Technische Universität Berlin
Institut für Mathematik

On adaptive finite element methods for the simulation of two-dimensional photonic crystals

Marine Froidevaux

Preprint 06-2018

Preprint-Reihe des Instituts für Mathematik
Technische Universität Berlin
http://www.math.tu-berlin.de/preprints

Preprint 06-2018 July 2018

On adaptive finite element methods for the simulation of two-dimensional photonic crystals[*]

MARINE FROIDEVAUX[‡]

Abstract. The first part of this paper is devoted to the modeling of wave propagation inside a perfect two-dimensional photonic crystal and the spectral analysis of the resulting eigenvalue problem. In the second part of the paper, we focus on a special case, where the eigenvalue problem is linear and Hermitian. We introduce a residual-based estimator approximating the algebraic error, i.e., the error in the computed eigenvector induced by iterative methods. The estimator for the algebraic error approximates the error in the same norm as the one used in already existing estimators for the discretization error, therefore enabling error balancing between both types of errors, and thus improving the efficiency of the adaptive finite element procedure.

Key words. Photonic crystals, error estimators, adaptive finite element methods

AMS subject classifications. 35P15, 78M10, 65F15

1. Introduction

Photonic crystals are composite materials, made of periodically arranged components, whose periodic structure affects the propagation of electromagnetic waves. These waves may propagate through or be reflected by the crystal depending on their frequencies ω and on their wave-vector k, but also on parameters such as the electrical permittivity ϵ of the constituent materials, or the geometry of the spatial arrangement of the crystal.

Photonic crystals creating so-called *band gaps* are of special interest for applications in engineering. A band gap denotes a range of frequencies at which no wave can propagate in any direction through the crystal. The design of photonic crystals for industrial applications requires the optimization of many geometrical and material parameters, in order to obtain large bad-gaps around certain frequencies specified by the application. For this reason, we are interested in developing accurate mathematical models and numerical methods which can enable the time-efficient simulation of these physical systems.

The mathematical modeling of a photonic crystal is based on the Maxwell equations. In the case of a perfect infinite crystal, the Maxwell equation lead to a sequence of parameter-dependant PDE *eigenvalue problems* (EVPs), which can be discretized via a conforming finite element method. One can show that eigenpairs of the discretized problem converge to the eigenpairs of the PDE EVP as one increases the number of basis functions in the finite element space, see e.g. [Hac92]. However, having time-efficient computations in mind, the number of basis functions in the finite element space, as well as the spatial localization, should be chosen in a way that minimizes the size of the discretized system, while warranting a certain accuracy. This kind of trade-off problems is typically addressed via adaptive

[‡] Institut für Mathematik, MA 4-5, Technische Universität Berlin, Strasse des 17. Juni 136, 10623 Berlin, Germany

E-mail address: froideva@math.tu-berlin.de.

[*] Research supported by the Einstein Center ECMath via project OT10 *Model Reduction for Nonlinear Parameter-Dependent Eigenvalue Problems in Photonic Crystals*.

finite element methods, and error estimators are the central tool enabling this adaptivity. In [GG12], a reliable and efficient error estimator was developed in order to estimate the error produced by the finite element method. However, for large-scaled problems it is computationally too expensive to solve the discrete EVPs exactly via direct methods. Instead, iterative methods are used, such as e.g. Lanczos' method, in order to solve the large-scale discrete problems up to a prescribed tolerance. In this setting, the error arising from the solution of discrete EVPs via iterative methods, also called the *algebraic* error, cannot be ignored as a contribution to the total numerical error. Moreover, as part of the adaptive process, it may be sufficient to solve the coarsely discrete EVPs with an accuracy of the magnitude of the discretization error. For these reasons, error estimators are not only needed to approximate the discretization error, but also to approximate the algebraic error. Moreover, both types of error estimators should be part of a unified framework, where all errors are measured in the same norm, so that meaningful comparisons can be made as part of error balancing procedures. For the Laplace eigenvalue problem, error balancing between the discretization and the algebraic errors was addressed in [CDM$^+$17] and a fully-adpative algorithm called *AFEMLA* algorithm was introduced in [MM11], and extensively discussed in [Mię10].

The mathematical description and simulation of photonic crystals has already been the subject of many research papers. Discretization via Nédélec finite elements for three-dimensional photonic crystals has been considered in [BCG06], while the same problem was solved in [HLM16] via a finite difference scheme followed by a Newton-type numerical method. For two-dimensional photonic crystals, a spectral analysis was performed in [GG12] for the special case where the electric permittivity is real and frequency-independent. The extensive spectral analysis of the non-Hermitian non-linear case can be found in [Eng10] and [ELT17]. Moreover, a linearization technique for the Drude-Lorentz model of the electrical permittivity was introduced in [EKE12].

In the three first sections of this paper, we review the modeling of a perfect photonic crystals based on the Maxwell equations which leads to the operator NLEVP. In the fourth section, we consider the case where the electric permittivity ϵ is a real-valued and frequency-independent function. We introduce the weak formulation of the resulting Hermitian linear EVP and analyse its properties. In particular we show that its spectrum consists in at most countably many real and positive eigenvalues with finite geometric multiplicities. In the fifth section, we develop a residual-based error estimator which approximate the algebraic error on the computed eigenvector in the norm defined by the sesquilinear forms of the weak formulation. Finally, the last section of this paper is devoted to a numerical experiment, where we simulate a two-dimensional photonic crystal and test the reliability, as well as the efficiency, of the algebraic error estimator.

2. Preliminaries on the Maxwell equations

2.1. **General Form.** The Maxwell equations in their general form and in SI units are stated as follows:

$$\left[\begin{array}{rl} -\partial_t D(x,t) + \nabla \times H(x,t) = J(x,t), & \text{(Ampère's circuital law)} \\ \nabla \cdot D(x,t) = \rho(x,t), & \text{(Gauss' law)} \\ \partial_t B(x,t) + \nabla \times E(x,t) = 0, & \text{(Faraday's law of induction)} \\ \nabla \cdot B(x,t) = 0, & \text{(Gauss' law for magnetism)} \end{array}\right.$$

where E, H, D, B and J are three-dimensional vector fields and ρ is a real scalar field, see e.g. [Jac99]. These fundamental equations come along with the following empirical constitutive equations:

$$\left[\begin{array}{l} D(x,t) = \epsilon_0 E(x,t) + \epsilon_0 \chi_E(x,t) * E(x,t), \\ B(x,t) = \mu_0 H(x,t) + \mu_0 \chi_M(x,t) * H(x,t), \end{array}\right.$$

where ϵ_0, μ_0 are strictly positive constants, χ_E, χ_M are causal transfer functions, and $*$ denotes a convolution in time, i.e.

$$\chi_E(x,t) * E(x,t) = \int\limits_{-\infty}^{\infty} \chi_E(x,t-\tau) E(x,\tau) \mathrm{d}\tau \overset{\text{causality}}{=} \int\limits_{-\infty}^{t} \chi_E(x,t-\tau) E(x,\tau) \mathrm{d}\tau.$$

Remark 2.1. In the remainder, we omit writing the explicit dependence of vector field on space, time and frequency when the dependence is clear from the context.

In the remainder, we will assume that no free current nor free charges are present in the system, i.e. $J \equiv 0$ and $\rho \equiv 0$, and that the materials are non magnetizable, i.e. $\chi_M \equiv 0$. Under these assumptions, the Maxwell equations simplify to the following system of equations:

$$\left[\begin{array}{ll} -\partial_t(\epsilon_0 E + \epsilon_0 \chi_E * E) + \nabla \times H = 0, & \partial_t(\mu_0 H) + \nabla \times E = 0, \\ \nabla \cdot (\epsilon_0 E + \epsilon_0 \chi_E * E) = 0, & \nabla \cdot (\mu_0 H) = 0. \end{array}\right. \tag{2.1}$$

2.2. **The eigenvalue problem.** We recall briefly the definition of the Fourier transform. Given a function $f(t) \in L^1(\mathbb{R})$, i.e., $f(t)$ is absolutely integrable on $\mathbb{R}$, its Fourier transform is defined by

$$\mathcal{F}\{f\}(\omega) := \int\limits_{\mathbb{R}} f(t) e^{-i\omega t} \, \mathrm{d}t.$$

Note that when the variable t represents time with units [s], then ω represents the *angular frequency* and has units [rad/s]. To shorten the notation, we write formally $\hat{f}(\omega) := \mathcal{F}\{f\}(\omega)$. On a subspace of $L^1(\mathbb{R})$, namely the space $\mathcal{S} \subset L^1(\mathbb{R})$ of so-called *Schwartz functions*, the Fourier transformation satisfies a number of important properties. A rigorous definition of $\mathcal{S}$ can be found e.g. in [Jan71], but it is sufficient for our purpose to think of it as the space of infinitely differentiable functions decaying rapidly at infinity. For any function $f \in \mathcal{S}$ it holds that $\hat{f} \in \mathcal{S}$ and the inverse Fourier transform can be defined by

$$\mathcal{F}^{-1}\{\hat{f}\}(t) = \frac{1}{2\pi} \int\limits_{\mathbb{R}} \hat{f}(\omega) e^{+i\omega t} d\omega = f(t).$$

Moreover, given two functions $f, g \in \mathcal{S}$, the following properties of the Fourier transform follow from its definition:

$$\mathcal{F}\{f * g\}(\omega) = \hat{f}(\omega)\hat{g}(\omega) \qquad \text{and} \qquad \mathcal{F}\{f'\}(\omega) = +i\omega \hat{f}(\omega), \tag{2.2}$$

where f' denotes the first-order derivative of f. The Fourier transform is not restricted to functions in $L^1(\mathbb{R})$ and can be extended to the space of *tempered distributions*, i.e., to the dual space of $\mathcal{S}$. For a rigorous definition of these concepts and for the proof of properties analogous to (2.2) satisfied by the Fourier transform of tempered distributions, see e.g. [Jan71].

It is physically meaningful to expect that E and H stay bounded at all times since the system contains no energy source. Therefore, we can assume E and H to be tempered distributions and we can apply a Fourier transformation in time to the Maxwell equations (2.1). This yields to the following constrained *eigenvalue problem (EVP)*:

$$\begin{aligned}
&\begin{pmatrix} 0 & \frac{1}{\epsilon(x,\omega)}\nabla\times \\ -\frac{1}{\mu_0}\nabla\times & 0 \end{pmatrix}\begin{pmatrix}\hat{E}(x,\omega)\\ \hat{H}(x,\omega)\end{pmatrix} = +i\omega\begin{pmatrix}\hat{E}(x,\omega)\\ \hat{H}(x,\omega)\end{pmatrix}\\
&\nabla\cdot(\epsilon(x,\omega)\hat{E}(x,\omega)) = 0,\\
&\nabla\cdot(\mu_0\hat{H}(x,\omega)) = 0,
\end{aligned} \tag{2.3}$$

where $\epsilon(x,\omega) \coloneqq \epsilon_0(1+\hat{\chi_E}(x,\omega))$. The first-order (order of derivation in space) EVP of dimension 6 in (2.3) can be turned into a second-order EVP of dimension 3 simply by substituing $\hat{E} = +\frac{1}{i\omega}\frac{1}{\epsilon}\nabla\times\hat{H}$ in the second line or, similarly, by substituing $\hat{H} = -\frac{1}{i\omega}\frac{1}{\mu_0}\nabla\times\hat{E}$ in the first line. This yields two equivalent formulations of the EVP:

$$\nabla\times\Big(\frac{1}{\epsilon(x,\omega)}\nabla\times\hat{H}(x,\omega)\Big) = \mu_0\omega^2\hat{H}(x,\omega) \tag{2.4}$$

and

$$\nabla\times\nabla\times\hat{E}(x,\omega) = \mu_0\epsilon(x,\omega)\omega^2\hat{E}(x,\omega). \tag{2.5}$$

Remark that in these two formulations, the constraints $\nabla\cdot(\epsilon\hat{E}) = 0$ and $\nabla\cdot(\mu_0\hat{H}) = 0$ are automatically satisfied since

$$\nabla\cdot(\nabla\times f)\equiv 0$$

for any function f twice continuously differentiable.

In order to simplify the equations, we restrict our solution space to *separable* electromagnetic fields, i.e. we take the ansatz

$$H(x,t) = \mathsf{H}(x)f(t) \qquad \text{and} \qquad E(x,t) = \mathsf{E}(x)g(t), \tag{2.6}$$

where f and g are two nonzero tempered distributions.

Incorporating the first ansatz into Equation (2.4) and simplifying the terms $\hat{f}(\omega)$ appearing on both sides of the equation yields

$$\nabla\times\Big(\frac{1}{\epsilon(x,\omega)}\nabla\times\mathsf{H}(x)\Big) = \mu_0\omega^2\mathsf{H}(x). \tag{2.7}$$

Similarly, Equation (2.5) can be rewritten as

$$\nabla\times\nabla\times\mathsf{E}(x) = \mu_0\epsilon(x,\omega)\omega^2\mathsf{E}(x). \tag{2.8}$$

2.2.1. *The two-dimensional case.* In this subsection, we assume all properties of the system to be periodic in a *two-dimensional* (2D) plane and to be constant along the direction perpendicular to this plane. Without loss of generality, we assume that the plane with periodic properties is spanned by x_1 and x_2, and that the the direction with constant properties is along x_3. Using the short notation $\partial_j = \frac{\partial}{\partial x_j}$ for $j = 1,2,3$, the assumption of constant material properties along x_3 writes $\partial_3\epsilon = 0$, and implies $\partial_3\mathsf{E} = \partial_3\mathsf{H} = 0$ since no parameter in Equations (2.7) and (2.8) depends on x_3.

Let us introduce some 2D equivalents of the standard *three-dimensional* (3D) differential operators, as well as some notation. We consider a 3D vector $u = (u_1,u_2,u_3)^T$, a 2D vector $\underline{v} = (v_1,v_2)^T$, and a scalar function w, and we define

$$\underline{u} \coloneqq \begin{pmatrix}u_1\\u_2\end{pmatrix}, \qquad \underline{v}^\perp \coloneqq \begin{pmatrix}-v_2\\v_1\end{pmatrix},$$

$$\underline{\mathrm{grad}}_{2D}w \coloneqq \begin{pmatrix}\partial_1 w\\ \partial_2 w\end{pmatrix}, \qquad \mathrm{div}_{2D}\,\underline{v} \coloneqq \partial_1 v_1 + \partial_2 v_2,$$

$$\underline{\mathrm{curl}}_{2D} w := -(\mathrm{grad}_{2D}\, w)^{\perp} = \begin{pmatrix} \partial_2 w \\ -\partial_1 w \end{pmatrix}, \qquad \mathrm{curl}_{2D}\, \underline{v} := -\,\mathrm{div}_{2D}(\underline{v}^{\perp}) = \partial_1 v_2 - \partial_2 v_1,$$

$$\mathrm{laplace}_{2D}\, w := \partial_1\partial_1 w + \partial_2\partial_2 w = -\,\mathrm{curl}_{2D}\, \underline{\mathrm{curl}}_{2D} w = \mathrm{div}_{2D}\, \underline{\mathrm{grad}}_{2D} w.$$

It follows that

$$\nabla \times u = \begin{pmatrix} \underline{\mathrm{curl}}_{2D} u_3 \\ \mathrm{curl}_{2D}\, \underline{u} \end{pmatrix} + \partial_3 \begin{pmatrix} \underline{u}^{\perp} \\ 0 \end{pmatrix}.$$

With these tools we can easily show that, in the 2D case, (2.7) becomes

$$\begin{pmatrix} \underline{\mathrm{curl}}_{2D} \frac{1}{\epsilon}\, \mathrm{curl}_{2D} \begin{pmatrix} \mathsf{H}_1 \\ \mathsf{H}_2 \end{pmatrix} \\ \mathrm{curl}_{2D}\, \frac{1}{\epsilon} \underline{\mathrm{curl}}_{2D}(\mathsf{H}_3) \end{pmatrix} = \mu_0 \omega^2 \mathsf{H}$$

and (2.8) becomes

$$\begin{pmatrix} \underline{\mathrm{curl}}_{2D}\, \mathrm{curl}_{2D} \begin{pmatrix} \mathsf{E}_1 \\ \mathsf{E}_2 \end{pmatrix} \\ \mathrm{laplace}_{2D}(\mathsf{E}_3) \end{pmatrix} = \mu_0 \epsilon \omega^2 \mathsf{E}.$$

Therefore the equations including H_3 and E_3 are decoupled from the rest of the equations. This decoupling can be best expressed in terms of the *transverse electric* (TE) and *transverse magnetic* (TM) modes, which we introduce now. Let us start from (2.3) again and add the effects of the continuous translational symmetry along x_3. We obtain

$$\begin{pmatrix} \frac{1}{\epsilon} \underline{\mathrm{curl}}_{2D}(\hat{H}_3) \\[2ex] \frac{1}{\epsilon}\, \mathrm{curl}_{2D} \begin{pmatrix} \hat{H}_1 \\ \hat{H}_2 \end{pmatrix} \\[2ex] \frac{1}{-\mu_0} \underline{\mathrm{curl}}_{2D}(\hat{E}_3) \\[2ex] \frac{1}{-\mu_0}\, \mathrm{curl}_{2D} \begin{pmatrix} \hat{E}_1 \\ \hat{E}_2 \end{pmatrix} \end{pmatrix} = i\omega \begin{pmatrix} \begin{pmatrix} \hat{E}_1 \\ \hat{E}_2 \end{pmatrix} \\[2ex] (\hat{E}_3) \\[2ex] \begin{pmatrix} \hat{H}_1 \\ \hat{H}_2 \end{pmatrix} \\[2ex] (\hat{H}_3) \end{pmatrix}$$

where the first and fourth block lines, which define the TE mode $(\hat{E}_1^T, \hat{E}_2^T, 0, 0, 0, \hat{H}_3^T)^T$, are decoupled from the second and third block lines, which define the TM mode $(0, \hat{H}_1^T, \hat{H}_2^T, \hat{E}_3^T, 0, 0)^T$. Combining the equations describing the TE mode yields

$$\begin{array}{l} \text{(2.9)} \\[4ex] \text{(2.10)} \end{array} \qquad \left[\begin{aligned} -\frac{1}{\mu_0}\, \mathrm{curl}_{2D}\, \frac{1}{i\omega} \frac{1}{\epsilon} \underline{\mathrm{curl}}_{2D} \hat{H}_3 &= i\omega \hat{H}_3, \\ \frac{1}{\epsilon} \underline{\mathrm{curl}}_{2D} \hat{H}_3 &= i\omega \begin{pmatrix} \hat{E}_1 \\ \hat{E}_2 \end{pmatrix}, \end{aligned} \right.$$

and combining the equations describing the TM mode yields

$$\begin{array}{l} \text{(2.11)} \\[4ex] \text{(2.12)} \end{array} \qquad \left[\begin{aligned} \frac{1}{\epsilon}\, \mathrm{curl}_{2D}\, \frac{1}{i\omega} \frac{1}{-\mu_0} \underline{\mathrm{curl}}_{2D} \hat{E}_3 &= i\omega \hat{E}_3, \\ \frac{1}{-\mu_0} \underline{\mathrm{curl}}_{2D} \hat{E}_3 &= i\omega \begin{pmatrix} \hat{H}_1 \\ \hat{H}_2 \end{pmatrix}. \end{aligned} \right.$$

It follows that the components $\hat{E}_1, \hat{E}_2, \hat{H}_1, \hat{H}_2$ can be easily computed from the solutions $\hat{H}_3, \hat{E}_3$ of (2.9) and (2.11).

Moreover, the constraints in (2.3) are automatically satisfied in the formulation of Equations (2.9)–(2.12) since

$$\nabla\cdot(\epsilon\hat{E}) = \nabla\cdot\begin{pmatrix} \frac{1}{i\omega}\underline{\mathrm{curl}}_{2D}\hat{H}_3 \\ \frac{1}{\mu_0\omega^2}\,\mathrm{curl}_{2D}\,\underline{\mathrm{curl}}_{2D}\hat{E}_3 \end{pmatrix} = \nabla\cdot\nabla\times\begin{pmatrix} \frac{1}{\mu_0\omega^2}\underline{\mathrm{curl}}_{2D}\hat{E}_3 \\ \frac{1}{i\omega}\hat{H}_3 \end{pmatrix} = 0$$

and

$$\nabla\cdot(\mu_0\hat{H}) = \nabla\cdot\begin{pmatrix} \frac{1}{-i\omega}\underline{\mathrm{curl}}_{2D}\hat{E}_3 \\ \mathrm{curl}_{2D}\,\frac{1}{\epsilon\omega^2}\underline{\mathrm{curl}}_{2D}\hat{H}_3 \end{pmatrix} = \nabla\cdot\nabla\times\begin{pmatrix} \frac{1}{\epsilon\omega^2}\underline{\mathrm{curl}}_{2D}\hat{H}_3 \\ \frac{-1}{i\omega}\hat{E}_3 \end{pmatrix} = 0.$$

Finally, note that with the ansatz (2.6), Equations (2.9) and (2.11) can be reformulated as follows

$$\begin{aligned} -\,\mathrm{div}_{2D}(\frac{1}{\epsilon}\underline{\mathrm{grad}}_{2D}\mathsf{H}_3) &= \mu_0\omega^2\mathsf{H}_3 \\ -\,\mathrm{laplace}_{2D}\,\mathsf{E}_3 &= \mu_0\epsilon\omega^2\mathsf{E}_3, \end{aligned}$$

or equivalently (for $\partial_3\epsilon = \partial_3\mathsf{H}_3 = \partial_3\mathsf{E}_3 \equiv 0$)

$$-\nabla\cdot(\frac{1}{\epsilon}\nabla\mathsf{H}_3) = \mu_0\omega^2\mathsf{H}_3 \tag{2.13}$$

$$-\Delta\mathsf{E}_3 = \mu_0\epsilon\omega^2\mathsf{E}_3. \tag{2.14}$$

The TE mode can be determined by the solution of (2.13) while the TM mode can be determined by the solution of (2.14).

2.3. **Models for the electric permittivity.**

It has been mentioned in Section 2.1 that we only consider nonmagnetic materials, i.e. the magnetic permeability in the whole system is constant and is equal to that of vacuum, namely μ_0. The electric permittivity ϵ however is assumed to differ from one material to the other but to stay constant in space within a given material. Mathematically this means that, if the crystal is composed of N_{mat} materials and we denote by Ω_j the domain filled with the j-th material, the electric permittivity of the crystal can be defined as follows

$$\epsilon(x,\omega) = \sum_{j=1}^{N_{\mathrm{mat}}} \epsilon_j(\omega)\chi_{\Omega_j}(x)$$

where $\epsilon_j(\omega)$ are constant functions in space and χ_{Ω_j} is the indicator function of Ω_j.

The functions $\epsilon_j(\omega)$ need to be approximated by empirical models. In the remainder of this subsection, we introduced two of the most commonly used models. Note that we only consider the dependency of the electric permittivity on the frequency and neglect the influence of other factors, such as e.g. the temperature.

2.3.1. *The frequency-independent model.*

The simplest model for $\epsilon_i(\omega)$ is to assume it to be constant in frequency, i.e.

$$\epsilon_i(\omega) = \epsilon_r^{(i)}\epsilon_0$$

where $\epsilon_r^{(i)}$ denotes the relative permittivity of the i-th material and ϵ_0 the electric permittivity of vacuum.

With this model, the EVPs (2.7), (2.8), (2.13), and (2.14) become linear in the eigenvalue $\lambda := \omega^2$.

2.3.2. *The Drude-Lorentz model.* We follow here the derivation of the Drude-Lorentz model presented in [Jac99, p.309-312]. We consider an electron bound to a kernel in a crystal and submitted to an electric field. We model the interaction between the electron and the kernel with a spring of stiffness k and of damping constant γ. We make the following approximations:

- The difference between the local and applied electric fields is neglectable, i.e. the material has low density;
- Magnetic forces are neglectable;
- The amplitudes of electron oscillations are small enough to evaluate the field at a constant average position $\bar{x}$;
- The applied field E is assumed to vary harmonically in time.

We denote by $q = -e$ the charge of the electron and by x its position (with reference point at the equilibrium point of the spring). Newton's equation of motion writes in this case

$$m\ddot{x} = -kx - \gamma\dot{x} + qE$$

After defining $\omega_0 \coloneqq \sqrt{k/m}$ we obtain

$$m(\ddot{x} + \frac{\gamma}{m}\dot{x} + \omega_0^2 x) = -eE. \tag{2.15}$$

Under the assumption that E is time-harmonic, i.e. with the ansatz $E(\bar{x}, t) = E(\bar{x})e^{i\omega t}$ for a certain ω in $\mathbb{C}$, the Fourier transformation of (2.15) yields to

$$\mathcal{F}\{x(t)\}(\nu) = \frac{2\pi e E(\bar{x})}{m} \frac{1}{\omega_0^2 - \nu^2 + i\nu\gamma} \delta(\omega - \nu) = C(\omega)\delta(\omega - \nu)$$

where $C(\omega) \coloneqq 2\pi e E(\bar{x})(m(\omega_0^2 - \omega^2 + i\omega\gamma))^{-1}$ is constant in ν. Therefore $x(t)$ is time harmonic with (angular) frequency ω as well and satifies

$$x(t) = -\frac{e}{m}\frac{1}{(\omega_0^2 - \omega^2) + i\omega\gamma} E(\bar{x}, t).$$

It follows that the dipole moment $p \coloneqq qx$ equals

$$p = \frac{e^2}{m}\frac{E}{\omega_0^2 - \omega^2 + i\omega\gamma}.$$

Let us now consider the dipole moment created by multiple electrons having possibly different binding frequencies. We denote by N the number of molecules per unit volume and by Z the number of electrons per molecule. We suppose that there are f_l electrons per molecule with binding frequency ω_l and damping constant γ_l, and therefore that $\sum_l f_l = Z$. We call p_l the dipole moment created by the f_l electrons corresponding to the binding frequency ω_l. Given the relationships

$$\frac{\epsilon}{\epsilon_0} = 1 + \hat{\chi}_E$$

and

$$N \sum_l f_l \hat{p}_l = \epsilon_0 \hat{\chi}_E \hat{E}$$

we deduce that

$$\frac{\epsilon}{\epsilon_0} = 1 + \frac{e^2 N}{m\epsilon_0}\sum_l \frac{f_l}{\omega_l^2 - \omega^2 + i\omega\gamma_l} \tag{2.16}$$

$$= 1 + \frac{e^2 N}{m\epsilon_0}\sum_l f_l \left[\frac{(\omega_l^2 - \omega^2)}{(\omega_l^2 - \omega^2)^2 + \omega^2\gamma_l^2} - i\frac{\omega\gamma_l}{(\omega_l^2 - \omega^2)^2 + \omega^2\gamma_l^2}\right]. \tag{2.17}$$

The damping constant γ_l are generally small compared to the binding frequencies ω_l, therefore ϵ is approximately real for almost all frequencies ω. We notice that the j-th term in the sum in (2.16) becomes purely imaginary when $\omega = \omega_l$. Moreover, at this point, the imaginary part becomes significantly larger in magnitude than at points away from the resonance.

2.3.3. *Physical properties of the electric permittivity.* In the previous section, we have seen that taking into account the damping from the crystal leads to a complex permittivity. Since the damping is a non-conservative force dissipating kinetic energy into other types of energy, e.g. heat or sound, a non-real permittivity corresponds to a dissipative system.

Actually, it will be discussed in subection 4.2 that the eigenvalue problems 2.8 and 2.7 are self-adjoint and have their spectrum (i.e. the possible values of ω in the ansatz) contained in $\mathbb{R}$ if and only if ϵ is a real function.

Under the assumption that E varies harmonically in time, i.e.

$$\begin{aligned} E(x,t) &= \Re\Big(\mathsf{E}(x)e^{i\omega t}\Big) = \Re\Big(\mathsf{E}(x)e^{i(\Re(\omega)+i\Im(\omega))t}\Big) \\ &= \mathsf{E}(x)e^{-\Im(\omega)t}\Re\Big(e^{i\Re(\omega)t}\Big), \end{aligned}$$

the complex part of the frequency describes the decay of the eigenmode over time. In this setting, in particular with this sign convention in the ansatz, ω can only be located in the upper half of the complex plane, since no energy source is present in the crystal.

3. The Floquet transformation & Bloch theorem

3.1. **Some fundamentals concepts in crystallography.** In this section we introduce some important concepts in the theory of periodic structures. The reader can refer to the standard textbook [Kit05] for a more extensive description of these concepts in a physical context. A detailed presentation of these same concepts in a mathematical context can be found in [Sim13].

An ideal crystal is formed of two or more components that are arranged in a periodic manner, i.e. as an infinite repetition of an identical pattern. In this paper we consider an ideal n-dimensional crystal composed of two different materials, with n=2 or n=3. Mathematically, one uses the concept of lattice to represent such a periodic structure, and multiple definitions of this concept can be found in the literature. We choose to use the following one for its mathematical precision:

Definition 3.1. Let $a_1, \ldots, a_n \in \mathbb{R}^n$ be linearly independent. Then the set

$$G := \{g \in \mathbb{R}^n \mid g = \sum_{i=1}^{n} \nu_i a_i \ \text{ for some } \nu_i \in \mathbb{Z}\}.$$

is called a *(Bravais) lattice* and the vectors $a_1, \ldots, a_n$ are called the *primitive lattice vectors.*

With this definition, it follows that any periodic structure can be expressed as a lattice of repeating objects [Sim13]. This statement is illustrated informally in Figure 3.1.

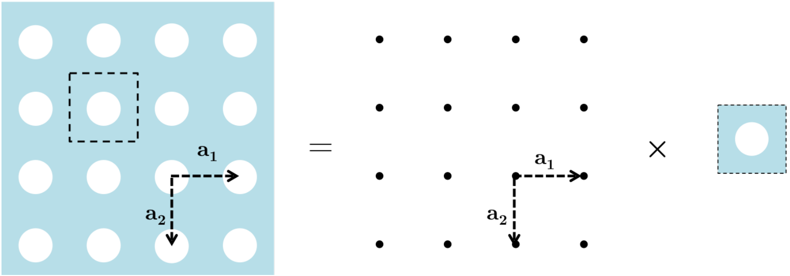


FIGURE 3.1. Decomposition of a periodic structure into a tiling of repetitive units located on a lattice.

Any lattice G is the result of the tiling of $\mathbb{R}^n$ by infinitely many identical tiles called *unit cells*, and a unit cell is called *primitive* if it contains exactly one lattice point. Equivalently, a primitive unit cell can be thought of as a domain in $\mathbb{R}^n$ of smallest volume that can tile G using only translations. In the remainder, we use a particular type of primitive unit cell, namely the Wigner-Seitz cell defined as follows:

Definition 3.2. Let G be a lattice in $\mathbb{R}^n$ and $g \in G$. The *Wigner-Seitz cell* of G at the point g is

$$\Omega_g := \{r \in \mathbb{R}^n \mid \|r - g\| < \|r - \tilde{g}\| \text{ for all } \tilde{g} \in G, \tilde{g} \neq g\},$$

where $\|\cdot\|$ is the standard Euclidean norm in $\mathbb{R}^n$. One may refer simply to the Wigner-Seitz Ω of G without mentioning any point g in order to denote tacitly the Wigner-Seitz cell centered around the origin $(0 \in G)$.

To any Bravais lattice G in $\mathbb{R}^n$ corresponds a reciprocal lattice G' defined by

$$G' := \{g' \in \mathbb{R}^n \mid \text{for all } g \in G, \text{ there exists } m \in \mathbb{Z} \text{ satisfying } \; g' \cdot g = 2\pi m\}.$$

It can be shown (see e.g. [Kit05]) that the primitive lattice vectors $b_1, \dots, b_n$ of G' can always be chosen in a such way that they satisfy the following constraints: For all $j, l = 1, \dots, n$,

$$b_j \cdot a_l = 2\pi\delta_{jl}, \tag{3.1}$$

where δ_{jl} denotes the Kronecker delta, and $a_1, \dots, a_n$ denote the primitive lattice vectors of G.

The Wigner-Seitz cell of G' is called the Brillouin zone of G. In addition to the discrete translational symmetries represented by the lattice vectors, a lattice G can be invariant under other sorts of symmetries, such as e.g. rotational or axial symmetries. In such a case, the Brillouin zone reduced by all the symmetries of the lattice is called the *irreducible Brillouin zone* $\mathcal{K}$ of G (see Figure 3.2).

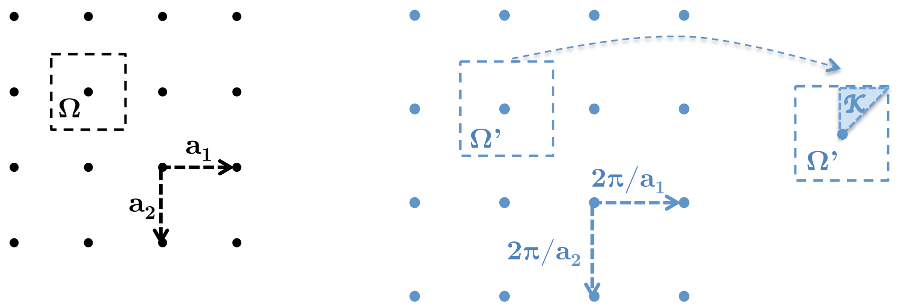
3.2. **Learning from the electron in a crystal.** In this susbsection we recall physical results historically arising from the analysis of an electron in a perfect crystal.

3.2.1. *Translational operators.* For any $\Delta x \in \mathbb{R}^n$ we denote by $T_{\Delta x} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ the translational operator shifting a function to the right by Δx, i.e.

$$T_{\Delta x} v(x) = v(x - \Delta x)$$

for all $v(x) \in L^2(\mathbb{R}^n)$. Its dual operator with respect to the inner product

$$(v, w)_{L^2(\mathbb{R}^n)} := \int_{\mathbb{R}^n} v(x) w(x) \, \mathrm{d}x \qquad \text{for all } v, w \in L^2(\mathbb{R}^n)$$



Square lattice G with lattice vectors a_1, a_2 and Wigner-Seitz cell Ω.

Reciprocal lattice G' of G with Brillouin zone Ω' of G.

Irreducible Brillouin zone 𝒦 of G.

FIGURE 3.2. Illustrations of some fundamental concepts in crystallography.

is

$$T^*_{\Delta x} := T_{-\Delta x},$$

since for all $v, w, \in L^2(\mathbb{R}^n)$ it holds

$$(T^*_{\Delta x}v, w) = (v, T_{\Delta x}w) = \int\limits_{\mathbb{R}^n} v(x)w(x - \Delta x)\,\mathrm{d}x = \int\limits_{\mathbb{R}^n} v(\tilde{x} + \Delta x)w(\tilde{x})\,\mathrm{d}\tilde{x} = (T_{-\Delta x}v, w).$$

Thus

$$T_{\Delta x}T^*_{\Delta x}v(x) = T_{\Delta x}v(x + \Delta x) = v(x)$$

for all $v \in L^2(\mathbb{R}^n)$ and therefore the translational operators is unitary. It follows in particular that its spectrum is contained on the unit circle in the complex plane.

3.2.2. *Bloch theorem.*

Lemma 3.3. *Given a n-dimensional lattice G with lattice vectors $\{a_j\}_{j=1}^n$, if a function $\psi \in L^2(\mathbb{R}^d)$ is an eigenfunction of all translational operators $\{T_{a_j}\}_{j=1}^n$, then ψ is a Bloch wave, i.e., it is of the form*

$$\psi(x) = e^{ik\cdot x}u_k(x)$$

for some periodic function $u_k(x)$ with periodicity vectors $\{a_j\}_{j=1}^n$ and for some vector k in the Brillouin zone Ω'.

Proof. If $\psi(x)$ is an eigenfunction of all translational operators $\{T_{a_j}\}_{j=1}^n$, then for all $j = 1, \ldots, n$ there exists a real constant $\theta_j \in [-\frac{1}{2}, \frac{1}{2})$ such that

$$\psi(x + a_j) = e^{2\pi i\theta_j}\psi(x),$$

since T_{a_j} is unitary. Let us denote by $\{b_j\}_{j=1}^n$ the lattice vectors of the reciprocal lattice G' satisfying Equation (3.1), and define

$$k := \sum_j^n \theta_j b_j.$$

Note that $k \in \Omega'$ per construction. Then the function

$$u_k(x) := e^{-ik\cdot x}\psi(x).$$

satisfies

$$u_k(x+a_j) = e^{-ik\cdot x}e^{-ik\cdot a_j}\psi(x+a_j) = e^{-ik\cdot x}e^{-i2\pi\theta_j}e^{i2\pi\theta_j}\psi(x) = e^{-ik\cdot x}\psi(x) = u_k(x)$$

for all $j = 1, \ldots, n$, which terminates the proof. □

Theorem 3.4 (Bloch theorem)**.** *The wavefunction describing the state of electron in a crystal can be expressed in a basis formed of Bloch waves. Moreover, these Bloch waves are energy eigenstates.*

Proof. Due to the discrete symmetry of the problem, the Hamiltonian and the translation operators commute with each others. Therefore they can be diagonalized in a common basis. This basis is thus formed of functions that are energy eigenstates and have the form of Bloch waves. The expansion of the wavefunction in this basis yields to a linear combination of Bloch waves, which terminates the proof. □

3.3. **Floquet transformation.** The Bloch theorem is very well-known in solid state physics. In the mathematics community, this result is more commonly known as a part of the Floquet theory. Eventhough the Bloch and the Floquet theories are analogous, their derivations differs a bit. In this subsection we recall the definition of the Floquet transformation, as well as some of the main results related to it.

Definition 3.5. For any function $u \in L^2(\mathbb{R}^n)$, we define the *Floquet transform of u with respect to the lattice G* as the function

$$\mathcal{F}\{u\}(k,x) := \sum_{g\in G} u(x-g)e^{-ik\cdot(x-g)}.$$

For the sake of brevity, we use the notation $u_k(x) := \mathcal{F}\{u\}(x,k)$. Some important properties follow directly from this definition:

- u_k is $G-$periodic in x and quasi-G'-periodic in k, i.e.
$$u_k(x+g) = u_k(x) \text{ for all } g \in G,$$
$$u_{k+g'}(x) = e^{-ig'\cdot x}u_k(x) \text{ for all } g' \in G',$$
- given a G-periodic function $\epsilon(x)$ such that $\epsilon u \in L^2(\mathbb{R}^n)$, it holds
$$\mathcal{F}\{\epsilon u\}(k,x) = \epsilon(x)u_k(x),$$
- for $j = 1, \ldots, n$,
$$\mathcal{F}\{\partial_j u\}(k,x) = [\partial_j + ik_j]u_k(x), \tag{3.2}$$
where ∂_j denotes the partial derivative with respect to the j-th component of x and k_j is the j-th component of k.

The following theorem ensures that the Floquet transformation is invertible when the codomain is properly defined in terms of the irreducible Brillouin zone $\mathcal{K}$ of the lattice G.

Theorem 3.6 ([Kuc01])**.** *Let us denote by $\mathbb{T}^n$ the n-dimensional torus corresponding to the Wigner-Seitz cell Ω of G with periodic boundary conditions. The Floquet transformation with respect to G, considered as a mapping*

$$\mathcal{F} : L^2(\mathbb{R}^n) \to L^2(\mathcal{K}, L^2(\mathbb{T}^n))$$
$$u \to (k \to u_k(\cdot)),$$

is isometric and its inverse is given by

$$u(x) = \mathcal{F}^{-1}\{u_k\}(x) = \frac{1}{|\mathcal{K}|}\int_{\mathcal{K}} u_k(x)e^{ik\cdot x}dk.$$

3.4. **Link between Bloch waves and Floquet transformations.** Similarly to the state of an electron in a crystal, the state of a photon in a crystal, represented by the electromagnetic field, can be expressed as a linear combination of Bloch waves. Indeed, let u denote either the electric field E or the magnetic field H. Then

$$u(x) = \mathcal{F}^{-1}\{\mathcal{F}\{u\}\}(x) = \int\limits_{\mathcal{K}} \frac{1}{|\mathcal{K}|} e^{ik\cdot x} u_k(x) dk = \int\limits_{\mathcal{K}} \frac{1}{|\mathcal{K}|} \psi_k(x) dk,$$

where $\psi_k(x) \coloneqq e^{ik\cdot x} u_k x$ is a Bloch wave.

3.5. **The Floquet transformation applied to the Maxwell equations.** In the remainder, $G \subset \mathbb{R}^n$ represents the lattice underlying the spatial distribution of ϵ, i.e. $\epsilon(x, \omega)$ is G-periodic in x. In this subsection, we apply a Floquet transformation with respect to G to the Maxwell EVPs for 3D and 2D crystals. In view of the property (3.2) we introduce the notation $\nabla_k \coloneqq \nabla + ik$ for any $k \in \mathbb{R}^3$ and $\nabla_k \coloneqq [\partial_1, \partial_2]^T + ik$ for any $k \in \mathbb{R}^2$.

3.5.1. *2D photonic crystals.* Applying the Floquet transformation to (2.13) and (2.14) for some k in the irreducible Brillouin zone $\mathcal{K} \subset \mathbb{R}^2$ gives

$$-\nabla_k \cdot (\frac{1}{\epsilon} \nabla_k \mathsf{H}_{3,k}) = \mu_0 \omega^2 \mathsf{H}_{3,k} \tag{3.3}$$

$$-\nabla_k \cdot \nabla_k \mathsf{E}_{3,k} = \mu_0 \epsilon \omega^2 \mathsf{E}_{3,k}, \tag{3.4}$$

respectively, where (3.3) and (3.4) are now two equations on the Wigner-Seitz cell Ω, with periodic boundary conditions on the functions $\mathsf{H}_{3,k}(x), \mathsf{E}_{3,k}(x)$ and $\epsilon(x)$.

3.5.2. *3D photonic crystals.* Similarly as in the 2D case, applying the Floquet transformation to (2.7) and (2.5) for some k in the irreducible Brillouin zone $\mathcal{K} \subset \mathbb{R}^3$ gives

$$\nabla_k \times \Big(\frac{1}{\epsilon(x,\omega)} \nabla_k \times \mathsf{H}_k(x) \Big) = \mu_0 \omega^2 \mathsf{H}_k(x)$$

$$\nabla_k \times \nabla_k \times \mathsf{E}_k(x) = \mu_0 \epsilon(x,\omega) \omega^2 \mathsf{E}_k(x).$$

respectively, which are now two equations on the Wigner-Seitz cell Ω, with periodic boundary conditions on the functions $\mathsf{H}_k(x), \mathsf{E}_k(x)$ and $\epsilon(x)$.

4. Finite Element Formulation of the EVP

In the remainder we focus our analysis on 2D photonic crystals. Equations (3.3) and (3.4) are very similar and so are the methods to solve them as well. Therefore, we concentrate the detailed analysis on (3.4) and keep in my mind that the analysis of (3.3) is analogous. Moreover, from now on, we consider the easiest empirical model for the electric permittivity, i.e. we assume that $\epsilon = \epsilon(x) = \sum_{j=1,2} \epsilon_j \mathcal{I}_{\Omega_j}$ is independent of the (angular) frequency ω. In this case, (3.4) depends only quadratically on ω, so we define the new variable $\lambda \coloneqq \omega^2$.

4.1. **Weak and discrete formulations.** We aim at solving the EVP (3.4) by means of the finite element method. To do so, we first define the weak formulation of the EVP, namely: For all $k \in \mathcal{K}$, find $(u, \lambda) \in V \times \mathbb{C}$ such that

$$a_k(u, v) = \lambda b_\epsilon(u, v) \qquad \text{for all } v \in V, \tag{4.1}$$

where $V \coloneqq H^1_{\text{per}}(\Omega) = H^1(\mathbb{T}^2)$ for $\mathbb{T}^2$ the 2-dimensional torus corresponding to the Wigner-Seitz cell Ω, and where $a : V \times V \to \mathbb{C}$ and $b_\epsilon : V \times V \to \mathbb{C}$ are defined by

$$a_k(u, v) \coloneqq \int\limits_{\Omega} \nabla_k u \cdot \overline{\nabla_k v} \, \mathrm{d}x$$

and, respectively,

$$b_\epsilon(u,v) := \int\limits_\Omega \mu_0 \epsilon(x) u\overline{v}\,\mathrm{d}x.$$

Remark 4.1. We have used the following integration by parts formula for functions in V:

$$\int_\Omega -\nabla\cdot\nabla_k u\overline{v}\,\mathrm{d}x = \int_\Omega \nabla_k u\cdot\overline{\nabla v}\,\mathrm{d}x - \int_{\partial\Omega}(\nabla_k u\cdot\hat{n})v\,\mathrm{d}s = \int_\Omega \nabla_k u\cdot\overline{\nabla v}\,\mathrm{d}x$$

where $\hat{n}$ denotes the the outward unit surface normal to $\partial\Omega$ and where $\int_{\partial\Omega}(\nabla_k u\cdot\hat{n})v\,\mathrm{d}s = 0$ is to be understood in the sense of traces. This yields

$$\int_\Omega -\nabla_k\cdot\nabla_k u\overline{v}\,\mathrm{d}x = \int_\Omega \nabla_k u\cdot\overline{\nabla v} - ik\cdot\nabla_k u\overline{v}\,\mathrm{d}x = \int_\Omega \nabla_k u\cdot\overline{\nabla_k v}\,\mathrm{d}x.$$

With the aim of solving Problem (4.1) numerically, we define a subspace $V_h \subset V$ of dimension $N<\infty$ and N basis functions $\{\phi_j\}_{j=1,\dots,N}$ such that

$$V_h = \mathrm{span}\{\phi_1,\dots\phi_N\}. \tag{4.2}$$

It follows that any $v_h\in V_h$ can be expressed as a linear combination of the basis functions, i.e. there exist scalars $v_1,\dots,v_N$ such that $v_h = \sum\limits_{j=1}^{N} v_j\phi_j$. Thus we can define a *discretized* problem corresponding to (4.1) as follows: For all $k\in\mathcal{K}$, find $(\lambda_h,u_h)\in\mathbb{C}\times V_h$ such that

$$A_k\mathbf{u} = \lambda_h\mu_0 B_\epsilon\mathbf{u}. \tag{4.3}$$

where $\mathbf{u} := [u_1,\dots,u_N]^T$, and A_k, B_ϵ are defined by $(A_k)_{j,i} := a_k(\phi_i,\phi_j)$ and $(B_\epsilon)_{j,i} := b_\epsilon(\phi_i,\phi_j)$, for $i,j=1,\dots,N$.

Before discussing about methods to solve Problem (4.1) and (4.3), let us analyze the properties of the linear forms and the corresponding matrices defining the EVPs.

4.2. **Properties of the EVPs.** Let us start this subsection with some comments and results concerning the considered vector spaces.

4.2.1. *Vector spaces considerations.* In the Hilbert spaces $H := L^2(\Omega)$ and $V = H^1(\Omega)\subset H$ we consider the standard norms

$$\|v\|_H := \int_\Omega |v(x)|^2\,\mathrm{d}x$$

for all $v\in H$ and

$$\|\cdot\|_V := \|\cdot\|_{H^1(\Omega)} = \Big(\|\nabla\cdot\|_H^2 + \|\cdot\|_H^2\Big)^{1/2}.$$

Remark 4.2. For any $u\in V$, it holds $\nabla u\in\big(L^2(\Omega)\big)^2$. Therefore the notation $\|\nabla u\|_{L^2(\Omega)}$ is a bit abusive. In fact, for any $w,w'\in\big(L^2(\Omega)\big)^2$ we use the conventional notation

$$(w,w')_{L^2(\Omega)} := \int_\Omega w(x)\cdot\overline{w'(x)}\,\mathrm{d}x$$

where $\cdot$ denotes the standard dot product, namely $w(x)\cdot\overline{w'(x)} := \sum_{i=1,2} w_i(x)\overline{w_i'(x)}$ for all $w(x),w'(x)\in\mathbb{C}^2$. Additionally,

$$\|w\|_{L^2(\Omega)}^2 := (w,w)_{L^2(\Omega)} = \int_\Omega |w|^2\,\mathrm{d}x$$

where $|w| := \Big(\sum_{i=1,2} w_i\overline{w_i}\Big)^{1/2}$ denotes the standard norm in $\mathbb{C}^2$.

When $\epsilon(x)$ is a real-valued function, we can define another inner product in H that will be more handy for the spectral analysis of the problem.

Lemma 4.3. *Let $\epsilon_0, \epsilon_\infty$ be some positive contants and let $\epsilon(x)$ be a real-valued function satisfying $0 < \epsilon_0 \leq \epsilon(x) \leq \epsilon_\infty$ for all $x \in \Omega$. Then the sesqui-linear form b_ϵ defines an inner product in H. Moreover, the norm $\|\cdot\|_{b_\epsilon}$ derived from this inner product is equivalent to $\|\cdot\|_H$ in H.*

Proof. The form b_ϵ possesses the following properties:

- Positivity: $b_\epsilon(u,u) = \int_\Omega \epsilon(x)|u|^2 \, \mathrm{d}x \geq \epsilon_0 \|u\|_H^2 \geq 0$ and $b_\epsilon(u,u) = 0$ if and only if $u \equiv 0$,
- Conjugate symmetry: $b_\epsilon(u,v) = \int_\Omega \epsilon(x) u\overline{v} \, \mathrm{d}x = \int_\Omega \overline{\epsilon(x) v} u \, \mathrm{d}x = \overline{b_\epsilon(u,v)}$ since $\epsilon(x) \in \mathbb{R}$,
- Linearity in the first argument: $b_\epsilon(\alpha u, v) = \alpha b_\epsilon(u,v)$ for all $\alpha \in \mathbb{C}$.

Moreover, it holds that $\epsilon_\infty \|u\|^2 \geq b_\epsilon(u,u) \geq \epsilon_0 \|u\|^2$, which shows the equivalence of the norms. □

In the remainder, we assume $\epsilon(x)$ to be a real-valued function satisfying $0 < \epsilon_0 \leq \epsilon(x) \leq \epsilon_\infty$ for all $x \in \Omega$. Hence, we can define the inclusion map $\mathcal{I}_\epsilon \colon V \to V^*$ by means of the Gelfand triple

$$V \hookrightarrow H \cong H^* \hookrightarrow V^*,$$

i.e., $\mathcal{I}_\epsilon = \iota_{H^* \to V^*} j_H \iota_{V \to H}$ where $j_H \colon H \to H^*$ is the Riesz isomorphism with respect to the inner product b_ϵ, $\iota_{V \to H}$ is the embedding of V into H, and $\iota_{H^* \to V^*} = (\iota_{V \to H})^*$. An important property of this particular Gelfand triple is that the imbedding of V into H is compact, see e.g. [Eng10] and references therein. It follows that the embedding of H^* into V^* is also compact [Hac92, Ch. 6].

4.2.2. *Definition of the operator equation.* In this subsection, we show that Problem (4.1) can be reformulated as an equivalent problem in the dual space V^*. e start with a lemma proving Young's inequality in the specific case of complex vectors.

Lemma 4.4 (Young's inequality)**.** *Consider $\mathbf{a}, \mathbf{b} \in \mathbb{C}^2$. Then, for every $\delta > 0$ we have an estimate of the form*

$$|\mathbf{a} \cdot \overline{\mathbf{b}} + \overline{\mathbf{a}} \cdot \mathbf{b}| \leq \frac{1}{\delta} |\mathbf{a}|^2 + \delta |\mathbf{b}|^2.$$

Proof. For any two vectors $\mathbf{c}, \mathbf{d} \in \mathbb{C}^2$, the following estimates hold:

$$0 \leq |\mathbf{c} + \mathbf{d}|^2 = (\mathbf{c} + \mathbf{d}) \cdot \overline{(\mathbf{c} + \mathbf{d})} = |\mathbf{c}|^2 + |\mathbf{d}|^2 + \mathbf{c} \cdot \overline{\mathbf{d}} + \overline{\mathbf{c}} \cdot \mathbf{d},$$
$$0 \leq |\mathbf{c} - \mathbf{d}|^2 = (\mathbf{c} - \mathbf{d}) \cdot \overline{(\mathbf{c} - \mathbf{d})} = |\mathbf{c}|^2 + |\mathbf{d}|^2 - \mathbf{c} \cdot \overline{\mathbf{d}} - \overline{\mathbf{c}} \cdot \mathbf{d}.$$

As a result, we have $|\mathbf{c} \cdot \overline{\mathbf{d}} + \overline{\mathbf{c}} \cdot \mathbf{d}| \leq |\mathbf{c}|^2 + |\mathbf{d}|^2$. The claim then follows by setting $\mathbf{c} = \mathbf{a}/\sqrt{\delta}$, and $\mathbf{d} = \mathbf{b}\sqrt{\delta}$. □

By means of the Young inequality, we can now show the boundedness of a_k:

Lemma 4.5. *For all $k \in \mathcal{K}$, the sesquilinear form $a_k(\cdot,\cdot)$ in (4.1) is bounded (or, equivalently, continuous) in V, with constant $C_M = 2\max(1, |k|^2)$. Moreover, a_k is Hermitian.*

Proof. By means of the Cauchy-Schwarz inequality we obtain for all $u, v \in V$ that

$$\begin{aligned} |a_k(u,v)| &\leq \int_\Omega |\nabla_k u(x)| \, |\nabla_k v(x)| \, \mathrm{d}x \\ &\leq \left(\int_\Omega |\nabla_k u(x)|^2 \, \mathrm{d}x \right)^{1/2} \left(\int_\Omega |\nabla_k v(x)|^2 \, \mathrm{d}x \right)^{1/2} \end{aligned}$$

Lemma 4.4 with $\delta = 1$ yields

$$\begin{aligned}\int_\Omega |\nabla_k u(x)|^2 \,\mathrm{d}x &\le \int_\Omega |\nabla u(x)|^2 + |ku(x)|^2 + \left| iku\overline{\nabla u(x)} + \overline{iku}\nabla u(x)\right|^2 \mathrm{d}x \\ &\le \|\nabla u\|_H^2 + |k|^2\|u\|_H^2 + \left(\|ku\|_H^2 + \|\nabla u\|_H^2\right).\end{aligned}$$

It follows that

$$|a_k(u,v)| \le C_M \|u\|_V \|v\|_V$$

where $C_M = 2\max(1, |k|^2)$. Moreover, it is straightforward to see that $a_k(u,v) = \overline{a_k(v,u)}$, which implies the Hermiticity of a_k. □

Hence, we can define an operator $\mathcal{A}_k$ mapping from the space V to its dual V^* by the equality

$$\langle \mathcal{A}_k u, \cdot\rangle_{V^*,V} := a_k(u,\cdot)$$

for all $u \in V$. The EVP (4.1) can then be reformulated as an equation in V^*:

$$\mathcal{A}_k u_k = \lambda \mathcal{I}_\epsilon u_k. \tag{4.4}$$

4.2.3. *Properties of the resolvent.* Instead of the eigenvalue problem (4.4) it is more convenient for the analysis to look at the equivalent shifted problem:

$$(\mathcal{A}_k + \sigma\mathcal{I}_\epsilon)u_k = (\lambda + \sigma)\mathcal{I}_\epsilon u_k$$

where $\sigma > 0$ is a fixed parameter. In this section we investigate the properties of the resolvant $(\mathcal{A}_k + \sigma\mathcal{I}_\epsilon)^{-1}$ of $\mathcal{A}_k$. To this end, we demonstrate that the form $a_{k,\sigma} := a_k + \sigma b_\epsilon$ corresponding to the operator $\mathcal{A}_{k,\sigma} := \mathcal{A}_k + \sigma\mathcal{I}_\epsilon$ is an elliptic sesquilinear form.

Lemma 4.6. *For any $\sigma > 0$, the sesquilinear form $a_{k,\sigma}$ is elliptic with constant*

$$C_m(\sigma) = \begin{cases} \min\{1, \sigma\epsilon_0\}, & \text{if } k = 0, \\ \min\left\{\frac{\sigma\varepsilon_0}{2|k|^2+\sigma\varepsilon_0}, \frac{\sigma}{2}\varepsilon_0\right\}, & \text{if } k \ne 0. \end{cases}$$

Proof. If $k = 0$, then $a_{0,\sigma}(u,u) = \int_\Omega |\nabla u|^2 + \sigma\epsilon(x)|u|^2 \,\mathrm{d}x \ge \min\{1, \sigma\epsilon_0\}\|u\|_V^2$. If $k \ne 0$, we observe that $a_{k,\sigma}(u,u) = \int_\Omega |(\nabla + ik)u|^2 + \sigma\epsilon(x)|u|^2 \,\mathrm{d}x \in \mathbb{R}^+$. Moreover,

$$a_{k,\sigma}(u,u) \ge \left(\|\nabla u\|_H^2 + (\sigma\epsilon_0 + |k|^2)\|u\|_H^2 - \int_\Omega \left|iku \cdot \overline{\nabla u} + \overline{iku}\cdot\nabla u\right| \,\mathrm{d}x \right)$$

and thanks to Lemma 4.4 we obtain :

$$\int_\Omega \left|iku \cdot \overline{\nabla u} + \overline{iku}\cdot\nabla u\right| \,\mathrm{d}x \le \frac{|k|^2\|u\|_H^2}{\delta} + \delta\|\nabla u\|_H^2$$

for any $\delta > 0$. Therefore,

$$a_{k,\sigma}(u,u) \ge (1-\delta)\|\nabla u\|_H^2 + \left(\sigma\varepsilon_0 + |k|^2 - \frac{|k|^2}{\delta}\right)\|u\|_H^2,$$

and choosing $\delta = \dfrac{|k|^2}{|k|^2 + \frac{\sigma}{2}\varepsilon_0} < 1$ yields

$$a_{k,\sigma}(u,u) = \underbrace{(1 - \frac{|k|^2}{|k|^2 + \frac{\sigma}{2}\varepsilon_0})}_{>0} \|\nabla u\|_H^2 + \underbrace{\frac{\sigma}{2}\varepsilon_0}_{>0} \|u\|_H^2 \ge C_m(\sigma)\|u\|_V^2$$

with $C_m(\sigma) = \min\left\{\frac{\sigma\varepsilon_0}{2|k|^2+\sigma\varepsilon_0}, \frac{\sigma}{2}\varepsilon_0\right\}$. □

Corollary 4.7. *For any $\sigma > 0$, the operator $\mathcal{A}_{k,\sigma} = (\mathcal{A}_k + \sigma\mathcal{I}_\epsilon)$ is invertible and the inverse is bounded with constant $C_m(\sigma)^{-1}$, i.e. ${\mathcal{A}_{k,\sigma}}^{-1} \in \mathcal{L}(V^*, V)$ satisfies*

$$\|{\mathcal{A}_{k,\sigma}}^{-1}\|_{V^*\to V} := \sup_{\substack{w\in V^* \\ \|w\|_{V^*}=1}} \|{\mathcal{A}_{k,\sigma}}^{-1} w\|_V \leq \frac{1}{C_m(\sigma)}.$$

Proof. This follows directly from the Lax-Milgram lemma, see e.g. [Hac92, Lem. 6.98]. □

Corollary 4.8. *For any $\sigma > 0$, the sesquilinear form $a_{k,\sigma}$ defines an inner product in V and the corresponding norm $\|v\|_{a_k+\sigma b_\epsilon} := \sqrt{a_k(v,v) + \sigma b_\epsilon(v,v)}$ is equivalent to the standard norm $\|v\|_V$.*

Proof. The positivity of $a_{k,\sigma}$ follows from Lemma 4.6, and the sesquilinearity as well as the conjugate symmetry are easily verified. Moreover Lemma 4.5 yields

$$\|v\|_{a_k+\sigma b_\epsilon} \leq \sqrt{C_M\|v\|_V^2 + \sigma b_\epsilon(v,v)} \leq \sqrt{C_M\|v\|_V^2 + \sigma\epsilon_\infty\|v\|_H^2} \leq \sqrt{C_M + \sigma\epsilon_\infty}\|v\|_V$$

and by Lemma 4.6

$$\|v\|_{a_k+\sigma b_\epsilon} \geq \sqrt{C_m(\sigma)}\|v\|_V$$

for all $v \in V$. □

Many important properties of the spectrum of $\mathcal{A}_k$ can now be proven by applying the Riesz-Schauder therory on the resolvant $\mathcal{R}_\sigma : V \to V$, $\mathcal{R} = \mathcal{A}_{k,\sigma}^{-1}\mathcal{I}_\epsilon$. These are gathered in the following lemma.

Lemma 4.9. *The spectrum of $\mathcal{A}_k$ consists in at most countably many real eigenvalues with finite geometrical multiplicities, which can only accumulate at infinity. Moreover, all eigenvalues are nonnegative.*

Proof. From the Hermiticity of a_k follows directly that $\mathcal{A}_k$ is Hemitian. Therefore all eigenvalues λ of $\mathcal{A}_k$ must be real since

$$\lambda = \frac{\langle\mathcal{A}_k u, u\rangle_{V^*,V}}{\langle\mathcal{I}_\epsilon u, u\rangle_{V^*,V}} = \frac{\overline{\langle\mathcal{A}_k u, u\rangle_{V^*,V}}}{\overline{\langle\mathcal{I}_\epsilon u, u\rangle_{V^*,V}}} = \overline{\lambda},$$

where u denotes the eigenvector corresponding to λ. The rest of the proof follows from [Hac92, Satz 6.108]. In particular,it is shown there that for all $\lambda \in \mathbb{C}$ one of the following alternatives holds:

(1) $(\mathcal{A} - \lambda\mathcal{I}_\epsilon)^{-1} \in \mathcal{L}(V^*, V)$,
(2) λ is an eigenvalue.

In Corollary 4.7 it was shown that $(\mathcal{A} - \lambda\mathcal{I}_\epsilon)^{-1}$ exists and is bounded for all $\lambda = -\sigma$, where $\sigma > 0$. Hence, all eigenvalues must be nonnegative. □

Corollary 4.10. *Let A_k and B_ϵ be the matrices in the discretized problem* (4.3)*. Then, A_k is Hermitian positive semi-definite and B_ϵ, as well as $(A_k + B_\epsilon)$, are Hermitian positive definite.*

Proof. It is straightforward to see that $(A_k)_{j,i} = a_k(\phi_i, \phi_j) = \overline{a_k(\phi_j, \phi_i)} = \overline{(A_k)_{i,j}}$ and similarly for B_ϵ. Moreover, by Lemma 4.6 it holds for all $\sigma > 0$ and for all $\mathbf{v} = [v_1, \dots, v_N]$ with $v_h = \sum_i^N v_i$

$$\mathbf{v}^* A_k \mathbf{v} = a(v_h, v_h) \geq C_m(\sigma)\|v_h\|_V - \sigma b_\epsilon(v_h, v_h) \to 0 \text{ as } \sigma \to 0,$$

where the superscript * denotes the conjugate transpose of a vector or a matrix Since this inequality holds for $\sigma > 0$ arbitrarily small, to follows that $\mathbf{v}^* A_k \mathbf{v} \geq 0$, and thus A_k is positive semi-definite. The rest of the proof follows easily. □

5. Error estimators

We wish to approximate the solution of the discretize problem (4.3) by using an iterative method. In the remainder, u denotes the exact eigenfunction of (4.1), u_h denotes the exact eigenfunction of the discretized problem (4.3), and $\widetilde{u}_h = \sum_{i=1}^{N} \widetilde{u}_i \phi_i$ denotes the approximated eigenfunction computed by m steps of an iterative method (e.g. Lanczos method or LOBPCG [GVL96, Kny01]).

In order to be able to evaluate the quality of the approximate solution $\widetilde{u}_h$ we need to construct an estimator for the error $\|u - \widetilde{u}_h\|_V$. We start by applying the triangular inequality

$$\|u - \widetilde{u}_h\|_V = \|u - u_h + u_h - \widetilde{u}_h\|_V \leq \|u - u_h\|_V + \|u_h - \widetilde{u}_h\|_V$$

and observe that the error can be separated into the *discretization* error $\|u - u_h\|_V$ and the *algebraic error* $\|u_h - \widetilde{u}_h\|_V$.

An estimator for the error introduced by the discretization of the linear EVP (3.4) to obtain (4.3) has been developed in [GG12]. In that paper it is shown that an error estimator η_{dis}, which can be computed a posteriori, gives a reliable and efficient approxmation of the error $\|u - \widetilde{u}_h\|_{a_k+\sigma b}$ where $\sigma = 1$. Since the norms $\|\cdot\|_V$ and $\|\cdot\|_{a_k+\sigma b}$ are equivalent (see Lemma 4.8), η_{dis} is also a reliable and efficient estimator of the error measured in the standard norm in V. However, the equivalence constants $C_m(\sigma)$ and C_M may be much bigger or much smaller than 1, depending on the values of σ and k.

In order to enable the error balancing between the discretization and the algebraic errors, we need an error estimator which can be compared to η_{dis} in a meaningful manner, i.e., an estimator which measures the algebraic error in the norm defined by the sesquilinear form $a_{k,\sigma}$ on V. This is the subject of the next subsection.

5.1. Algebraic error.

We aim at finding a bound for the algebraic errors $\|u_h - \tilde{u}_h\|_V$ and $\|u_h - \tilde{u}_h\|_{a_k+\sigma b}$, which depend on the convergence of an iterative numerical method. The convergence analysis of such methods is typically based on the concepts of angle between vectors. While the definition of the angle between *real* vectors with respect to a scalar product is clear, there exist multiple nonequivalent concepts of angle between *complex* vectors. Each one of these angles has different interesting properties. We start by introducing the Euclidean angle between complex vectors, which satisfies the law of cosines:

Definition 5.1. Let v and w be two vectors in a complex Hilbert space V and let $(\cdot,\cdot)_{\star}$ as well as $\|\cdot\|_{\star}$ denote an inner product in V and its corresponding norm, respectively. The *(real-valued) Euclidean* angle between v and w with respect to the $\star$-inner-product is defined as the quantity $\angle_{\star}^{(E)}(v,w) \in [0,\pi]$ satisfying

$$\angle_{\star}^{(E)}(v,w) = \operatorname{acos} \frac{\Re\big((v,w)_{\star}\big)}{\|v\|_{\star}\|w\|_{\star}}.$$

The Euclidean angle indeed satisfies the law of cosines since for all $v, w \in V$ it holds

$$\begin{aligned}\|v - w\|_{\star}^2 &= (v - w, v - w)_{\star} = \|v\|_{\star}^2 + \|w\|_{\star}^2 - (v,w)_{\star} - (w,v)_{\star} \\ &= \|v\|_{\star}^2 + \|w\|_{\star}^2 - 2\Re\big((v,w)_{\star}\big) = \|v\|_{\star}^2 + \|w\|_{\star}^2 - 2\cos\angle_{\star}^{(E)}(v,w)\|v\|_{\star}\|w\|_{\star}.\end{aligned}$$

The Euclidean angle inherits other properties from the real case, e.g. the angle between vectors pointing in opposite directions is π since $\angle_{\star}^{(E)}(v,-v) = \operatorname{acos}(-1) = \pi$. However, the Euclidean angle is not invariant under scaling since, for $\alpha \in \mathbb{C}$, $\angle_{\star}^{(E)}(\alpha v, w)$ is generally not equal to $\angle_{\star}^{(E)}(v,w)$.

On the other hand, one can define another angle between complex vectors which is invariant under scaling:

Definition 5.2. Let $v, w \in V$, $(\cdot,\cdot)_\star$ and $\|\cdot\|_\star$ be as in Definition 5.1. The *Hermitian* angle between v and w with respect to the $\star$-inner-product is defined as the quantity $\angle_\star^{(H)}(v,w) \in [0, \frac{\pi}{2}]$ satisfying:

$$\angle_\star^{(H)}(v,w) = \operatorname{acos} \frac{|(v,w)_\star|}{\|v\|_\star \|w\|_\star}.$$

The link between both angles can be expressed as follows:

$$\begin{aligned}
\inf_{\substack{v_\theta \in \operatorname{span}\{v\} \\ w_\theta \in \operatorname{span}\{w\}}} \angle_\star^{(E)}(v_\theta, w_\theta) &= \inf_{\substack{\theta_1 \in [0,2\pi) \\ \theta_2 \in [0,2\pi)}} \angle_\star^{(E)}(e^{i\theta_1} v, e^{i\theta_2} w) = \operatorname{acos} \sup_{\theta \in [0,2\pi)} \frac{\Re\Big(e^{i\theta}(v,w)_\star\Big)}{\|v\|_\star \|w\|_\star} \\
&= \operatorname{acos} \frac{1}{\|v\|_\star \|w\|_\star} \Re\Big(\frac{(w,v)_\star}{|(v,w)_\star|}(v,w)_\star\Big) \qquad (5.1) \\
&= \operatorname{acos} \frac{|(v,w)_\star|}{\|v\|_\star \|w\|_\star} = \angle_\star^{(H)}(v,w).
\end{aligned}$$

Therefore, the Hermitian angle between v and w can be thought of as the Euclidean angle between the subspaces spanned by v and w. Moreover, the Hermitian and Euclidean angles between two vectors v, w are equal if the vectors have the *same orientation in* V, i.e. if they satisfy the condition

$$(v,w)_\star \in \mathbb{R}^+.$$

In the remainder, we consider the algebraic error $\|u_h - \tilde{u}_h\|_\star$ where the inner product $(\cdot,\cdot)_\star$ represents either the standard inner product $(\cdot,\cdot)_V$ in V or the inner product $(\cdot,\cdot)_{a_k+\sigma b_\epsilon}$ linked to the weak form (4.1). Moreover, both vectors u_h and $\tilde{u}_h$ are supposed to be normalized and have the same orientation in V (with respect to the considered inner product), i.e.

$$\|u_h\|_\star = \|\tilde{u}_h\|_\star = 1 \qquad \text{and} \qquad (u_h, \tilde{u}_h)_\star \in \mathbb{R}^+.$$

Under these conditions, the law of cosines combined with the following trigonometric identity

$$\cos^2(\frac{\phi}{2}) = \left(\frac{e^{i\frac{\phi}{2}} + e^{-i\frac{\phi}{2}}}{2}\right)^2 = \frac{e^{i\phi} + e^{-i\phi} + 2}{4} = \frac{1}{2}(\cos(\phi) + 1), \qquad \text{for all } \phi \in [0, 2\pi),$$

yields

$$\begin{aligned}
\|u_h - \tilde{u}_h\|_\star^2 &= 2 - 2\cos\angle_\star^{(E)}(u_h, \tilde{u}_h) = 4\Big(1 - \frac{1}{2}\big(\cos\angle_\star^{(E)}(v,w) + 1\big)\Big) \\
&= 4\big(1 - \cos^2\frac{\angle_\star^{(E)}(v,w)}{2}\big) = 4\sin^2\frac{\angle_\star^{(E)}(v,w)}{2} = 4\sin^2\frac{\angle_\star^{(H)}(v,w)}{2}.
\end{aligned}$$

Therefore, we can bound the algebraic error by a term proportionnal to the sinus of the Hermitian angle:

$$\|u_h - \tilde{u}_h\|_\star = 2\sin\frac{\angle_\star^{(H)}(u_h, \tilde{u}_h)}{2} \leq \sqrt{2}\sin\angle_\star^{(H)}(u_h, \tilde{u}_h). \qquad (5.2)$$

For small angles, the following expression based on Taylor expansions gives a more accurate estimation of the error:

$$\begin{aligned}\|u_h - \tilde{u}_h\|_\star &= 2\sin\frac{\angle_\star^{(H)}(u_h,\tilde{u}_h)}{2} = 2\sum_{j=0}^{\infty}\frac{(-1)^j}{(2j+1)!}\Big(\frac{\angle_\star^{(H)}(u_h,\tilde{u}_h)}{2}\Big)^{2j+1}\\ &= \sum_{j=0}^{\infty}\frac{(-1)^j}{(2j+1)!}\Big(\angle_\star^{(H)}(u_h,\tilde{u}_h)\Big)^{2j+1} + \sum_{j=1}^{\infty}\frac{(-1)^j}{(2j+1)!}(1-\frac{1}{2^{2j}})\Big(\angle_\star^{(H)}(u_h,\tilde{u}_h)\Big)^{2j+1}\\ &= \sin\angle_\star^{(H)}(u_h,\tilde{u}_h) + \mathcal{O}(\angle_\star^{(H)}(u_h,\tilde{u}_h)^3) \approx \sin\angle_\star^{(H)}(u_h,\tilde{u}_h).\end{aligned}$$

5.1.1. *Equivalence of continuous and discrete norms.* In this subsection we establish equivalences between the norms of functions u_h, v_h in the finite dimensional space V_h defined in (4.2) and the corresponding vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$ expressed in the basis formed by the functions $\{\phi_i\}_{i=1}^N$. In the remainder of this paper, we assume $\sigma > 0$.

For any $u_h, v_h \in V_h$ it holds

$$(u_h, v_h)_V = \sum_{i=1}^{N}\sum_{j=1}^{N}\overline{v_j}\int_\Omega \nabla\phi_i\cdot\overline{\nabla\phi_j} + \phi_i\overline{\phi_j}\,\mathrm{d}x\ u_i = \mathbf{v}^*\Phi_V\mathbf{u}$$

and

$$\|u_h\|_V^2 = \mathbf{u}_h\Phi_V\mathbf{u} =: \|\mathbf{u}\|_{\Phi_V}^2,$$

where $(\Phi_V)_{i,j} := (\phi_j, \phi_i)_V$. Moreover,

$$(u_h, v_h)_{b_\epsilon} = \sum_{i=1}^{N}\sum_{j=1}^{N}\overline{v_j}\int_\Omega \epsilon(x)\phi_i\overline{\phi_j}\,\mathrm{d}x\ u_i = \mathbf{v}^* B_\epsilon\mathbf{u},$$

$$\|u_h\|_{b_\epsilon}^2 = \mathbf{u}^* B_\epsilon\mathbf{u} = \|\mathbf{u}\|_{B_\epsilon}^2$$

and

$$(u_h, v_h)_{a_k+\sigma b_\epsilon}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}\overline{v_j}\int_\Omega \nabla_k\phi_i\cdot\overline{\nabla_k\phi_j} + \sigma\epsilon\phi_i\overline{\phi_j}\,\mathrm{d}x\ u_i = \mathbf{v}^*(A_k+\sigma B_\epsilon)\mathbf{u},$$

$$\|u_h\|_{a_k+\sigma b_\epsilon}^2 = \mathbf{u}^*(A_k+\sigma B_\epsilon)\mathbf{u} = \|\mathbf{u}\|_{(A_k+\sigma B_\epsilon)}^2.$$

Similarly, a discrete norm that is equivalent to a norm in V^* can be found.

Lemma 5.3. *For any $r \in V^*$ such that $r = \mathcal{I}_\epsilon r_h$ for some $r_h = \sum_{j=1}^{N}\mathbf{r}_j\phi_j \in V_h$, it holds*

$$\|r\|_{V^*} = \|B_\epsilon\mathbf{r}\|_{(A_k+\sigma B_\epsilon)^{-1}} = \|\mathbf{r}\|_{B_\epsilon(A_k+\sigma B_\epsilon)^{-1}B_\epsilon},$$

where

$$\|f\|_{V^*} := \sup_{\substack{v\in V,\\ \|v\|_{a_k+\sigma b_\epsilon}=1}} |f(v)|$$

for any $f \in V^*$, and where $\mathbf{r} = [r_1, \ldots, r_N]$.

Proof. For any $v \in V$ we denote by $\pi_h(v)$ the orthogonal projection of v onto V_h with orthogonality in the inner product $(\cdot,\cdot)_{b_\epsilon}$. Then it holds

$$\begin{aligned}
\|r\|_{V^*} &= \sup_{\substack{v\in V,\\ \|v\|_{(a_k+\sigma b_\epsilon)}=1}} |(r_h, v)_{b_\epsilon}| = \sup_{\substack{v\in V,\\ \|v\|_{(a_k+\sigma b_\epsilon)}=1}} |(r_h, v-\pi_h(v))_{b_\epsilon} + (r_h, \pi_h(v))_{b_\epsilon}| \\
&= \sup_{\substack{v\in V,\\ \|v\|_{(a_k+\sigma b_\epsilon)}=1}} |(r_h, \pi_h(v))_{b_\epsilon}| = \sup_{\substack{v_h\in V_h,\\ \|v_h\|_{(a_k+\sigma b_\epsilon)}=1}} |(r_h, v_h)_{b_\epsilon}| \\
&= \sup_{\substack{\mathbf{v}\in \mathbb{C}^N,\\ \|\mathbf{v}\|_{(A_k+\sigma B_\epsilon)}=1}} |\mathbf{r}^* B_\epsilon \mathbf{v}| = \sup_{\substack{\mathbf{v}\in \mathbb{C}^N,\\ \|\mathbf{v}\|_{(A_k+\sigma B_\epsilon)}=1}} |\mathbf{r}^* B_\epsilon (A_k+\sigma B_\epsilon)^{-1}(A_k+\sigma B_\epsilon)\mathbf{v}|.
\end{aligned}$$

For the sake of brevity, we introduce the notation $A_{k,\sigma} := A_k + \sigma B_\epsilon$ and $\mathbf{b} := A_{k,\sigma}^{-1} B_\epsilon \mathbf{r}$. Using the Cauchy-Schwarz inequality [Bre11, Ch. 5.1] yields

$$\|r\|_{V^*} = \sup_{\substack{\mathbf{v}\in \mathbb{C}^N,\\ \|\mathbf{v}\|_{A_{k,\sigma}}=1}} |\mathbf{b}^* A_{k,\sigma}\mathbf{v}| \le \sup_{\substack{\mathbf{v}\in \mathbb{C}^N,\\ \|\mathbf{v}\|_{A_{k,\sigma}}=1}} \|\mathbf{b}\|_{A_{k,\sigma}} \|\mathbf{v}\|_{A_{k,\sigma}} = \sup_{\substack{\mathbf{v}\in \mathbb{C}^N,\\ \|\mathbf{v}\|_{A_{k,\sigma}}=1}} \|\mathbf{b}\|_{A_{k,\sigma}} = \|\mathbf{b}\|_{A_{k,\sigma}}.$$

The upper bound is attained by setting $\mathbf{v} = \frac{\mathbf{b}}{\|\mathbf{b}\|_{A_{k,\sigma}}}$ which satisfies the constraint $\|\mathbf{v}\|_{A_{k,\sigma}} = 1$. Therefore

$$\|r\|_{V^*} = \|\mathbf{b}\|_{A_{k,\sigma}} = \left\|A_{k,\sigma}^{-1} B_\epsilon \mathbf{r}\right\|_{A_{k,\sigma}} = \|B_\epsilon \mathbf{r}\|_{A_{k,\sigma}^{-1}} = \|\mathbf{r}\|_{B_\epsilon (A_k+\sigma B_\epsilon)^{-1} B_\epsilon}. \qquad \square$$

We show that the trigonometric functions of angles between vectors in V_h can also be expressed in terms of the corresponding discrete vectors in $\mathbb{C}^N$.

Theorem 5.4. *Let $v_h, w_h \in V_h \subset V$ with $v_h = \sum_{i=1}^{N} v_i \phi_i$, $w_h = \sum_{i=1}^{N} w_i \phi_i$ and $\mathbf{v} = [v_1, \dots, v_N]^T$, $\mathbf{w} = [w_1, \dots, w_N]^T$. Then*

$$\cos \angle_V^{(H)}(v_h, w_h) = \cos \angle_{\Phi_V}^{(H)}(\mathbf{v}, \mathbf{w})$$

and

$$\cos \angle_{a_k+\sigma b_\epsilon}^{(H)}(v_h, w_h) = \cos \angle_{A_k+\sigma B_\epsilon}^{(H)}(\mathbf{v}, \mathbf{w})$$

where

$$\angle_M^{(H)}(\mathbf{v}, \mathbf{w}) := \operatorname{acos} \frac{|\mathbf{w}^* M \mathbf{v}|}{\sqrt{\mathbf{v}^* M \mathbf{v}}\sqrt{\mathbf{w}^* M \mathbf{w}}} \in \left[0, \frac{\pi}{2}\right]$$

for any matrix M.

Proof.

$$\begin{aligned}
\cos \angle_V^{(H)}(v_h, w_h) &= \frac{|(v_h, w_h)_V|}{\|v_h\|_V \|w_h\|_V} = \frac{|\sum_{i,j} \overline{w_i} (\phi_j, \phi_i)_V v_j|}{\sqrt{\sum_{i,j} \overline{v_i} (\phi_j, \phi_i)_V v_j}\sqrt{\sum_{i,j} \overline{w_i} (\phi_j, \phi_i)_V w_j}} \\
&= \frac{|\mathbf{w}^* \Phi_V \mathbf{v}|}{\sqrt{\mathbf{v}^* \Phi_V \mathbf{v}}\sqrt{\mathbf{w}^* \Phi_V \mathbf{w}}}
\end{aligned}$$

and similarly for the inner product $a_k + \sigma b_\epsilon$ on V. $\square$

Corollary 5.5. *Let $v_h, w_h, \mathbf{v}, \mathbf{w}$ as in Theorem 5.4. Then*

$$\sin \angle_V^{(H)}(v_h, w_h) = \sin \angle_{\Phi_V}^{(H)}(\mathbf{v}, \mathbf{w})$$

and

$$\sin \angle_{A_k+\sigma B_\epsilon}^{(H)}(v_h, w_h) = \sin \angle_{a_k+\sigma b_\epsilon}^{(H)}(\mathbf{v}, \mathbf{w}).$$

5.1.2. *Sine theorems.* In order to obtain a useful bound on the algebraic error, we need a computable bound on $\sin \angle^{(H)}_{\Phi_V}(\mathbf{v},\mathbf{w})$ or $\sin \angle^{(H)}_{A_k+\sigma B_\epsilon}(\mathbf{v},\mathbf{w})$. First, we recall a result for the (non-generalized) linear eigenvalue problem.

Theorem 5.6 (The sine theorem)**.** *Let $M \in \mathbb{C}^{N\times N}$ be a Hermitian positive definite matrix defining an inner product $(\mathbf{v_1},\mathbf{v_2})_M$ such that*

$$(\mathbf{v_1},\mathbf{v_2})_M := \mathbf{v_1}^* M \mathbf{v_2}$$

for all $\mathbf{v_1},\mathbf{v_1} \in \mathbb{C}^N$. Let $A \in \mathbb{C}^{N\times N}$ be a Hermitian matrix with respect to this inner product. We denote by $\{\lambda_i\}_{i=1}^N$ the eigenvalues of A and by $\{\mathbf{u}^{(i)}\}_{i=1}^N$ the corresponding normalized eigenvectors of A in the norm $\|\cdot\|_M$ induced by $(\cdot,\cdot)_M$. Given any vector $\mathbf{v} \in \mathbb{C}^N$ and any scalar $\widetilde{\lambda} \in \mathbb{C}$, we order the eigenpairs $(\lambda_i,\mathbf{u}^{(i)})$ such that $|\lambda_1 - \widetilde{\lambda}| \le |\lambda_2 - \widetilde{\lambda}| \le \ldots \le |\lambda_N - \widetilde{\lambda}|$ and we define the residual $\mathbf{r} := A\widetilde{\mathbf{v}} - \widetilde{\lambda}\widetilde{\mathbf{v}}$ where $\widetilde{\mathbf{v}} := \alpha\mathbf{v}$ is the scaled vector which is normalized and lies in the same directions as $\mathbf{u}^{(1)}$, i.e.

$$\|\widetilde{\mathbf{v}}\|_M = 1, \qquad (\widetilde{\mathbf{v}},\mathbf{u}^{(1)})_M \in \mathbb{R}^+.$$

Then it holds

$$\sin \angle^{(H)}_M(\widetilde{\mathbf{v}},\mathbf{u}^{(1)}) \le \frac{\|\mathbf{r}\|_M}{\delta}.$$

where $\delta := |\lambda_2 - \widetilde{\lambda}|$.

Proof. We follow the proof of Theorem 11.7.1 in [Par98] which proves the result for real symmetric matrices and we extend it to the case of complex Hermitian matrices.

First recall that the eigenvectors of A form an orthogonal basis of $\mathbb{C}^n$ since A is Hermitian (see the *spectral theorem* in [Par98, Ch. 1.4]). Therefore we can expand

$$\widetilde{\mathbf{v}} = \sum_{j=1}^N (\widetilde{\mathbf{v}},\mathbf{u}^{(j)})_M \mathbf{u}^{(j)},$$

which implies

$$\begin{aligned}\mathbf{r} &= (A-\widetilde{\lambda})(\widetilde{\mathbf{v}},\mathbf{u}^{(1)})_M\mathbf{u}^{(1)} + (A-\widetilde{\lambda})\sum_{j=2}^N (\widetilde{\mathbf{v}},\mathbf{u}^{(j)})_M\mathbf{u}^{(j)} \\ &= (\lambda_1-\widetilde{\lambda})(\widetilde{\mathbf{v}},\mathbf{u}^{(1)})_M\mathbf{u}^{(1)} + \sum_{j=2}^N (\lambda_j-\widetilde{\lambda})(\widetilde{\mathbf{v}},\mathbf{u}^{(j)})_M\mathbf{u}^{(j)}.\end{aligned}$$

Since the eigenvectors are orthogonal to each other in the inner product $(\cdot,\cdot)_M$, we can apply the Pythagoreus theorem, which yields:

$$\begin{aligned}\|\mathbf{r}\|_M^2 &= |\lambda_1-\widetilde{\lambda}|^2|(\widetilde{\mathbf{v}},\mathbf{u}^{(1)})_M|^2 + \sum_{j=2}^N |\lambda_j-\widetilde{\lambda}|^2|(\widetilde{\mathbf{v}},\mathbf{u}^{(j)})_M|^2 \\ &\ge |\lambda_1-\widetilde{\lambda}|^2\cos^2\angle^{(H)}_M(\widetilde{\mathbf{v}},\mathbf{u}^{(1)}) + |\lambda_2-\widetilde{\lambda}|^2\sum_{j=2}^N |(\widetilde{\mathbf{v}},\mathbf{u}^{(j)})_M|^2 \\ &= |\lambda_1-\widetilde{\lambda}|^2\cos^2\angle^{(H)}_M(\widetilde{\mathbf{v}},\mathbf{u}^{(1)}) + |\lambda_2-\widetilde{\lambda}|^2(\|\widetilde{\mathbf{v}}\|^2 - |(\widetilde{\mathbf{v}},\mathbf{u}^{(1)})_M|^2) \\ &= |\lambda_1-\widetilde{\lambda}|^2\cos^2\angle^{(H)}_M(\widetilde{\mathbf{v}},\mathbf{u}^{(1)}) + |\lambda_2-\widetilde{\lambda}|^2\sin^2\angle^{(H)}_M(\widetilde{\mathbf{v}},\mathbf{u}^{(1)}) \\ &\ge \delta^2\sin^2\angle^{(H)}_M(\widetilde{\mathbf{v}},\mathbf{u}^{(1)}).\end{aligned}$$

□

In order to apply Theorem 5.6 to Problem (4.3) we need to reformulate the latter as a standard (non-generalized) EVP. Since the matrix B_ϵ in (4.3) is Hermitian positive definite, it is also invertible and the EVP can be equivalently written as

$$B_\epsilon^{-1} A_k u = \lambda u.$$

Corollary 5.7. *Let A_k, B_ϵ be the Hermitian matrices defining the EVP* (4.3)*. We denote by $\{\lambda_i\}_{i=1}^N$ the eigenvalues of the pencil (A_k, B_ϵ) and by $\{\mathbf{u}^{(i)}\}_{i=1}^N$ the corresponding normalized eigenvectors of (A_k, B_ϵ) in the norm $\|\cdot\|_{A_k+\sigma B_\epsilon}$ where $\sigma > 0$. Given any vector $\mathbf{v} \in \mathbb{C}^N$ and any scalar $\widetilde{\lambda} \in \mathbb{C}$, we order the eigenpairs $(\lambda_i, \mathbf{u}^{(i)})$ such that $|\lambda_1 - \widetilde{\lambda}| \le |\lambda_2 - \widetilde{\lambda}| \le \ldots \le |\lambda_N - \widetilde{\lambda}|$ and we define the residual $\mathbf{r} := A_k\widetilde{\mathbf{v}} - \widetilde{\lambda} B_\epsilon \widetilde{\mathbf{v}}$ where $\widetilde{\mathbf{v}} := \alpha\mathbf{v}$ is the scaled vector which is normalized and lies in the same directions as $\mathbf{u}^{(1)}$, i.e.*

$$\|\widetilde{\mathbf{v}}\|_{A_k+\sigma B_\epsilon} = 1, \qquad (\widetilde{\mathbf{v}}, \mathbf{u}^{(1)})_{A_k+\sigma B_\epsilon} \in \mathbb{R}^+.$$

Then it holds

$$\sin \angle^{(H)}_{A_k+\sigma B_\epsilon}(\widetilde{\mathbf{v}}, \mathbf{u}^{(1)}) \le \frac{\|B_\epsilon^{-1}\mathbf{r}\|_{A_k+\sigma B_\epsilon}}{\delta} = \frac{\|\mathbf{r}\|_{B_\epsilon^{-1}(A_k+\sigma B_\epsilon)B_\epsilon^{-1}}}{\delta}. \tag{5.3}$$

where $\delta := |\lambda_2 - \widetilde{\lambda}|$.

Proof. Observe that the matrix $B_\epsilon^{-1}A_k$ is Hermitian in the inner product defined by $A_k + \sigma B_\epsilon$ since for any $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{C}^N$

$$\begin{aligned}(B_\epsilon^{-1}A_k\mathbf{v}_1, \mathbf{v}_2)_{A_k+\sigma B_\epsilon} &= \mathbf{v}_1^* A_k B_\epsilon^{-1}(A_k+\sigma B_\epsilon)\mathbf{v}_2 = \mathbf{v}_1^* A_k B_\epsilon^{-1} A_k \mathbf{v}_2 + \sigma \mathbf{v}_1^* A_k \mathbf{v}_2 \\ &= \mathbf{v}_1^*(A_k+\sigma B_\epsilon) B_\epsilon^{-1} A_k \mathbf{v}_2 = (\mathbf{v}_1, B_\epsilon^{-1} A_k \mathbf{v}_2)_{(A_k+\sigma B_\epsilon)}.\end{aligned}$$

Therefore Theorem 5.6 can be applied with $A := B_\epsilon^{-1}A_k$, $M := (A_k + \sigma B_\epsilon)$. □

Alternatively, one can follow [HL06, Section 2] and solve the EVP (4.3) with a shift-invert method. Recall that for any real positive shift $\sigma > 0$, the matrix $(A_k + \sigma B_\epsilon)$ is invertible, see Corollary 4.7. Therefore (4.3) is equivalent to

$$(A_k + \sigma B_\epsilon)^{-1} B_\epsilon \mathbf{u} = \nu \mathbf{u}, \tag{5.4}$$

where the eigenvalues of both problems are related via the equation

$$\nu = \frac{1}{\lambda + \sigma}.$$

The matrix $(A_k + \sigma B_\epsilon)^{-1}B_\epsilon$ is Hermitian in the $(A_k + \sigma B_\epsilon)$-inner product since for all $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{C}^N$ it holds

$$\begin{aligned}\big((A_k+\sigma B_\epsilon)^{-1}B_\epsilon\mathbf{v}_1, \mathbf{v}_2\big)_{A_k+\sigma B_\epsilon} &= \mathbf{v}_1^* B_\epsilon (A_k+\sigma B_\epsilon)^{-1}(A_k+\sigma B_\epsilon)\mathbf{v}_2 = \mathbf{v}_1^* B_\epsilon \mathbf{v}_2 \\ &= \big(\mathbf{v}_1, (A_k+\sigma B_\epsilon)^{-1}B_\epsilon\mathbf{v}_2\big)_{A_k+\sigma B_\epsilon}\end{aligned}$$

We can now apply Theorem 5.6 to (5.4).

Corollary 5.8. *Let A_k, B_ϵ be the Hermitian matrices defining the EVP* (4.3) *and $\sigma > 0$ be a real and positive scalar shift. We denote by $\{\lambda_i\}_{i=1}^N$ the eigenvalues of the Hermitian pencil (A_k, B_ϵ) and by $\{\mathbf{u}^{(i)}\}_{i=1}^N$ the corresponding normalized eigenvectors of (A_k, B_ϵ). Given any vector $\mathbf{v} \in \mathbb{C}^N$ and any scalar $\widetilde{\lambda} \in \mathbb{C}$, we order the eigenpairs $(\lambda_i, \mathbf{u}^{(i)})$ such that*

$$\left|\frac{1}{\lambda_1+\sigma} - \frac{1}{\widetilde{\lambda}+\sigma}\right| \le \left|\frac{1}{\lambda_2+\sigma} - \frac{1}{\widetilde{\lambda}+\sigma}\right| \le \ldots \le \left|\frac{1}{\lambda_N+\sigma} - \frac{1}{\widetilde{\lambda}+\sigma}\right|,$$

and we define the residual $\mathbf{r} := A_k\widetilde{\mathbf{v}} - \widetilde{\lambda}B_\epsilon\widetilde{\mathbf{v}}$*, where* $\widetilde{\mathbf{v}} := \alpha\mathbf{v}$ *is normalized and lies in the same directions as* $\mathbf{u}^{(1)}$*, i.e.*

$$\|\widetilde{\mathbf{v}}\|_{A_k+\sigma B_\epsilon} = 1, \qquad (\widetilde{\mathbf{v}}, \mathbf{u}^{(1)})_{A_k+\sigma B_\epsilon} \in \mathbb{R}^+.$$

Then it holds

$$\sin\angle^{(H)}_{A_k+\sigma B_\epsilon}(\mathbf{u}^{(1)},\widetilde{\mathbf{v}}) \le \frac{|\lambda_2+\sigma|}{|\lambda_2-\widetilde{\lambda}|}\|\mathbf{r}\|_{(A_k+\sigma B_\epsilon)^{-1}}. \tag{5.5}$$

Proof. Applying Theorem 5.6 with $A := (A_k+\sigma B_\epsilon)^{-1}B_\epsilon$, $M = A_k+\sigma B_\epsilon$ and $\widetilde{\nu} := (\widetilde{\lambda}+\sigma)^{-1}$ yields

$$\begin{aligned}\sin\angle^{(H)}_{A_k+\sigma B_\epsilon}(\mathbf{u}^{(1)},\widetilde{\mathbf{v}}) &\le \frac{\|(A_k+\sigma B_\epsilon)^{-1}B_\epsilon\widetilde{\mathbf{v}} - \widetilde{\nu}\widetilde{\mathbf{v}}\|_{(A_k+\sigma B_\epsilon)}}{|\widetilde{\nu}-\nu_2|} \\ &= \frac{|\widetilde{\nu}|}{|\widetilde{\nu}-\nu_2|}\|\widetilde{\nu}^{-1}B_\epsilon\widetilde{\mathbf{v}} - (A_k+\sigma B_\epsilon)\widetilde{\mathbf{v}}\|_{(A_k+\sigma B_\epsilon)^{-1}} \\ &= \frac{|\lambda_2+\sigma|}{|\lambda_2-\widetilde{\lambda}|}\|\widetilde{\lambda}B_\epsilon\widetilde{\mathbf{v}} - A_k\widetilde{\mathbf{v}}\|_{(A_k+\sigma B_\epsilon)^{-1}}.\end{aligned}$$ □

Remark 5.9. Observe that the norm of the residual $\mathbf{r}$ in the upper bound (5.5) can be rewritten as

$$\|\mathbf{r}\|_{(A_k+\sigma B_\epsilon)^{-1}} = \|\widetilde{\lambda}B_\epsilon\widetilde{\mathbf{u}} - A_k\widetilde{\mathbf{u}}\|_{(A_k+\sigma B_\epsilon)^{-1}} = \|\widetilde{\lambda}\widetilde{\mathbf{u}} - B_\epsilon^{-1}A_k\widetilde{\mathbf{u}}\|_{B_\epsilon(A_k+\sigma B_\epsilon)^{-1}B_\epsilon}.$$

Therefore, according to Lemma 5.3, $\|\mathbf{r}\|_{(A_k+\sigma B_\epsilon)^{-1}}$ can be seen as the norm of

$$\operatorname{res}(\widetilde{\lambda},\widetilde{\mathbf{u}}) := \mathcal{I}_\epsilon\left(\sum_{i=1}^{N}\left(B_\epsilon^{-1}\mathbf{r}\right)_i\phi_i\right) \tag{5.6}$$

in the dual space V^*, i.e.

$$\|\mathbf{r}\|_{(A_k+\sigma B_\epsilon)^{-1}} = \|\operatorname{res}(\widetilde{\lambda},\widetilde{\mathbf{u}})\|_{V^*}$$

Note that (5.6) indeed defines a residual, since for all $v_h = \sum_{i=1}^N \mathbf{v}_i\phi_i \in V_h$ it holds that

$$\begin{aligned}\langle\operatorname{res}(\widetilde{\lambda},\widetilde{\mathbf{u}}), v_h\rangle_{V^*,V} &= \int_\Omega \epsilon(x)\sum_{i=1}^{N}\left(B_\epsilon^{-1}\mathbf{r}\right)_i\phi_i\overline{v_h}\,\mathrm{d}x = \sum_{i,j=1}^{N}\left(B_\epsilon^{-1}\mathbf{r}\right)_i\overline{\mathbf{v}_j}\int_\Omega\epsilon(x)\phi_i\overline{\phi_j}\,\mathrm{d}x \\ &= \sum_{i,j=1}^{N}\left(B_\epsilon^{-1}\mathbf{r}\right)_i\overline{\mathbf{v}_j}(B_\epsilon)_{j,i} = \mathbf{v}^*B_\epsilon B_\epsilon^{-1}\mathbf{r} = \mathbf{v}^*\mathbf{r} = \mathbf{v}^*(A_k\widetilde{\mathbf{u}} - \widetilde{\lambda}B_\epsilon\widetilde{\mathbf{u}}) \\ &= a_k(\widetilde{u}_h, v_h) - \widetilde{\lambda}b_\epsilon(\widetilde{u}_h, v_h).\end{aligned}$$

where $\widetilde{u}_h := \sum_{j=1}^N \widetilde{\mathbf{u}}_j\phi_j$

Remark 5.10. Since $(A_k+\sigma B_\epsilon)$ is Hermitian positive definite we can factorize it with help of the Choleski decomposition $(A_k+\sigma B_\epsilon) = LL^*$ where L is invertible. Therefore the term appearing on the right side of (5.3) can be rewritten as

$$\|B_\epsilon^{-1}A_k\widetilde{\mathbf{v}} - \widetilde{\lambda}\widetilde{\mathbf{v}}\|_{A_k+\sigma B_\epsilon} = \|L^*B_\epsilon^{-1}A_kL^{-*}\widetilde{\mathbf{w}} - \widetilde{\lambda}\widetilde{\mathbf{w}}\|$$

with $L^*\widetilde{\mathbf{v}} = \widetilde{\mathbf{w}}$. Observe that the matrix $L^*B_\epsilon^{-1}A_kL^{-*}$ is Hermitian since

$$\begin{aligned}(L^*B_\epsilon^{-1}A_kL^{-*})^* &= L^{-1}A_kB_\epsilon^{-1}L = L^{-1}A_kB_\epsilon^{-1}(A_k+\sigma B_\epsilon)L^{-*}L^{-1}L \\ &= L^{-1}(A_k+\sigma B_\epsilon)B_\epsilon^{-1}A_kL^{-*} = L^*B_\epsilon^{-1}A_kL^{-*}.\end{aligned}$$

Therefore, if $\widetilde{\mathbf{w}}$ is the approximation of the eigenvector corresponding to the smallest eigenvalue of $L^*B_\epsilon^{-1}A_kL^{-*}$ returned after m steps of an inverse Lanczos or LOBPCG method, then $\|L^*B_\epsilon^{-1}A_kL^{-*}\widetilde{\mathbf{w}} - \widetilde{\lambda}\widetilde{\mathbf{w}}\|$ is directly available as a by-product of each iteration, see e.g. [GVL96, Thm. 9.1.2] and [Kny01].

Similarly the term on the right hand side of (5.5) can be rewritten as

$$\|\widetilde{\lambda}B_\epsilon\widetilde{\mathbf{v}} - A_k\widetilde{\mathbf{v}}\|_{(A_k+\sigma B_\epsilon)^{-1}} = \|L^{-1}A_kL^{-*}\widetilde{\mathbf{w}} - \widetilde{\lambda}L^{-1}B_\epsilon L^{-*}\widetilde{\mathbf{w}}\|$$

and can be obtained as a by-product of a LOBPCG iteration on the Hermitian pencil $(L^{-1}A_kL^{-*}, L^{-1}B_\epsilon L^{-*})$.

Remark 5.11. In Lanczos' method and LOBPCG, the eigenvalue converges faster than the eigenvector, therefore it is reasonable to approximate $\lambda \approx \widetilde{\lambda}$ and $\lambda_2 \approx \widetilde{\lambda}_2$ to compute the error bound.

All results of this section can be summarized in the following theorem:

Theorem 5.12. *Let (A_k, B_ϵ) be the matrix pencil defining the discretized problem* (4.3) *and let $\mathbf{u}$ be the eigenvector corresponding to the smallest eigenvalue of (A_k, B_ϵ). We write L the triangular matrix satisfying the Choleski decompostion $LL^* = A_k + \sigma B_\epsilon$ for some shift $\sigma > 0$. We denote by $\widetilde{\mathbf{w}} =: L^*\widetilde{\mathbf{u}}$ the approximation of $L^*\mathbf{u}$ obtained after m iterations of the LOBPCG method applied to the pencil $(L^{-1}A_kL^{-*}, L^{-1}B_\epsilon L^{-*})$, and assume $\mathbf{u}$ and $\widetilde{\mathbf{u}}$ to be scaled in such a way that*

$$\|\mathbf{u}\|_{A_k+\sigma B_\epsilon} = \|\widetilde{\mathbf{u}}\|_{A_k+\sigma B_\epsilon} = 1, \qquad (\widetilde{\mathbf{u}}, \mathbf{u})_{A_k+\sigma B_\epsilon} \in \mathbb{R}^+.$$

Let $\widetilde{\lambda}$ and $\widetilde{\lambda}_2$ be the approximations of the eigenvalue λ closest to σ and, respectively, of the eigenvalue λ_2 second closest t σ, of Problem (4.3) *obtained from the LOBPCG iterations. Then the distance between the functions $u_h = \sum_{j=1}^N \mathbf{u}_j\phi_j \in V_h$ and $\widetilde{u}_h = \sum_{j=1}^N \widetilde{\mathbf{u}}_j\phi_j \in V_h$ can be approximately bounded by*

$$\begin{aligned}\|u_h - \tilde{u}_h\|_V &\leq \sqrt{C_m(\sigma)}^{-1}\|u_h - \tilde{u}_h\|_{a_k+\sigma b_\epsilon} = \sqrt{C_m(\sigma)}^{-1}\sin\angle^{(H)}_{a_k+\sigma b_\epsilon}(u_h, \widetilde{u}_h) + \mathcal{O}(\angle^{(H)}_{a_k+\sigma b_\epsilon}(u_h, \widetilde{u}_h)^3)\\ &\approx \sqrt{C_m(\sigma)}^{-1}\sin\angle^{(H)}_{a_k+\sigma b_\epsilon}(u_h, \widetilde{u}_h) \leq \sqrt{C_m(\sigma)}^{-1}\frac{|\lambda_2+\sigma|}{|\lambda_2-\widetilde{\lambda}|}\|A_k\widetilde{\mathbf{u}} - \widetilde{\lambda}B_\epsilon\widetilde{\mathbf{u}}\|_{(A_k+\sigma B_\epsilon)^{-1}}\\ &\approx \sqrt{C_m(\sigma)}^{-1}\frac{|\widetilde{\lambda}_2+\sigma|}{|\widetilde{\lambda}_2-\widetilde{\lambda}|}\|L^{-1}A_kL^{-*}\widetilde{\mathbf{w}} - \widetilde{\lambda}L^{-1}B_\epsilon L^{-*}\widetilde{\mathbf{w}}\|\end{aligned}$$

when $\angle^{(H)}_{a_k+\sigma b_\epsilon}(u_h, \widetilde{u}_h)$ is a small angle.

5.2. **Numerical experiments.** In this section, we compare the error estimator

$$\eta_{alg} := \frac{|\widetilde{\lambda}_2+\sigma|}{|\widetilde{\lambda}_2-\widetilde{\lambda}|}\|L^{-1}A_kL^{-*}\widetilde{\mathbf{w}} - \widetilde{\lambda}L^{-1}B_\epsilon L^{-*}\widetilde{\mathbf{w}}\| \tag{5.7}$$

to the error $\|\mathbf{w}_{eig} - \widetilde{\mathbf{w}}\| \approx \|u_h - \tilde{u}_h\|_{a_k+\sigma b_\epsilon}$ where $\mathbf{w}_{eig}$ is the approximated eigenvector returned by the MATLAB function `eig` and is therefore assumed here to be a very good approximation of the exact eigenvector $\mathbf{w} = L^*\mathbf{u}$. We observe that η_{alg} is composed of a factor $\rho = |\widetilde{\lambda}_2 + \sigma|/|\widetilde{\lambda}_2 - \widetilde{\lambda}|$ depending only on the approximations of the eigenvalues, and of a factor corresponding to the norm or the residual. In this section, we denote with a superscript $^{(m)}$ a quantity computed with the approximations $\widetilde{\lambda}, \widetilde{\lambda}_2, \widetilde{\mathbf{w}}$ obtained after m iterations.

We have tested the error estimator (5.7) on a problem of the form (4.3) with $\epsilon(x, \omega)$ as in subsection 2.3.1, Ω being a square cell of length $l_\Omega = 500 \cdot 10^{-9}$m, $\Omega_1 \subset \Omega$ being a centered disc of radius $0.2l_\Omega$, $\epsilon_1(\omega) \equiv \epsilon_0 = 8.854187817 \cdot 10^{-6}$ μF/m, $\epsilon_2 \equiv 8.9\epsilon_0$. We have set $\sigma = 1$, and Figures 5.1 and 5.2 correspond to $k = [0.6888, 6.2832]^T \cdot 10^6$ m^{-1}. The problem was discretized via the C^{++} library Concepts [FL02] with quadratic Lagrange basis functions. The dimension of the resulting matrices is 896×896. We solved the Hermitian generalized EVP with the LOBPCG algorithm.

On Figure 5.1 we see that the error estimator does not give reliable estimation of the error during the first dozen of iterations, i.e. when the iterative algorithm is far from convergence. However, after approximately 25 iterations, the error estimator starts mimicking the decrease of the error. We call this behavior *asymptotic*, as opposed to the *transient* behavior happening in the first steps of the iterative process. We observe that there exists a constant shift M between the graph of the error estimator and the one of the error in the asymptotic regime, which corresponds to approximately M =13 iterations.

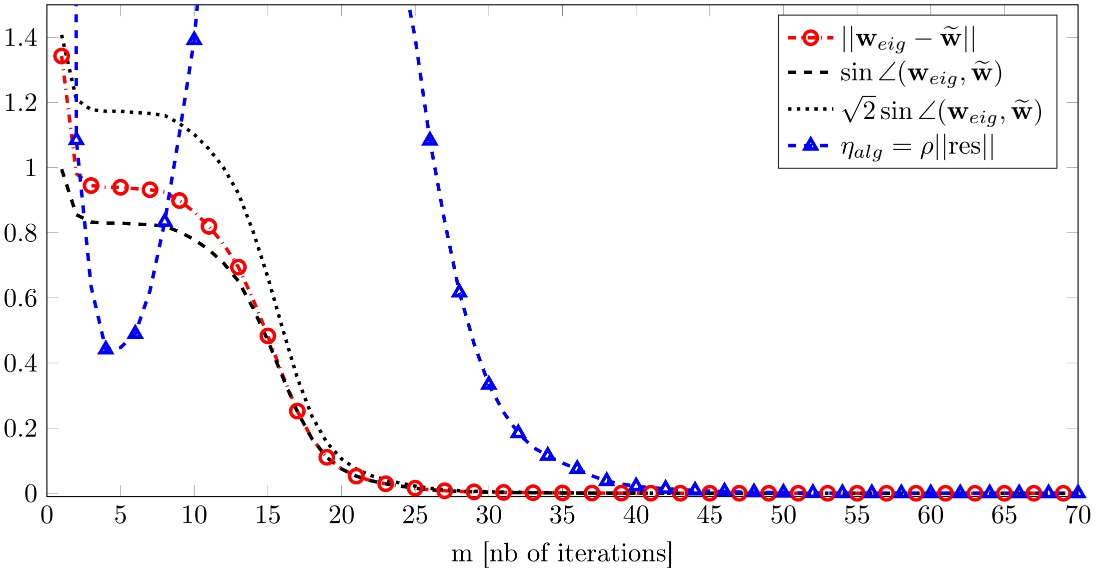


FIGURE 5.1. Comparison of the error with the error estimator

On Figure 5.2 we see that the error, the error estimator, and the norm of the residual indeed decrease at the same speed, and that the beginning of the asymptotic regime correlates with the stabilization of the factor ρ at $m \approx$. This correlation is made clearer in Figure 5.3, where the rates of change of the factors ρ and $\|res\|$ are plotted. Indeed, we see that the biggest m at which $r_\rho^{(m)} := \frac{|\rho^{(m)} - \rho^{(m-1)}|}{\rho^{(m)}}$ is bigger than $r_{res}^{(m)} := \frac{|\|res^{(m)}\| - \|res^{(m-1)}\||}{\|res^{(m)}\|}$ is also $m \approx 20$. In the range $m > 20$, $r_\rho^{(m)}$ becomes significantly smaller than $r_{res}^{(m)}$.

Note that $r_\rho^{(m)}$ and $r_{res}^{(m)}$ can be cheaply evaluated at each iteration step as a complement of the error estimator η_{alg}. Comparing both rates of change gives a way to decide whether the error estimator is reliable or not. The efficiency of the error estimator, however, depends on the shift M, and it is for the moment not clear how to estimate this quantity.

We have made a qualitatively similar observations for other values of k on the boundary of the irreducible Brillouin zone. However, the shift M observed between the error and the error estimator in the asymptotic domain does not stay constant with k but varies within the range $6 - 20$ iterations.

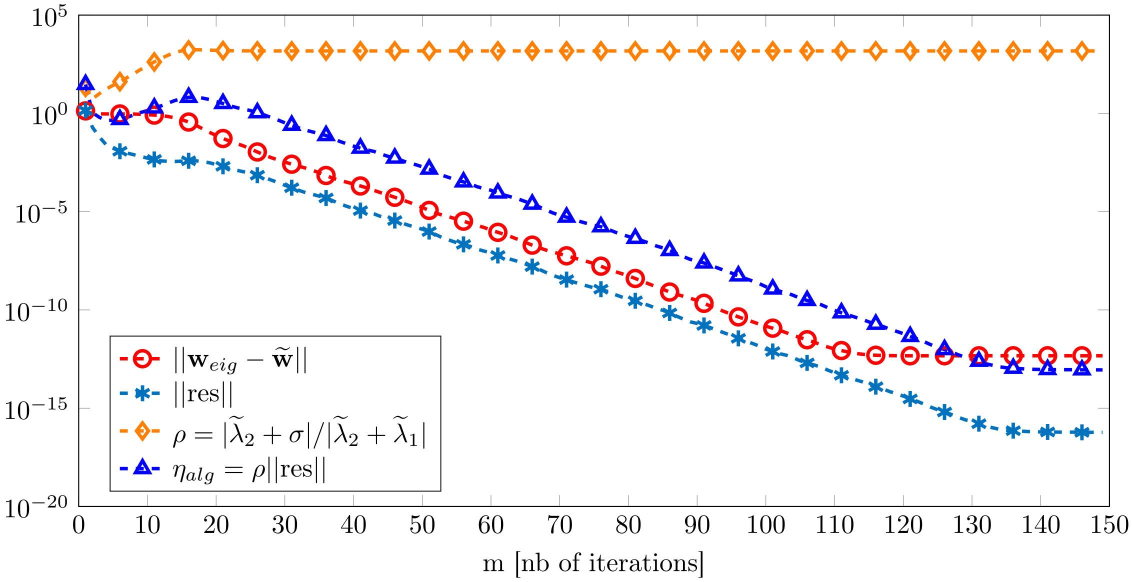


FIGURE 5.2. Evolution of the factors composing the error estimator

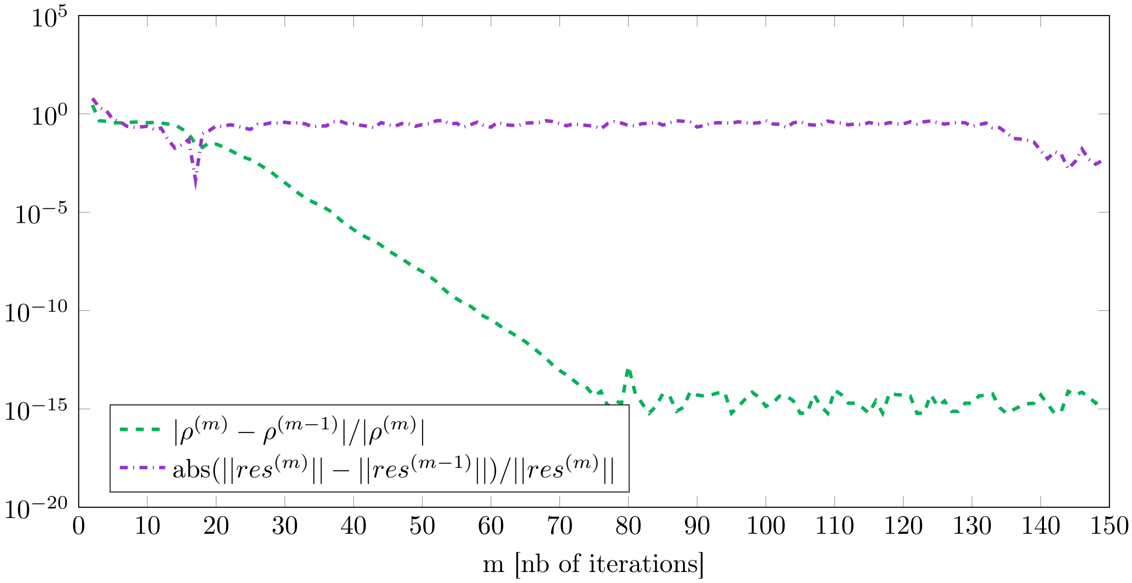


FIGURE 5.3. Rates of change of the factors composing the error estimator

6. CONCLUSION

We have developed a residual-based estimator which approximates the algebraic error resulting from the solution of the discrete generalized linear EVP by an iterative method. The error estimator approximates the error in the norm arising from the sesquilinear forms defining the EVP, and can therefore by combined in an error balancing procedure with already existing estimators of the discretization error. We have shown how to heuristically determine whether enough iterations have been computed for the error estimator to be reliable.

References

[BCG06] D. Boffi, M. Conforti, and L. Gastaldi. Modified edge finite elements for photonic crystals. *Numer. Math.*, 105(2):249–266, 2006.

[Bre11] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, New York, 2011.

[CDM$^+$17] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralí k. Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: conforming approximations. *SIAM J. Numer. Anal.*, 55(5):2228–2254, 2017.

[EKE12] C. Effenberger, D. Kressner, and C. Engström. Linearization techniques for band structure calculations in absorbing photonic crystals. *Int. J. Numer. Meth. Engn*, 89(2):180–191, 2012.

[ELT17] Christian Engström, Heinz Langer, and Christiane Tretter. Rational eigenvalue problems and applications to photonic crystals. *J. Math. Anal. Appl.*, 445(1):240–279, 2017.

[Eng10] C. Engström. On the spectrum of a holomorphic operator-valued function with applications to absorptive photonic crystals. *Math. Models Methods Appl. Sci.*, 20(8):1319–1341, 2010.

[FL02] P. Frauenfelder and C. Lage. Concepts - An Object-Oriented Software Package for Partial Differential Equations. *ESAIM: M2AN*, 36(5):937–951, 2002.

[GG12] S. Giani and I. G. Graham. Adaptive finite element methods for computing band gaps in photonic crystals. *Numer. Math.*, 121(1):31–64, 2012.

[GVL96] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[Hac92] W. Hackbusch. *Elliptic Differential Equations: Theory and Numerical Treatment*. Springer, Berlin, 1992.

[HL06] U. L. Hetmaniuk and R. B. Lehoucq. Uniform accuracy of eigenpairs from a shift-invert Lanczos method. *SIAM J. Matrix Anal. Appl.*, 28(4):927–948, 2006.

[HLM16] T.-M. Huang, W.-W. Lin, and V. Mehrmann. A Newton-type method with non-equivalence deflation for nonlinear eigenvalue problems arising in photonic crystal modeling. *SIAM J. Sci. Comput.*, 38(2):B191–B218, 2016.

[Jac99] J. D. Jackson. *Classical electrodynamics*. John Wiley & Sons, Inc., New York-London-Sydney, third edition, 1999.

[Jan71] L. Jantscher. *Distributionen*. Walter de Gruyter & Co, Berlin, 1971.

[Kit05] C. Kittel. *Introduction to Solid State Physics*. John Wiley & Sons, Inc., Hoboken, NJ, eighth edition, 2005.

[Kny01] A. V. Knyazev. Toward the optimal preconditioned eigensolver: locally optimal block preconditioned conjugate gradient method. *SIAM J. Sci. Comput.*, 23(2):517–541, 2001. Copper Mountain Conference (2000).

[Kuc01] P. Kuchment. *The Mathematics of Photonic Crystals*, chapter 7, pages 207–272. SIAM, Philadelphia, PA, 2001.

[Mię10] A. Międlar. *Inexact Adaptive Finite Element Methods for Elliptic PDE Eigenvalue Problems*. PhD thesis, TU Berlin, 2010.

[MM11] V. Mehrmann and V. Miedlar. Adaptive computation of smallest eigenvalues of self-adjoint elliptic partial differential equations. *Numer. Linear Algebra Appl.*, 18(3):387–409, 2011.

[Par98] B. N. Parlett. *The symmetric eigenvalue problem*. SIAM, Philadelphia, PA, 1998.

[Sim13] S. H. Simon. *The Oxford Solid State Basics*. Oxford University Press, New York, 2013.