	180	30	-	-	-	-	-
17	204	30	-	-	-	-	-
19	228	30	-	-	-	-	-
21	252	30	41.720	0.092	0	1	-2
23	276	30	41.242	0.099	0	1	-2
25	300	30	44.669	0.093	0	1	-2
27	324	30	49.287	0.097	0	1	-2
29	348	30	52.129	0.097	0	1	-2
31	372	30	55.512	0.093	0	1	-2
33	396	30	59.139	0.091	0	1	-2
35	420	30	62.785	0.092	0	1	-2
37	444	30	66.664	0.095	0	1	-2
39	468	30	70.073	0.089	0	1	-2
41	492	30	72.139	0.090	0	1	-2
43	516	30	76.210	0.099	0	1	-2

Table 5.12: Calculation for the same parameters as in Table 5.11, but with parameters ω chosen around initial guesses $\tilde{\omega}$ (calculated by solving a GEP). The variable `nparam` δ indicates the amount of parameters per δ–environment around each $\tilde{\omega}$, where $\delta = 0.005$.

Acronyms

BZ Brillouin zone. 6–8, 16, 21, 22, 25, 38, 47, 48, 52, 59, 61, 65

CN Chern number. 59–74

FEM finite element method. 15, 51, 52, 59–61, 66, 67, 70–73

FPC first principal calculation. 51, 59–61, 72–74

GEP general eigenvalue problem. i, 15, 25, 29, 44, 52–56, 59, 64, 74

LOBPCG locally optimal block preconditioned conjugate gradient. i, 26–28, 53

ONB orthonormal basis. 22, 25, 31–33, 35, 36, 38

PC photonic crystal. 3, 5–7, 13, 52, 56, 59

QEP quadratic eigenvalue problem. i, 15, 25, 28, 29, 31, 44, 52–56, 59, 64, 65, 68, 69, 72

RB reduced basis. 41–44, 51, 53–62, 66, 67, 70–74

SOAR second-order Arnoldi. 31, 32, 34, 38

TE transversal-electric. 7

TM transversal-magnetic. 7, 8

TOAR two-level orthogonal Arnoldi. 31, 34, 38, 53

WLA Wilson loop approach. 51, 59–62, 65–69, 71, 72

YIG Yittrium-Iron-Garnet. 3, 51, 52

Bibliography

[1] F. N. Alfio Quarteroni, Andrea Manzoni. *Reduced Basis Methods for Partial Differential Equations*. Springer Cham, 1st ed. edition, 2015.

[2] I. Babuška and J. Osborn. Eigenvalue problems. In *Finite Element Methods (Part 1)*, volume 2 of *Handbook of Numerical Analysis*, pages 641–787. Elsevier, 1991.

[3] Z. Bai and Y. Su. Soar: A second-order arnoldi method for the solution of the quadratic eigenvalue problem. *SIAM J. Matrix Analysis Applications*, 26:640–659, 01 2005.

[4] M. V. Berry. Quantal phase factors accompanying adiabatic changes. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 392(1802):45–57, 1984.

[5] D. Boffi. Finite element approximation of eigenvalue problems. *Acta Numerica*, 19:1–120, 2010.

[6] K. Goudarzi, H. G. Maragheh, and M. Lee. Calculation of the berry curvature and chern number of topological photonic crystals. *Journal of the Korean Physical Society*, 81(5):386–390, Sep 2022.

[7] F. Haldane and S. Raghu. Possible realization of directional optical waveguides in photonic crystals with broken time-reversal symmetry. *Physical review letters*, 100:013904, 02 2008.

[8] J. D. Joannopoulos, S. G. Johnson, J. N. Winn, and R. D. Meade. *Photonic Crystals: Molding the Flow of Light (Second Edition)*. Princeton University Press, 2 edition, 2008.

[9] A. Knyazev. A preconditioned conjugate gradient method for eigenvalue problems and its implementation in a subspace. *Proc. Workshop, Oberwolfach/Germ.1990, ISNM 96*, 5, 01 1991.

[10] A. V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM Journal on Scientific Computing*, 23(2):517–541, 2001.

[11] A. V. Knyazev and K. Neymeyr. Efficient solution of symmetric eigenvalue problems using multigrid preconditioners in the locally optimal block conjugate gradient method. *Electron. Trans. Numer. Anal*, 15:38–55, 2003.

[12] D. Lu, Y. Su, and Z. Bai. Stability analysis of the two-level orthogonal arnoldi procedure. *SIAM Journal on Matrix Analysis and Applications*, 37:195–214, 01 2016.

[13] L. Lu, J. D. Joannopoulos, and M. Soljačić. Topological photonics. *Nature Photonics*, 8(11):821–829, oct 2014.

[14] D. M. Pozar. *Microwave engineering; 3rd ed.* Wiley, Hoboken, NJ, 2005.

[15] C. Scheiber, A. Schultschik, O. Bíró, and R. Dyczij-Edlinger. A model order reduction method for efficient band structure calculations of photonic crystals. *IEEE Transactions on Magnetics*, 47(5):1534–1537, 2011.

[16] J. Schöberl. Netgen an advancing front 2d/3d-mesh generator based on abstract rules. *Computing and Visualization in Science*, 1(1):41–52, Jul 1997.

[17] J. Schöberl. C++11 implementation of finite elements in ngsolve. *Institute for analysis and scientific computing, Vienna University of Technology*, 09 2014.

[18] D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. den Nijs. Quantized hall conductance in a two-dimensional periodic potential. *Phys. Rev. Lett.*, 49:405–408, Aug 1982.

[19] D. Vanderbilt. *Berry Phases in Electronic Structure Theory: Electric Polarization, Orbital Magnetization and Topological Insulators*. Cambridge University Press, 2018.

[20] H.-X. Wang, G.-Y. Guo, and J.-H. Jiang. Band topology in classical waves: Wilson-loop approach to topological numbers and fragile topology. *New Journal of Physics*, 21(9):093029, sep 2019.

[21] Z. Wang, Y. D. Chong, J. D. Joannopoulos, and M. Soljačić. Reflection-free one-way edge modes in a gyromagnetic photonic crystal. *Phys. Rev. Lett.*, 100:013905, Jan 2008.

[22] H. Weng, R. Yu, X. Hu, X. Dai, and Z. Fang. Quantum anomalous hall effect and related topological electronic states. *Advances in Physics*, 64(3):227–282, may 2015.

[23] W. Xiao and J. Sun. A novel method for band structure calculation of photoniccrystals with frequency-dependent permittivities. *Journal of the Optical Society of America A*, 38, 03 2021.

[24] R. Zhao, G.-D. Xie, M. L. N. Chen, Z. Lan, Z. Huang, and W. E. I. Sha. First-principle calculation of chern number in gyrotropic photonic crystals. *Optics Express*, 28(4):4638, feb 2020.