Karlsruher Institut für Technologie

Fachbereich Mathematik

Adaptivity in Bandstructure Calculations of Photonic Crystals

Axel Krämer

Von der Fakultät für Mathematik
des Karlsruher Institut für Technologie
zur Verleihung des akademischen Grades
Doktor der Naturwissenschaften
(Doctor rerum naturalium, Dr. rer. nat.)
genehmigte Dissertation

1. Gutachter: Prof. Dr. Willy Dörfler

2. Gutachter: PD Dr. Nicolas Neuß

Vollzug der Promotion: 26. Januar 2011

Acknowledgements

I am very grateful to Deutsche Forschungsgemeinschaft for the PhD scholarship. I enjoyed my work in the GRK 1294 and the working environment was excellent.

I would especially like to thank Prof. Dr. Willy Dörfler for supervising and supporting my thesis, many helpful discussions, and thought-provoking impulses. I am grateful for his contributions to my work and for keeping me up to date with recent developments in my field of research. I would also like to thank PD Dr. Nicolas Neuß for his advice and for always being available to discuss problems.

Last but not least, I would like to thank my colleagues from the GRK and the KIT for the nice working atmosphere and for helpful discussions.

Contents

Preface

Photonic crystals are refractive materials with a certain periodic structure in one, two or three linearly independent dimensions. The behaviour of light in such media strongly depends on its frequency. At so called “forbidden frequencies” lying in the band gap of a particular photonic crystal no wave propagation is possible. Such effects allow for applications in photonics and optics. For the prediction of photonic crystal properties one relies on a model of an infinite crystal with perfect periodicity. By the Floquet-Bloch transformation the Maxwell eigenvalue problem for the propagating frequencies in an infinite domain is reformulated into a set of eigenvalue problems in the elementary cell, parameterised by the quasi-momemtum $\mathbf{k}$. The relation between quasimomentum and eigenfrequencies is the well-known band structure.

The aim of this thesis is to develop adaptive techniques to deal with the family of eigenvalue problems. In one or two space dimensions there is a lot more that can be said about the problem compared with the three-dimensional case. In Section 1 of Chapter II we pose the problem in more detail. We then focus on the two-dimensional elliptic eigenvalue problem and develop a convergent algorithm for fixed $\mathbf{k}$ and a chosen eigenvalue (Section 2 of Chapter II). In the same Chapter, in Section 3, we investigate what can be done in a posteriori error estimation when the dielectricity function has jumps that are not aligned with the discretisation. This is a difficult mathematical topic that has not been treated successfully so far. The last theoretic task of Chapter II is in Section 4. It is to develop an algorithm to perform the entire band structure calculations adaptively and use as little as possible computer resources, i.e. a suitable combination of a discrete set of Floquet parameters where the eigenvalue problem is solved and an adaptive finite element mesh for the eigenvalue problems.

In three dimensions there are many more difficulties involved in each eigenvalue problem because the associated operator is no longer elliptic. In Chapter III, we consider the corresponding boundary value problem and give a brief review of what methods can be used to discretize the Maxwell equations in three dimensions using H^1 conforming finite elements. Our main results concerning the a posteriori error estimation are valid for the full $3d$ curl curl system. However, the a priori results are only stated for the $2d$ curl curl system for which we perform the numerical experiments. In $3d$ these results would be more involved.

CHAPTER I

Introduction

1. Problem statement

1.1. Maxwell's equations. The propagation of light inside a photonic crystal is governed by Maxwell's equations, which in the absence of free charges and currents is the following system of partial differential equations

$$\begin{aligned} \nabla \times \mathbf{E}(\mathbf{x},t) &= -\frac{1}{c}\frac{\partial \mathbf{B}(\mathbf{x},t)}{\partial t}, \\ \nabla \times \mathbf{H}(\mathbf{x},t) &= \frac{1}{c}\frac{\partial \mathbf{D}(\mathbf{x},t)}{\partial t}, \\ \nabla\cdot\mathbf{B}(\mathbf{x},t) &= 0, \\ \nabla\cdot\mathbf{D}(\mathbf{x},t) &= 0, \end{aligned} \tag{1.1}$$

where $\mathbf{E}$ is the electric field, $\mathbf{H}$ is the magnetic field, $\mathbf{D}$ and $\mathbf{B}$ are the displacement and magnetic induction fields, respectively, and c is the vacuum speed of light. All vector fields are functions from $\mathbb{R}^3 \times \mathbb{R}$ to $\mathbb{R}^3$. Henceforth a bold font denotes a vector quantity.

For a physical explanation and a detailed problem statement, we refer to [**25**]. We consider only the case of a *linear* medium. For such a medium there are two linear *constituitive laws*, that relate $\mathbf{D}$ to $\mathbf{E}$ and $\mathbf{B}$ to $\mathbf{H}$ as

$$\begin{aligned} \mathbf{D}(\mathbf{x},t) &= \epsilon(\mathbf{x})\ \mathbf{E}(\mathbf{x},t) \\ \mathbf{B}(\mathbf{x},t) &= \mu(\mathbf{x})\ \mathbf{H}(\mathbf{x},t). \end{aligned} \tag{1.2}$$

In this way, any linear dielectrical material is determined by two properties, the *electric permittivity* ϵ and the *magnetic permeability* μ.

We assume that the medium has no frequency dependence (*material dispersion*), which means that ϵ and μ are functions of position $\mathbf{x}$ only. To keep things simple, we furthermore restrict to the case of *isotropic* and *lossless* media, which amounts to saying that ϵ and μ are scalar real-valued positive functions.

Since the common choice of materials for photonic crystals do not possess any magnetic property, we assume that $\mu = 1$. Inserting the relations (1.2) into (1.1), we obtain a simplified form, where we only keep two out of four vector fields:

$$\begin{aligned}
\nabla \times \mathbf{E}(\mathbf{x},t) &= -\frac{1}{c}\frac{\partial \mathbf{H}(\mathbf{x},t)}{\partial t}, \\
\nabla \times \mathbf{H}(\mathbf{x},t) &= \frac{1}{c}\,\epsilon(\mathbf{x})\frac{\partial \mathbf{E}(\mathbf{x},t)}{\partial t}, \\
\nabla\cdot\mathbf{B}(\mathbf{x},t) &= 0, \\
\nabla\cdot\Big(\epsilon(\mathbf{x})\;\mathbf{E}(\mathbf{x},t)\Big) &= 0,
\end{aligned} \tag{1.3}$$

From the time-dependent problem we go to the *time-harmonic* form. This assumption simplifies the problem, but does not restrict the generality of the approach, because from Fourier analysis it follows that any solution can be modeled by harmonic modes. *Monochromatic* light of frequency ω can be modeled by

$$\begin{aligned}
\mathbf{E}(\mathbf{x},t) &= \Re\Big\{e^{\mathrm{i}\omega t}\tilde{\mathbf{E}}(\mathbf{x})\Big\}, \\
\mathbf{H}(\mathbf{x},t) &= \Re\Big\{e^{\mathrm{i}\omega t}\tilde{\mathbf{H}}(\mathbf{x})\Big\},
\end{aligned} \tag{1.4}$$

where $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{H}}$ are complex vector fields. Substituting (1.4) in (1.3) we obtain a system of differential equations describing the propagation of light of frequency ω in a photonic crystal:

$$\begin{aligned}
\nabla \times \tilde{\mathbf{E}}(\mathbf{x}) &= -\frac{\mathrm{i}\omega}{c}\tilde{\mathbf{H}}(\mathbf{x}), \\
\nabla \times \tilde{\mathbf{H}}(\mathbf{x}) &= \frac{\mathrm{i}\omega}{c}\,\epsilon(\mathbf{x})\tilde{\mathbf{E}}(\mathbf{x}), \\
\nabla\cdot\tilde{\mathbf{H}}(\mathbf{x}) &= 0, \\
\nabla\cdot\Big(\epsilon(\mathbf{x})\;\tilde{\mathbf{E}}(\mathbf{x})\Big) &= 0.
\end{aligned} \tag{1.5}$$

System (1.5) is time-independent. Any frequency ω, such that (1.5) possesses a nontrivial solution, is allowed to travel through the crystal. On the other hand, light at a frequency, for which (1.5) does not possess a nonvanishing solution, cannot travel through the crystal.

1.2. Polarized modes. In Chapter II, we will deal with $2d$ photonic crystals. Here, we briefly illustrate how the system (1.5) can be simplified when the electric permittivity function ϵ only depends on the first two components of $\mathbf{x} = (x, y, z)$, such that we can write $\epsilon = \epsilon(x, y)$.

It is straightforward to conclude, that in this case the quantities $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{H}}$ in (1.5) will also only depend on x and y. The system (1.5) thus becomes

$$\begin{aligned} \nabla \times \tilde{\mathbf{E}}(x,y) &= -\frac{\mathrm{i}\omega}{c}\tilde{\mathbf{H}}(x,y), \\ \nabla \times \tilde{\mathbf{H}}(x,y) &= \frac{\mathrm{i}\omega}{c}\,\epsilon(x,y)\tilde{\mathbf{E}}(x,y), \\ \nabla\cdot\tilde{\mathbf{H}}(x,y) &= 0, \\ \nabla\cdot\Big(\epsilon(x,y)\,\tilde{\mathbf{E}}(x,y)\Big) &= 0. \end{aligned} \tag{1.6}$$

Now, we would like to show how the system (1.6) naturally splits into two disjoint subproblems, called TM and TE modes.

1.2.1. *TM mode.* Substituting in (1.6) the first into the second equation we obtain:

$$\nabla \times \Big(\nabla \times \tilde{\mathbf{E}}(x,y)\Big) = -\frac{\omega^2}{c^2}\,\epsilon(x,y)\tilde{\mathbf{E}}(x,y). \tag{1.7}$$

Denoting the components of $\tilde{\mathbf{E}}$ by (E_1, E_2, E_3) and remembering that each one of these components only depends on x and y, we can write the system (1.7) as

$$\begin{aligned} (E_2)_{yx} - (E_1)_{yy} &= \tfrac{\omega^2}{c^2}\,\epsilon\,E_1, \\ -(E_2)_{xx} + (E_1)_{xy} &= \tfrac{\omega^2}{c^2}\,\epsilon\,E_2, \\ -(E_3)_{xx} - (E_3)_{yy} &= \tfrac{\omega^2}{c^2}\,\epsilon\,E_3, \end{aligned} \tag{1.8}$$

where the subscripts x and y mean differentiation along the indicated axes. The last equation of (1.8) does not depend on the other two equations. We call this equation the TM mode and after the substitution $\lambda = \frac{\omega^2}{c^2}$ it can be written as

$$\mathcal{L}\,E_3 := -\Delta E_3 = \lambda\,\epsilon\,E_3. \tag{1.9}$$

1.2.2. *TE mode.* To obtain the TE mode we again start with the system (1.6). This time we substitute the second equation into the first one to obtain:

$$\nabla \times \Big(\epsilon(x,y)^{-1}\nabla \times \tilde{\mathbf{H}}(x,y)\Big) = -\frac{\omega^2}{c^2}\tilde{\mathbf{H}}(x,y). \tag{1.10}$$

This system again possesses an equation that is independent of the other two equations. This time it can be written as

$$-\nabla\cdot(\epsilon^{-1}\nabla H_3) = \lambda H_3, \tag{1.11}$$

where H_3 is the 3rd component of $\tilde{\mathbf{H}}$ and $\lambda = \frac{\omega^2}{c^2}$.

1.3. Periodic media and the band structure of a photonic crystal. Photonic crystals are *periodic structures*. We assume that a crystal is unbounded and occupies the whole space $\mathbb{R}^3$, which is an abstraction that is justified, when we assume that the whole crystal is very large compared to the size of a periodicity cell.

A useful overview of mathematical approaches concerning photonic crystals is given in [**27**].

Let $d \in \{1,2,3\}$ and suppose that there exist linearly independent vectors $\mathbf{r}_1, \ldots, \mathbf{r}_d$, such that

$$\epsilon(\mathbf{x}+\mathbf{r}_j) = \epsilon(\mathbf{x}) \quad \text{for all } x \in \mathbb{R}^3, \quad j = 1, \ldots, d.$$

Then the medium is called a *d-dimensional periodic medium* (photonic crystal) and the vectors $\mathbf{r}_j$ with minimal lengths are called the *primitive vectors.* For such a medium one may define the d-dimensional *Bravais lattice*

$$\Lambda := \left\{ \sum_{j=1}^{d} l_j \; \mathbf{r}_j \;\mid\; l_1, \ldots, l_d \in \mathbb{Z} \right\}.$$

A d-dimensional domain Ω is called *fundamental domain* if for any $\mathbf{x} \in \mathbb{R}^d$ there exists $\mathbf{a} \in \Lambda$ such that either $\mathbf{a}$ is unique and $\mathbf{x}+\mathbf{a} \in \Omega$, or $\mathbf{a}$ is not unique and $\mathbf{x}+\mathbf{a} \in \partial\Omega$. So we may write

$$\mathbb{R}^d = \bigcup_{\mathbf{a}\in\Lambda} \left(\bar{\Omega} + \mathbf{a}\right),$$

where for two distinct $\mathbf{a}_1, \mathbf{a}_2 \in \Lambda$, $\bar{\Omega}+\mathbf{a}_1$ and $\bar{\Omega}+\mathbf{a}_2$ may only have a non-empty intersection along their boundaries. We now define *primitive reciprocal vectors* $\hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_d$, such that

$$\mathbf{r}_i \cdot \hat{\mathbf{r}}_j = 2\pi\delta_{ij} \;\; \text{for all } i,j \in \{1, \ldots, d\},$$

and the d-dimensional *reciprocal lattice* by

$$\hat{\Lambda} := \left\{ \sum_{j=1}^{d} l_j \; \hat{\mathbf{r}}_j \;\mid\; l_1, \ldots, l_d \in \mathbb{Z} \right\}.$$

The domain B that consists of $\mathbf{k} \in \mathbb{R}^d$ which are closer to the origin than to any other $\hat{\mathbf{a}} \in \hat{\Lambda}$ is called the *Brillouin zone.*

In Chapter II we focus on the solution of the equation (1.9). This is quite a challenging task since the domain of the problem is the whole space $\mathbb{R}^2$. A standard tool for the analysis of partial differential equations with periodic coefficients is the Floquet transform (see [**26**]). The result of an application of this theory is, that the spectrum of the unbounded operator $\mathcal{L}$ from equation (1.9) can be represented as the union of spectra of differential operators $\mathcal{L}_{\text{per}}(\mathbf{k})$ on bounded domains

$$\sigma\left(\mathcal{L}\right) = \bigcup_{\mathbf{k}\in\overline{B}} \sigma\left(\mathcal{L}_{\text{per}}(\mathbf{k})\right),$$

where B stands for the Brillouin zone and $\mathcal{L}_{\text{per}}(\mathbf{k})$ is a $\mathbf{k}$-dependent partial differential operator on a fixed domain. Its eigenpairs satisfy:

$$\begin{aligned} \mathcal{L}_{\text{per}}(\mathbf{k})E_3 = -(\nabla + \mathrm{i}\mathbf{k}) \cdot (\nabla + \mathrm{i}\mathbf{k})E_3 &= \lambda\, \epsilon\, E_3 \quad \text{in } \Omega, \\ E_3(\mathbf{x}+\mathbf{a}) &= E_3(\mathbf{x}), \quad \mathbf{a} \in \Lambda,\ \mathbf{x}, \mathbf{x}+\mathbf{a} \in \partial\Omega, \end{aligned} \tag{1.12}$$

where Ω is a fundamental domain. For each $\mathbf{k} \in B$, $\mathcal{L}_{\text{per}}(\mathbf{k})$ is a self-adjoint operator with a discrete and positive spectrum

$$0 < \lambda_1(\mathbf{k}) \leq \ldots \leq \lambda_j(\mathbf{k}) \to \infty, \quad \text{as } j \to \infty \tag{1.13}$$

and finite dimensional eigenspaces (*cf.* [**21**]). Furthermore, for each $j \in \mathbb{N}$ an eigenvalue $\lambda_j(\mathbf{k})$ of the operator $\mathcal{L}_{\text{per}}(\mathbf{k})$ is a continuous function of the parameter $\mathbf{k} \in B$ (*cf.* [**26**]). $\lambda_j(\cdot)$ is called a *band function* and its graph is called a *band*. For $j \in \mathbb{N}$ we define the j-th band $I_j := \left\{\lambda_j(\mathbf{k}) \mid \mathbf{k} \in \bar{B}\right\}$. Since $\bar{B}$ is compact and connected and the operator $\mathcal{L}_{\text{per}}(\mathbf{k})$ is symmetric, I_j is a compact real interval. This gives another representation of the spectrum:

$$\sigma(\mathcal{L}) = \bigcup_{j \in \mathbb{N}} I_j. \tag{1.14}$$

CHAPTER II

Adaptive bandstructure calculations

1. Approximating the bandstructure

We have seen in the preceding chapter, that the spectrum of the unbounded operator $\mathcal{L}$ (introduced in equation (1.9) in Chapter I) is given by the union of all values of all bandfunctions (*cf.* (1.14) in Chapter I). Given $\mathbf{k} \in B$ we are seeking solutions $(\lambda_j(\mathbf{k}), u_{\mathbf{k},j})$ to the following eigenvalue problem

$$\begin{aligned} -(\nabla + \mathrm{i}\mathbf{k}) \cdot (\nabla + \mathrm{i}\mathbf{k}) u_{\mathbf{k},j} &= \lambda_j(\mathbf{k}) \epsilon u_{\mathbf{k},j} \quad \text{in } \Omega, \\ u_{k,j}(\mathbf{x} + \mathbf{a}) &= u_{k,j}(\mathbf{x}), \qquad \mathbf{a} \in \Lambda,\ \mathbf{x},\ \mathbf{x} + \mathbf{a} \in \partial\Omega. \end{aligned} \tag{1.1}$$

Ω is regarded as a torus, i.e. opposite sides are identified with each other. We are considering several "band functions" $j \in M$ (with $M := \{1, \ldots, 4\}$, for instance) at the same time. We are interested in choosing a sufficiently large and well distributed discrete set of parameters $\{\mathbf{k}_i\}_{i=1}^{N_k} \subset B$ hand in hand with a good adaptive mesh for the eigenvalue problem such that

$$ERROR := \max_{j \in M} \|\lambda_j(\cdot) - \lambda_{j,\text{num}}(\cdot)\|_{C^0(B)} \leq TOL, \tag{1.2}$$

where $\lambda_{j,\text{num}}(\cdot)$ suitably interpolates different approximated eigenvalues $\lambda_{j,n}(\mathbf{k}_i)$ on the discrete set $\{\mathbf{k}_i\}_{i=1}^{N_k}$ and $\lambda_j(\cdot)$ stands for the exact band function.

2. The shifted Laplace eigenvalue problem for fixed k

In this section we focus on the approximation of the eigenpairs of equation (1.1) for fixed $\mathbf{k}$ by means of the finite element method. For this purpose, we first need some preparation and some notation.

2.1. Functional spaces, norms, and notation.

Definition 2.1. Let Ω be a bounded, simply-connected Lipschitz-domain. We define the following spaces of infinitely differentiable functions:

$$C^\infty(\bar{\Omega}) := \left\{ f : \Omega \to \mathbb{C} \mid \forall\, m, n \in \{0\} \cup \mathbb{N},\ \exists\ \frac{\partial^{m+n} f}{\partial x^m \partial y^n} \text{ is continuous} \right\},$$

$$C_c^\infty(\bar{\Omega}) := \left\{ f \in C^\infty(\bar{\Omega}) \ \mid\ \overline{\operatorname{supp}(f)} \subset K \text{ compact in } \Omega \right\}.$$

For a polygonal domain Ω that has an even number of sides, such that opposite sides have the same length and orientation, we additionally define

$$C_1^\infty(\bar{\Omega}) := \left\{ f \in C^\infty(\bar{\Omega}) \mid f(\mathbf{x}+\mathbf{a}) = f(\mathbf{x}),\ \mathbf{a} \in \Lambda,\ \mathbf{x}, \mathbf{x}+\mathbf{a} \in \partial\Omega \right\}.$$

Furthermore, we define the following scalar products

$$(f,g)_{L^2} = \int_\Omega f\bar{g},$$

$$(f,g)_{H^1} = \int_\Omega \nabla f \cdot \overline{\nabla g}, + (f,g)_{L^2}$$

that induce the norms $||\cdot||_{L^2}$ and $||\cdot||_{H^1}$, which we also denote by $||\cdot||$ and $||\cdot||_1$, respectively. Finally, we define

$$H^1_{\text{per}}(\Omega) := \overline{C_1^\infty(\bar{\Omega})}^{||\cdot||_{H^1}}.$$

We use standard notation for all Sobolev spaces $W^{s,p}(\Omega)$ and their associated norms and seminorms, see e.g. [**1**]. For $p = 2$ we denote $H^s(\Omega) = W^{s,p}(\Omega)$ and $H_0^1(\Omega) := \overline{C_c^\infty(\bar{\Omega})}^{||\cdot||_{H^1}}$. Furthermore, we denote $||\cdot||_{s,\Omega} = ||\cdot||_{s,2,\Omega}$ and $||\cdot||_\Omega = ||\cdot||_{0,2,\Omega}$. Throughout the thesis, we shall use C to denote a generic positive constant which may stand for different values at its different appearances. At times, we write $A \lesssim B$ to denote $A \leq CB$, for some constant C that is independent of mesh parameters.

2.2. Weak formulation. We define the following bilinear forms

$$\begin{aligned} a_{\mathbf{k}}(u,v) &:= \int_\Omega (\nabla + \mathrm{i}\mathbf{k})u \cdot \overline{(\nabla + \mathrm{i}\mathbf{k})v} && \text{for all } u, v \in H^1_{\text{per}}(\Omega), \\ a_{\mathbf{k},\sigma}(u,v) &:= a_{\mathbf{k}}(u,v) + \sigma b(u,v) && \text{for all } u, v \in H^1_{\text{per}}(\Omega), \\ b(u,v) &:= \int_\Omega \epsilon\ u\ \overline{v} && \text{for all } u, v \in L^2(\Omega), \end{aligned} \tag{2.1}$$

where ϵ is assumed to be a piecewise smooth function that is bounded from above and below by positive constants $\underline{\epsilon}$, $\bar{\epsilon}$

$$\underline{\epsilon} \leq \epsilon(x) \leq \bar{\epsilon} \quad \text{for all } x \in \Omega. \tag{2.2}$$

The norms associated to the bilinear forms, we denote in the following way

$$|||v|||_{a,\mathbf{k},\sigma} := \{a_{\mathbf{k},\sigma}(v,v)\}^{\frac{1}{2}}, \tag{2.3}$$

$$||v||_b := \{b(v,v)\}^{\frac{1}{2}}. \tag{2.4}$$

For fixed $\mathbf{k}$, the weak formulation of problem (1.1) is the following problem.

<u>Problem</u> 2.2. Seek eigenpairs $(\lambda_j, u_j) \in \mathbb{C} \times H^1_{\text{per}}(\Omega)$, with $||u_j||_b = 1$ and

$$a_{\mathbf{k}}(u_j, v) = \lambda_j\ b(u_j, v) \quad \text{for all } v \in H^1_{\text{per}}(\Omega). \tag{2.5}$$

We also define a problem, where the spectrum is shifted by σ. In this way we are able to deal with a coercive bilinear form on the left-hand side.

<u>Problem</u> 2.3. Seek eigenpairs $(\zeta_j, u_j) \in \mathbb{C} \times H^1_{\text{per}}(\Omega)$, with $||u_j||_b = 1$ and

$$a_{\mathbf{k},\sigma}(u_j, v) = \zeta_j \; b(u_j, v) \quad \text{for all } v \in H^1_{\text{per}}(\Omega). \tag{2.6}$$

Note that the shift σ defines the relation $(\zeta_j, u_j) = (\lambda_j + \sigma, u_j)$, which is a one-to-one relation between the spectra of Problems 2.2 and 2.3.

2.3. The finite element method and interpolation estimates. We let $\mathcal{T}_n,\ n = 1, 2, \ldots$, denote a family of triangular meshes on Ω, such that

$$\Omega = \bigcup_{T \in \mathcal{T}_n} T. \tag{2.7}$$

In other words, we assume that Ω has a polygonal boundary. We assume that for each n, $\mathcal{T}_{n+1}$ is refinement of $\mathcal{T}_n$. For a typical element $T \in \mathcal{T}_n$, its diameter is denoted by $h_{T,n}$. We denote the maximal diameter in a triangulation by $h_n^{\max} := \max_{T \in \mathcal{T}_n} h_{T,n}$. h_n denotes the function whose restriction to an element T is $h_{T,n}$. All the meshes are assumed to be conforming and shape regular in the meaning explained in [**10**]. On any mesh $\mathcal{T}_n$ we denote by $V_n \in C^0(\Omega) \cap H^1_{\text{per}}(\Omega)$ the finite dimensional space of piecewise polynomials, i.e.

$$V_n := \left\{ v_h \in C^0(\Omega) \cap H^1_{\text{per}}(\Omega) \;\mid\; v_{h|T} \in \mathcal{P}_r(T) \text{ for all } T \in \mathcal{T}_h \right\}. \tag{2.8}$$

Unless specified otherwise, we are dealing with linear polynomials in this chapter, i.e. $r = 1$ in the above definition.

We denote by $\mathcal{F}_n$ the set of all the edges of the elements in $\mathcal{T}_n$. Moreover, we denote by h_F the length of an edge $F \in \mathcal{F}_n$. We use the following notation to denote element patches:

$$\omega(T) := \bigcup \left\{ T' \in \mathcal{T}_n \;\mid\; T' \cap T \neq \emptyset \right\}, \tag{2.9}$$

$$\omega(F) := \bigcup \left\{ T' \in \mathcal{T}_n \;\mid\; T' \cap F \neq \emptyset \right\}. \tag{2.10}$$

Assumption 2.4. There exists an interpolation operator $\Pi_h :\ H^1_{\text{per}}(\Omega) \to V_n$ with the following properties:

$$\|v - \Pi_h v\|_{0,T} \leq C h_T \, |v|_{1,\omega(T)} \tag{2.11}$$

and

$$\|v - \Pi_h v\|_{0,F} \leq C h_F^{\frac{1}{2}} \, |v|_{1,\omega(F)} \,. \tag{2.12}$$

Furthermore we assume that the following stability estimate holds

$$\|v - \Pi_h v\|_{H^l(\Omega)} \leq C \, |v|_{H^l(\Omega)} \;\;,\;\; 0 \leq l < \frac{3}{2}. \tag{2.13}$$

One possible choice of an interpolation operator that fulfills Assumption 2.4 would be the interpolation operator introduced by Clément in [**12**].

Remark 2.5. The Scott-Zhang interpolation operator introduced in [**33**] satisfies (2.11) and (2.12), however it satisfies (2.13) only for $\frac{1}{2} \leq l$. It is on the other hand a projection on V_n in the sense that:

$$v_n = \Pi_h v_n \quad \text{for all } v_n \in V_n. \tag{2.14}$$

We define the discrete solutions to Problems 2.2 and 2.3 as follows.

Problem 2.6. Seek eigenpairs $(\lambda_{j,n}, u_{j,n}) \in \mathbb{C} \times V_n$, with $||u_{j,n}||_b = 1$ and

$$a_{\mathbf{k}}(u_{j,n}, v) = \lambda_{j,n}\ b(u_{j,n}, v) \quad \text{for all } v \in V_n. \tag{2.15}$$

Problem 2.7. Seek eigenpairs $(\zeta_{j,n}, u_{j,n}) \in \mathbb{C} \times V_n$, with $||u_{j,n}||_b = 1$ and

$$a_{\mathbf{k},\sigma}(u_{j,n}, v) = \zeta_{j,n}\ b(u_{j,n}, v) \quad \text{for all } v \in V_n. \tag{2.16}$$

2.4. A priori analysis. The following result, which we formulate as a Lemma, is obvious.

Lemma 2.8 (Continuity). *The bilinear form $a_{\mathbf{k}}(u,v)$ and $a_{\mathbf{k},\sigma}(u,v)$, respectively, introduced in (2.1) are continuous in the sense that*

$$a_{\mathbf{k}}(u,v) \leq C_a \, ||u||_1 \, ||v||_1 \quad \forall u, v \in H^1_{\text{per}}(\Omega), \tag{2.17}$$

$$a_{\mathbf{k},\sigma}(u,v) \leq C_{a,\sigma} \, ||u||_1 \, ||v||_1 \quad \forall u, v \in H^1_{\text{per}}(\Omega). \tag{2.18}$$

Lemma 2.9 (Coercivity). *With the choice $\sigma = (\max_{\mathbf{k} \in B} \frac{|\mathbf{k}|^2}{\underline{\epsilon}}) + 1$ and the norm $|||u|||_{a,\mathbf{k},\sigma} = a_{\mathbf{k}}(u,u) + \sigma b(u,u)$, we have the following coercivity estimate*

$$|||u|||_{a,\mathbf{k},\sigma} \geq c\, \|u\|_1 \quad \forall\ \mathbf{k} \in B \quad \forall u \in H^1_{\text{per}}(\Omega), \tag{2.19}$$

where $\underline{\epsilon} = \min_{x\in\Omega} \epsilon(x)$ and $c = \min\left\{\frac{1}{2}, \underline{\epsilon}\right\}$.

Proof. We readily calculate

$$\begin{aligned} a_{\mathbf{k}}(u,u) := \int_\Omega (\nabla + \mathrm{i}\mathbf{k})u \cdot \overline{(\nabla + \mathrm{i}\mathbf{k})u} = \int_\Omega |\nabla u|^2 + |\mathbf{k}|^2 \int_\Omega |u|^2 \\ + 2\int_\Omega \Re\left\{\mathrm{i}u\mathbf{k} \cdot \overline{\nabla u}\right\}. \end{aligned} \tag{2.20}$$

It can easily be seen that

$$\Re\left\{\mathrm{i}u\mathbf{k} \cdot \overline{\nabla u}\right\} \leq |u| \left|\mathbf{k} \cdot \overline{\nabla u}\right| \leq |u| \, |\mathbf{k}| \, |\nabla u| \,, \tag{2.21}$$

and thus

$$\int_\Omega \Re\left\{\mathrm{i}u\mathbf{k} \cdot \overline{\nabla u}\right\} \leq \frac{|u|_1^2}{2} + 2\,|\mathbf{k}|^2 \, \|u\|^2 \,, \tag{2.22}$$

where we have used the arithmetic-geometric mean inequality $2\alpha\beta \leq \delta\alpha^2 + \frac{\beta^2}{\delta}$ with $\delta = \frac{1}{2}$.

Remembering that $\sigma = (\max_{\mathbf{k}\in B} \frac{|k|^2}{\underline{\epsilon}}) + 1$ we see that

$$\begin{aligned} a_{\mathbf{k}}(u,u) + \sigma b(u,u) &\geq \frac{|u|_1^2}{2} + \left(-\left|\mathbf{k}\right|^2 + \sigma\underline{\epsilon}\right)\left\|u\right\|^2 \\ &\geq \min\left\{\frac{1}{2}, \underline{\epsilon}\right\}\left\|u\right\|_1^2 . \end{aligned} \tag{2.23}$$

□

Applying the Lax-Milgram lemma, we can deduce that for sufficiently large σ there is a uniquely defined solution operator $T : L^2(\Omega) \to H^1_{\text{per}}(\Omega)$, such that

$$\forall f \in L^2(\Omega), \quad a_{\mathbf{k},\sigma}(Tf, v) = b(f, v), \quad \text{for all } v \in H^1_{\text{per}}(\Omega). \tag{2.24}$$

Since the imbedding $H^1_{\text{per}}(\Omega) \subset L^2(\Omega)$ is compact, we can conclude that the solution operator T in (2.24) is a compact operator. Since T is furthermore a self-adjoint operator, we can apply the spectral theorem for compact self-adjoint operators to conclude that T has a positive discrete spectrum and the eigenspaces to these eigenvalues are finite dimensional. To each eigenpair (μ, u) of T there corresponds an eigenpair (μ^{-1}, u) of (2.3), so that we can conclude that (2.3) also has a positive and discrete spectrum. Because the spectrum of (2.2) is a shifted version of the spectrum of (2.3), it is also discrete with finite dimensional eigenspaces.

We make an additional assumption:

Assumption 2.10 (H^{1+s} regularity). There exists $s \in (0, 1]$ and there exists a solution

$$v \in H^1_{\text{per}}(\Omega) \cap H^{1+s}(\Omega) \tag{2.25}$$

for any $f \in L^2(\Omega)$ to the following boundary value problem:

$$\begin{aligned} -(\nabla + \mathrm{i}\mathbf{k}) \cdot (\nabla + \mathrm{i}\mathbf{k})v(\mathbf{x}) &= f(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \Omega, \\ v(\mathbf{x} + \mathbf{a}) &= v(\mathbf{x}) \quad \mathbf{a} \in \Lambda;\ \mathbf{x},\ \mathbf{x} + \mathbf{a} \in \partial\Omega. \end{aligned} \tag{2.26}$$

This solution is not unique in general.

Theorem 2.11 (Convergence with respect to uniform mesh size). *Let s be as given in Assumption 2.10. Let h_n be sufficiently small and consider the eigenvalue $\zeta_l \in \mathbb{C}$ of Problem 2.3. If $(\zeta_{l,n}, u_{l,n}) \in \mathbb{C} \times V_n$ is a computed eigenpair of Problem 2.7 with $\zeta_{l,n}$ converging to ζ_l, then the following statements hold:*

$$0 \leq \zeta_{l,n} - \zeta_l \lesssim (h_n^{\max})^{2s} \tag{2.27}$$

$$\|u_l - u_{l,n}\|_{L^2(\Omega)} \lesssim \ (h_n^{\max})^s \, |||u_l - u_{l,n}|||_{a,\mathbf{k},\sigma} \, . \tag{2.28}$$

If in addition $u_{l,n} \in H^{m+1}(\Omega)$ with $m \geq 2$ then we have the following estimates:

$$|||u_l - u_{l,n}|||_{a,\mathbf{k},\sigma} \lesssim (h_n^{\max})^{\min\{m,r\}} \, \|u_l\|_{m+1} \, , \tag{2.29}$$

$$|\zeta - \zeta_h| \lesssim (h_n^{\max})^{2\min\{m,r\}} \, \|u_j\|_{m+1,\Omega} \, , \tag{2.30}$$

where r is the degree of the polynomials in the piecewise polynomial space V_n introduced in (2.8)*.*

Proof. These are standard results that can be adapted to our case very easily. The main tools can be found in [**2**]. □

Remark 2.12. Undoing the effect of the shift of the spectrum by σ, we obtain the same kind of result for the convergence of the eigenpairs of Problem 2.2 to the eigenpairs of Problem 2.6. We refer to [**21**].

Theorem 2.13 (Relation between eigenvalue and eigenfunction error). *Let $(\zeta_{j,h}, u_{j,h}) \in \mathbb{R} \times V_h$ be a solution of Problem 2.7 that approximates the eigenpair $(\zeta_j, u_j) \in \mathbb{R} \times H^1_{\text{per}}(\Omega)$ of Problem 2.3. Then we have the following estimate:*

$$|\zeta_j - \zeta_{j,h}| \lesssim a_{\mathbf{k},\sigma}(u_j - u_{j,h}, u_j - u_{j,h}). \tag{2.31}$$

Proof. This is a standard result that can for instance be found in [**21**]. □

2.5. A posteriori error estimation. Before we can formulate and prove the main results of this section, we first need to define the error estimator. This requires some notation.

Definition 2.14 (Jump of a function). For any sufficiently regular function $g : \Omega \to \mathbb{C}$ and for any edge $F \in \mathcal{F}$ that is common for the elements T and T', we define *the jump of g along F* by

$$[g]_F\,(x) = \lim_{\tilde{x} \in T,\, \tilde{x} \to x} g(\tilde{x}) - \lim_{\tilde{x} \in T',\, \tilde{x} \to x} g(\tilde{x}). \tag{2.32}$$

Definition 2.15 (Residual error estimator). On each element $T \in \mathcal{T}_n$ the following *local error indicator* is defined for an eigenpair $(\zeta_{l,n}, u_{l,n}) \in \mathbb{C} \times V_n$ of Problem 2.7:

$$\begin{aligned} \eta^2_{\mathcal{T}_n}((u_{l,n}, \zeta_{l,n}), T) := \ & h_T^2 \, \|(\nabla + \mathrm{i}\mathbf{k}) \cdot (\nabla + \mathrm{i}\mathbf{k}) u_{l,n} + \zeta_{l,n} \epsilon u_{l,n}\|_T^2 \\ & + \frac{1}{2} \sum_{F \subset \partial T, F \in \mathcal{F}_n} h_F \left\| [\partial_{\mathbf{n}} u_{l,n} + \mathrm{i}\mathbf{k} u_{l,n}]_F \right\|_F^2 . \end{aligned} \tag{2.33}$$

On the whole domain Ω the error can be estimated by means of the following estimator:

$$\eta^2_{\mathcal{T}_n}((u_{l,n},\zeta_{l,n}),\Omega) := \sum_{T\in\mathcal{T}_n} \eta^2_{\mathcal{T}_n}((u_{l,n},\zeta_{l,n}),T). \tag{2.34}$$

Definition 2.16. For $g \in L^2(\Omega)$ we define the *data oscillation* as

$$\operatorname{osc}(g,\mathcal{T}_n) := \left(\sum_{T\in\mathcal{T}_n} ||h_{T,n}(g-\bar{g}_T)||^2\right)^{\frac{1}{2}}, \tag{2.35}$$

where $\bar{g}_T$ is a polynomial approximation of g on T.

Theorem 2.17 (Reliability of the estimator). *Let $(\zeta_{j,n},u_{j,n}) \in \mathbb{C}\times H^1_{\text{per}}(\Omega)$ be a simple eigenpair of Problem 2.7. Then we have the following error estimate:*

$$|||u_j - u_{j,n}|||_{a,\mathbf{k},\sigma} \lesssim \eta_{\mathcal{T}_n}\left(\left(\zeta_{j,n},u_{j,n}\right),\Omega\right) + (h_n^{\max})^s\,|||u_j-u_{j,n}|||_{a,\mathbf{k},\sigma}\,, \tag{2.36}$$

where s is from Assumption 3.1, and u_j is the eigenfunction number j of Problem 2.3.

Proof. We readily calculate for arbitrary $v \in H^1_{\text{per}}(\Omega)$ and arbitrary $v_h \in V_h$, that

$$\begin{aligned} a_{\mathbf{k},\sigma}(u_j-u_{j,n},v) = \zeta_j\ b(u_j,v)& \\ &+\sum_{T\in\mathcal{T}}\int_T \left((\nabla+\mathrm{i}\mathbf{k})\cdot(\nabla+\mathrm{i}\mathbf{k})u_{j,n}+\epsilon\zeta_{j,n}u_{j,n}\right)\overline{(v-v_h)} \\ &-\sum_{T\in\mathcal{T}}\int_{\partial T}\partial_n(u_{j,n}+\mathrm{i}\mathbf{k}u_{j,n})\overline{(v-v_h)}-\zeta_{j,n}b(u_{j,n},v). \end{aligned} \tag{2.37}$$

It is equivalent to

$$\begin{aligned} a_{\mathbf{k},\sigma}(u_j-u_{j,n},v) = &\sum_{T\in\mathcal{T}}\int_T \left((\nabla+\mathrm{i}\mathbf{k})\cdot(\nabla+\mathrm{i}\mathbf{k})u_{j,n}+\epsilon\zeta_{j,n}u_{j,n}\right)\overline{(v-v_h)} \\ &-\frac{1}{2}\sum_{T\in\mathcal{T}}\int_{\partial T}\left[\partial_{\mathbf{n}}(u_{j,n}+\mathrm{i}\mathbf{k}u_{j,n})\right]\overline{(v-v_h)} \\ &+\zeta_j\ b(u_j,v)-\zeta_{j,n}b(u_{j,n},v). \end{aligned} \tag{2.38}$$

We choose $v := u - u_{j,n}$ and define $v_h := \Pi_h v$, where $\Pi_h : \ H^1_{\text{per}}(\Omega) \to V_n$ is the interpolation operator satisfying Assumption 2.4. This yields:

$$\begin{aligned}
&|||u_j - u_{j,n}|||^2_{a,\mathbf{k},\sigma} \\
&\lesssim \left(\sum_{T \in \mathcal{T}} \eta_T^2 \right)^{\frac{1}{2}} |u_j - u_{j,n}|_{H^1(\Omega)} \\
&\quad + \zeta_j \ b(u_j, u_j - u_{j,n}) - \zeta_{j,n} b(u_{j,n}, u_j - u_{j,n}) \\
&\lesssim \left(\sum_{T \in \mathcal{T}} \eta_T^2 \right)^{\frac{1}{2}} |||u_j - u_{j,n}|||_{a,\mathbf{k},\sigma} + (\zeta_j + \zeta_{j,n}) \left(1 + b\left(u_j, u_{j,n}\right)\right) \\
&= \left(\sum_{T \in \mathcal{T}} \eta_T^2 \right)^{\frac{1}{2}} |||u_j - u_{j,n}|||_{a,\mathbf{k},\sigma} \\
&\quad + \frac{\zeta_j + \zeta_{j,n}}{2} \ b(u_j - u_{j,n}, u_j - u_{j,n}) \\
&\lesssim \left(\sum_{T \in \mathcal{T}} \eta_T^2 \right)^{\frac{1}{2}} |||u_j - u_{j,n}|||_{a,\mathbf{k},\sigma} + (h_n^{\max})^{2s} \, |||u_j - u_{j,n}|||_{a,\mathbf{k},\sigma} \, .
\end{aligned} \tag{2.39}$$

Here we used that $|u_j - u_{j,n}|_{H^1(\Omega)} \leq |||u_j - u_{j,n}|||_{a,\mathbf{k},\sigma}$, and that $b(u_j, u_j) = b(u_{j,n}, u_{j,n}) = 1$. We additionaly used the fact that the bilinear form $b(\cdot, \cdot)$ is hermitian and in the last step we employed (2.28) from Theorem 2.11. Dividing the equation by $|||u_j - u_{j,n}|||_{a,\mathbf{k},\sigma}$ leads to the result. □

Remark 2.18. The second summand in (2.36) is a higher order term. If $h_n^{\max} < \frac{1}{2}$ independent of n, then we can absorb the second term in (2.36) to get

$$|||u_j - u_{j,n}|||_{a,\mathbf{k},\sigma} \lesssim \ \eta_{\mathcal{T}_h} \left(\left(\zeta_{j,n}, u_{j,n} \right), \Omega \right) \tag{2.40}$$

independent of n.

2.6. A convergent adaptive algorithm for the eigenvalue problem. In this section we propose a convergent adaptive algorithm of the form

$$\mathsf{SOLVE} \to \mathsf{ESTIMATE} \to \mathsf{MARK} \to \mathsf{REFINE} \tag{2.41}$$

We will explain the different modules shortly.

We achieve convergence of the algorithm that we describe below by extending the results from [**17**], which rely on a similar result for the corresponding boundary value problem (cf. [**9**]), to our case of photonic crystals. An oscillation term as in [**22**] is merely used in the analysis, however, we do not need to refine according to this quantity as was proposed in that source.

Now we describe the different modules put forth in (2.41) that make up the algorithm.

2.6.1. *The Module* SOLVE. We do not discuss the step SOLVE in detail. We simply assume that the finite dimesnional Problem 2.7 is computed exactly, i.e. for given $l \in \mathbb{N}$ the eigenpair $(\zeta_{l,n}, u_{l,n})$ is computed via exact linear algebra and exact integration. In Section 3 we will investigate the effect of solving Problem 2.7 employing quadrature schemes, though not in conjunction with a convergent adaptive algorithm.

2.6.2. *The Module* ESTIMATE. We assume that given a triangulation $\mathcal{T}_n$ and the solution $(\zeta_{l,n}, u_{l,n}) \in \mathbb{R} \times V_n$ of Problem 2.7, this module outputs the local error indicators as defined in (2.33):

$$\{\eta_{\mathcal{T}_n}((u_{l,n}, \zeta_{l,n}), T)\}_{T \in \mathcal{T}_n} = \mathsf{ESTIMATE}((\zeta_{l,n}, u_{l,n}), \mathcal{T}_n). \tag{2.42}$$

2.6.3. *The Module* MARK. Given a grid $\mathcal{T}_n$, the set of indicators $\{\eta_T\}_{T \in \mathcal{T}_n}$, and a parameter $\theta \in (0,1)$, we suppose that MARK outputs a subset of marked elements $\mathcal{M} \in \mathcal{T}_n$, i.e.

$$\mathcal{M} = \mathsf{MARK}((\zeta_{l,n}, u_{l,n}), \mathcal{T}_n, \theta), \tag{2.43}$$

such that

$$\sum_{T \in \mathcal{M}} \eta^2_{\mathcal{T}_n}((u_{l,n}, \zeta_{l,n}), T) \geq \theta \eta^2_{\mathcal{T}_n}((u_{l,n}, \zeta_{l,n}), \Omega). \tag{2.44}$$

This strategy was introduced in Dörfler [**18**].

2.6.4. *The Module* REFINE. We do not require the procedure REFINE to satisfy the Interior Node Property as in [**29**], for instance. Instead, we assume that a procedure REFINE is at our disposal, as described in [**9**] and [**17**]. It is some iterative or recursive bisection of elements with the minimal condition on the refinement condition that each element that is marked is bisected at least once.

The discussion above results in the following Algorithm.

Algorithm 2.19. (Adaptive convergent FEM to solve for (ζ_l, u_l))

```
Choose $0 < \theta < 1$.
Choose an initial mesh $\mathcal{T}_0$.
SOLVE the eigenvalue Problem 2.3 on $\mathcal{T}_0$ to find the eigenpair $(\lambda_{l,0}, u_{l,0})$.
Let $n := 0$
ESTIMATE: Compute the local error indicators $\{\eta_{\mathcal{T}_n}((u_{l,n}, \zeta_{l,n}), T)\}_{T \in \mathcal{T}_n}$ from (2.33).
Mark a set $\mathcal{M} \subset \mathcal{T}_n$ according to the the module MARK from Section 2.6.3 using $\theta$.
Refine $\mathcal{T}_n$ according to the procedure REFINE described in Section 2.6.4 to get a new conforming triangulation $\mathcal{T}_{n+1}$.
Solve Problem 2.3 on $\mathcal{T}_{n+1}$ to find the eigenpair $(\zeta_{l,n+1}, u_{l,n+1})$.
```

9: Let $n := n + 1$ and go to Step 5.

The above algorithm is fairly standard. We estimate according to a standard residual error estimator and refine a number of elements in each step that contribute a certain fraction of the overall estimated error. That the discrete eigenpair in Algorithm 2.19 actually converges to the corresponding eigenpair of Problem 2.3 will be formulated in the next theorem, which is taken from [**17**]. Remark 2.9 of the mentioned article clearly states that the result can be used for our eigenvalue problem by employing a shift as was done in Problem 2.3. The proof given in [**17**] uses a similar result for an adaptive convergent algorithm to solve the boundary value problem that corresponds to the eigenvalue problem, which is given in [**9**]. The authors of [**9**] were the first to prove convergence of a standard adaptive algorithm without the need to refine according to an oscillation term similar to the one introduced in (2.35), when g is replaced by some terms appearing in the estimator, and without any additional assumptions, such as a saturation assumption for instance. Though no refinement needs to be done according to an oscillation term, these kinds of terms still play some role in the convergence analysis.

THEOREM 2.20 (Convergence of the adaptive finite element method). *Let (ζ_l, u_l) be some simple eigenpair of Problem 2.7 and $\{(\zeta_{l,n}, u_{l,n})\}_{n\in\mathbb{N}_0}$ be the sequence of finite element solutions produced by Algorithm 2.19. Then there exist constants $\gamma > 0$ and $\alpha \in (0, 1)$, depending only on the shape regularity of the meshes, c and C_a, from the continuity and coercivity estimates, respectively, and the parameter θ used by Algorithm 2.19, such that for two consecutive iterates n and $n + 1$, we have*

$$|||u_l - u_{l,n+1}|||^2_{a,\mathbf{k},\sigma} + \gamma\eta^2_{n+1} \leq \alpha^2 \left(|||u_l - u_{l,n}|||^2_{a,\mathbf{k},\sigma} + \gamma\eta^2_n\right). \tag{2.45}$$

PROOF. The proof can be found in [**17**]. □

3. Investigating the effect of quadrature

In Section 2 we assumed exact integration of all the integrals that had to be evaluated in the solution process and in the a posteriori error estimation. Simple quadrature formulas can be used if the finite element mesh, which we unlike in Section 2.3 denote by $\mathcal{T}_h$, is aligned with the jumps of the function ϵ, i.e. if the restriction of ϵ to any element $T \in \mathcal{T}_h$ is a smooth function. We use in this section the index h (instead of n) as the mesh size of the triangulation, so that the discrete space denoted earlier as V_n now is denoted as V_h. $h_n^{\max}$ now simply is denoted by h.

We would like to study the effect of evaluating the integrals by means of quadrature both in the assembly of the finite element matrices and in the error estimation. We distinguish the cases where the mesh is aligned with the jumps of ϵ and where it is not.

Our notation will be similar to the one used in [**10**] and [**5**]. Throughout this section we assume that the integral appearing in the bilinear form $b(\cdot,\cdot)$ is decomposed into its different contributions from different elements. On each element we use a quadrature rule of the following form to evaluate integrals

$$\int_T \phi \approx \sum_{l=1}^{L} \omega_{l,T}\phi(\mathbf{x}_{l,T}), \tag{3.1}$$

where we assume that the weights $\omega_{l,T}$ and quadrature nodes $\mathbf{x}_{l,T}$ are induced by a quadrature rule on a reference triangle (see [**10**] for details). We furthermore assume that there exists $t \geq 1$, such that for each $T \in \mathcal{T}_h$ there holds:

$$\int_T \phi = \sum_{l=1}^{L} \omega_{l,T}\phi(\mathbf{x}_{l,T}) \quad \forall \phi \in \mathcal{P}_t(T), \tag{3.2}$$

where $\mathcal{P}_t(T)$ denotes the space of polynomials of degree at most t. We then say that the quadrature scheme has degree of precision t.

We are still interested in Problem 2.2 of the previous Section. That is why we study the approximation of eigenpairs of Problem 2.3 by eigenpairs of Problem 2.7, when the integrals that appear in Problem 2.7 are evaluated numerically. Actually, instead of assuming that we approximate any given eigenpair (ζ_l, u_l) by solving Problem 2.7, we need to do a convergence analysis for the following discrete problem.

<u>Problem</u> 3.1. Seek eigenpairs $(\tilde{\zeta}_{j,h}, \tilde{u}_{j,h}) \in \mathbb{C} \times V_h$, with $||u_{j,h}||_{\tilde{b}_h} = 1$ and

$$a_{\mathbf{k},\sigma}(\tilde{u}_{j,h}, v_h) = \tilde{\zeta}_{j,h}\ \tilde{b}_h(\tilde{u}_{j,h}, v_h) \quad \text{for all } v_h \in V_h, \tag{3.3}$$

where for $u_h, v_h \in V_h$

$$\tilde{b}_h(u_h, v_h) := \sum_{T\in\mathcal{T}_h}\sum_{l=1}^{L} \omega_{l,T}\epsilon(\mathbf{x}_{l,T})u_h(\mathbf{x}_{l,T})\overline{v_h(\mathbf{x}_{l,T})}, \tag{3.4}$$

and

$$||v_h||_{\tilde{b}_h} := \left\{\tilde{b}_h(v_h, v_h)\right\}^{\frac{1}{2}}, \tag{3.5}$$

and the weights $\omega_{l,T}$ and quadrature nodes $\mathbf{x}_{l,T}$ are chosen, such that the scheme has degree of precision t.

To accomodate an easy analyis we introduce a function $\tilde{\epsilon}$ as an approximation to ϵ. The restriction of $\tilde{\epsilon}$ to an element $T \in \mathcal{T}_h$ will be denoted by $\tilde{\epsilon}_T$ and is defined as the polynomial of least possible degree, such that

$$\int_T \tilde{\epsilon}_T\ u_h\ \overline{v_h} := \sum_{l=1}^{L} \omega_{l,T}\epsilon(\mathbf{x}_{l,T})u_h(\mathbf{x}_{l,T})\overline{v_h(\mathbf{x}_{l,T})} \quad \text{for all } u_h, v_h \in V_h. \tag{3.6}$$

Since u_h and v_h are from the discrete space V_h, the point evaluations on the right-hand side of equation (3.6) make sense. From this definition the following property of the piecewise polynomial $\tilde{\epsilon}$ follows:

$$\begin{aligned}\tilde{b}_h(u_h, v_h) &= \sum_{T\in\mathfrak{T}_h}\sum_{l=1}^{L}\omega_{l,T}\epsilon(\mathbf{x}_{l,T})u_h(\mathbf{x}_{l,T})\overline{v_h(\mathbf{x}_{l,T})} \\ &= \int_\Omega \tilde{\epsilon}\; u_h\; \overline{v_h} \quad \text{for all } u_h, v_h \in V_h.\end{aligned} \tag{3.7}$$

Since we have assumed that the quadrature scheme that is employed has degree of precision $t \geq 1$, the following fact easily follows, which we formulate as an assumption.

Assumption 3.2. The following convergence in $L^\infty(\Omega)$ for the function $\tilde{\epsilon}$, defined in (3.7), towards ϵ holds:

$$\lim_{h\to 0} ||\epsilon - \tilde{\epsilon}||_{L^\infty(\Omega)} = 0. \tag{3.8}$$

3.1. A priori estimates. The a priori analysis in the absence of exact integration uses the same tools as in the standard a priori analyis for eigenvalue problems that can be found in [**2**] and [**3**]. The analysis has been carried out for smooth coefficients in [**5**] and [**6**]. We quickly summarize the results corresponding to our problem and show what can be achieved in the case of discontinuous coefficients.

We recall the definition of the solution operator T to a corresponding boundary value problem from (2.24) in Section 2.4. We state it once more and define similar operators for the discrete Problems 2.7 and 3.1.

Definition 3.3. Let the operators T, T_h and $\tilde{T}_h$ be defined as in [**3**] with the obvious modifications, i.e. T, T_h: $H^1_{\text{per}}(\Omega)\rightarrow H^1_{\text{per}}(\Omega)$ are defined by

$$\begin{cases} Tf \in H^1_{\text{per}}(\Omega) \\ a_{\mathbf{k},\sigma}(Tf, v) = b(f, v) \quad \forall v \in H^1_{\text{per}}(\Omega), \end{cases} \tag{3.9}$$

$$\begin{cases} T_h f \in V_h \subset H^1_{\text{per}}(\Omega) \\ a_{\mathbf{k},\sigma}(T_h f, v_h) = b(f, v_h) \quad \forall v_h \in V_h, \end{cases} \tag{3.10}$$

and $\tilde{T}_h : V_h \rightarrow V_h$ is defined by:

$$\begin{cases} \tilde{T}_h f \in V_h \subset H^1_{\text{per}}(\Omega) \\ a_{\mathbf{k},\sigma}(\tilde{T}_h f, v_h) = \tilde{b}(f, v_h) \quad \forall v_h \in V_h \end{cases} \tag{3.11}$$

The eigenvalues of T are the reciprocals of the eigenvalues of Problem 2.3, and T and Problem 2.3 have the same eigenfunctions. Similarly, the non-zero eigenvalues of T_h and $\tilde{T}_h$ respectively, are the reciprocals of those of Problems 2.7 and 3.1, respectively, and they have the same eigenfunctions. In our analysis of eigenvalue error with quadrature we will compare T, T_h and $\tilde{T}_h$. T will be viewed as an operator on $H^1_{\text{per}}(\Omega)$ equipped with the norm $|||\cdot|||_{a,\mathbf{k},\sigma}$ and $\tilde{T}_h$ as an operator on V_h. T_h will be viewed as an operator either on $H^1_{\text{per}}(\Omega)$ equipped with the norm $|||\cdot|||_{a,\mathbf{k},\sigma}$ or as an operator on V_h. In Assumption 3.5 and Theorem 3.10 the operators T, T_h and $\tilde{T}_h$ will be viewed as acting on the same spaces as described above, though equipped with the $L^2(\Omega)$ norm.

Lemma 3.4 (Convergence of $\tilde{T}_h$ to T). *There holds*

$$\lim_{h\to 0} \left|\left|\left|T - \tilde{T}_h\right|\right|\right|_{a,\mathbf{k},\sigma,V_h} = 0, \tag{3.12}$$

where for any linear mapping $A_h : V_h \to H^1_0(\Omega)$ with $V_h \subset H^1_0(\Omega)$:

$$|||A_h|||_{a,\mathbf{k},\sigma,V_h} = \sup_{g\in V_h} \frac{|||A_h g|||_{a,\mathbf{k},\sigma}}{|||g|||_{a,\mathbf{k},\sigma}}. \tag{3.13}$$

Proof. As in [**5**] and [**10**], we apply the first Strang lemma. Since the bilinear form $a_{\mathbf{k},\sigma}(\cdot,\cdot)$ is not approximated, we can conclude that

$$\left|\left|\left|Tf - \tilde{T}_h f\right|\right|\right|_{a,\mathbf{k},\sigma} \lesssim \inf_{v_h\in V_h} ||f - v_h||_{H^1(\Omega)} + \inf_{w_h\in V_h} \frac{\left|b(v_h,w_h) - \tilde{b}_h(v_h,w_h)\right|}{|||w_h|||_{a,\mathbf{k},\sigma}}, \tag{3.14}$$

for arbitrary $f \in V_h$ arbitrary. Choosing $v_h := \Pi_h f$ as the Scott–Zhang interpolation operator (*cf.* Remark 2.5) we furthermore estimate for arbitrary $w \in V_h$:

$$\begin{aligned} \left|b(\Pi_h f, w_h) - \tilde{b}_h(\Pi_h f, w_h)\right| &\le \left|\int_\Omega (\epsilon - \tilde{\epsilon})\ \Pi_h f\ \overline{w_h}\right| \\ &\lesssim ||\epsilon - \tilde{\epsilon}||_{L^\infty(\Omega)} ||\Pi_h f||_{L^2(\Omega)} ||w_h||_{L^2(\Omega)} \\ &\lesssim ||\epsilon - \tilde{\epsilon}||_{L^\infty(\Omega)} ||\Pi_h f||_{H^1(\Omega)} ||w_h||_{L^2(\Omega)} \\ &\lesssim ||\epsilon - \tilde{\epsilon}||_{L^\infty(\Omega)} ||f||_{H^1(\Omega)} ||w_h||_{L^2(\Omega)}, \end{aligned} \tag{3.15}$$

where we have used the stability of the interpolation Π_h (2.13). Since Π_h is a projection onto V_h (*cf.* (2.14)) we conclude that

$$\left|\left|\left|Tf - \tilde{T}_h f\right|\right|\right|_{a,\mathbf{k},\sigma} \lesssim ||\epsilon - \tilde{\epsilon}||_{L^\infty(\Omega)} ||f||_{H^1(\Omega)}. \tag{3.16}$$

Thus

$$\left|\left|\left|T - \tilde{T}_h\right|\right|\right|_{a,\mathbf{k},\sigma,V_h} \lesssim ||\epsilon - \tilde{\epsilon}||_{L^\infty(\Omega)}, \tag{3.17}$$

and we conclude that (3.12) holds from the fact that

$$\lim_{h\to 0} ||\epsilon - \tilde{\epsilon}||_{L^\infty(\Omega)} = 0, \tag{3.18}$$

that we have assumed in Assumption 3.2. □

Assumption 3.5. We assume the following convergence result, when we consider the operators T and $\tilde{T}_h$ for $L^2(\Omega)$ norms:

$$\lim_{h\to 0} \left|\left|T - \tilde{T}_h\right|\right|_{0,V_h} = 0, \tag{3.19}$$

where for any linear mapping $A_h : V_h \to H_0^1(\Omega)$ with $V_h \subset H_0^1(\Omega)$:

$$||A_h||_{0,V_h} = \sup_{g\in V_h} \frac{||A_h g||_{L^2(\Omega)}}{||g||_{L^2(\Omega)}}. \tag{3.20}$$

Let $F \in \rho(T)$, the resolvent set of T, be a closed set. Then we additionaly assume that

$$\left|\left|\left(z - \tilde{T}_h\right)^{-1}\right|\right|_{0,V_h} \leq C, \quad \forall z \in F, \quad \forall \text{ small } h. \tag{3.21}$$

The following result, which can be found in [**5**], is also valid in the periodic case that we are considering.

Theorem 3.6 (Convergence result in the case of quadrature and smooth coefficients). *Let $m \in \mathbb{N}$, and $(\zeta_j, u_j) \in \mathbb{C} \times \left(H^1_{\text{per}}(\Omega) \cap H^{m+1}(\Omega)\right)$ be the solution of Problem 2.3. Let $(\tilde{\zeta}_{j,h}, \tilde{u}_{j,h}) \in \mathbb{C} \times V_h$ be its approximation, the solution of Problem 3.1, where the degree of precision of the quadrature scheme employed in the bilinear form $\tilde{b}_h(\cdot,\cdot)$ in equation (3.4) is supposed to be at least $t = 2m - 1$. Let furthermore $\epsilon\in C^\infty_{\text{per}}(\bar{\Omega})$. Then for h sufficiently small we have the following estimate:*

$$|||u - \tilde{u}_{j,h}|||_{a,\mathbf{k},\sigma} \leq Ch^{\min\{m,r\}}, \tag{3.22}$$

and

$$\left|\zeta_j - \tilde{\zeta}_{j,h}\right| \leq Ch^{2\min\{m,r\}}, \tag{3.23}$$

where r is the degree of the polynomials in the piecewise polynomial space V_h.

Now we wish to make an analysis in the case of discontinuous coefficients. For simplicity, we restrict ourselves to the case of simple eigenvalues.

Theorem 3.7 (Convergence result in the case of quadrature and discontinuous coefficients). *Let $(\tilde{\zeta}_{j,h}, \tilde{u}_{j,h}) \in \mathbb{C} \times V_h$ be an eigenpair of Problem 3.1 that approximates a simple eigenpair $(\zeta_j, u_j) \in \mathbb{C} \times \left(H^1_{\text{per}}(\Omega) \cap H^{m+1}(\Omega)\right)$ of Problem 2.3. Then for h sufficiently small we have the following estimate:*

$$|||u_j - \tilde{u}_{j,h}|||_{a,\mathbf{k},\sigma} \lesssim \; h^{\min\{m,r\}} + \|(\epsilon - \tilde{\epsilon})E_h(\zeta_{j,h})u_j\|_{L^2(\Omega)}\,, \tag{3.24}$$

where r is the degree of the polynomials in the piecewise polynomial space V_h and $E_h(\zeta_{j,h}) : H^1_{\text{per}}(\Omega) \to H^1_{\text{per}}(\Omega)$ is the spectral projection onto the eigenspace of $\zeta_{j,h}$:

$$E_h(\zeta_{j,h}) = \frac{1}{2\pi \text{i}} \int_{\Gamma_j} (z - T_h)^{-1} dz. \tag{3.25}$$

Proof. The proof uses the ideas of [**2**]. Recall the definitions (3.9), (3.10) and (3.11). Let Γ_j be a circle of radius $r\left(\Gamma_j\right)$ in the complex plane centered at $\mu_j = \zeta_j^{-1}$ and enclosing no other eigenvalues of T. Then for h sufficiently small $\tilde{\mu}_{j,h} = (\tilde{\zeta}_{j,h})^{-1}$ but no other eigenvalues of $\tilde{T}_h$ are contained in Γ_j and

$$E(\zeta_j) = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_j} (z-T)^{-1} dz, \tag{3.26}$$

$$\tilde{E}_h(\tilde{\zeta}_{j,h}) = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_j} (z-\tilde{T}_h)^{-1} dz, \tag{3.27}$$

where $E(\zeta_j) : H^1_{\mathrm{per}}(\Omega) \rightarrow H^1_{\mathrm{per}}(\Omega)$ and $\tilde{E}_h(\tilde{\zeta}_{j,h}) : V_h \rightarrow V_h$ are the spectral projections onto the eigenspaces of ζ_j, and $\tilde{\zeta}_{j,h}$, respectively. Using (3.26), (3.25) and (3.27) we calculate

$$\begin{aligned}
&\left|\left|\left| u_j - \tilde{E}_h(\tilde{\zeta}_{j,h}) E_h(\zeta_{j,h}) u_j \right|\right|\right|_{a,\mathbf{k},\sigma} \\
&\lesssim |||E(\zeta_j)u_j - E_h(\zeta_{j,h})u_j|||_{a,\mathbf{k},\sigma} \\
&\quad + \left|\left|\left| E_h(\zeta_{j,h})u_j - \tilde{E}_h(\tilde{\zeta}_{j,h}) E_h(\zeta_{j,h}) u_j \right|\right|\right|_{a,\mathbf{k},\sigma} \\
&\lesssim \left|\left|\left| \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_j} \left[(z-T)^{-1} - (z-T_h)^{-1}\right] u_j dz \right|\right|\right|_{a,\mathbf{k},\sigma} \\
&\quad + \left|\left|\left| \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_j} \left[(z-T_h)^{-1} - (z-\tilde{T}_h)^{-1}\right] E_h(\zeta_{j,h}) u_j dz \right|\right|\right|_{a,\mathbf{k},\sigma} \\
&\lesssim \left|\left|\left| \frac{1}{2\pi} \int_{\Gamma_j} (z-T_h)^{-1}(T-T_h)(z-T)^{-1} u_j dz \right|\right|\right|_{a,\mathbf{k},\sigma} \\
&\quad + \left|\left|\left| \frac{1}{2\pi} \int_{\Gamma_j} (z-\tilde{T}_h)^{-1}(T_h-\tilde{T}_h)(z-T_h)^{-1} E_h(\zeta_{j,h}) u_j dz \right|\right|\right|_{a,\mathbf{k},\sigma} \\
&\lesssim \left|\left|\left| \int_{\Gamma_j} (z-T_h)^{-1}(T-T_h) \frac{u_j}{z-\mu_j} dz \right|\right|\right|_{a,\mathbf{k},\sigma} \\
&\quad + \left|\left|\left| \int_{\Gamma_j} (z-\tilde{T}_h)^{-1}(T_h-\tilde{T}_h)(z-T_h)^{-1} E_h(\zeta_{j,h}) u_j dz \right|\right|\right|_{a,\mathbf{k},\sigma} .
\end{aligned} \tag{3.28}$$

We continue the estimate by

$$\begin{aligned}
&\left|\left|\left|u_j - \tilde{E}(\tilde{\zeta}_{j,h})E_h(\zeta_{j,h})u_j\right|\right|\right|_{a,\mathbf{k},\sigma} \\
\leq\ & r\left(\Gamma_j\right) \sup_{z\in\Gamma_j,0<h} \left|\left|\left|(z-T_h)^{-1}\right|\right|\right|_{a,\mathbf{k},\sigma,H_0^1(\Omega)} \frac{\left|\left|\left|(T-T_h)u_j\right|\right|\right|_{a,\mathbf{k},\sigma}}{r\left(\Gamma_j\right)} \\
&+ r\left(\Gamma_j\right) \sup_{z\in\Gamma_j,0<h} \left|\left|\left|(z-\tilde{T}_h)^{-1}\right|\right|\right|_{a,\mathbf{k},\sigma,V_h} \frac{\left|\left|\left|(T-\tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|\right|_{a,\mathbf{k},\sigma}}{r\left(\Gamma_j\right)} \\
=\ & \mu_j \sup_{x\in\Gamma_j,0<h} \left|\left|\left|(z-T_h)^{-1}\right|\right|\right|_{a,\mathbf{k},\sigma,H_0^1(\Omega)} \left|\left|\left|(I-P_h)u_j\right|\right|\right|_{a,\mathbf{k},\sigma} \\
&+ r\left(\Gamma_j\right) \sup_{z\in\Gamma_j,0<h} \left|\left|\left|(z-\tilde{T}_h)^{-1}\right|\right|\right|_{a,\mathbf{k},\sigma,V_h} \frac{\left|\left|\left|(T_h-\tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|\right|_{a,\mathbf{k},\sigma}}{r\left(\Gamma_j\right)}
\end{aligned} \tag{3.29}$$

In the last equality we used the relation $(T-T_h)u_j = (I-P_h)Tu_j = \mu_j(I-P_h)u_j$, where $P_h : H_0^1(\Omega) \to V_h$ is the Rayleigh-Ritz projection, i.e.

$$a_{\mathbf{k},\sigma}(u,v_h) = a_{\mathbf{k},\sigma}(P_h u, v_h) \quad \forall v_h \in V_h. \tag{3.30}$$

Now $\lim\limits_{h\to 0}\left|\left|\left|T-\tilde{T}_h\right|\right|\right|_{a,\mathbf{k},\sigma,V_h} = 0$ from Lemma 3.4 implies that (*cf.* (3.7) in [**5**])

$$\sup_{z\in\Gamma_j,0<h} \left|\left|\left|(z-\tilde{T}_h)^{-1}\right|\right|\right|_{a,\mathbf{k},\sigma,V_h} < \infty. \tag{3.31}$$

Furthermore, we have that (*cf.* [**2**])

$$\sup_{z\in\Gamma_j,0<h} \left|\left|\left|(z-T_h)^{-1}\right|\right|\right|_{a,\mathbf{k},\sigma,H_0^1(\Omega)} < \infty. \tag{3.32}$$

Hence we conclude that

$$\begin{aligned}
&\left|\left|\left|u_j - \tilde{E}_h(\tilde{\zeta}_{j,h})E_h(\zeta_{j,h})u_j\right|\right|\right|_{a,\mathbf{k},\sigma} \\
&\qquad\lesssim \left|\left|\left|(I-P_h)u_j\right|\right|\right|_{a,\mathbf{k},\sigma} + \left|\left|\left|(T_h-\tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|\right|_{a,\mathbf{k},\sigma}.
\end{aligned} \tag{3.33}$$

In order to achieve the desired estimate, we now have to estimate the term

$$\left|\left|\left|(T_h-\tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|\right|_{a,\mathbf{k},\sigma}. \tag{3.34}$$

Observing that $(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j \in V_h$ we are able to conclude that

$$\begin{aligned}
&\left|\left|\left|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|\right|^2_{a,\mathbf{k},\sigma} \\
&\quad \lesssim a_{\mathbf{k},\sigma}((T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j, (T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j) \\
&\quad = b(E_h(\zeta_{j,h})u_j, (T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j) \\
&\qquad - \tilde{b}_h(E_h(\zeta_{j,h})u_j, (T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j) \\
&\quad = \int_\Omega (\epsilon - \tilde{\epsilon})E_h(\zeta_{j,h})u_j((T_h - \tilde{T}_h)u_j) \\
&\quad \leq C \left\|(\epsilon - \tilde{\epsilon})E_h(\zeta_{j,h})u_j\right\|_{L^2(\Omega)} \left\|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right\|_{L^2(\Omega)}.
\end{aligned} \tag{3.35}$$

Since $\left\|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right\|_{L^2(\Omega)} \leq \left|\left|\left|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|\right|_{a,\mathbf{k},\sigma}$ we thus end up with the estimate

$$\left|\left|\left|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|\right|_{a,\mathbf{k},\sigma} \leq C \left\|(\epsilon - \tilde{\epsilon})E_h(\zeta_{j,h})u_j\right\|_{L^2(\Omega)}. \tag{3.36}$$

Combining (3.33) and (3.36) we achieve the estimate

$$|||u_j - u_{j,h}|||_{a,\mathbf{k},\sigma} \lesssim |||(I - P_h)u_j|||_{a,\mathbf{k},\sigma} + \left\|(\epsilon - \tilde{\epsilon})E_h(\zeta_{j,h})u_j\right\|_{L^2(\Omega)} \tag{3.37}$$

Hence we conclude that:

$$|||u_j - u_{j,h}|||_{a,\mathbf{k},\sigma} \lesssim\ h^m + \left\|(\epsilon - \tilde{\epsilon})E_h(\zeta_{j,h})u_j\right\|_{L^2(\Omega)}. \tag{3.38}$$

□

REMARK 3.8. In one space dimension we can achieve a more refined estimation: As

$$\begin{aligned}
\left\|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right\|_{L^\infty(\Omega)} &\lesssim \left\|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right\|_{H^1(\Omega)} \\
&\lesssim \left|\left|\left|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|\right|_{a,\mathbf{k},\sigma},
\end{aligned} \tag{3.39}$$

we achieve

$$\left|\left|\left|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|\right|_{a,\mathbf{k},\sigma} \lesssim \left\|(\epsilon - \tilde{\epsilon})E_h(\zeta_{j,h})u_j\right\|_{L^1(\Omega)} \tag{3.40}$$

instead of estimate (3.36). In the special case that ϵ only has jumps in a finite number of intervals, say T_i, $i = 1, \ldots, N$, and stays constant everywhere else we can furthermore estimate to get:

$$
\begin{aligned}
&\left\|(\epsilon-\tilde{\epsilon})E_h(\zeta_{j,h})u_j\right\|_{L^1(\Omega)}\\
&\qquad\le \sum_{i=1,\dots,N}\left\|(\epsilon-\tilde{\epsilon})E_h(\zeta_{j,h})u_j\right\|_{L^1(T_i)}\\
&\qquad\le \sum_{i=1,\dots,N}\left\|\epsilon-\tilde{\epsilon}\right\|_{L^\infty(T)}\left\|E_h(\zeta_{j,h})u_j\right\|_{L^\infty(T)}|T_i|\\
&\qquad\le Ch.
\end{aligned}
\tag{3.41}
$$

Theorem 3.9 (Relation between eigenvalue and eigenfunction error in the case of quadrature). *Let $(\tilde{\zeta}_{j,h},\tilde{u}_{j,h})\in\mathbb{R}\times V_h$ be a solution of Problem 3.1 that approximates the eigenpair $(\zeta_j,u_j)\in\mathbb{R}\times H^1_{\text{per}}(\Omega)$ of Problem 2.3. Then we have the following estimate:*

$$
\begin{aligned}
\left|\zeta_j-\tilde{\zeta}_{j,h}\right| \le & a_{\mathbf{k},\sigma}(u_j-\tilde{u}_{j,h},u_j-\tilde{u}_{j,h})\\
&\qquad + \zeta_j b(u_j-\tilde{u}_{j,h},u_j-\tilde{u}_{j,h})+\zeta_j\int_\Omega|\tilde{\epsilon}-\epsilon|\,|\tilde{u}_{j,h}|^2.
\end{aligned}
\tag{3.42}
$$

Proof. A simple calculation yields:

$$
\begin{aligned}
a_{\mathbf{k},\sigma}(u_j-\tilde{u}_{j,h},u_j-\tilde{u}_{j,h}) =&\zeta_j+\tilde{\zeta}_{j,h}-a(u_j,\tilde{u}_{j,h})-\overline{a_{\mathbf{k},\sigma}(u_j,\tilde{u}_{j,h})}\\
=&\tilde{\zeta}_{j,h}-\zeta_j+2\zeta_j-2\Re\left\{a_{\mathbf{k},\sigma}(u_j,\tilde{u}_{j,h})\right\}\\
=&\tilde{\zeta}_{j,h}-\zeta_j+2\zeta_j-2\Re\left\{\zeta_j\ b(u_j,\tilde{u}_{j,h})\right\}\\
=&\tilde{\zeta}_{j,h}-\zeta_j+\zeta_j\left(2-2\Re\left\{\ b(u_j,\tilde{u}_{j,h})\right\}\right)\\
=&\tilde{\zeta}_{j,h}-\zeta_j+\zeta_j b(u_j-\tilde{u}_{j,h},u_j-\tilde{u}_{j,h})\\
&\quad+\zeta_j\int_\Omega(\epsilon-\tilde{\epsilon})\,|\tilde{u}_{j,h}|^2.
\end{aligned}
\tag{3.43}
$$

From this we conclude that (3.42) holds. □

Theorem 3.10 ($L^2(\Omega)$ estimate of the eigenfunctions). *Let $(\tilde{\zeta}_{j,h},\tilde{u}_{j,h})\in\mathbb{C}\times V_h$ be an eigenpair of Problem 3.1 that approximates a simple eigenpair $(\zeta_j,u_j)\in\mathbb{C}\times\left(H^1_{\text{per}}(\Omega)\cap H^{m+1}(\Omega)\right)$ of Problem 2.3. Then for h sufficiently small we have the following estimate:*

$$
\left\|u_j-\tilde{u}_{j,h}\right\|_{L^2(\Omega)}\lesssim\ h^{\min\{m,r\}+1}+\left\|(\epsilon-\tilde{\epsilon})E_h(\zeta_{j,h})u_j\right\|_{L^2(\Omega)},
\tag{3.44}
$$

where r is the degree of the polynomials in the piecewise polynomial space V_h and $E_h(\zeta_{j,h}):H^1_{\text{per}}(\Omega)\to H^1_{\text{per}}(\Omega)$ is the spectral projection onto the eigenspace of $\zeta_{j,h}$:

$$
E_h(\zeta_{j,h})=\frac{1}{2\pi\mathrm{i}}\int_{\Gamma_j}(z-T_h)^{-1}dz.
\tag{3.45}
$$

Proof. The proof is quite similar to the proof of Theorem 3.7. This time the operators T and T_h from (3.9) and (3.10), respectively, are considered from $L^2(\Omega)$

into $L^2(\Omega)$. $\tilde{T}_h$ from (3.11) will still be considered as acting on V_h. We recall the spectral projections from (3.26) and (3.27),

$$E(\zeta_j) = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_j} (z - T)^{-1} dz,$$

$$\tilde{E}_h(\tilde{\zeta}_{j,h}) = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_j} (z - \tilde{T}_h)^{-1} dz.$$

$\tilde{E}_h(\zeta_{j,h})$ is a map from V_h to V_h and $E(\zeta_j)$ acts on $H^1_{\mathrm{per}}(\Omega)$. Using similar steps as we did in Theorem 3.7 we arrive at the estimate

$$\begin{aligned}
&\left|\left|u_j - \tilde{E}_h(\tilde{\zeta}_{j,h})E_h(\zeta_{j,h})u_j\right|\right|_{L^2(\Omega)} \\
&\quad \leq r\left(\Gamma_j\right) \sup_{z \in \Gamma_j, 0<h} \left|\left|(z - T_h)^{-1}\right|\right|_{L^2(\Omega)\to L^2(\Omega)} \frac{\left|\left|(T - T_h)u_j\right|\right|_{L^2(\Omega)}}{r\left(\Gamma_j\right)} \\
&\qquad + r\left(\Gamma_j\right) \sup_{z \in \Gamma_j, 0<h} \left|\left|(z - \tilde{T}_h)^{-1}\right|\right|_{0,V_h} \frac{\left|\left|(T - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|_{L^2(\Omega)}}{r\left(\Gamma_j\right)} \\
&\quad = \mu_j \sup_{z \in \Gamma_j, 0<h} \left|\left|(z - T_h)^{-1}\right|\right|_{L^2(\Omega)\to L^2(\Omega)} \left|\left|(I - P_h)u_j\right|\right|_{L^2(\Omega)} \\
&\qquad + r\left(\Gamma_j\right) \sup_{z \in \Gamma_j, 0<h} \left|\left|(z - \tilde{T}_h)^{-1}\right|\right|_{0,V_h} \frac{\left|\left|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|_{L^2(\Omega)}}{r\left(\Gamma_j\right)}.
\end{aligned} \tag{3.46}$$

From Assumption 3.5 we can deduce that

$$\sup_{z \in \Gamma_j, 0<h} \left|\left|(z - \tilde{T}_h)^{-1}\right|\right|_{0,V_h} < \infty. \tag{3.47}$$

The argument that

$$\sup_{z \in \Gamma_j, 0<h} \left|\left|(z - T_h)^{-1}\right|\right|_{L^2(\Omega)\to L^2(\Omega)}, \tag{3.48}$$

i.e. $(z - T_h)^{-1}$ regarded as operator from $L^2(\Omega)$ to $L^2(\Omega)$, is bounded can be found in [**2**]. Hence we conclude that

$$\begin{aligned}
&\left|\left|u_j - \tilde{E}_h(\tilde{\zeta}_{j,h})E_h(\zeta_{j,h})u_j\right|\right|_{L^2(\Omega)} \\
&\qquad \lesssim \left|\left|(I - P_h)u_j\right|\right|_{L^2(\Omega)} + \left|\left|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|_{L^2(\Omega)},
\end{aligned} \tag{3.49}$$

where $P_h : H^1_0(\Omega) \to V_h$ again denotes the Rayleigh-Ritz projection satisfying (3.30). In order to achieve the desired estimate, we now have to estimate the term

$$\left|\left|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|_{L^2(\Omega)}.$$

Observing that $(T_h - \tilde{T}_h)u_j \in V_h$ and that

$$\left|\left|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|_{L^2(\Omega)} \leq \left|\left|\left|(T_h - \tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|\right|_{a,\mathbf{k},\sigma}$$

we are able to conclude that

$$
\begin{aligned}
&\left|\left|(T_h-\tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|^2_{L^2(\Omega)}\\
&\lesssim\ a_{\mathbf{k},\sigma}((T_h-\tilde{T}_h)E_h(\zeta_{j,h})u_j,(T_h-\tilde{T}_h)E_h(\zeta_{j,h})u_j)\\
&=\ b(E_h(\zeta_{j,h})u_j,(T_h-\tilde{T}_h)E_h(\zeta_{j,h})u_j)\\
&\quad-\tilde{b}_h(E_h(\zeta_{j,h})u_j,(T_h-\tilde{T}_h)E_h(\zeta_{j,h})u_j)\\
&=\ \textstyle\int_\Omega(\epsilon-\tilde{\epsilon})E_h(\zeta_{j,h})u_j((T_h-\tilde{T}_h)u_j)\\
&\leq\ C\left\|(\epsilon-\tilde{\epsilon})E_h(\zeta_{j,h})u_j\right\|_{L^2(\Omega)}\left|\left|(T_h-\tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|_{L^2(\Omega)}.
\end{aligned}
\tag{3.50}
$$

Dividing by $\left|\left|(T_h-\tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|_{L^2(\Omega)}$ we thus end up with the estimate

$$
\left|\left|(T_h-\tilde{T}_h)E_h(\zeta_{j,h})u_j\right|\right|_{L^2(\Omega)}\lesssim\ \left\|(\epsilon-\tilde{\epsilon})E_h(\zeta_{j,h})u_j\right\|_{L^2(\Omega)}.
\tag{3.51}
$$

Combining (3.49) and (3.51) we achieve the estimate

$$
||u_j-u_{j,h}||_{L^2(\Omega)}\lesssim||(I-P_h)u_j||_{L^2(\Omega)}+\left\|(\epsilon-\tilde{\epsilon})E_h(\zeta_{j,h})u_j\right\|_{L^2(\Omega)}.
\tag{3.52}
$$

We conclude the proof by noting that $||(I-P_h)E_h(\zeta_{j,h})u_j||_{L^2(\Omega)}\lesssim h^{\min\{m,r\}+1}$. □

Theorem 3.11 (Convergence result in the case of quadrature and discontinuous coefficients II). *Let $(\tilde{\zeta}_{j,h},\tilde{u}_{j,h})\in\mathbb{C}\times V_h$ be an eigenpair of Problem 3.1 that approximates a simple eigenpair $(\zeta_j,u_j)\in\mathbb{C}\times H^1_{\text{per}}(\Omega)$ of Problem 2.3. Then for h sufficiently small we have the following estimate:*

$$
\left|\zeta_j-\tilde{\zeta}_{j,h}\right|\lesssim\ h^{2\min\{m,r\}}+\left\|(\epsilon-\tilde{\epsilon})E_h(\zeta_{j,h})u_j\right\|^2_{L^2(\Omega)}+\zeta_j\int_\Omega|\tilde{\epsilon}-\epsilon|\,|\tilde{u}_{j,h}|^2,
\tag{3.53}
$$

where $E_h(\zeta_{j,h})$ is the spectral projection onto the eigenspace of $\zeta_{j,h}$ introduced in (3.25).

Proof. We start with the estimate (3.42).

$$
\begin{aligned}
\left|\zeta_j-\tilde{\zeta}_{j,h}\right|\leq&a_{\mathbf{k},\sigma}(u_j-\tilde{u}_{j,h},u_j-\tilde{u}_{j,h})\\
&+\zeta_jb(u_j-\tilde{u}_{j,h},u_j-\tilde{u}_{j,h})+\zeta_j\int_\Omega|\tilde{\epsilon}-\epsilon|\,|\tilde{u}_{j,h}|^2.
\end{aligned}
\tag{3.54}
$$

The first summand on the right hand side is the square of the quantitiy that was estimated in Theorem 3.7. Because ϵ is bounded from above and below by positive constants, the second term is equivalent to $||u_j-\tilde{u}_{j,h}||^2_{L^2(\Omega)}$, the square root of which was estimated in Theorem 3.10. □

Remark 3.12. We would like to mention that when we employ simple quadrature rules to evaluate the integrals in the assembly of the matrices and the discontinuities of the dielectricity function ϵ are not resolved by the mesh, the convergence for the eigenvalues need not be faster than the convergence of the eigenfunctions in energy norm. This is in contrast to Theorem 2.11 and even to Theorem 3.6 where the convergence of the eigenvalues is twice as fast as the convergence of the eigenfunctions.

We wish to illustrate this fact for a one-dimensional example, where ϵ only jumps in two intervals and linear elements are employed. We have discovered in Remark 3.8 that in this special setting the energy error of the eigenfunctions can be estimated to converge at least at the speed of

$$|||u_j - \tilde{u}_{j,h}|||_{a,\mathbf{k},\sigma} \lesssim h + \|(\epsilon - \tilde{\epsilon})E_h(\zeta_{j,h})u_j\|_{L^1(\Omega)} \lesssim h \tag{3.55}$$

Starting with Theorem 3.11 and doing similar manipulations as in Remark 3.8, we can conclude that

$$\left|\zeta_j - \tilde{\zeta}_{j,h}\right| \lesssim h^2 + \|(\epsilon - \tilde{\epsilon})E_h(\zeta_{j,h})u_j\|^2_{L^2(\Omega)} + \int_\Omega |\tilde{\epsilon} - \epsilon| \, |\tilde{u}_{j,h}|^2 \lesssim h, \tag{3.56}$$

since $||E_h(\zeta_{j,h})u_j||_{L^\infty(\Omega)}$ and $||u_{j,h}||_{L^\infty(\Omega)}$ are bounded in our special situation.

The inaccuracies in the assembly of the matrices due to quadrature, which are rather significant, thus equally affect the calculated eigenfunctions and eigenvalues. We could observe this fact also in numerical experiments (see Section 5.1 for details).

3.2. A posteriori estimates.

We wish to show a result similar to the one in Theorem 2.17 but without assuming exact integration in the solution procedure. This comes at the expense of some additional terms.

Theorem 3.13 (Reliability of the estimator). *Let $(\tilde{\zeta}_{j,h}, \tilde{u}_{j,h}) \in \mathbb{C} \times H^1_{\mathrm{per}}(\Omega)$ be a simple eigenpair of Problem 3.1. Then for h sufficiently small we have the following estimate:*

$$|||u_j - \tilde{u}_{j,h}|||_{a,\mathbf{k},\sigma} \lesssim \eta_{\mathfrak{T}_h}\left((\zeta_{j,h}, \tilde{u}_{j,h}), \Omega\right) + ||\epsilon - \tilde{\epsilon}||_{L^2(\Omega)} + \zeta_j \int_\Omega |\tilde{\epsilon} - \epsilon| \, |\tilde{u}_{j,h}|^2 + h^2, \tag{3.57}$$

where $\eta_{\mathfrak{T}_h}\left((\tilde{\zeta}_{j,h}, \tilde{u}_{j,h}), \Omega\right)$ has been defined in Definition 2.15.

Proof. We readily calculate for arbitrary $v \in H^1_{\mathrm{per}}(\Omega)$ and arbitrary $v_h \in V_h$, that

$$\begin{aligned} a_{\mathbf{k},\sigma}(u_j - \tilde{u}_{j,h}, v) = \zeta_j \ b(u_j, v) & \\ & + \sum_{T \in \mathfrak{T}} \int_T \Big((\nabla + \mathrm{i}\mathbf{k}) \cdot (\nabla + \mathrm{i}\mathbf{k})\tilde{u}_{j,h} + \tilde{\epsilon}\tilde{\zeta}_{j,h}\tilde{u}_{j,h}\Big) \overline{(v - v_h)} \\ & - \sum_{T \in \mathfrak{T}} \int_{\partial T} \partial_n(\tilde{u}_h + \mathrm{i}\mathbf{k}\tilde{u}_h)\overline{(v - v_h)} - \tilde{\zeta}_{j,h}\tilde{b}(\tilde{u}_{j,h}, v). \end{aligned} \tag{3.58}$$

It is equivalent to

$$\begin{aligned}a_{\mathbf{k},\sigma}(u_j-\tilde{u}_{j,h},v) &= \sum_{T\in\mathfrak{T}}\int_T \Big((\nabla+\mathrm{i}\mathbf{k})\cdot(\nabla+\mathrm{i}\mathbf{k})\tilde{u}_{j,h}+\tilde{\epsilon}\tilde{\zeta}_{j,h}\tilde{u}_{j,h}\Big)\overline{(v-v_h)}\\ &-\frac{1}{2}\sum_{T\in\mathfrak{T}}\int_{\partial T}[\partial_{\mathbf{n}}(\tilde{u}_{j,h}+\mathrm{i}\mathbf{k}\tilde{u}_h)]\,\overline{(v-v_h)}\\ &+\int_\Omega(\epsilon\zeta_j u_j-\tilde{\epsilon}\tilde{\zeta}_{j,h}\tilde{u}_{j,h})\overline{v}.\end{aligned}\tag{3.59}$$

We are going to choose $v := u-\tilde{u}_{j,h}$ and define $v_h := \Pi_h v$, where $\Pi_h:\ H^1_{\mathrm{per}}(\Omega)\to V_h$ is the interpolation operator satisfying Assumption 2.4. This yields:

$$\begin{aligned}a_{\mathbf{k},\sigma}(u_j-\tilde{u}_{j,h},v) &\lesssim \left(\sum_{T\in\mathfrak{T}}\eta_T^2\right)^{\frac{1}{2}}|v|_{H^1(\Omega)}+\int_\Omega(\epsilon\zeta_j u_j-\tilde{\epsilon}\tilde{\zeta}_{j,h}\tilde{u}_{j,h})\overline{v}\\ &\lesssim \left(\sum_{T\in\mathfrak{T}}\eta_T^2\right)^{\frac{1}{2}}|||v|||_{a,\mathbf{k},\sigma}+\int_\Omega\epsilon(\zeta_j u_j-\tilde{\zeta}_{j,h}\tilde{u}_{j,h})\overline{v}\\ &\quad+\int_\Omega(\epsilon-\tilde{\epsilon})\tilde{\zeta}_{j,h}\tilde{u}_{j,h}\overline{v}\\ &\lesssim \left(\sum_{T\in\mathfrak{T}}\eta_T^2\right)^{\frac{1}{2}}|||v|||_{a,\mathbf{k},\sigma}+\int_\Omega\epsilon\zeta_j(u_j-\tilde{u}_{j,h})\overline{v}\\ &\quad+\int_\Omega\epsilon(\zeta_j-\tilde{\zeta}_{j,h})\tilde{u}_{j,h}\overline{v}+\int_\Omega(\epsilon-\tilde{\epsilon})\tilde{\zeta}_{j,h}\tilde{u}_{j,h}\overline{v}.\end{aligned}\tag{3.60}$$

Here we used that $|v|_{H^1(\Omega)}\lesssim|||v|||_{a,\mathbf{k},\sigma}$. Now we employ Theorems 3.9 and 3.10. This leads to:

$$\begin{aligned}|||u_j-\tilde{u}_{j,h}|||^2_{a,\mathbf{k},\sigma} &\lesssim \left(\sum_{T\in\mathfrak{T}}\eta_T^2\right)^{\frac{1}{2}}|||u_j-\tilde{u}_{j,h}|||_{a,\mathbf{k},\sigma}\\ &+\Big(h^{\min\{m,r\}+1}+\|(\epsilon-\tilde{\epsilon})E_h(\zeta_{j,h})u_j\|_{L^2(\Omega)}\Big)\|u_j-\tilde{u}_{j,h}\|_{L^2(\Omega)}\\ &+\left(h^{2\min\{m,r\}}+\|(\epsilon-\tilde{\epsilon})E_h(\zeta_{j,h})u_j\|^2_{L^2(\Omega)}+\int_\Omega|\tilde{\epsilon}-\epsilon|\,|\tilde{u}_{j,h}|^2\right)\\ &\times\|\tilde{u}_{j,h}\|_{L^2(\Omega)}\|u_j-\tilde{u}_{j,h}\|_{L^2(\Omega)}\\ &+\tilde{\zeta}_{j,h}\|(\epsilon-\tilde{\epsilon})\tilde{u}_{j,h}\|_{L^2(\Omega)}\|u_j-\tilde{u}_{j,h}\|_{L^2(\Omega)}.\end{aligned}\tag{3.61}$$

Using the fact that both $\|\tilde{u}_{j,h}\|_{L^2(\Omega)}$ and $\|E_h(\zeta_{j,h})u_j\|_{L^2(\Omega)}$ are bounded as well as the fact that $\|u_j-\tilde{u}_{j,h}\|_{L^2(\Omega)}\lesssim|||u_j-\tilde{u}_{j,h}|||_{a,\mathbf{k},\sigma}$ we can deduce the desired result by dividing by the latter quantity and using the fact that h is small. □

Remark 3.14. One conclusion that could be drawn from Theorem 3.13 is that in order to perform a successful a posteriori analysis of the error in the eigenfunctions in energy norm, we should have some element-wise contributions of the terms $||\epsilon - \tilde{\epsilon}||_{L^2(\Omega)}$ and $\int_\Omega |\tilde{\epsilon} - \epsilon| \, |\tilde{u}_{j,h}|^2$ in the estimator. The terms should be approximated by some easier expressions. If we were to assume that we could evaluate $||\epsilon - \tilde{\epsilon}||_{L^2(T)}$ and $\int_T |\tilde{\epsilon} - \epsilon| \, |\tilde{u}_{j,h}|^2$ exactly, then there would be no point in our whole analysis since in that case the integrals in the assembly of the finite element matrices, which involve similar expressions, should be evaluated exactly in the first place.

Hence we recommend to include terms of the form

$$h_T \left(\bar{\epsilon} - \underline{\epsilon}\right) \tag{3.62}$$

in the estimator $\eta^2_{\mathcal{T}_n}((\tilde{u}_{j,h}, \tilde{\zeta}_{j,h}), T)$, where $\bar{\epsilon}$ and $\underline{\epsilon}$ are from (2.2), in the case of the eigenvalue problem when the discontinuities of ϵ are not resolved by the mesh.

4. An algorithm to calculate adaptively in both Ω and the Brillouin zone

It is the aim of this section to use the results from Sections 2.5 and 2.6 as building bricks in an algorithm to compute an entire band structure adaptively. The ultimate goal is to reduce the quantity *ERROR* from (1.2) using as little as possible computational resources. We recall this quantity here,

$$ERROR := \max_{j \in M} \|\lambda_j(\cdot) - \lambda_{j,\text{num}}(\cdot)\|_{C^0(B)}\,, \tag{4.1}$$

which shall be reduced starting from a combination of a uniform triangulation $\mathcal{T}_0$ in Ω and a set of discrete points $\mathbb{K}_0 := \{\mathbf{k}_i\}_{i=1}^{N_0}$ uniformly distributed in the Brillouin zone. In the course of the calculations we work with only one mesh in Ω, the refinement procedure of which will be quite similar as the one in Algorithm 2.19. Before each refinement in Ω, though, we accumulate contributions to the set marked for refinement from the a posteriori error estimates for the different eigenvalue problems for the different values of the discrete set of parameters $\{\mathbf{k}_i\}_{i=1}^{N_0}$. In this way, we have a routine that solves the eigenvalue problem for various parameters simultaneously using adaptivity in Ω.

In the Brillouin zone, we also create a mesh from the set of discrete points $\mathbb{K}_0 := \{\mathbf{k}_i\}_{i=1}^{N_0}$. If $N_{0,\text{I}}$ is the number of elements of the initial mesh in the Brillouin zone, we define an array $\{I_j\}_{j=1}^{N_{0,\text{I}}}$ with zero entries. The zeros indicate that we have not evaluated yet, if the parameters are chosen dense enough. Elements of the mesh, that will not be refined further will be set to 1 in the algorithm. In the course of the calculations, since we do not know the true band functions $\lambda_j(\cdot)$, we estimate the expression (4.1) by means of further band function evaluations, i.e. further eigenvalue problems to be solved.

Algorithm 4.1. (Adaptive bandstructure calculations)

```
1: Choose TOL $> 0$ .
2: Choose $0 < \theta < 1$.
3: Set $n$:=0.
4: Pick any initial mesh $\mathcal{T}_0$ and initial set of discrete points $\mathbb{K}_0 := \{\mathbf{k}_i\}_{i=1}^{N_0} \subset B$,
   which are the nodes of a conforming triangulation of $B$.
5: Solve for eigenpairs $\{(\zeta_{l,0,\mathbf{k}_i}, u_{l,0,\mathbf{k}_i})\}_{\mathbf{k}_i \in \mathbb{K}_0, l \in M}$ using Algorithm 4.2 with $\mathcal{T}_0$, $\mathbb{K}_0$
   and $\theta$ as input.
6: Define array $\mathrm{I} = \{\mathrm{I}_j\}_{j=1}^{N_{0,\mathrm{I}}}$ of size $N_{0,\mathrm{I}}$ with zero entries.
7: while $\{i : I_i = 0\}$ is not empty do
8:    Set $\mathbb{V} = \emptyset$.
9:    for $j \in \{i : \ \mathrm{I}_i \neq 1\}$ do
10:      Compute the centers $e_{j,1}$, $e_{j,2}$ and $e_{j,3}$ of all edges of $I_j$ and add them to a
         list $\mathbb{V} := \mathbb{V} \cup e_{j,1} \cup e_{j,2} \cup e_{j,3}$.
11:   end for
12:   Create new elements of the mesh in $B$ by joining all elements from $\mathbb{V}$. The
      old elements are removed from the list I, and the newly created elements are
      added to the list with value 0.
13:   Solve for eigenpairs $\{(\zeta_{l,n,\mathbf{k}_i}, u_{l,n,\mathbf{k}_i})\}_{\mathbf{k}_i \in \mathbb{V}, l \in M}$ with triangulation $\mathcal{T}_n$ using Al-
      gorithm 4.2 with $\mathcal{T}_n$ and $\mathbb{V}$ as input.
14:   for $i \in \{j : \ \mathrm{I}_j \neq 1\}$ do
15:      if $\mathrm{I}_i$ is such that for any $\mathbf{k}_i \in \mathbb{V}$ in the closure of the element $\mathrm{I}_i$ the expression
         $\zeta_{l,n,\mathbf{k}_i}$ from step 13 differs at most TOL from the linear interpolation, which
         is the unique linear function joining the eigenvalues previously calculated
         for the boundary nodes of the element, for all $l \in M$ then
16:         Set $\mathrm{I}_i := 1$.
17:      end if
18:   end for
19: end while
```

Algorithm 4.2. (Adaptive solution of the eigenvalue problem for various $\mathbf{k}_i$ at once)

```
1: Input: Initial mesh $\mathcal{T}_0$ on $\Omega$, set of discrete points $\mathbb{K}$ in the Brillouin zone $B$, $n$,
   $\theta$.
2: Solve Problem 2.7 on $\mathcal{T}_0$ for each $\mathbf{k}_i \in \mathbb{K}_0$ to find the eigenpairs
   $\{(\zeta_{l,0,\mathbf{k}_i}, u_{l,0,\mathbf{k}_i})\}_{\mathbf{k}_i \in \mathbb{K}_0, l \in M}$ .
3: Compute the local error indicators
   $\{\eta_{\mathcal{T}_0,i}((\zeta_{l,0,\mathbf{k}_i}, u_{l,0,\mathbf{k}_i}), T)\}_{T \in \mathcal{T}_0} := \{\eta_{\mathcal{T}_0}((\zeta_{l,0,\mathbf{k}_i}, u_{l,0,\mathbf{k}_i}), T)\}_{T \in \mathcal{T}_0}$ for all $\mathbf{k}_i \in \mathbb{K}$ using
   the expression (2.33).
```

```
4:  while $\exists\ i$ such that $\eta_{\mathcal{T}_n,i}((\zeta_{l,n,\mathbf{k}_i}, u_{l,n,\mathbf{k}_i}), \Omega) \geq \sqrt{\mathrm{TOL}}$ do
5:    for $i$ such that $\eta_{\mathcal{T}_n,i}((\zeta_{l,n,\mathbf{k}_i}, u_{l,n,\mathbf{k}_i}), \Omega) \geq \sqrt{\mathrm{TOL}}$ do
6:      Mark a set $\mathcal{M}_{n,\mathbf{k}_i} \subset \mathcal{T}_n$ according to the marking strategy described in
        Section 2.6.3 using $\theta$.
7:    end for
8:    Define $\mathcal{M} := \bigcup_{\mathbf{k}_i \in \mathbb{K}} \mathcal{M}_{n,\mathbf{k}_i}$
9:    Refine $\mathcal{T}_n$ according to the procedure described in Section 2.6.4 to get a new
      conforming triangulation $\mathcal{T}_{n+1}$.
10:   for $\mathbf{k}_i \in \mathbb{K}$ do
11:     Solve the eigenvalue problem (2.15) on $\mathcal{T}_{n+1}$ to find the eigenpair $(\zeta_{l,n+1,\mathbf{k}_i}, u_{l,n+1,\mathbf{k}_i})$.
12:     Compute the local error indicators
        $\{\eta_{\mathcal{T}_{n+1},i}((\zeta_{l,n+1,\mathbf{k}_i}, u_{l,n+1,\mathbf{k}_i}), T)\}_{T \in \mathcal{T}_{n+1}} := \{\eta_{\mathcal{T}_{n+1}}((\zeta_{l,n+1,\mathbf{k}_i}, u_{l,n+1,\mathbf{k}_i}), T)\}_{T \in \mathcal{T}_{n+1}}$
        using the expression from (2.33).
13:   end for
14:   Let $n := n + 1$.
15: end while
16: Output: Eigenpairs $\{(\zeta_{l,n,\mathbf{k}_i}, u_{l,n,\mathbf{k}_i})\}_{\mathbf{k}_i \in \mathbb{K}_0, l \in M}$, triangulation $\mathcal{T}_n$
```

Remark 4.3. Algorithm 4.2 terminates due to the convergence result in Theorem 2.20, if we start from a sufficiently fine mesh in Ω. The reason is that due to the contraction property in that theorem we can reach any prescribed tolerance for the error in energy norm of an eigenfunction in a finite number of steps. The fact that we consider several distinct eigenfunctions at the same time does not pose any problems, because we consider the union of all refinements as would occur for independent applications of Algorithm 2.19, which is convergent. Although Céa's Lemma is not valid in the case of eigenvalue problems, any additional refinement for one eingenvalue problem that results from the other problems cannot worsen the approximation significantly in the regime where we started with small initial step size in Ω. In steps 4 and 5 of Algorithm 4.2, we require that the estimator be smaller than $\sqrt{\mathrm{TOL}}$ because due to Lemma 2.13 this is a quite reasonable requirement if we want to have the error in the eigenvalues to be less than TOL.

Algorithm 4.1 terminates when the set $\{i : I_i = 0\}$ in step 7 is empty. Since the band functions are highly regular functions of the parameter $\mathbf{k}$, it is justified to assume that at some point, as the mesh size decreases in the Brillouin zone B, the linear interpolation that occurs in step 15 of the algorithm will be a good approximation of the true band function, which in turn is sufficiently well approximated in discrete points by Algorithm 4.2. Then the different I_i in step 16 will be set to one and the set the set $\{i : I_i = 0\}$ will be empty.

5. Numerical results

5.1. Results concerning the effect of quadrature. We first wish to illustrate numerically the findings at the end of Remark 3.8, that the error in the eigenvalues does not converge asymptotically as the square of the error in the eigenfunctions in the energy norm.

This can be easily seen in a one-dimensional example, that follows. In an example, where the dielectricity function ϵ has a finite number of discontinuities in Ω, we were able to estimate in equations (3.55) and (3.56), that we have convergence in the energy norm of the eigenfunctions and in the absolute values of the eigenvalues of order 1 with respect to the mesh size h.

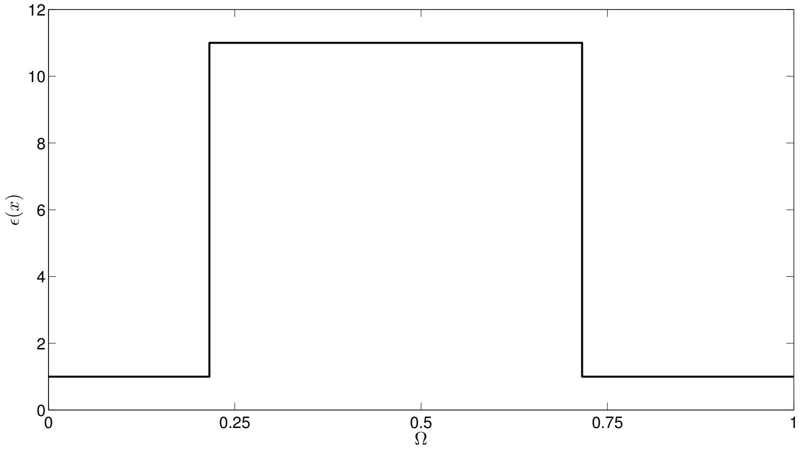


FIGURE 5.1. Dielectricity funciton ϵ that is used for the experiments.

Our examples serve the purpose to show that, in general, we cannot expect better convergence behaviour for the eigenvalues than for the eigenfunctions with certainty. Because the dielectricity function ϵ possesses some symmetry, we distorted the initial grid just a bit in order not to have cancellation of different errors. In Figures 5.2 and 5.3 we plot the different errors for uniform refinements starting from two different randomly created distorted initial grids. If we do not distort the initial grid, or if the dielectricity function ϵ is continuous on Ω, than we observe (Figure 5.4) that the rate of convergence of the eigenvalues is twice as fast as that for the eigenfunctions in energy norm.

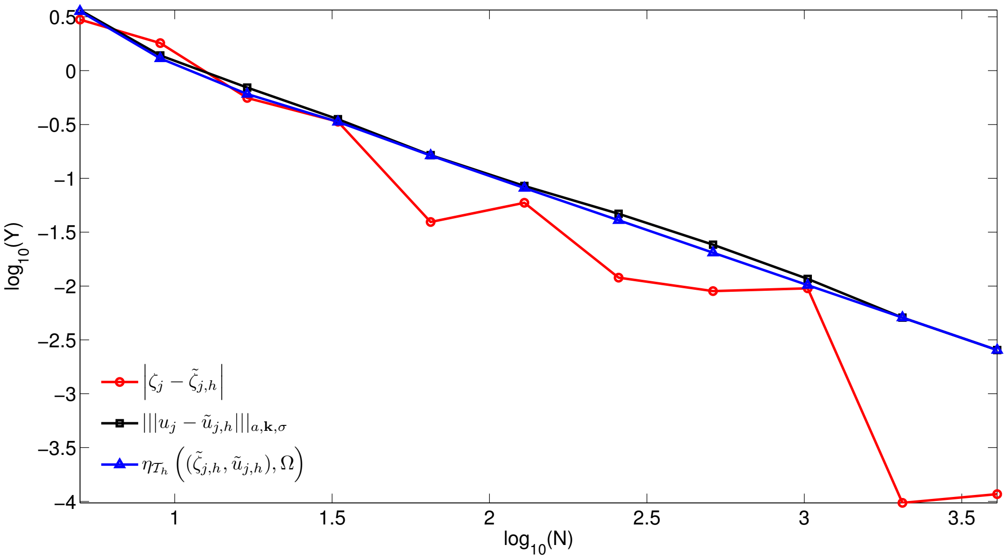


Figure 5.2. First Example: Number of unknowns vs. errors for the eigenvalue and eigenfunction (energy error and estimated error) for sequences of uniformly refined grids starting from a distorted grid. The quantities that are displayed in the legend are to be taken as Y and their respective logarithms are the y-values in the plot.

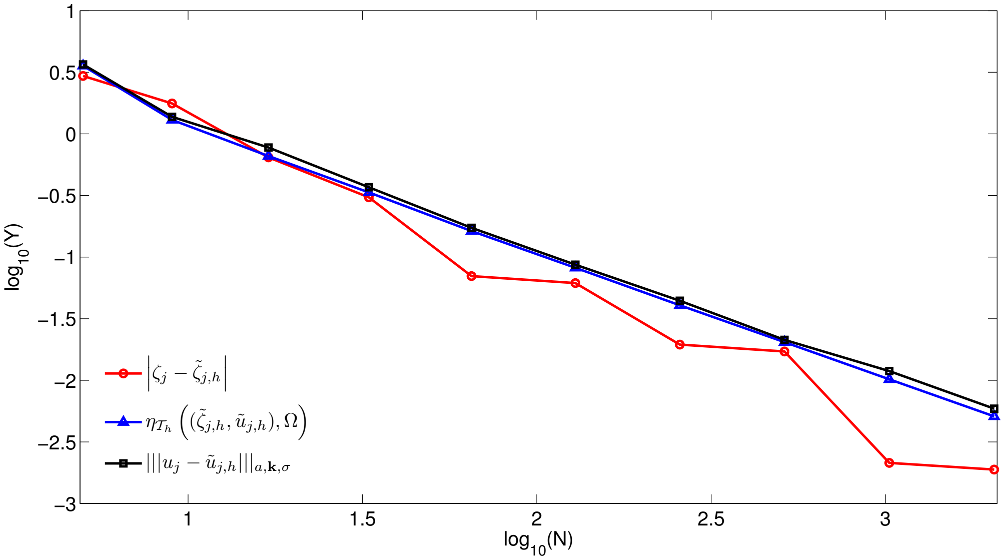


Figure 5.3. Second Example: Number of unknowns vs. errors for the eigenvalue and eigenfunction (energy error and estimated error) for sequences of uniformly refined grids starting from a distorted grid. The quantities that are displayed in the legend are to be taken as Y and their respective logarithms are the y-values in the plot.

If we apply Algorithm 2.19 from Section 2.6 to solve an eigenvalue adaptively for fixed parameter $\mathbf{k}$, where the dielectricity function ϵ is discontinuous and using

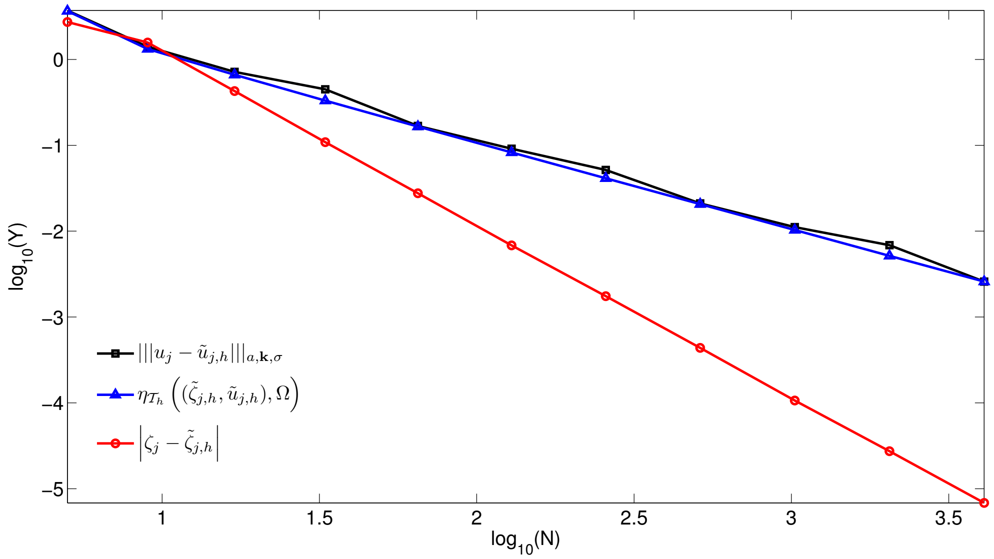


FIGURE 5.4. Third Example: Number of unknowns vs. errors for the eigenvalue and eigenfunction (energy error and estimated error) for sequences of uniformly refined grids. The initial grid is not distorted. The quantities that are displayed in the legend are to be taken as Y and their respective logarithms are the y-values in the plot.

a simple quadrature rule but without additional terms in the estimator, we get very bad results as displayed in Figure 5.5. We notice that the error in eigenvalues (the red line in Figure 5.5) behaves rather eratically. In the course of the adaptive refinements the error due to quadrature on some meshes is of the same scale as the error due to discretization while on other meshes it is much larger. We conclude that the error due to quadrature cannot be neglected in the a posteriori analysis and that terms as suggested in Remark 3.14 should be added to the a posteriori estimator. Thus, assuming exact integration when developing a theory limits the range of examples to which it can be applied.

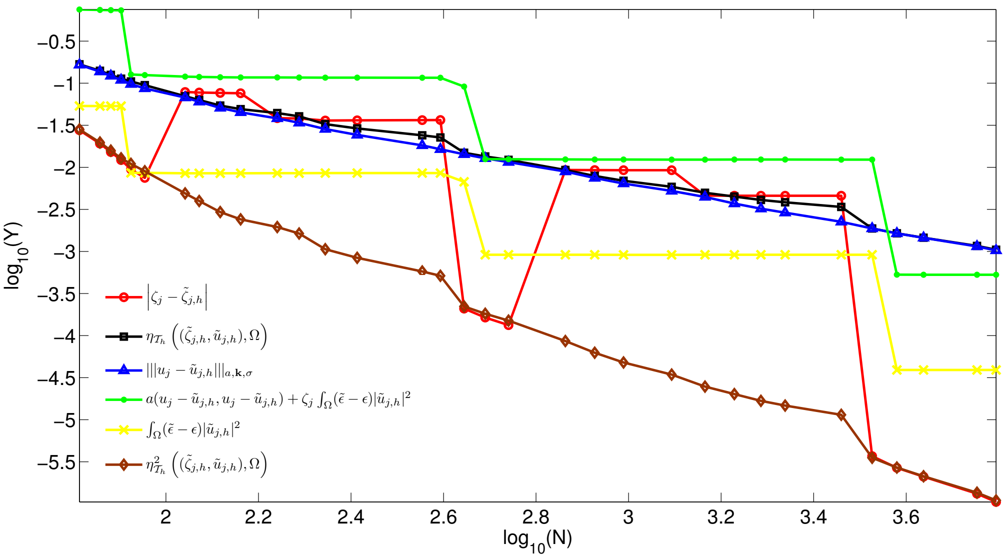


FIGURE 5.5. Adaptive refinements for discontinuous ϵ that is not resolved by the mesh: Number of unknowns vs. errors for the eigenvalue and eigenfunction (energy error and estimated error) for the solutions of Algorithm 2.19 but solving Problem 3.1 instead of Problem 2.7 in each step 8 of Algorithm 2.19. We additionally display some quantities that appear in equation (3.43). The quantities that are displayed in the legend are to be taken as Y and their respective logarithms are the y-values in the plot.

5.2. Results concerning the adaptive band structure calculations from Section 4. First we have a look at what can be achieved with uniform meshes (both in Ω and in B). We are dealing with a one-dimensional example for no other reason than that it can be implemented easily. Table 5.1 lists the *ERROR* quantity defined in (4.1) for different combinations of uniform meshes in Ω and in B. In the example that is used for the calculations, ϵ attains the values 1 and 11 on equal parts of the domain Ω and is piecewise constant with only two jumps. No jumps occur in the interior of any element.

Ω	32	64	128	256	512	1024	2048
B							
8	7.81e-01	3.88e-01	2.91e-01	3.26e-01	3.36e-01	3.39e-01	3.40e-01
16	8.25e-01	2.06e-01	1.22e-01	1.62e-01	1.72e-01	1.75e-01	1.76e-01
32	8.73e-01	2.18e-01	5.44e-02	4.91e-02	5.95e-02	6.21e-02	6.27e-02
64	8.96e-01	2.23e-01	5.57e-02	1.39e-02	1.43e-02	1.69e-02	1.76e-02
128	9.04e-01	2.25e-01	5.61e-02	1.40e-02	3.51e-03	3.73e-03	4.39e-03
256	9.06e-01	2.25e-01	5.62e-02	1.41e-02	3.51e-03	8.78e-04	9.42e-04

TABLE 5.1. Error in the eigenvalue approximation on uniform meshes.

We also display the results graphically:

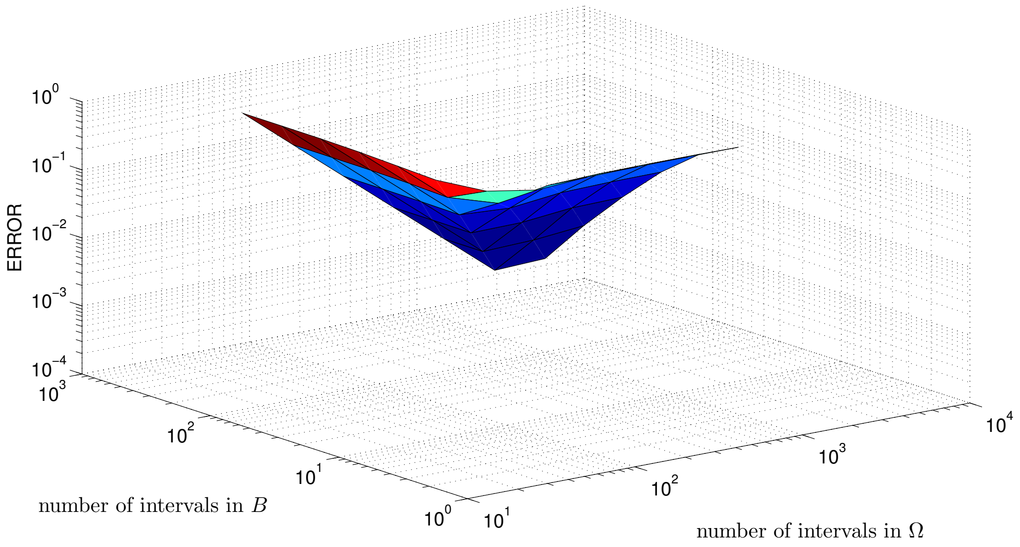


FIGURE 5.6. Error on different uniform meshes.

We realize that care should be taken when deciding how fine the meshes in Ω and in B should be. It does not help to approximate the eigenvalues very well for each parameter $\mathbf{k}$, but only use a coarse mesh in B or vice versa. We realize that a prudent choice of the fineness of the meshes leads to good results with a reasonable amount of work.

Now we turn our attention to what kinds of results we achieve implementing our algorithm. The results of the algorithm described above can be seen in the black diamonds in the graph below.

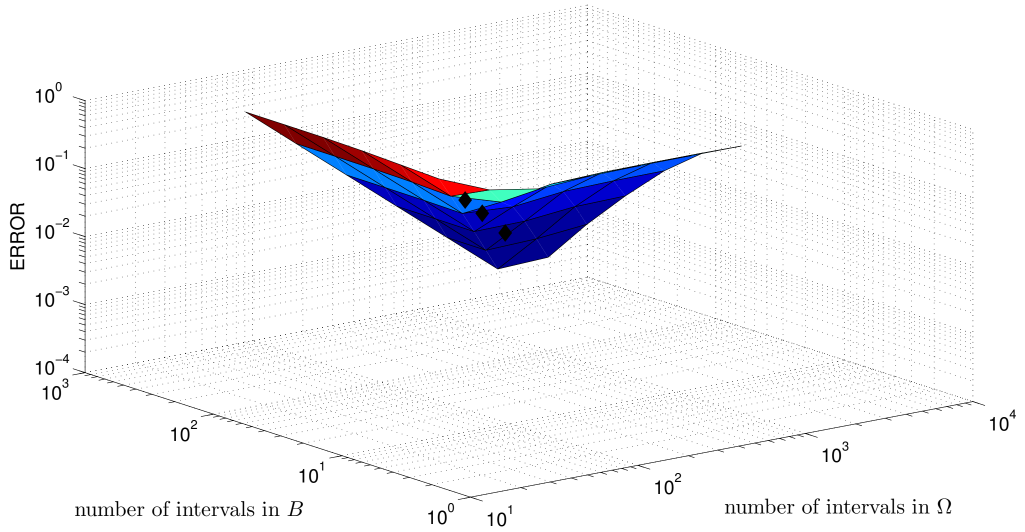


Figure 5.7. Here we compare the results of Algorithm 4.1 (black diamonds) with the previously displayed results for uniform refinements.

In order to make it easier to draw some conclusions we now plot the error against only one quantity, which we think is a good measure of the amount of work or time necessary to solve the eigenvalue problem, namely the square root of the product of the numbers of intervals in Ω and in B. The red and blue lines in the graph correspond to the numbers that are printed in the same color in Table 1. The black line corresponds to data from Table 1 between the red and blue diagonals. Once again we have plotted the results of the adaptive strategy as in the previous graphic as black diamonds.

We realize that the adaptive algorithm automatically finds good combinations of meshes in Ω and in B. Since the eigenfunctions in the example are relatively smooth functions, it only leads to results that are a bit better than those obtained for the

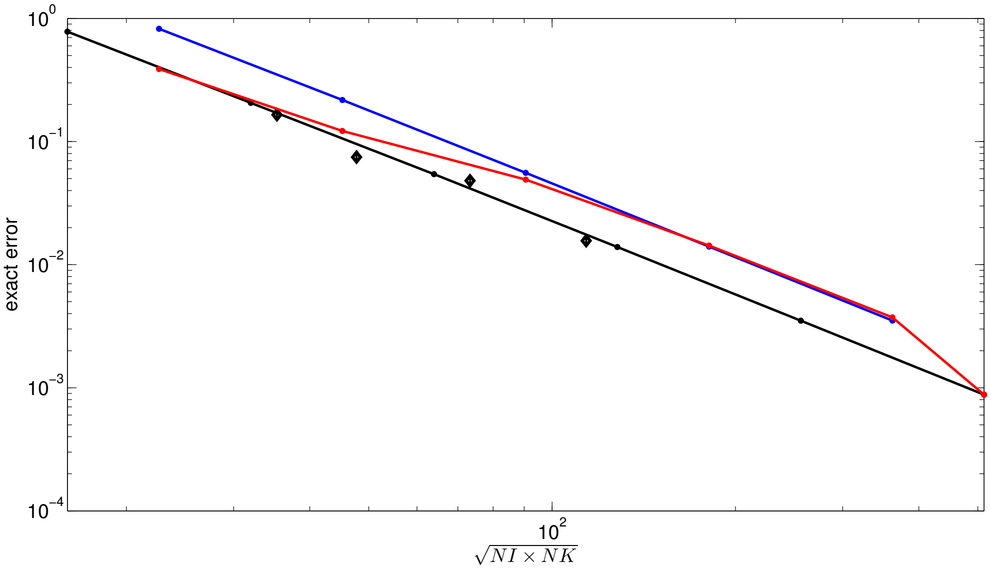


Figure 5.8. Visualization of the coloured data points of Table 5.1. The diamonds are results from adaptive calculations.

optimal combinations of uniform meshes. Nonetheless, it is more effective to use one automatic procedure than to solve the problem on many different combinations of uniform meshes and then guess which of the bandstructures is the most accurate.

CHAPTER III

Adaptive methods for the 2d curl curl problem

The most interesting case of course are 3d photonic crystals. In this case, the problem does not split any more into two disjoint elliptic problems. There are lots of difficulties involved in solving the curl curl equation adaptively. That is why we focus on this issue in this section. A first step in this direction is to study the following boundary value problem:

$$\begin{aligned} \nabla \times \nabla \times \mathbf{E} &= \mathbf{f} && \text{in } \Omega, \\ \nabla \cdot \mathbf{E} &= 0 && \text{in } \Omega, \\ \mathbf{E} \times \mathbf{n} &= \mathbf{0} && \text{on } \partial\Omega, \end{aligned} \tag{0.1}$$

where $\mathbf{f} \in \mathbf{L}^2(\Omega)$ with $\nabla \cdot \mathbf{f} \equiv 0$.

The a posteriori estimates apply to this general $3d$ case. As we have restricted the numeric experiments to the $2d$ curl curl equation, the a priori estimates that we cite for the *weighted regularization method* that we introduce in Section 2.2 are only for this case. In two space dimensions we distinguish between a scalar-valued curl that maps from $L^2(\Omega)^2$ to $L^2(\Omega)$

$$\nabla \times \mathbf{v} = \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2}, \tag{0.2}$$

and a vector-valued curl that maps from $L^2(\Omega)$ to $\left[L^2(\Omega)\right]^2$

$$\nabla \times \phi = \begin{pmatrix} \dfrac{\partial \phi}{\partial x_2} \\ -\dfrac{\partial \phi}{\partial x_1} \end{pmatrix}. \tag{0.3}$$

We assume that Ω is a polygonal domain. If the domain Ω is non-convex, which amounts to considering a domain that has finitely many corners with an interior angle $\omega_{\mathbf{a}} > \pi$, the set of all these non-convex corners $\mathbf{a} \in \partial\Omega$ will be denoted by $\mathcal{A}$.

1. Discretizing the curl curl problem — edge elements vs. nodal elements

Let us first state some spaces that are necessary for the discussion of the problem. We write vector-valued functions in bold as well as the corresponding spaces. Thus, $\mathbf{L}^2(\Omega)$ means $L^2(\Omega)^2$ or $L^2(\Omega)^3$, depending on whether we are in the two- or three-dimensional setting. However, we do not make any notational distinction between the scalar products in $L^2(\Omega)$ and $\mathbf{L}^2(\Omega)$. The following space, equipped with its natural norm, will be crucial when dealing with Maxwell's equations:

$$\mathbf{H}(\mathrm{curl},\Omega) := \left\{\mathbf{E} \in \mathbf{L}^2(\Omega) : \nabla \times \mathbf{E} \in \mathbf{L}^2(\Omega)\right\}. \tag{1.1}$$

We also use a corresponding space with zero trace of the tangential component on the boundary,

$$\mathbf{H}_0(\mathrm{curl},\Omega) := \{\mathbf{E} \in \mathbf{H}(\mathrm{curl},\Omega) : \mathbf{E} \times \mathbf{n} = \mathbf{0} \text{ on } \partial\Omega\}. \tag{1.2}$$

On $\mathbf{H}_0(\mathrm{curl},\Omega)$, the weak formulation of the problem

$$\mathbf{E} \in \mathbf{H}_0(\mathrm{curl},\Omega) \quad \forall \mathbf{v} \in \mathbf{H}_0(\mathrm{curl},\Omega) \qquad \int_\Omega \nabla \times \mathbf{E} \cdot \nabla \times \mathbf{v} = \int_\Omega \mathbf{f} \cdot \mathbf{v} \tag{1.3}$$

is not coercive since any gradient field is in the kernel of the operator associated with the bilinear form on the left hand side of the equality. Hence the Lax-Milgram lemma cannot be applied. There are several ways to circumvent this difficulty.

One way to proceed is to recall that we are only looking for divergence-free solutions. If we define

$$\mathbf{H}_0^0(\mathrm{curl},\Omega) := \{\mathbf{E} \in \mathbf{H}_0(\mathrm{curl},\Omega) : \nabla \times \mathbf{E} = 0 \text{ in } \Omega\}, \tag{1.4}$$

and

$$\mathbf{Z}_0(\Omega) := \left\{\mathbf{E} \in \mathbf{H}_0(\mathrm{curl},\Omega) : (\mathbf{E},\mathbf{z}) = 0 \ \forall \mathbf{z} \in \mathbf{H}_0^0(\mathrm{curl},\Omega)\right\}, \tag{1.5}$$

then we are in position to quote a Poincaré–Friedrichs-type inequality, Corollary 4.4 in [**24**]:

Lemma 1.1. *There is a constant $C > 0$ depending only on Ω, such that*

$$||\mathbf{E}||_\Omega \leq ||\nabla \times \mathbf{E}||_\Omega \quad \textit{for all} \ \ \mathbf{E} \in \mathbf{Z}_0(\Omega). \tag{1.6}$$

This lemma is one step in the analysis of the existence and the uniqueness of a solution to both the Maxwell source problem and the Maxwell eigenvalue problem.

A well-known strategy for finite element computations of the Maxwell equations is to use one of the two families of edge elements due to Nédélec (*cf.* [**30**] and [**31**]). These elements are conforming in $\mathbf{H}(\mathrm{curl},\Omega)$ since they are made up of piecewise polynomial functions whose tangential components are continuous across elements. The normal component is allowed to jump across elements. On a discrete level these edge elements admit a splitting into a large kernel space and a space where the Maxwell solutions are approximated. For instance, for linear elements of the first family of Nédélec edge elements the splitting takes the form (*cf.* (3.5) in [**34**])

$$\mathbf{N}_0^h \cap \mathbf{H}_0^0(\mathrm{curl},\Omega) = \nabla V_0^h, \tag{1.7}$$

where $\mathbf{N}_0^h$ denotes the linear edge elements conforming in $\mathbf{H}_0(\mathrm{curl},\Omega)$ and V_0^h denotes the piecewise linear finite element subspace of $H_0^1(\Omega)$. This splitting is a key ingredient in the stability analysis of the discrete counterpart of (1.3) and ultimately in the convergence analysis of finite element methods that employ edge elements. Unfortunately, when approximating the Maxwell eigenvalue problem by means of the edge elements, we calculate many zero eigenvalues that correspond to eigenfunctions in the kernel of the curl-operator, $\mathbf{N}_0^h \cap \mathbf{H}_0^0(\mathrm{curl},\Omega)$. For an illustration we refer to Figure 4 in [**15**], for instance.

When we use nodal elements, on the other hand, to approximate the Maxwell equations, there is no splitting similar to the one in (1.7) that we could use for the analysis. Usually some sort of regularization is performed such that the formulation includes $\nabla\cdot\mathbf{E}$ in some term. One possibility would be to introduce the space

$$\mathbf{X}_N := \mathbf{X}_N[L^2(\Omega)] := \big\{\mathbf{E} \in \mathbf{H}_0(\mathrm{curl},\Omega) : \nabla\cdot\mathbf{E} \in \mathbf{L}^2(\Omega)\big\} \tag{1.8}$$

and seek solutions

$$\begin{aligned} &\mathbf{E} \in \mathbf{X}_N \quad \forall \mathbf{v} \in \mathbf{X}_N \\ &\qquad \int_\Omega \{\nabla\times\mathbf{E}\cdot\nabla\times\mathbf{v} + s\nabla\cdot\mathbf{E}\ \nabla\cdot\mathbf{v}\} = \int_\Omega \mathbf{f}\cdot\mathbf{v}, \end{aligned} \tag{1.9}$$

where we have an additional term that integrates in $L^2(\Omega)$ the product of the divergence of E and the divergence of the test function and $s > 0$ is a penalty parameter that can be chosen. Formulation (1.9) can be discretized using nodal finite elements. The discrete space that consists of piecewise polynomials is required to be curl and div conforming. Hence each function in the finite element space is continuous across interfaces and contained in $\mathbf{H}^1(\Omega)$. If Ω is convex this choice of a discrete space is dense in $\mathbf{X}_N[L^2(\Omega)]$. But if Ω is not convex, this is no longer the case (*cf.* [**13**]).

A better solution thus seems to be to stabilize the divergence in an intermediate space Y between $L^2(\Omega)$ and $H^{-1}(\Omega)$ and pose the problem in

$$\mathbf{X}_N[Y] := \{\mathbf{E} \in \mathbf{H}_0(\mathrm{curl},\Omega) : \nabla\cdot\mathbf{E} \in Y\} \tag{1.10}$$

with this choice of Y. The problem then reads as follows:

$$\begin{aligned} &\mathbf{E} \in \mathbf{X}_N[Y] \quad \forall \mathbf{v} \in \mathbf{X}_N[Y] \\ &\qquad \int_\Omega \nabla\times\mathbf{E}\cdot\nabla\times\mathbf{v} + s\,(\nabla\cdot\mathbf{E},\nabla\cdot\mathbf{v})_Y = \int_\Omega \mathbf{f}\cdot\mathbf{v}, \end{aligned} \tag{1.11}$$

where $(\cdot,\cdot)_Y$ denotes the scalar product in Y and s is a parameter to be chosen. The approach taken by Badia and Codina in [**4**] corresponds to the choice $Y := H^{-1}(\Omega)$. Bonito and Guermond in [**7**] choose $Y := H^{-\alpha}(\Omega)$, $\alpha \in (\frac{1}{2},1)$, in the case of the eigenvalue problem. The approach by Costabel and Dauge on the other hand makes the choice $Y := L^2_{d,\gamma}(\Omega)$, a weighted Sobolev space.

2. Two different approaches using nodal elements

We have chosen to focus on solving system (0.1) using nodal elements, which is a rather new area of research. To our knowledge, there are no results concerning a posteriori error estimation for these methods available in the literature. We will focus on this issue in Sections 4 and 5. For now, we introduce the discrete spaces that are used for both methods.

As we did in Section 2.3 of Chapter II, we define a space of piecewise polynomial functions on a triangulation $\mathcal{T}_h$ (this time conforming in $H^1(\Omega)$ and not in $H^1_{\mathrm{per}}(\Omega)$),

$$Q_h := \left\{ v_h \in C^0(\Omega) \;\middle|\; v_{h|T} \in \mathcal{P}_r(T) \text{ for all } T \in \mathcal{T}_h \right\}, \tag{2.1}$$

where $\mathcal{P}_r(T)$ stands for the space of polynomials less than or equal to $r > 0$.

We consider this kind of H^1-conforming finite element space for every component of vectorial fields:

$$\mathbf{X}_h := \left(Q_h\right)^d \cap \mathbf{H}_0(\mathrm{curl}, \Omega). \tag{2.2}$$

The degree r of the polynomials at hand will be specified later. For the methods of Badia and Codina [**4**] and Bonito and Guermond [**7**], which we henceforth will refer to as *negative Sobolev penalty discretizations* we additionaly need another scalar finite element space

$$M_h := Q_h/\mathbb{R}. \tag{2.3}$$

Again, the polynomial degree for the space M_h will be specified later. We recall from Section 2.3 that for a typical element T, its diameter is denoted by h_T. We denote the maximal diameter in a triangluation by $h^{\max} := \max_{T\in\mathcal{T}} h_T$. h denotes a piecewise constant function whose restriction to an element T is h_T.

As in Section 2.3 of Chapter II we assume the existence of some standard interpolation operators (*cf.* [**12**], [**33**]).

Assumption 2.1. There exist interpolation operators $\Pi_h :\; H^1_0(\Omega) \to M_h$ and $\boldsymbol{\Pi}_h :\; H^1(\Omega) \to \mathbf{X}_h$ with the following properties:

$$\left\| v - \Pi_h v \right\|_{0,T} \le C h_T \left| v \right|_{1,\omega(T)}, \tag{2.4}$$

$$\left\| \mathbf{w} - \boldsymbol{\Pi}_h \mathbf{w} \right\|_{0,T} \le C h_T \left| \mathbf{w} \right|_{1,\omega(T)}, \tag{2.5}$$

and

$$\left\| v - \Pi_h v \right\|_{0,F} \le C h_F^{\frac{1}{2}} \left| v \right|_{1,\omega(F)}, \tag{2.6}$$

$$\left\| \mathbf{w} - \boldsymbol{\Pi}_h \mathbf{w} \right\|_{0,F} \le C h_F^{\frac{1}{2}} \left| \mathbf{w} \right|_{1,\omega(F)}. \tag{2.7}$$

Furthermore, we assume that the following stability estimates hold

$$\|v-\Pi_h v\|_{H^l(\Omega)} \le C\,|v|_{H^l(\Omega)}\ ,\ 0\le l<\frac{3}{2}, \tag{2.8}$$

$$\|\mathbf{w}-\mathbf{\Pi}_h \mathbf{w}\|_{H^l(\Omega)} \le C\,|\mathbf{w}|_{H^l(\Omega)}\ ,\ 0\le l<\frac{3}{2}. \tag{2.9}$$

2.1. Negative Sobolev space penalty discretization. Both Badia and Codina [**4**] as well as Bonito and Guermond [**7**] independently suggested to solve (0.1) by means of the following discrete scheme: Find $(\mathbf{E}_h, p_h) \in \mathbf{X}_h \times M_h$, such that

$$\begin{aligned} &\int_\Omega \left\{\nabla\times\mathbf{E}_h\cdot\nabla\times\mathbf{v}_h+\nabla p_h\cdot\mathbf{v}_h+h^2\nabla\cdot\mathbf{E}_h\nabla\cdot\mathbf{v}_h\right\}=\int_\Omega \mathbf{f}\cdot\mathbf{v}_h,\\ &\int_\Omega \{-\mathbf{E}_h\cdot\nabla q_h+\nabla p_h\cdot\nabla q_h\}=0 \end{aligned} \tag{2.10}$$

for all $(\mathbf{v}_h, q_h) \in \mathbf{X}_h \times M_h$. Bonito and Guermond [**7**] more generally consider the discretizations

$$\begin{aligned} &\int_\Omega \left\{\nabla\times\mathbf{E}_h\cdot\nabla\times\mathbf{v}_h+\nabla p_h\cdot\mathbf{v}_h+h^{2\alpha}\nabla\cdot\mathbf{E}_h\nabla\cdot\mathbf{v}_h\right\}=\int_\Omega \mathbf{f}\cdot\mathbf{v}_h,\\ &\int_\Omega \left\{-\mathbf{E}_h\cdot\nabla q_h+h^{2(1-\alpha)}\nabla p_h\cdot\nabla q_h\right\}=0 \end{aligned} \tag{2.11}$$

for all $(\mathbf{v}_h, q_h) \in \mathbf{X}_h \times M_h$, where $\alpha \in (\frac{1}{2}, 1]$. In what follows, we will call the scheme (2.11) the $H^{-\alpha}$ *penalty discretization*. The motivation why (2.10) is a good scheme to discretize (0.1) is quite different in Badia and Codina [**4**] compared with the one given in Bonito and Guermond [**7**]. We briefly summarize the viewpoints.

In [**4**] the point of view is that the Maxwell problem (0.1) first is recast as a saddle point problem with a Lagrange multiplier $p \in H_0^1(\Omega)$:

$$\begin{aligned} \nabla\times\nabla\times\mathbf{E}-\nabla p&=\mathbf{f} && \text{in } \Omega,\\ -\nabla\cdot\mathbf{E}&=0 && \text{in } \Omega,\\ \mathbf{E}\times\mathbf{n}&=0 && \text{on } \partial\Omega. \end{aligned} \tag{2.12}$$

For the weak formulation of this problem, the inf-sup condition

$$\inf_{(\mathbf{E},p)\in\mathbf{H}_0(\mathrm{curl},\Omega)\times H_0^1(\Omega)}\ \sup_{(\mathbf{v},q)\in\mathbf{H}_0(\mathrm{curl},\Omega)\times H_0^1(\Omega)} \frac{\int_\Omega\{\nabla\times\mathbf{E}\cdot\nabla\times\mathbf{v}-\nabla p\cdot\mathbf{v}+\mathbf{E}\cdot\nabla q\}}{|||\mathbf{E},p|||\ |||\mathbf{v},q|||}\ge\beta>0 \tag{2.13}$$

is satisfied, where

$$|||\mathbf{v},q|||=||\mathbf{v}||_{\mathbf{H}(\mathrm{curl},\Omega)}+||q||_{H^1(\Omega)}\,. \tag{2.14}$$

The problem thus possesses a unique solution. For numerical purposes this form is augmented. Otherwise the discrete counterpart would not satisfy an inf-sup condition. The augmented form reads as follows:

$$\begin{aligned} &\nabla\times\nabla\times\mathbf{E}-\nabla p=\mathbf{f} && \text{in } \Omega,\\ &-\nabla\cdot\mathbf{E}-\Delta p=0 && \text{in } \Omega,\\ &\mathbf{E}\times\mathbf{n}=0 && \text{on } \partial\Omega. \end{aligned}\tag{2.15}$$

Its weak formulation is:

$$\begin{aligned} \mathbf{E}\in\mathbf{H}_0(\mathrm{curl},\Omega)\quad \forall\mathbf{v}\in\mathbf{H}_0(\mathrm{curl},\Omega)\quad &\int_\Omega\{\nabla\times\mathbf{E}\cdot\nabla\times\mathbf{v}-\nabla p\cdot\mathbf{v}\}=\int_\Omega\mathbf{f}\cdot\mathbf{v},\\ \forall q\in H_0^1(\Omega)\quad &\int_\Omega\{\mathbf{E}\cdot\nabla q+\nabla p\cdot\nabla q\}=0. \end{aligned}\tag{2.16}$$

If we multiply the last line by -1 and substitute $\nabla\phi=-\nabla p$, then we have exactly (2.10) with $\nabla\phi$ instead of ∇p_h and apart from an additional stabilization term that is motivated both theoretically and numerically in [**4**].

The point of view taken in [**7**] is that we choose $Y:=H^{-\alpha}(\Omega),\ \alpha\in(\frac{1}{2},1]$, in

$$\begin{aligned} &\mathbf{E}\in\mathbf{X}_N[Y]\ \forall\mathbf{v}\in\mathbf{X}_N[Y]\\ &\qquad\int_\Omega\nabla\times\mathbf{E}\cdot\nabla\times\mathbf{v}+(\nabla\cdot\mathbf{E},\nabla\cdot\mathbf{v})_Y=\int_\Omega\mathbf{f}\cdot\mathbf{v}. \end{aligned}\tag{2.17}$$

For the choice $\alpha=1$ this results in

$$\begin{aligned} &\mathbf{E}\in\mathbf{X}_N[H^{-1}(\Omega)]\quad\forall\mathbf{v}\in\mathbf{X}_N[H^{-1}(\Omega)]\\ &\qquad\int_\Omega\nabla\times\mathbf{E}\cdot\nabla\times\mathbf{v}+(\nabla\cdot\mathbf{E},\nabla\cdot\mathbf{v})_{H^{-1}(\Omega)}=\int_\Omega\mathbf{f}\cdot\mathbf{v}. \end{aligned}\tag{2.18}$$

In what follows we use the following definition of the $H^{-1}(\Omega)$-scalar product

$$(\cdot,\cdot)_{H^{-1}(\Omega)}=\left\langle\cdot,(-\Delta)^{-1}\cdot\right\rangle,\tag{2.19}$$

where $\langle\cdot,\cdot\rangle$ denotes the $H^{-1}(\Omega)-H_0^1(\Omega)$ duality pairing. If for an arbitrary vector field $\mathbf{E}\in\mathbf{L}^2(\Omega)$ we let $p(\mathbf{E})\in H_0^1(\Omega)$ be so that

$$\Delta p(\mathbf{E})=\nabla\cdot\mathbf{E},\tag{2.20}$$

then the following identity holds:

$$(\nabla\cdot\mathbf{v},\nabla\cdot\mathbf{E})_{H^{-1}(\Omega)}=\int_\Omega\nabla p\cdot\mathbf{v}.\tag{2.21}$$

That is why we look for solutions to

$$\begin{aligned} &\mathbf{E}\in\mathbf{X}_N[H^{-1}(\Omega)]\quad\forall\mathbf{v}\in\mathbf{X}_N[H^{-1}(\Omega)]\\ &\qquad\int_\Omega\{\nabla\times\mathbf{E}\cdot\nabla\times\mathbf{v}+\nabla p\cdot\mathbf{v}\}=\int_\Omega\mathbf{f}\cdot\mathbf{v}, \end{aligned}\tag{2.22}$$

where p satisfies (2.20), or as expressed weakly

$$\int_\Omega \nabla p \cdot \nabla q = \int_\Omega \mathbf{E} \cdot \nabla q \quad \forall q \in H_0^1(\Omega). \tag{2.23}$$

We realize that we have again found (2.16) and the need for the stabilization term $h^2 \nabla\cdot\mathbf{E}_h \nabla\cdot\mathbf{v}_h$ is explained in [**7**].

2.2. Weighted regularization method. The idea of the weighted regularization method due to Costabel and Dauge [**16**] is to choose Y in (2.17) as a weighted L^2 space.

$$L^2_{d,\gamma}(\Omega) := \left\{\phi \in L^2_{\text{loc}}(\Omega) : d^\gamma \phi \in L^2(\Omega)\right\}, \tag{2.24}$$

where d is a function, that behaves locally like the distance to the nearest non-convex corner $\mathbf{a} \in \mathcal{A}$ of the domain Ω, and the choice $0 \le \gamma \le 1$ ensures that

$$L^2(\Omega) \subset L^2_{d,\gamma}(\Omega) \subset H^{-1}(\Omega). \tag{2.25}$$

The Maxwell problem is then posed over the space

$$\mathbf{X}_N[L^2_{d,\gamma}(\Omega)] = \left\{\mathbf{v} \in \mathbf{H}_0(\text{curl},\Omega) : \nabla\cdot\mathbf{v} \in L^2_{d,\gamma}(\Omega)\right\}. \tag{2.26}$$

In [**16**] it is shown that for $\gamma \in (\gamma_{\min}, 1]$ the operator associated with the weak formulation of (0.1): Find $u \in \mathbf{X}_N[L^2_{d,\gamma}(\Omega)]$, such that

$$\int_\Omega \left\{\nabla \times \mathbf{u} \cdot \nabla \times \mathbf{v} + d^{2\gamma} \nabla\cdot\mathbf{u} \nabla\cdot\mathbf{v}\right\} = \int_\Omega \mathbf{f} \cdot \mathbf{v} \quad \text{for all } \mathbf{v} \in \mathbf{X}_N[L^2_{d,\gamma}(\Omega)] \tag{2.27}$$

is elliptic, the subspace $\mathbf{H}^1(\Omega) \cap \mathbf{H}_0(\text{curl},\Omega)$ is dense in $\mathbf{X}_N[L^2_{d,\gamma}(\Omega)]$ and the solution of (2.27) coincides with the solution of (0.1). $\gamma_{\min}$ depends on the domain Ω. According to [**8**] and [**11**], we have the following equivalence of norms on the space $\mathbf{X}_N[L^2_{d,\gamma}(\Omega)]$:

$$||\mathbf{v}||_{\mathbf{X}_N[L^2_{d,\gamma}(\Omega)]} := ||\mathbf{v}|| + ||\nabla \times \mathbf{v}|| + ||d^\gamma \nabla\cdot\mathbf{v}|| \sim ||\nabla \times \mathbf{v}|| + ||d^\gamma \nabla\cdot\mathbf{v}||, \tag{2.28}$$

where two norms $||\cdot||_A$ and $||\cdot||_B$ are equivalent in the sense $||\cdot||_A \sim ||\cdot||_B$ if both $||\cdot||_A \lesssim ||\cdot||_B$ and $||\cdot||_B \lesssim ||\cdot||_A$ for all elements from the Hilbert space. The discrete version of equation (2.27) can simply be stated as: Find $\mathbf{u}_h \in \mathbf{X}_h$, such that

$$\int_\Omega \left\{\nabla \times \mathbf{u}_h \cdot \nabla \times \mathbf{v}_h + d^{2\gamma} \nabla\cdot\mathbf{u}_h \nabla\cdot\mathbf{v}_h\right\} = \int_\Omega \mathbf{f} \cdot \mathbf{v}_h \quad \text{for all } \mathbf{v}_h \in \mathbf{X}_h. \tag{2.29}$$

3. A priori results from literature

In order to prove convergence in energy norm, both Costabel and Dauge in [**16**] and Badia and Codina in [**4**] make an assumption that the finite element space $\mathbf{X}_h$ contains gradients of another finite element space, which in turn possesses good approximation properties.

Assumption 3.1. There exists a finite element space G_h defined over the mesh partition $\mathcal{T}_h$, such that $\nabla\phi_h \in \mathbf{X}_h$ for any function $\phi_h \in G_h$. Furthermore, this space satisfies the approximation property

$$\inf_{\phi_h \in G_h} ||\phi - \phi_h||_{H^s(\omega)} \lesssim (h^{\max})^{t-s} ||\phi - \phi_h||_{H^t(\omega)}, \tag{3.1}$$

in any bounded set $\omega \subset \Omega$, $\phi \in H^t(\omega)$ and $0 \le s \le t \le r+1$.

Remark 3.2. In [**4**], several examples of elements are listed for which Assumption 3.1 holds true. For instance, it is known to hold in dimension 2 for $r \ge 4$. In this case we can take G_h as the finite element space obtained for the Argyris triangle. G_h could also be constructed by using the Bogner-Fox-Schmidt triangle, which is $r \ge 2$. In order to do this, the triangulation $\mathcal{T}_h$ should admit a coarser mesh of macroelements. Under the same kind of restriction on the mesh topology, the discrete space recently introduced in [**35**], based on the Powell-Sabin interpolant, makes true Assumption 3.1 in both 2 and 3 dimensions for $r \ge 1$.

3.1. Negative Sobolev space penalty discretization.

Here we would like to summarize the pointwise convergence results from Badia and Codina [**4**] and Bonito and Guermond [**7**] for the discrete solution of (2.10) and of (2.11), respectively, to the solution of (0.1).

From Bonito and Guermond [**7**] we cite the following convergence result for the error measured in $L^2(\Omega)$ norm.

Theorem 3.3 (Convergence in $L^2(\Omega)$ norm). *Let $r \ge 1$ be the polynomial degree of the space $\mathbf{X}_h$. Then the solution $\mathbf{E}_h$ of (2.11) converges to the solution $\mathbf{E}$ of (0.1) and the following estimate is valid*

$$||\mathbf{E} - \mathbf{E}_h||_{L^2(\Omega)} \le C \, (h^{\max})^{\left(\alpha - \frac{1}{2} - \frac{\alpha}{2(r+1)}\right)^-}, \tag{3.2}$$

where the notation b^- denotes any real number strictly smaller than b.

In order to state the convergence result from Badia and Codina [**4**], we use the following regular decomposition.

Lemma 3.4 (Regular decomposition). *Let $\mathbf{E}$ be the solution of the continuous problem (0.1). Then it can be decomposed into a regular and a singular part.*

$$\mathbf{E} = \mathbf{w} + \nabla\phi, \tag{3.3}$$

where $\mathbf{w} \in H^{1+\lambda}(\Omega)^3 \cap \mathbf{H}_0(\mathrm{curl}, \Omega)$ *and* $\phi \in H^1_0(\Omega) \cap H^{1+\lambda}(\Omega)$ *for some real number* $\lambda > \min_{\mathbf{a}\in\mathcal{A}} \frac{\pi}{\omega_{\mathbf{a}}}$*. If* $\mathcal{A}$ *is empty the lemma holds with* $\lambda := 1$*.*

The lemma is stated in [**4**] and is a result of the deep analysis about the singularities of the Maxwell problem due to Costabel and Dauge [**14**].

Theorem 3.5 (Convergence in $|||\cdot|||$ norm). *Let Assumption 3.1 be satisfied. Then the solution* $(\mathbf{E}_h, p_h)$ *of (2.11) converges to the solution* $(\mathbf{E}, p)$ *of (2.12) and the following estimate holds*

$$|||\mathbf{E} - \mathbf{E}_h, p - p_h||| \leq C\,(h^{\max})^t\,||w||_{H^{1+t}(\Omega)^3} + (h^{\max})^{t-\epsilon}\,||\phi||_{H^{1+t}(\Omega)} \tag{3.4}$$

for any $\epsilon \in]0, t - \frac{1}{2}[$ *and for* $t = \min\{\lambda, r\}$*, with* λ *from Lemma 3.4.*

Remark 3.6. We stress the fact, that the formulation in both [**7**] and [**4**] is developed for uniform meshes. The parameter h in the formulation is a quantitiy that is uniform on each discretization. All arguments carried out in both papers can be generalized to non-uniform meshes and all proofs work also with a piecewise constant function h that attains the value of the diameter on each element.

3.2. Weighted regularization method. In order to state the a priori results for the weighted regularization method we need to specify which space $L^2_{d,\gamma}(\Omega)$ and $\mathbf{X}_N[L^2_{d,\gamma}(\Omega)]$ we are using. The weak formulation (2.27) depends on the choice of γ which in turn depends on the domain Ω.

For a convex domain Ω we can choose $d \equiv 1$. If the domain Ω is non-convex on the other hand, we let d denote the distance to the non-convex corners $\mathbf{a} \in \mathcal{A}$: $d(\mathbf{x}) = \mathrm{dist}(\mathbf{x}, \cup_{\mathbf{a}\in\mathcal{A}}\omega_{\mathbf{a}})$. The weight function d^γ thus behaves locally as the distance function to a non-convex corner, raised to the power $\gamma > 0$. It is a non-negative function that is bounded from above and beolow by a strictly positive constant outside a neighbourhood of $\mathcal{A}$.

From Theorem 7.4 in [**16**] we have the following result:

Theorem 3.7 (Convergence in $\mathbf{X}_N[L^2_{d,\gamma}(\Omega)]$ norm). *Let Assumption 3.1 be satisfied. If* γ *is chosen such that* $\delta^{Dir} := \min_{\mathbf{a}\in\mathcal{A}} 1 - \frac{\pi}{\omega_{\mathbf{a}}} < \gamma \leq 1$ *for all* $\mathbf{a} \in \mathcal{A}$*, then*

$$||\mathbf{E} - \mathbf{E}_h||_{\mathbf{X}_N[L^2_{d,\gamma}(\Omega)]} \leq C\,(h^{\max})^{\min\left\{r,\ \lambda^{Neu}-\epsilon,\ \gamma-\delta^{Dir}-\epsilon\right\}}\,||\mathbf{f}||_{\mathbf{L}^2(\Omega)}\,, \tag{3.5}$$

where $\lambda^{Neu} := \min_{\mathbf{a}\in\mathcal{A}} \frac{\pi}{\omega_{\mathbf{a}}}$ *is the minimum singularity exponent for the Neumann Laplace operator and* r *is the degree of the polynomials in* $\mathbf{X}_h$*. If* $\mathcal{A}$ *is empty we can make the choice* $\delta^{Dir} := 0$ *and the theorem holds with* $\lambda^{Neu} := 1$*.*

4. Estimator for the H^{-1} discretization

We develop an a posteriori error estimator for equation (2.10), that is equation (2.11) with the choice $\alpha = 1$. We then show that it is a reliable estimator. We define the element residuals

$$\begin{aligned}\left(\eta_T^{(1)}\right)^2 &:= h_T^2 \,||\mathbf{f} - \nabla\times\nabla\times\mathbf{E}_h - \nabla p_h||_T^2\,,\\ \left(\eta_T^{(2)}\right)^2 &:= h_T^2 \,||\nabla\cdot\mathbf{E}_h||_T^2\,,\\ \left(\eta_T^{(3)}\right)^2 &:= h_T^2 \,||\Delta p_h||_T^2\,,\end{aligned}\tag{4.1}$$

and jump residuals

$$\begin{aligned}\left(\eta_F^{(1)}\right)^2 &:= h_F \,||[\nabla\times\mathbf{E}_h\times\mathbf{n}]||_F^2\\ \left(\eta_F^{(2)}\right)^2 &:= h_F \,||[\partial_{\mathbf{n}} p_h]||_F^2\,.\end{aligned}\tag{4.2}$$

Furthermore we define

$$\begin{aligned}\eta(T)^2 := \;& \left(\eta_T^{(1)}\right)^2 + \left(\eta_T^{(2)}\right)^2 + \left(\eta_T^{(3)}\right)^2\\ &+\frac{1}{2}\sum_{F\subset\partial T, F\in\mathfrak{F}_h}\left(\eta_F^{(1)}\right)^2 + \frac{1}{2}\sum_{F\subset\partial T, F\in\mathfrak{F}_h}\left(\eta_F^{(2)}\right)^2,\end{aligned}\tag{4.3}$$

and

$$\eta(\Omega)^2 := \sum_{T\in\mathfrak{T}_h}\eta_T^2\tag{4.4}$$

and show reliability of the estimator in a special situation.

Definition 4.1. On $\mathbf{X}_h \times M_h$ we define the discrete norm as:

$$|||\mathbf{v}_h, p_h|||_h = ||\nabla\times\mathbf{v}_h|| + \left(\sum_{T\in\mathfrak{T}} h_T^2\,||\nabla\cdot\mathbf{v}_h||_T^2\right)^{\frac{1}{2}} + ||\nabla p_h||\,.\tag{4.5}$$

Lemma 4.2. *For the difference between the discrete solution* $(\mathbf{E}_h, p_h) \in \mathbf{X}_h \times M_h$ *in (2.10) and the solution* $(\mathbf{E}, p) \in \mathbf{H}_0(\mathrm{curl}, \Omega) \cap H_0^1(\Omega)$ *of the continuous problem* (2.12) *there holds*

$$|||\mathbf{E}-\mathbf{E}_h, p-p_h||| \le C\,|||\mathbf{E}-\mathbf{E}_h, p-p_h|||_h\tag{4.6}$$

independent of h.

Proof. The proof is similar to the proof of Lemma 3.3 in [**4**], where the result has been shown for the discrete solution alone and not for the differences $\mathbf{E} - \mathbf{E}_h$ and $p - p_h$. Since $\mathbf{X}_h \times M_h \subset \mathbf{H}_0(\mathrm{curl}, \Omega) \times H_0^1(\Omega)$, due to the inf-sup condition

(2.13) we know that

$$\sup_{(\mathbf{v},q)\in\mathbf{H}_0(\mathrm{curl},\Omega)\times H_0^1(\Omega)} \frac{\int_\Omega \{\nabla\times(\mathbf{E}-\mathbf{E}_h)\cdot\nabla\times\mathbf{v}-\nabla(p-p_h)\cdot\mathbf{v}+(\mathbf{E}-\mathbf{E}_h)\cdot\nabla q\}}{|||\mathbf{E}-\mathbf{E}_h,p-p_h|||\;|||\mathbf{v},q|||}\geq\beta>0. \tag{4.7}$$

We now do some manipulations with the numerator

$$\begin{aligned}
&\int_\Omega \{\nabla\times(\mathbf{E}-\mathbf{E}_h)\cdot\nabla\times\mathbf{v}-\nabla(p-p_h)\cdot\mathbf{v}+(\mathbf{E}-\mathbf{E}_h)\cdot\nabla q\}\\
= &\int_\Omega \{\nabla\times(\mathbf{E}-\mathbf{E}_h)\cdot\nabla\times\mathbf{v}-\nabla(p-p_h)\cdot\mathbf{v}+(\mathbf{E}-\mathbf{E}_h)\cdot\nabla(q-\mathbf{\Pi}_h q)\}\\
&+\int_\Omega(\mathbf{E}-\mathbf{E}_h)\cdot\nabla(\Pi_h q)\\
\lesssim\ & ||\nabla\times(\mathbf{E}-\mathbf{E}_h)||\,||\nabla\times\mathbf{v}||+||\nabla(p-p_h)||\,||\mathbf{v}||\\
&+\sum_{T\in\mathcal{T}_h} h_T\,||\nabla\cdot(\mathbf{E}-\mathbf{E}_h)||_T\,||\nabla q||_{\omega(T)}+||\nabla(p-p_h)||\,||\nabla(\Pi_h q)||\\
\lesssim\ & |||\mathbf{E}-\mathbf{E}_h,p-p_h|||_h\,|||\mathbf{v},q|||\,,
\end{aligned} \tag{4.8}$$

where to estimate the second integral after the equality sign, we used the fact that $\nabla\cdot\mathbf{E}=0$ as well as the fact that the discrete solution $(\mathbf{E}_h,p_h)$ solves (2.10). Π_h is the interpolation operator from Assumption 2.1 and we use its continuity in $H^1(\Omega)$ (2.8). When employing (4.8) in (4.7), cancelling $|||\mathbf{v},q|||$ and multiplying by $|||\mathbf{E}-\mathbf{E}_h,p-p_h|||$ we arrive at the result that we claimed was true. □

LEMMA 4.3 (Decomposition lemma). *The following decomposition is possible for the difference between the solution* $\mathbf{E}$ *of the continuous problem (0.1) and the solution* $\mathbf{E}_h$ *of the discrete problem (2.10)*

$$\mathbf{E}-\mathbf{E}_h=\mathbf{w}+\nabla\phi \tag{4.9}$$

with $\mathbf{w}\in\mathbf{H}^1(\Omega)$ *and* $\phi\in H_0^1(\Omega)$ *and the estimate*

$$||\mathbf{w}||_1+||\phi||_1\leq||\mathbf{E}-\mathbf{E}_h||_{\mathbf{H}(\mathrm{curl},\Omega)}\,. \tag{4.10}$$

PROOF. The result can be found in [**24**] (Lemma 2.4). □

ASSUMPTION 4.4 (Interpolation/Approximation by a C^1 regular functions). There exists an interpolation operator $\widetilde{\Pi}_h : H_0^1(\Omega)\mapsto H_0^1(\Omega)\cap C^1(\Omega)\cap V_{r+1}$, where V_{r+1} denotes a standard finite element of piecewise polynomials of degree at most $r+1$ and r is the degree of the space $\mathbf{X}_h$, such that there holds

$$\begin{aligned}
\left|\left|\phi-\widetilde{\Pi}_h\phi\right|\right|_T &\leq h_T^1\,|\phi|_{1,\omega(T)}\\
\left|\left|\phi-\widetilde{\Pi}_h\phi\right|\right|_F &\leq h_F^{\frac{1}{2}}\,|\phi|_{1,\omega(F)}\,.
\end{aligned} \tag{4.11}$$

Additionally, we assume that we have the following stability estimate

$$\left|\left|\nabla\widetilde{\Pi}_h\phi\right|\right|_\Omega \leq ||\nabla\phi||_\Omega\,. \tag{4.12}$$

Theorem 4.5 (Reliability of the estimator). *Let Assumption 4.4 be true. Then, if $h_T \leq 1$ for all $T \in \mathcal{T}_h$, the a posteriori error estimator (4.4) is reliable in the sense that*

$$|||\mathbf{E}-\mathbf{E}_h, p-p_h||| \leq C\eta(\Omega), \tag{4.13}$$

where $(\mathbf{E}_h, p_h) \in \mathbf{X}_h \times M_h$ denotes the solution of the discrete equation (2.10) and $(\mathbf{E}, p) \in \mathbf{H}_0(\mathrm{curl}, \Omega) \times H_0^1(\Omega)$ denotes the solution of the weak formulation (2.12)*.*

Proof. We start our calculation using (4.6) of Lemma 4.2.

$$\begin{aligned}
&|||\mathbf{E}-\mathbf{E}_h, p-p_h|||^2\\
&\quad\lesssim |||\mathbf{E}-\mathbf{E}_h, p-p_h|||_h^2\\
&\quad= (\mathbf{f}, \mathbf{E}-\mathbf{E}_h) - (\nabla\times\mathbf{E}_h, \nabla\times(\mathbf{E}-\mathbf{E}_h))\\
&\qquad - \sum_{T\in\mathcal{T}} h_T^2\,(\nabla\cdot\mathbf{E}_h, \nabla\cdot(\mathbf{E}-\mathbf{E}_h))_T - (\nabla p_h, \nabla(p-p_h))\\
&\quad= (\mathbf{f}, \mathbf{E}-\mathbf{E}_h) - (\nabla\times\mathbf{E}_h, \nabla\times(\mathbf{E}-\mathbf{E}_h))\\
&\qquad - \sum_{T\in\mathcal{T}} h_T^2\,(\nabla\cdot\mathbf{E}_h, \nabla\cdot(\mathbf{E}-\mathbf{E}_h))_T - (\nabla p_h, \mathbf{E}-\mathbf{E}_h)\\
&\qquad + (\mathbf{E}_h, \nabla(p-p_h)) - (\nabla p_h, \nabla(p-p_h))\\
&\quad= (\mathbf{f}, \mathbf{E}-\mathbf{E}_h-\mathbf{v}_h) - (\nabla\times\mathbf{E}_h, \nabla\times(\mathbf{E}-\mathbf{E}_h-\mathbf{v}_h))\\
&\qquad - \sum_{T\in\mathcal{T}} h_T^2\,(\nabla\cdot\mathbf{E}_h, \nabla\cdot(\mathbf{E}-\mathbf{E}_h-\mathbf{v}_h))_T - (\nabla p_h, \mathbf{E}-\mathbf{E}_h-\mathbf{v}_h)\\
&\qquad + (\mathbf{E}_h, \nabla(p-p_h-q_h)) - (\nabla p_h, \nabla(p-p_h-q_h))\,,
\end{aligned} \tag{4.14}$$

where we have used the fact that $\nabla\cdot\mathbf{f} \equiv 0$, $\nabla\cdot\mathbf{E} \equiv 0$ and $\nabla p \equiv 0$ and $\mathbf{v}_h$ and q_h are arbitrary functions from the finite element spaces. We now apply the decomposition of Lemma 4.3 and make the following choices: $\mathbf{v}_h := \mathbf{\Pi}_h\mathbf{w} + \nabla\widetilde{\Pi}_h\phi$ and $q_h := \Pi_h(p-p_h)$, where Π_h and $\mathbf{\Pi}_h$ are from Assumption 2.1 and $\widetilde{\Pi}_h$ is from Assumption

4.4. This leads to

$$\begin{aligned}
&|||\mathbf{E}-\mathbf{E}_h, p-p_h|||^2\\
&\quad\lesssim (\mathbf{f}, \mathbf{w}-\mathbf{\Pi}_h\mathbf{w}) - (\nabla\times\mathbf{E}_h, \nabla\times(\mathbf{w}-\mathbf{\Pi}_h\mathbf{w}))\\
&\qquad - \sum_{T\in\mathfrak{T}} h_T^2 \left(\nabla\cdot\mathbf{E}_h, \nabla\cdot(\mathbf{w}+\nabla\phi-\mathbf{\Pi}_h\mathbf{w}-\nabla\widetilde{\Pi}_h\phi)\right)_T\\
&\qquad - (\nabla p_h, \mathbf{w}+\nabla\phi-\mathbf{\Pi}_h\mathbf{w}-\nabla\widetilde{\Pi}_h\phi) + (\mathbf{E}_h, \nabla(p-p_h-\Pi_h(p-p_h)))\\
&\qquad - (\nabla p_h, \nabla(p-p_h-\Pi_h(p-p_h)))\\
&\quad= \sum_{T\in\mathfrak{T}}(\mathbf{f}-\nabla\times\nabla\times\mathbf{E}_h, \mathbf{w}-\mathbf{\Pi}_h\mathbf{w})_T + \sum_{F\in\mathfrak{F}}([\nabla\times\mathbf{E}_h\times\mathbf{n}], \mathbf{w}-\mathbf{\Pi}_h\mathbf{w})_F\\
&\qquad - \sum_{T\in\mathfrak{T}} h_T^2 \left(\nabla\cdot\mathbf{E}_h, \nabla\cdot(\mathbf{w}-\mathbf{\Pi}_h\mathbf{w}+\nabla\phi-\nabla\widetilde{\Pi}_h\phi)\right)_T\\
&\qquad - (\nabla p_h, \mathbf{w}-\mathbf{\Pi}_h\mathbf{w}+\nabla\phi-\nabla\widetilde{\Pi}_h\phi) + (\mathbf{E}_h, \nabla(p-p_h-\Pi_h(p-p_h)))\\
&\qquad - (\nabla p_h, \nabla(p-p_h-\Pi_h(p-p_h)))\,.
\end{aligned}\tag{4.15}$$

Using the Cauchy-Schwarz inequality we estimate

$$\begin{aligned}
&|||\mathbf{E}-\mathbf{E}_h, p-p_h|||^2\\
&\quad\lesssim \sum_{T\in\mathfrak{T}} ||f-\nabla\times\nabla\times\mathbf{E}_h-\nabla p_h||_T\,||\mathbf{w}-\mathbf{\Pi}_h\mathbf{w}||_T\\
&\qquad + \sum_{F\in\mathfrak{F}} ||[\nabla\times\mathbf{E}_h\times\mathbf{n}]||_F\,||\mathbf{w}-\mathbf{\Pi}_h\mathbf{w}||_F\\
&\qquad + \sum_{T\in\mathfrak{T}} h_T^2\,||\nabla\cdot\mathbf{E}_h||_T\,||\mathbf{w}||_{1,\omega(T)}\\
&\qquad + \sum_{T\in\mathfrak{T}} h_T\,||\nabla\cdot\mathbf{E}_h||_T \left(h_T\,||\Delta\phi||_T + h_T\left|\left|\Delta\widetilde{\Pi}_h\phi\right|\right|_T\right)\\
&\qquad + \sum_{T\in\mathfrak{T}} h_T\,||\Delta p_h||_T\,||\phi||_{1,\omega(T)}\\
&\qquad + \sum_{T\in\mathfrak{T}} h_T\,||\Delta p_h-\nabla\cdot\mathbf{E}_h||_T\,||p-p_h||_{1,\omega(T)}\\
&\qquad + \sum_{F\in\mathfrak{F}} ||[\partial_{\mathbf{n}} p_h]||_F\left|\left|\phi-\widetilde{\Pi}_h\phi\right|\right|_F\\
&\qquad + \sum_{F\in\mathfrak{F}} ||[\partial_{\mathbf{n}} p_h]||_F\,||p-p_h-\Pi_h(p-p_h)||_F\,.
\end{aligned}\tag{4.16}$$

Inserting $\nabla\phi = \mathbf{E}-\mathbf{E_h}-\mathbf{w}$ according to the decomposition Lemma 4.3, that we used, or rather inserting $\Delta\phi = -\nabla\cdot\mathbf{E_h}-\nabla\cdot\mathbf{w}$, because $\nabla\cdot\mathbf{E}\equiv 0$, and applying the inverse estimate (*cf.* [**4**], [**10**])

$$h_T\left|\left|\Delta\widetilde{\Pi}_h\phi\right|\right|_T \lesssim \left|\left|\nabla\widetilde{\Pi}_h\phi\right|\right|_T \tag{4.17}$$

followed by the stability estimate (4.12), we arrive at the following inequality

$$
\begin{aligned}
&|||\mathbf{E}-\mathbf{E}_h, p-p_h|||^2 \\
&\quad\lesssim \sum_{T\in\mathfrak{T}} ||f-\nabla\times\nabla\times\mathbf{E}_h-\nabla p_h||_T\, ||\mathbf{w}-\mathbf{\Pi}_h\mathbf{w}||_T \\
&\quad+\sum_{F\in\mathfrak{F}} ||[\nabla\times\mathbf{E}_h\times\mathbf{n}]||_F\, ||\mathbf{w}-\mathbf{\Pi}_h\mathbf{w}||_F \\
&\quad+\sum_{T\in\mathfrak{T}} h_T^2\, ||\nabla\cdot\mathbf{E}_h||_T\, ||\mathbf{w}||_{1,\omega(T)} \\
&\quad+\sum_{T\in\mathfrak{T}} h_T\, ||\nabla\cdot\mathbf{E}_h||_T \Big(h_T\, ||\nabla\cdot\mathbf{E}_h||_T + h_T\, ||\mathbf{w}||_T + ||\phi||_{1,T}\Big) \\
&\quad+\sum_{T\in\mathfrak{T}} h_T\, ||\Delta p_h||_T\, ||\phi||_{1,\omega(T)} \\
&\quad+\sum_{T\in\mathfrak{T}} h_T\, ||\Delta p_h-\nabla\cdot\mathbf{E}_h||_T\, ||p-p_h||_{1,\omega(T)} \\
&\quad+\sum_{F\in\mathfrak{F}} ||[\partial_{\mathbf{n}}p_h]||_F \left|\left|\phi-\widetilde{\Pi}_h\phi\right|\right|_F + \sum_{F\in\mathfrak{F}} ||[\partial_{\mathbf{n}}p_h]||_F\, ||p-p_h-\Pi_h(p-p_h)||_F\,.
\end{aligned}
\tag{4.18}
$$

Using the triangle inequality on the term $||\Delta p_h-\nabla\cdot\mathbf{E}_h||_T$, and then employing (4.10), Assumption 2.1 and the trace inequality we conclude that

$$
\begin{aligned}
&|||\mathbf{E}-\mathbf{E}_h, p-p_h|||^2 \\
&\quad\lesssim \left\{\left(\sum_{T\in\mathfrak{T}} h_T^2\, ||\mathbf{f}-\nabla\times\nabla\times\mathbf{E}_h-\nabla p_h||_T\right)^{\frac{1}{2}}\right. \\
&\qquad+\left(\sum_{T\in\mathfrak{T}} h_T^4\, ||\nabla\cdot\mathbf{E}_h||_T^2\right)^{\frac{1}{2}} + \left(\sum_{T\in\mathfrak{T}} h_T^2\, ||\Delta p_h||_T^2\right)^{\frac{1}{2}} \\
&\qquad+\left(\sum_{F\in\mathfrak{F}} h_F\, ||[\nabla\times\mathbf{E}_h\times\mathbf{n}]||_F^2\right)^{\frac{1}{2}} \\
&\qquad\left.+\left(\sum_{F\in\mathfrak{F}} h_F\, ||[\partial_{\mathbf{n}}p_h]||_F^2\right)^{\frac{1}{2}}\right\} \times |||\mathbf{E}-\mathbf{E}_h, p-p_h||| \\
&\qquad+\sum_{T\in\mathfrak{T}} h_T^2\, ||\nabla\cdot\mathbf{E}_h||_T^2\,.
\end{aligned}
\tag{4.19}
$$

Since we assume, that $h_T \leq 1$ for all $T \in \mathcal{T}_h$, we conclude that

$$\begin{aligned} &|||\mathbf{E}-\mathbf{E}_h, p-p_h||| \\ &\quad \lesssim \left(\sum_{T\in\mathcal{T}} h_T^2 \, ||\mathbf{f} - \nabla\times\nabla\times\mathbf{E}_h - \nabla p_h||_T\right)^{\frac{1}{2}} \\ &\quad\quad + \left(\sum_{T\in\mathcal{T}} h_T^2 \, ||\nabla\cdot\mathbf{E}_h||_T^2\right)^{\frac{1}{2}} + \left(\sum_{T\in\mathcal{T}} h_T^2 \, ||\Delta p_h||_T\right)^{\frac{1}{2}} \\ &\quad\quad + \left(\sum_{F\in\mathcal{F}} h_F \, ||[\nabla\times\mathbf{E}_h\times\mathbf{n}]||_F^2\right)^{\frac{1}{2}} \\ &\quad\quad + \left(\sum_{F\in\mathcal{F}} h_F \, ||[\partial_{\mathbf{n}} p_h]||_F^2\right)^{\frac{1}{2}} \\ &\quad \lesssim \eta(\Omega). \end{aligned} \tag{4.20}$$

□

5. Adaptivity for the weighted residual method

In this section we derive an a posteriori error estimator for the discrete scheme (2.29). We then resort to the ideas in [**9**] to show the reduction of a combined quantity of the error measured in energy norm and the estimator. This could be used to design a convergent adaptive algorithm similar to the one introduced in Algorithm 2.19 of Chapter II with the modules

$$\mathsf{SOLVE} \rightarrow \mathsf{ESTIMATE} \rightarrow \mathsf{MARK} \rightarrow \mathsf{REFINE}. \tag{5.1}$$

As in Section 2.3 of Chapter II, we let $\mathcal{T}_n,\ n = 1, 2, \ldots$, denote a family of triangular meshes on Ω. We assume that for each n, $\mathcal{T}_{n+1}$ is refinement of $\mathcal{T}_n$. We then consider finite element spaces as introduced in (2.2) and denote them by $\mathbf{X}_n$, where the index denotes which member of the family we are dealing with.

Lemma 5.1 (Orthogonality relation). *There holds:*

$$\begin{aligned} &||\nabla\times(\mathbf{E}-\mathbf{E}_n)||^2 + ||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_n)||^2 \\ &\quad = ||\nabla\times(\mathbf{E}-\mathbf{E}_{n+1})||^2 + ||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_{n+1})||^2 \\ &\quad\quad + ||\nabla\times(\mathbf{E}_{n+1}-\mathbf{E}_n)||^2 + ||d^\gamma\nabla\cdot(\mathbf{E}_{n+1}-\mathbf{E}_n)||^2 . \end{aligned} \tag{5.2}$$

Proof. Since $\mathbf{X}_n \subset \mathbf{X}_{n+1}$ we can deduce that

$$\begin{aligned} \int_\Omega \{&\nabla\times(\mathbf{E}-\mathbf{E}_{n+1}) \cdot \nabla\times(\mathbf{E}_n-\mathbf{E}_{n+1}) \\ &+ d^{2\gamma}\nabla\cdot(\mathbf{E}-\mathbf{E}_{n+1})\nabla\cdot(\mathbf{E}_n-\mathbf{E}_{n+1})\} = 0. \end{aligned} \tag{5.3}$$

by substracting equation (2.29) with the choice $\mathbf{v}_h := \mathbf{E}_n - \mathbf{E}_{n+1}$ from equation (2.27) with the choice $\mathbf{v} := \mathbf{E}_n - \mathbf{E}_{n+1}$. In other words, $\mathbf{E}_n - \mathbf{E}_{n+1}$ is orthogonal to $\mathbf{E}_n - \mathbf{E}_{n+1}$ in the scalar product

$$(\mathbf{g}, \mathbf{h}) := (\nabla \times \mathbf{g}, \nabla \times \mathbf{h}) + (d^{2\gamma} \nabla \cdot \mathbf{g}, \nabla \cdot \mathbf{h}). \tag{5.4}$$

From this we can conclude the orthogonality relation (5.2). □

We define the element residuals

$$\begin{aligned} \left(\eta_{\mathfrak{T}_n}(\mathbf{E}_n, T)_T^{(1)}\right)^2 &:= h_T^2 \,||\mathbf{f} - \nabla \times \nabla \times \mathbf{E}_n||_T^2 \\ \left(\eta_{\mathfrak{T}_n}(\mathbf{E}_n, T)_T^{(2)}\right)^2 &:= ||d^\gamma \nabla \cdot \mathbf{E}_n||_T^2 \end{aligned} \tag{5.5}$$

and jump residual

$$\left(\eta_{\mathfrak{T}_n}(\mathbf{E}_n, F)_F^{(1)}\right)^2 := h_F \,||[\nabla \times \mathbf{E}_n \times \mathbf{n}]||_F^2 \tag{5.6}$$

Furthermore we define

$$\begin{aligned} \eta_{\mathfrak{T}_n}(\mathbf{E}_n, T)^2 \; &:= \left(\eta_{\mathfrak{T}_n}(\mathbf{E}_n, T)_T^{(1)}\right)^2 + \left(\eta_{\mathfrak{T}_n}(\mathbf{E}_n, T)_T^{(2)}\right)^2 \\ &\quad + \frac{1}{2} \sum_{F \subset \partial T, F \in \mathfrak{F}_n} \left(\eta_{\mathfrak{T}_n}(\mathbf{E}_n, F)_F^{(1)}\right)^2, \end{aligned} \tag{5.7}$$

and

$$\eta_{\mathfrak{T}_n}(\mathbf{E}_n, \Omega)^2 := \sum_{T \in \mathfrak{T}_n} \eta_{\mathfrak{T}_n}(\mathbf{E}_n, T)^2. \tag{5.8}$$

In order to be able to show that the estimator is reliable we need the following decomposition as Theorem 1.2 of [**16**] (*cf.* also Theorem 2.3 in [**11**]).

Lemma 5.2 (Decomposition lemma). *Let $\mathbf{E} \in \mathbf{X}_N[L^2_{d,\gamma}(\Omega)]$ denote the solution of the weak formulation* (2.27) *and let $\mathbf{E}_n \in \mathbf{X}_n$ be the solution of the discrete equation (2.29). Then the following decomposition is possible*

$$\mathbf{E} - \mathbf{E}_h = \mathbf{w} + \nabla \phi, \tag{5.9}$$

with $\mathbf{w} \in \mathbf{H}^1(\Omega)$ and $\phi \in H^1_0(\Omega)$ and the estimate

$$||\mathbf{w}||_1 + ||\phi||_1 + ||d^\gamma \Delta \phi|| \le ||\mathbf{E} - \mathbf{E}_n||_{\mathbf{H}(\text{curl},\Omega)}. \tag{5.10}$$

Theorem 5.3 (Reliability of the estimator). *The a posteriori error estimator (5.8) is reliable in the sense that*

$$||\mathbf{E} - \mathbf{E}_n||_{\mathbf{X}_N[L^2_{d,\gamma}(\Omega)]} \le C \eta_{\mathfrak{T}_n}(\mathbf{E}_n, \Omega), \tag{5.11}$$

where $\mathbf{E}_n \in \mathbf{X}_n$ denotes the solution of the discrete equation (2.29) and $\mathbf{E} \in \mathbf{X}_N[L^2_{d,\gamma}(\Omega)]$ denotes the solution of the weak formulation (2.27).

Proof. We start our calculation using (2.28) to get

$$
\begin{aligned}
&||\mathbf{E}-\mathbf{E}_n||^2_{\mathbf{X}_N[L^2_{d,\gamma}(\Omega)]}\\
&\qquad \lesssim ||\nabla\times(\mathbf{E}-\mathbf{E}_n)||^2+||d^{\gamma}\nabla\cdot(\mathbf{E}-\mathbf{E}_n)||^2\\
&\qquad =(\mathbf{f},\mathbf{E}-\mathbf{E}_n)-(\nabla\times\mathbf{E}_h,\nabla\times(\mathbf{E}-\mathbf{E}_n))\\
&\qquad\quad -\left(d^{2\gamma}\nabla\cdot\mathbf{E}_h,\nabla\cdot(\mathbf{E}-\mathbf{E}_n)\right)\\
&\qquad =(\mathbf{f},\mathbf{E}-\mathbf{E}_n-\mathbf{v}_h)-(\nabla\times\mathbf{E}_n,\nabla\times(\mathbf{E}-\mathbf{E}_n-\mathbf{v}_h))\\
&\qquad\quad -\left(d^{2\gamma}\nabla\cdot\mathbf{E}_n,\nabla\cdot(\mathbf{E}-\mathbf{E}_n-\mathbf{v}_h)\right),
\end{aligned}
\tag{5.12}
$$

where $\mathbf{v}_h$ is an arbitrary function from the finite element spaces $\mathbf{X}_n$. We now apply the decomposition of Lemma 5.2 and decompose $\mathbf{E}-\mathbf{E}_n=\mathbf{w}+\nabla\phi$. Then we make the choice $\mathbf{v}_h:=\mathbf{\Pi}_h\mathbf{w}$. This leads to

$$
\begin{aligned}
&||\mathbf{E}-\mathbf{E}_n||^2_{\mathbf{X}_N[L^2_{d,\gamma}(\Omega)]}\\
&\qquad \lesssim (\mathbf{f},\mathbf{w}-\mathbf{\Pi}_h\mathbf{w})-(\nabla\times\mathbf{E}_n,\nabla\times(\mathbf{w}-\mathbf{\Pi}_h\mathbf{w}))\\
&\qquad\quad -\left(d^{2\gamma}\nabla\cdot\mathbf{E}_n,\nabla\cdot(\mathbf{w}+\nabla\phi-\mathbf{\Pi}_h\mathbf{w})\right)\\
&\qquad =\sum_{T\in\mathcal{T}_n}(\mathbf{f}-\nabla\times\nabla\times\mathbf{E}_n,\mathbf{w}-\mathbf{\Pi}_h\mathbf{w})_T\\
&\qquad\quad +\sum_{F\in\mathcal{F}_n}\left([\nabla\times\mathbf{E}_n\times\mathbf{n}],\mathbf{w}-\mathbf{\Pi}_h\mathbf{w}\right)\\
&\qquad\quad -\left(d^{2\gamma}\nabla\cdot\mathbf{E}_n,\nabla\cdot(\mathbf{w}+\nabla\phi-\mathbf{\Pi}_h\mathbf{w})\right).
\end{aligned}
\tag{5.13}
$$

Using the Cauchy-Schwarz inequality we estimate

$$
\begin{aligned}
&||\mathbf{E}-\mathbf{E}_n||^2_{\mathbf{X}_N[L^2_{d,\gamma}(\Omega)]}\\
&\qquad \lesssim \sum_{T\in\mathcal{T}}||\mathbf{f}-\nabla\times\nabla\times\mathbf{E}_h||_T\,||\mathbf{w}-\mathbf{\Pi}_h\mathbf{w}||_T\\
&\qquad\quad +\sum_{F\in\mathcal{F}}||[\nabla\times\mathbf{E}_h\times\mathbf{n}]||_F\,||\mathbf{w}-\mathbf{\Pi}_h\mathbf{w}||_F\\
&\qquad\quad +\left(\sum_{T\in\mathcal{T}_n}||d^{\gamma}\nabla\cdot\mathbf{E}_h||^2_T\right)^{\frac{1}{2}}||\mathbf{w}||_{1,\Omega}+\left(\sum_{T\in\mathcal{T}_n}||d^{\gamma}\nabla\cdot\mathbf{E}_h||^2_T\right)^{\frac{1}{2}}||d^{\gamma}\Delta\phi||\,.
\end{aligned}
\tag{5.14}
$$

Using the interpolation estimate and (5.10) we conclude the proof. □

Remark 5.4. Due to the equivalence of norms in (2.28), we also have reliability in another norm: The a posteriori error estimator (5.8) is reliable in the sense that

$$
||\nabla\times(\mathbf{E}-\mathbf{E}_n)||^2+||d^{\gamma}\nabla\cdot(\mathbf{E}-\mathbf{E}_n)||^2\le C_1\eta_{\mathcal{T}_n}(\mathbf{E}_n,\Omega),
\tag{5.15}
$$

where $\mathbf{E}_n\in\mathbf{X}_n$ denotes the solution of the discrete equation (2.29) and $\mathbf{E}\in\mathbf{X}_N[L^2_{d,\gamma}(\Omega)]$ denotes the solution of the weak formulation (2.27). The constant C_1 will be used in the proof of Theorem 5.6.

Now suppose that we are examining an algorithm

$$\mathsf{SOLVE} \rightarrow \mathsf{ESTIMATE} \rightarrow \mathsf{MARK} \rightarrow \mathsf{REFINE}, \tag{5.16}$$

as described in Section 2.6 of Chapter II, but modified to our situation. That is, in each step the marking is done in such a way that

$$\eta_{\mathcal{T}_n}(\mathbf{E}_n, \mathcal{M}) \geq \theta \eta_{\mathcal{T}_n}(\mathbf{E}_n, \mathcal{T}_n). \tag{5.17}$$

We will prove a contraction property for a combined quantity of the error measured in energy norm and the estimator in Theorem 5.6, which is the main ingredient for showing convergence of the adaptive algorithm in the case of the Laplace equation in [**9**].

Lemma 5.5 (Estimator reduction). *Let $\mathbf{E}_n$ and $\mathbf{E}_{n+1}$ be two consecutive solutions of the algorithm described above. Then there holds:*

$$\begin{aligned}
&\eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_{n+1}, \Omega) \\
&\qquad \leq (1+\delta)\left\{\eta^2_{\mathcal{T}_n}(\mathbf{E}_n, \Omega) - \lambda \eta^2_{\mathcal{T}_n}(\mathbf{E}_n, \mathcal{M})\right\} \\
&\qquad\quad (1+\delta^{-1})\,\Lambda\left(||\nabla \times (\mathbf{E}_{n+1} - \mathbf{E}_n)||^2 + ||d^\gamma \nabla\cdot(\mathbf{E}_{n+1} - \mathbf{E}_n)||^2\right),
\end{aligned} \tag{5.18}$$

where $\lambda := 1 - 2^{-\frac{1}{d}} > 0$, d denotes the dimension and the constant Λ only depends on the shape regularity of the mesh and the polynomial degree.

Proof. The main ideas of the proof were developed in [**9**]. First, for an arbitrary $T \in \mathcal{T}_{n+1}$ we use the definitions (5.5), (5.6) and (5.7) to calculate

$$\begin{aligned}
\eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_{n+1}, T)^2 &= \left(\eta_{\mathcal{T}_n}(\mathbf{E}_{n+1}, T)^{(1)}_T\right)^2 + \left(\eta_{\mathcal{T}_n}(\mathbf{E}_{n+1}, T)^{(2)}_T\right)^2 \\
&\quad + \frac{1}{2} \sum_{F \subset \partial T, F \in \mathcal{F}_n} \left(\eta_{\mathcal{T}_n}(\mathbf{E}_{n+1}, F)^{(1)}_F\right)^2 \\
&= h_T^2\,||\mathbf{f} - \nabla \times \nabla \times \mathbf{E}_{n+1}||^2_T + ||d^\gamma \nabla\cdot \mathbf{E}_{n+1}||^2_T \\
&\quad + \frac{1}{2} \sum_{F \subset \partial T, F \in \mathcal{F}_n} h_F\,||[\nabla \times \mathbf{E}_{n+1} \times \mathbf{n}]||^2_F \\
&\leq h_T^2\,||\mathbf{f} - \nabla \times \nabla \times \mathbf{E}_n||^2_T \\
&\quad + h_T^2\,||\nabla \times \nabla \times (\mathbf{E}_n - \mathbf{E}_{n+1})||^2_T \\
&\quad + ||d^\gamma \nabla\cdot \mathbf{E}_n||^2_T + ||d^\gamma \nabla\cdot(\mathbf{E}_n - \mathbf{E}_{n+1})||^2_T \\
&\quad + \frac{1}{2} \sum_{F \subset \partial T, F \in \mathcal{F}_n} h_F\,||[\nabla \times \mathbf{E}_n \times \mathbf{n}]||^2_F \\
&\quad + \frac{1}{2} \sum_{F \subset \partial T, F \in \mathcal{F}_n} h_F\,||[\nabla \times (\mathbf{E}_n - \mathbf{E}_{n+1}) \times \mathbf{n}]||^2_F\,,
\end{aligned} \tag{5.19}$$

where we have used the triangle inequality. Applying inverse estimates to the terms involving the differences $\mathbf{E}_n - \mathbf{E}_{n+1}$ and again using the definitions (5.5), (5.6) and (5.7) we conclude that

$$\begin{aligned} & \eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_{n+1}, T)^2 \\ \leq\ & \eta_{\mathcal{T}_{n+1}}(\mathbf{E}_n, T)^2 \\ & +\Lambda \left(||\nabla \times (\mathbf{E}_{n+1} - \mathbf{E}_n)||^2_{\omega(T)} + ||d^\gamma \nabla \cdot (\mathbf{E}_{n+1} - \mathbf{E}_n)||^2_{\omega(T)} \right) \end{aligned} \tag{5.20}$$

Using Young's inequality with parameter δ, summing over all elements $T \in \mathcal{T}_{n+1}$ and using the finite overlap property of the patches $\omega(T)$ we arrive at the following estimate

$$\begin{aligned} & \eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_{n+1}, \mathcal{T}_{n+1}) \\ & \quad \leq (1+\delta)\, \eta_{\mathcal{T}_{n+1}}(\mathbf{E}_n, T)^2 \\ & \qquad (1+\delta^{-1})\, \Lambda \left(||\nabla \times (\mathbf{E}_{n+1} - \mathbf{E}_n)||^2 + ||d^\gamma \nabla \cdot (\mathbf{E}_{n+1} - \mathbf{E}_n)||^2 \right), \end{aligned} \tag{5.21}$$

with a new constant Λ that only depends on the shape regularity of the mesh and the polynomial degree. For a marked element $T \in \mathcal{M} \subset \mathcal{T}_n$, we set

$$\mathcal{T}_{n+1,T} := \{ T' \in \mathcal{T}_{n+1} \ \mid \ T' \subset T \}.$$

Since $\mathbf{E}_n \in \mathbf{X}_n$, we see that

$$\left(\eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_n, F)^{(1)}_F \right)^2 := 0 \tag{5.22}$$

on sides F in the interior of T. We then obtain

$$\sum_{T' \in \mathcal{T}_{n+1,T}} \eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_n, T') \leq 2^{-\frac{1}{d}} \eta^2_{\mathcal{T}_n}(\mathbf{E}_n, T), \tag{5.23}$$

because refinement by bisection implies $h_{T'} = |T'|^{-\frac{1}{d}} \leq \left(2^{-1}|T|\right)^{-\frac{1}{d}} \leq 2^{-\frac{1}{d}} h_T$ for all $T' \in \mathcal{T}_{n+1,T}$. For an element $T \in \mathcal{T}_n \setminus \mathcal{M}$, on the other hand, Remark 2.1 in [**9**] yields $\eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_n, T) \leq \eta^2_{\mathcal{T}_n}(\mathbf{E}_n, T)$. Hence, summing over all $T \in \mathcal{T}_{n+1}$ we arrive at

$$\begin{aligned} & \eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_n, \Omega) \\ \leq\ & \eta^2_{\mathcal{T}_n}(\mathbf{E}_n, \Omega \setminus \mathcal{M}) - 2^{-\frac{1}{d}} \eta^2_{\mathcal{T}_n}(\mathbf{E}_n, \mathcal{M}) \\ =\ & \eta^2_{\mathcal{T}_n}(\mathbf{E}_n, \Omega) - \lambda \eta^2_{\mathcal{T}_n}(\mathbf{E}_n, \mathcal{M}). \end{aligned} \tag{5.24}$$

From this together with (5.21), the assertion finally follows. □

Theorem 5.6 (Contraction property). *Let* $\mathbf{E}_n$ *and* $\mathbf{E}_{n+1}$ *be two consecutive solutions of the algorithm described above. Then there exists* $\zeta > 0$ *and* $0 < \alpha < 1$, *depending only on the shape regularity of the meshes and the parameter* θ *from the* *marking strategy, such that for any two consecutive iterates n and $n+1$, we have*

$$\begin{aligned} &||\nabla\times(\mathbf{E}-\mathbf{E}_{n+1})||^2+||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_{n+1})||^2+\zeta\eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_{n+1},\Omega)\\ \leq\ &\alpha^2\left(||\nabla\times(\mathbf{E}-\mathbf{E}_n)||^2+||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_n)||^2+\zeta\eta^2_{\mathcal{T}_n}(\mathbf{E}_n,\Omega)\right).\end{aligned}\tag{5.25}$$

Proof. We start by employing Lemma 5.1 to get

$$\begin{aligned} &||\nabla\times(\mathbf{E}-\mathbf{E}_{n+1})||^2+||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_{n+1})||^2+\zeta\eta^2_{\mathcal{T}_{k+1}}(\mathbf{E}_{n+1},\Omega)\\ =&||\nabla\times(\mathbf{E}-\mathbf{E}_n)||^2+||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_n)||^2\\ &-||\nabla\times(\mathbf{E}_{n+1}-\mathbf{E}_n)||^2-||d^\gamma\nabla\cdot(\mathbf{E}_{n+1}-\mathbf{E}_n)||^2+\zeta\eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_{n+1},\Omega)\end{aligned}\tag{5.26}$$

Now we employ Lemma 5.5 to find:

$$\begin{aligned} &||\nabla\times(\mathbf{E}-\mathbf{E}_{n+1})||^2+||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_{n+1})||^2+\zeta\eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_{n+1},\Omega)\\ \leq&||\nabla\times(\mathbf{E}-\mathbf{E}_n)||^2+||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_n)||^2\\ &-||\nabla\times(\mathbf{E}_{n+1}-\mathbf{E}_n)||^2-||d^\gamma\nabla\cdot(\mathbf{E}_{n+1}-\mathbf{E}_n)||^2\\ &+\zeta\,(1+\delta)\left\{\eta^2_{\mathcal{T}_n}(\mathbf{E}_k,\Omega)-\lambda\eta^2_{\mathcal{T}_k}(\mathbf{E}_k,\mathcal{M})\right\}\\ &+\zeta\left(1+\delta^{-1}\right)\Lambda\left(||\nabla\times(\mathbf{E}_{n+1}-\mathbf{E}_n)||^2+||d^\gamma\nabla\cdot(\mathbf{E}_{n+1}-\mathbf{E}_n)||^2\right).\end{aligned}\tag{5.27}$$

Choosing ζ dependent on δ to be

$$\zeta:=\frac{1}{\left(1+\delta^{-1}\right)\Lambda}\quad\Longleftrightarrow\quad\zeta(1+\delta)=\frac{\delta}{\Lambda},\tag{5.28}$$

we achieve that all the terms involving the difference $\mathbf{E}_{n+1}-\mathbf{E}_n$ cancel each other out:

$$\begin{aligned} &||\nabla\times(\mathbf{E}-\mathbf{E}_{n+1})||^2+||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_{n+1})||^2+\zeta\eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_{n+1},\Omega)\\ \leq&||\nabla\times(\mathbf{E}-\mathbf{E}_n)||^2+||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_n)||^2\\ &+\zeta\,(1+\delta)\,\eta^2_{\mathcal{T}_n}(\mathbf{E}_n,\Omega)-\zeta\,(1+\delta)\,\lambda\eta^2_{\mathcal{T}_n}(\mathbf{E}_n,\mathcal{M})\end{aligned}\tag{5.29}$$

Invoking the marking strategy (5.17) we deduce

$$\begin{aligned} &||\nabla\times(\mathbf{E}-\mathbf{E}_{n+1})||^2+||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_{n+1})||^2+\zeta\eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_{n+1},\Omega)\\ \leq&||\nabla\times(\mathbf{E}-\mathbf{E}_n)||^2+||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_n)||^2\\ &+\zeta\,(1+\delta)\,\eta^2_{\mathcal{T}_n}(\mathbf{E}_n,\Omega)-\zeta\,(1+\delta)\,\lambda\theta^2\eta^2_{\mathcal{T}_n}(\mathbf{E}_n,\Omega)\end{aligned}\tag{5.30}$$

We rewrite this equality as follows with any $\kappa\in(0,1)$

$$\begin{aligned} &||\nabla\times(\mathbf{E}-\mathbf{E}_{n+1})||^2+||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_{n+1})||^2+\zeta\eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_{n+1},\Omega)\\ \leq&||\nabla\times(\mathbf{E}-\mathbf{E}_n)||^2+||d^\gamma\nabla\cdot(\mathbf{E}-\mathbf{E}_n)||^2\\ &+\zeta\,(1+\delta)\,\eta^2_{\mathcal{T}_n}(\mathbf{E}_n,\Omega)-\kappa\,\zeta\,(1+\delta)\,\lambda\theta^2\eta^2_{\mathcal{T}_n}(\mathbf{E}_n,\Omega)\\ &-(1-\kappa)\zeta\,(1+\delta)\,\lambda\theta^2\eta^2_{\mathcal{T}_n}(\mathbf{E}_n,\Omega).\end{aligned}\tag{5.31}$$

We apply the reliability of the estimator from Theorem 5.3 and replace ζ by the choice made in (5.28) earlier to get

$$\begin{aligned}
&||\nabla \times (\mathbf{E} - \mathbf{E}_{n+1})||^2 + ||d^\gamma \nabla \cdot (\mathbf{E} - \mathbf{E}_{n+1})||^2 + \zeta \eta^2_{\mathcal{T}_{n+1}}(\mathbf{E}_{n+1}, \Omega) \\
\le\ & \alpha_1^2 \left(||\nabla \times (\mathbf{E} - \mathbf{E}_n)||^2 + ||d^\gamma \nabla \cdot (\mathbf{E} - \mathbf{E}_n)||^2 \right) \\
&+ \alpha_2^2 \zeta\, \eta^2_{\mathcal{T}_n}(\mathbf{E}_n, \Omega),
\end{aligned} \tag{5.32}$$

with

$$\alpha_1^2 = 1 - \kappa \frac{\lambda \theta^2}{C_1 \Lambda} \delta, \qquad \alpha_2^2 = (1 + \delta) \left(1 - (1 - \kappa) \lambda \theta^2 \right). \tag{5.33}$$

Now choosing δ small enough yields

$$\alpha^2 := \max \left\{ \alpha_1^2, \alpha_2^2 \right\} < 1, \tag{5.34}$$

which is the desired result. □

6. Numerical results

In this section we test the methods described in this chapter to see how they work in practice. We report convergence rates for uniform refinements both for the weighted regularization method and when stabilizing the divergence term in a negative Sobolev space. We additionally display the results of the estimator for uniform refinements. Afterwards we adaptively refine according to the estimators to show that we can improve the effectiveness of the calculations.

Examples:
1 Torus $[0,1]^2$-periodic, smooth solution.
2 Square $[0,1]^2$, $\mathbf{n} \times \mathbf{E} = 0$, smooth solution.
3 L–shaped domain, $\mathbf{n} \times \mathbf{E} = 0$, singular solution.
This is an interesting example. The L-shaped domain is given as $\Omega = [0,1]^2 \setminus ([0,+1] \times [-1,0])$. The example has already been studied in both [**16**], [**7**] and [**4**] for uniform refinements. We reproduce these results and then apply our adaptive approach.

We consider the following boundary value problem:

$$\begin{aligned} \nabla \times \nabla \times \mathbf{E} &= \mathbf{0} \qquad \text{in } \Omega, \\ \nabla \cdot \mathbf{E} &= 0 \qquad \text{in } \Omega, \\ \mathbf{E} \times \mathbf{n} &= \mathbf{G} \times \mathbf{n} \qquad \text{on } \partial\Omega, \end{aligned} \tag{6.1}$$

where

$$\mathbf{G} = \frac{2}{3} r^{-\frac{1}{3}} \begin{pmatrix} -\sin(\frac{\theta}{3}) \\ \cos(\frac{\theta}{3}) \end{pmatrix}, \tag{6.2}$$

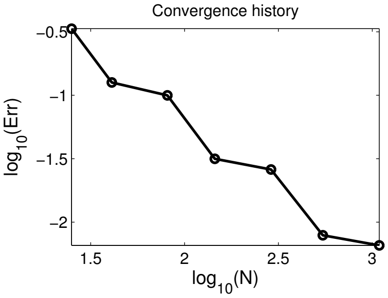
where (r,θ) are the polar coordinates centered at the re-entrant corner of the domain. The solution to the above problem is $\mathbf{E} = \nabla\phi$, where $\phi(r,\theta) = r^{\frac{2}{3}} \sin(\frac{2}{3}\theta)$, and $\mathbf{E} \in \mathbf{H}^{\frac{2}{3}}(\Omega)$.

6.1. Experimental order of convergence (eoc) for the weighted regularisation method.

$$\int_\Omega \{\nabla \times \mathbf{E}_h \cdot \nabla \times \mathbf{v}_h + d^{2\gamma} \nabla \cdot \mathbf{E}_h \nabla \cdot \mathbf{v}_h - \sigma_1 \nabla p_h \cdot \mathbf{v}_h\} = \int_\Omega \mathbf{f} \cdot \mathbf{v}_h$$

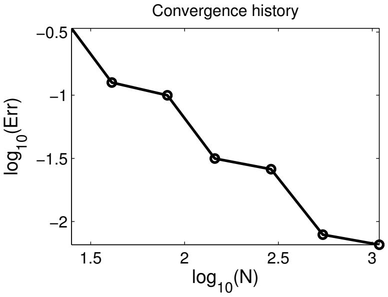
$$\int_\Omega \{-\sigma_1 \mathbf{E}_h \cdot \nabla q_h + \sigma_1 \sigma_2 \nabla q_h \cdot \nabla q_h\} = 0$$

6.1.1. *Example 2.* $\sigma_1, \sigma_2 \in \{0, 1\}$, $\gamma = 0.9$.



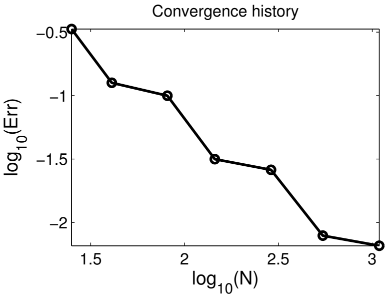
N	L^2-err	eoc
25	3.35_{-1}	
41	1.26_{-1}	
81	9.98_{-2}	1.98_0
145	3.15_{-2}	2.15_0
289	2.60_{-2}	2.07_0
545	7.87_{-3}	2.07_0
1089	6.56_{-3}	2.05_0

Ex=2, pd=1, sig1=0, sig2=0, $\gamma = 0.9$, imax=7.



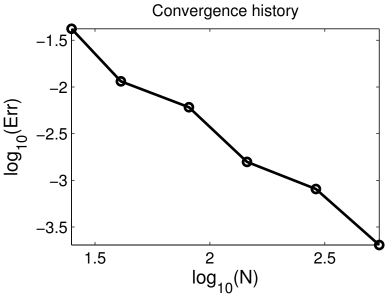
N	L^2-err	eoc
25	3.38_{-1}	
41	1.26_{-1}	
81	9.98_{-2}	1.99_0
145	3.15_{-2}	2.15_0
289	2.60_{-2}	2.07_0
545	7.87_{-3}	2.07_0
1089	6.56_{-3}	2.05_0

Ex=2, pd=1, sig1=1, sig2=0, $\gamma = 0.9$, imax=7.



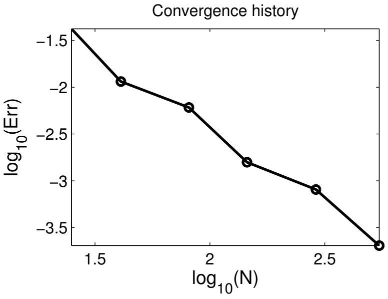
N	L^2-err	eoc
25	3.35_{-1}	
41	1.26_{-1}	
81	9.98_{-2}	1.98_0
145	3.15_{-2}	2.15_0
289	2.60_{-2}	2.07_0
545	7.87_{-3}	2.07_0
1089	6.56_{-3}	2.05_0

Ex=2, pd=1, sig1=1, sig2=1, $\gamma = 0.9$, imax=7.



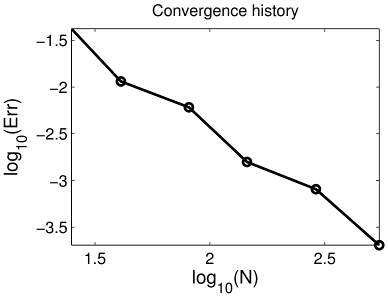
N	L^2-err	eoc
25	4.20_{-2}	
41	1.15_{-2}	
81	6.07_{-3}	3.21_0
145	1.58_{-3}	3.11_0
289	8.07_{-4}	3.14_0
545	2.03_{-4}	3.08_0

Ex=2, pd=2, sig1=0, sig2=0, $\gamma = 0.9$, imax=7.



N	L^2-err	eoc
25	4.20_{-2}	
41	1.15_{-2}	
81	6.07_{-3}	3.21_0
145	1.58_{-3}	3.11_0
289	8.07_{-4}	3.14_0
545	2.03_{-4}	3.08_0

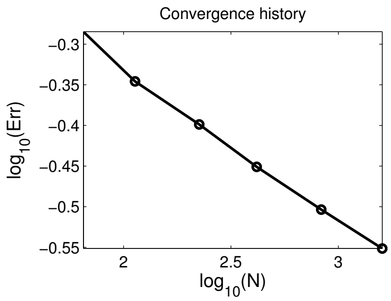
Ex=2, pd=2, sig1=1, sig2=0, $\gamma = 0.9$, imax=7.



N	L^2-err	eoc
25	4.20_{-2}	
41	1.15_{-2}	
81	6.07_{-3}	3.21_0
145	1.58_{-3}	3.11_0
289	8.07_{-4}	3.14_0
545	2.03_{-4}	3.08_0

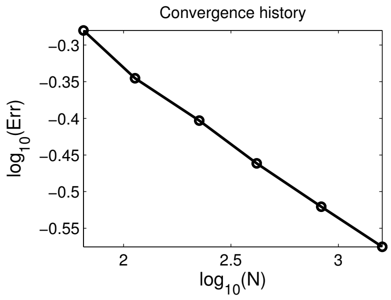
Ex=2, pd=2, sig1=1, sig2=1, $\gamma = 0.9$, imax=7.

6.1.2. *Example 3.* $\sigma_1, \sigma_2 \in \{0, 1\}$, $\gamma = 0.9$.



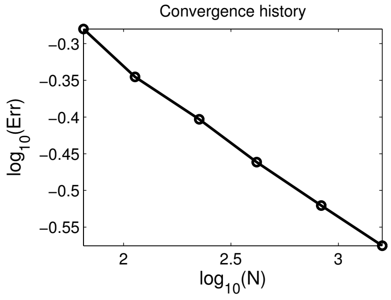
N	L^2-err	eoc
65	5.19_{-1}	
113	4.51_{-1}	
225	3.99_{-1}	0.42_{0}
417	3.54_{-1}	0.37_{0}
833	3.14_{-1}	0.37_{0}
1601	2.81_{-1}	0.34_{0}

Ex=3, pd=1, sig1=0, sig2=0, $\gamma = 0.9$, imax=6.



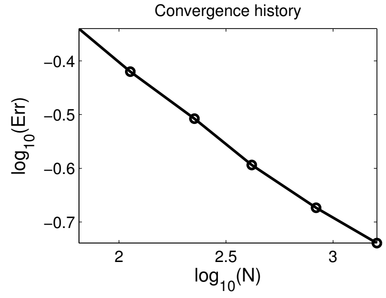
N	L^2-err	eoc
65	5.25_{-1}	
113	4.52_{-1}	
225	3.95_{-1}	0.45_{0}
417	3.46_{-1}	0.41_{0}
833	3.02_{-1}	0.41_{0}
1601	2.66_{-1}	0.39_{0}

Ex=3, pd=1, sig1=1, sig2=0, $\gamma = 0.9$, imax=6.



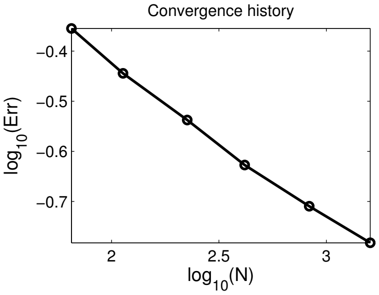
N	L^2-err	eoc
65	5.25_{-1}	
113	4.52_{-1}	
225	3.95_{-1}	0.45_{0}
417	3.46_{-1}	0.41_{0}
833	3.02_{-1}	0.41_{0}
1601	2.66_{-1}	0.39_{0}

Ex=3, pd=1, sig1=1, sig2=1, $\gamma = 0.9$, imax=6.



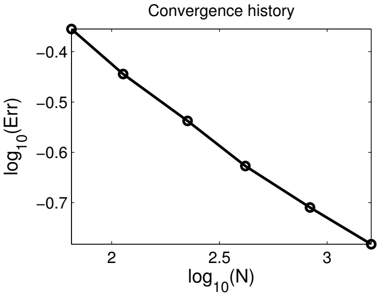
N	L^2-err	eoc
65	4.57_{-1}	
113	3.80_{-1}	
225	3.11_{-1}	0.62_{0}
417	2.55_{-1}	0.61_{0}
833	2.12_{-1}	0.58_{0}
1601	1.82_{-1}	0.50_{0}

Ex=3, pd=2, sig1=0, sig2=0, $\gamma = 0.9$, imax=6.



N	L^2-err	eoc
65	4.42_{-1}	
113	3.59_{-1}	
225	2.90_{-1}	0.67_0
417	2.36_{-1}	0.64_0
833	1.95_{-1}	0.60_0
1601	1.65_{-1}	0.53_0

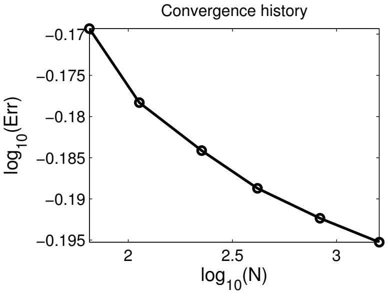
Ex=3, pd=2, sig1=1, sig2=0, $\gamma = 0.9$, imax=6.



N	L^2-err	eoc
65	4.42_{-1}	
113	3.59_{-1}	
225	2.90_{-1}	0.67_0
417	2.36_{-1}	0.64_0
833	1.95_{-1}	0.60_0
1601	1.65_{-1}	0.53_0

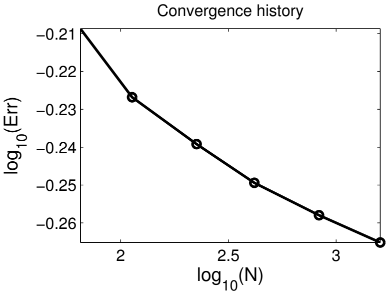
Ex=3, pd=2, sig1=1, sig2=1, $\gamma = 0.9$, imax=6.

6.1.3. *Example 3 — Dependence on γ.* $\sigma_1 = 0, \sigma_2 = 0$, $\gamma \in [0, 1]$.



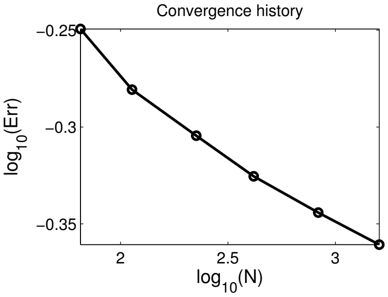
N	L^2-err	eoc
65	6.77_{-1}	
113	6.63_{-1}	
225	6.54_{-1}	0.05_0
417	6.48_{-1}	0.04_0
833	6.42_{-1}	0.03_0
1601	6.38_{-1}	0.02_0

Ex=3, pd=2, sig1=0, sig2=0, $\gamma = 0$, imax=6.



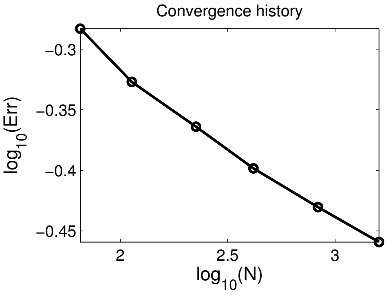
N	L^2-err	eoc
65	6.18_{-1}	
113	5.93_{-1}	
225	5.77_{-1}	0.11_0
417	5.63_{-1}	0.08_0
833	5.52_{-1}	0.07_0
1601	5.43_{-1}	0.05_0

Ex=3, pd=2, sig1=0, sig2=0, $\gamma = 0.1$, imax=6.



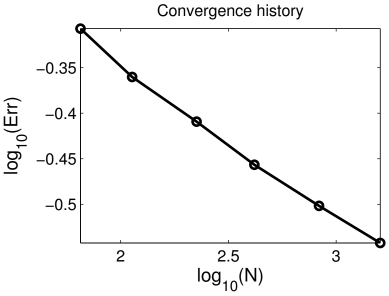
N	L^2-err	eoc
65	5.63_{-1}	
113	5.24_{-1}	
225	4.96_{-1}	0.20_{0}
417	4.73_{-1}	0.16_{0}
833	4.53_{-1}	0.14_{0}
1601	4.36_{-1}	0.12_{0}

Ex=3, pd=2, sig1=0, sig2=0, $\gamma = 0.2$, imax=6.



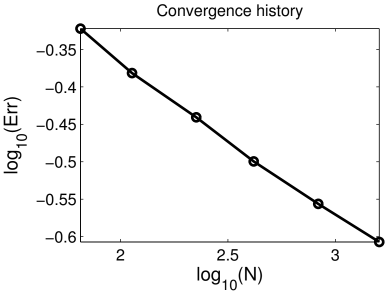
N	L^2-err	eoc
65	5.21_{-1}	
113	4.71_{-1}	
225	4.33_{-1}	0.30_{0}
417	4.00_{-1}	0.25_{0}
833	3.71_{-1}	0.23_{0}
1601	3.47_{-1}	0.21_{0}

Ex=3, pd=2, sig1=0, sig2=0, $\gamma = 0.3$, imax=6.



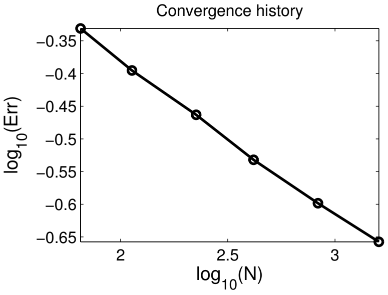
N	L^2-err	eoc
65	4.93_{-1}	
113	4.36_{-1}	
225	3.90_{-1}	0.38_{0}
417	3.50_{-1}	0.34_{0}
833	3.15_{-1}	0.32_{0}
1601	2.87_{-1}	0.29_{0}

Ex=3, pd=2, sig1=0, sig2=0, $\gamma = 0.4$, imax=6.



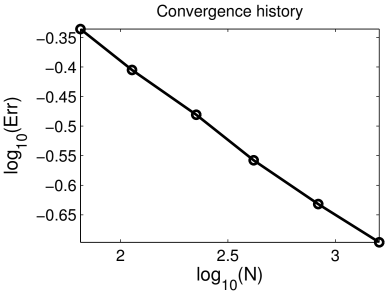
N	L^2-err	eoc
65	4.76_{-1}	
113	4.15_{-1}	
225	3.63_{-1}	0.44_{0}
417	3.17_{-1}	0.42_{0}
833	2.78_{-1}	0.41_{0}
1601	2.47_{-1}	0.37_{0}

Ex=3, pd=2, sig1=0, sig2=0, $\gamma = 0.5$, imax=6.



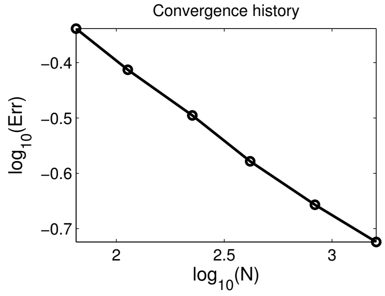
N	L^2-err	eoc
65	4.67_{-1}	
113	4.02_{-1}	
225	3.44_{-1}	0.49_{0}
417	2.94_{-1}	0.48_{0}
833	2.52_{-1}	0.48_{0}
1601	2.20_{-1}	0.43_{0}

Ex=3, pd=2, sig1=0, sig2=0, $\gamma = 0.6$, imax=6.



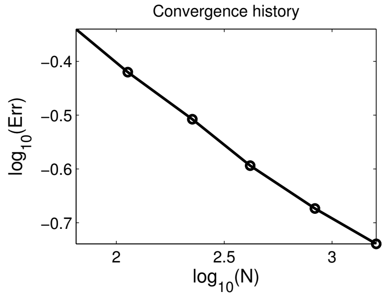
N	L^2-err	eoc
65	4.61_{-1}	
113	3.94_{-1}	
225	3.31_{-1}	0.54_{0}
417	2.77_{-1}	0.54_{0}
833	2.33_{-1}	0.53_{0}
1601	2.01_{-1}	0.47_{0}

Ex=3, pd=2, sig1=0, sig2=0, $\gamma = 0.7$, imax=6.



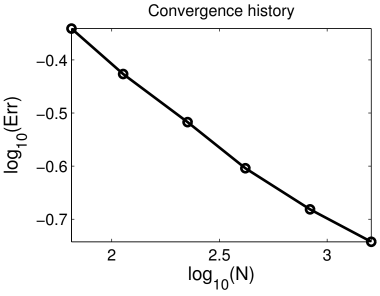
N	L^2-err	eoc
65	4.59_{-1}	
113	3.86_{-1}	
225	3.20_{-1}	0.58_{0}
417	2.64_{-1}	0.58_{0}
833	2.20_{-1}	0.57_{0}
1601	1.89_{-1}	0.50_{0}

Ex=3, pd=2, sig1=0, sig2=0, $\gamma = 0.8$, imax=6.



N	L^2-err	eoc
65	4.57_{-1}	
113	3.80_{-1}	
225	3.11_{-1}	0.62_{0}
417	2.55_{-1}	0.61_{0}
833	2.12_{-1}	0.58_{0}
1601	1.82_{-1}	0.50_{0}

Ex=3, pd=2, sig1=0, sig2=0, $\gamma = 0.9$, imax=6.



N	L^2-err	eoc
65	4.56_{-1}	
113	3.75_{-1}	
225	3.04_{-1}	0.65_0
417	2.49_{-1}	0.63_0
833	2.08_{-1}	0.58_0
1601	1.81_{-1}	0.47_0

Ex=3, pd=2, sig1=0, sig2=0, $\gamma = 1$, imax=6.

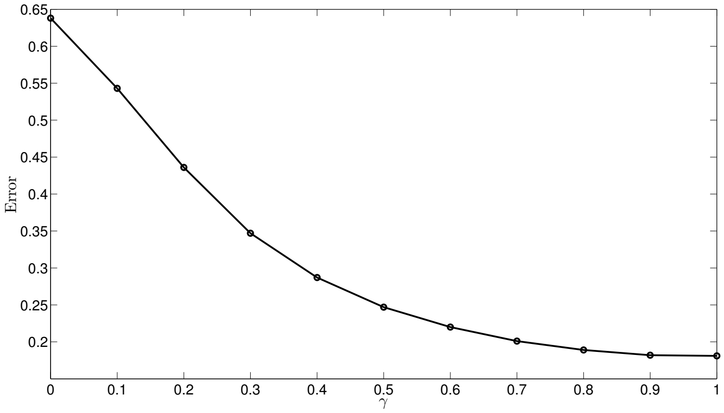


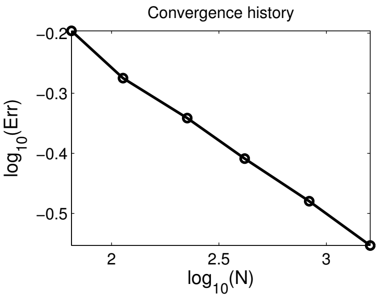
FIGURE 6.1. Dependence of the error on the choice of γ. The y-values are the errors for the approximation on a uniform mesh with polynomial degree 2 and 1601 degrees of freedom.

6.2. Eoc for the $H^{-\alpha}$ regularisation.

$$\int_\Omega \{\nabla \times \mathbf{E}_h \cdot \nabla \times \mathbf{v}_h + h^{2\alpha} \nabla \cdot \mathbf{E}_h \nabla \cdot \mathbf{v}_h - \nabla p_h \cdot \mathbf{v}_h\} = \int_\Omega \mathbf{f} \cdot \mathbf{v}_h$$

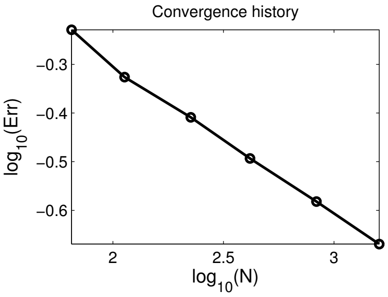
$$\int_\Omega \{-\mathbf{E}_h \cdot \nabla q_h + h^{2(1-\alpha)} \nabla q_h \cdot \nabla q_h\} = 0$$

6.2.1. *Example 3.* $\alpha \in (\frac{1}{2}, 1)$.



N	L^2-err	eoc
65	6.37_{-1}	
113	5.31_{-1}	
225	4.56_{-1}	0.54_0
417	3.90_{-1}	0.47_0
833	3.31_{-1}	0.49_0
1601	2.80_{-1}	0.50_0

Ex=3, pd=1, $\alpha = 0.6$, imax=6.



N	L^2-err	eoc
65	5.90_{-1}	
113	4.72_{-1}	
225	3.90_{-1}	0.66_0
417	3.21_{-1}	0.59_0
833	2.62_{-1}	0.61_0
1601	2.14_{-1}	0.60_0

Ex=3, pd=1, $\alpha = 0.7$, imax=6.

6.3. The estimator with uniform refinements. For uniform refinements we achieve the following convergence:

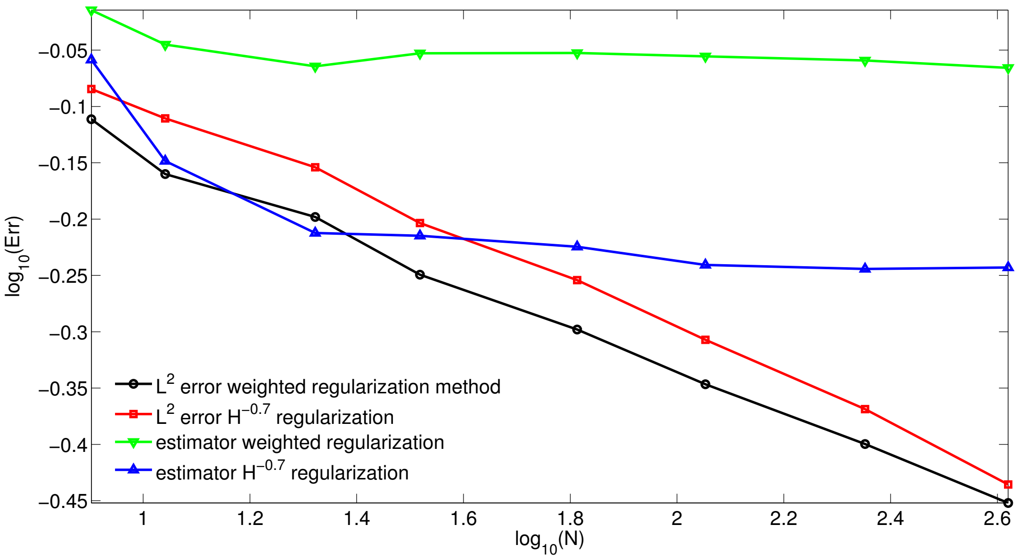


FIGURE 6.2. Results for uniform refinements.

N	L^2 error	eoc	estimator	eoc
8	8.23_{-1}		8.74_{-1}	
11	7.75_{-1}	0.38_0	7.10_{-1}	1.30_0
21	7.01_{-1}	0.31_0	6.13_{-1}	0.46_0
33	6.26_{-1}	0.50_0	6.10_{-1}	0.02_0
65	5.57_{-1}	0.34_0	5.96_{-1}	0.07_0
113	4.93_{-1}	0.44_0	5.75_{-1}	0.13_0
225	4.28_{-1}	0.41_0	5.70_{-1}	0.02_0
417	3.67_{-1}	0.50_0	5.72_{-1}	-0.01_0

TABLE 6.1. Uniform refinements for the $H^{-0.7}$ penalty regularization.

N	L^2 error	eoc	estimator	eoc
8	7.74_{-1}		9.67_{-1}	
11	6.92_{-1}	0.70_0	9.01_{-1}	0.44_0
21	6.34_{-1}	0.27_0	8.62_{-1}	0.14_0
33	5.63_{-1}	0.52_0	8.85_{-1}	-0.12_0
65	5.04_{-1}	0.33_0	8.86_{-1}	-0.00_0
113	4.50_{-1}	0.40_0	8.80_{-1}	0.03_0
225	3.98_{-1}	0.36_0	8.73_{-1}	0.02_0
417	3.53_{-1}	0.39_0	8.59_{-1}	0.05_0

TABLE 6.2. Uniform refinements with the weighted regularization method with penalty parameter 0.9.

6.4. Adaptive results for the Negative Sovolev Space penalty discretization.

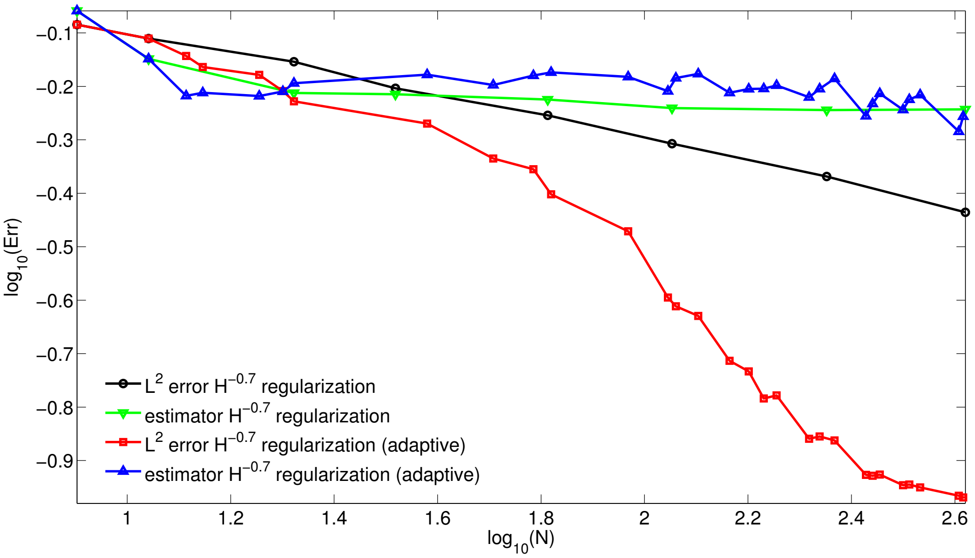


FIGURE 6.3. Adaptive refinements for the $H^{-0.7}$ penalty regularization.

We realize that although the estimator overestimates the error, it is a very good indicator. When we refine adaptively according to this indicator, we get much smaller errors for the same number of degrees of freedom.

6.5. Adaptive results for the Weighted Regularization Method.

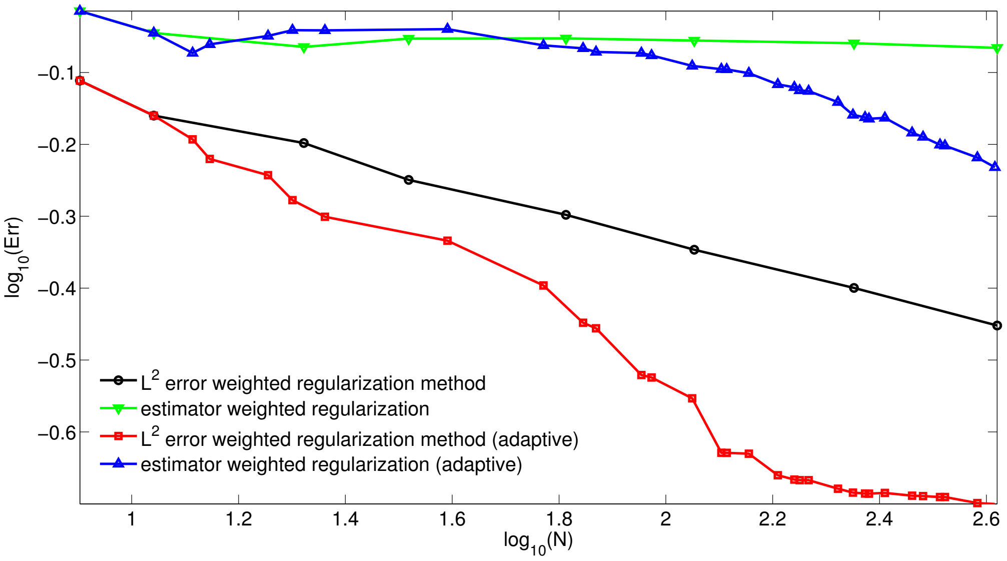


FIGURE 6.4. Adaptive refinements for the weighted regularization method.

We realize that also for the weighted regularization method we have found a good error indicator.

Bibliography

[1] R. A. Adams *Sobolev Spaces*, Academic Press, New York (1975)

[2] I. Babuška, J. E. Osborn *Finite Element-Galerkin Approximation of the Eigenvalues and Eigenvectors of Selfadjoint Problems*, Mathematics of Computation, Vol. 52, No. 186 (Apr., 1989), pp. 275-297

[3] I. Babuška, J. E. Osborn *Eigenvalue problems*, in: Clarlet, P.G., Lions, J.L. (eds.) Handbook of numerical analysis, vol. II. Finite Element Methods. Amsterdam: North Holland

[4] S. Badia, R. Codina *A nodal-based finite element approximation of the Maxwell problem suitable for singular solutions*, Manuscript (to appear in Numerische Mathematik).

[5] U. Banerjee, J. E. Osborn *Estimation of the Effect of Numerical Integration in Finite Element Eigenvalue Approximation*, Numer. Math. 56, 735-762 (1990)

[6] U. Banerjee *A note on the effect of numerical quadrature in finite element eigenvalue approximation*, Numer. Math. 61, 145-152 (1992)

[7] A. Bonito, J.-L. Guermond *Approximation of the eigenvalue problem for time harmonic Maxwell system by continuous Lagrange finite elements*, Report IAMCS 2009–121, Texas A&M (2009).

[8] A. Buffa, P. Ciarlet, Jr., E. Jamelot *Solving Electromagnetic Eigenvalue Problems in Polyhedral Domains with Nodal Finite Elements*, Numer. Math. 113, 497-518 (2009)

[9] J. M. Cascon, C. Kreuzer, R. H. Nochetto, K. G. Siebert *Quasi-Optimal Convergence Rate for an Adaptive Finite Element Method*, SIAM J. Numer. Anal., Vol. 46, Issue 5, pp. 2524-2550 (2008)

[10] P.G. Ciarlet *The finite element method for elliptic problems*, Amsterdam: North Holland, 1978

[11] P. Ciarlet, Jr., F. Lefèvre, S. Lohrengel, S. Nicaise *Weighted Regularization for Composite Materials in Electromagnetism*, Mathematical Modelling and Numerical Analysis, Vol. 44, 75-108 (2010)

[12] P. Clément *Approximation by Finite Element Functions Using Local Regularization*, RAIRO, 9, R-2 (1975), 77-84.

[13] M. Costabel *A Coercive Bilinear Form for Maxwell's Equations*, Journal of Mathematical Analysis and Applications Vol. 157, No. 2 (1991), 527-541.

[14] M. Costabel, M. Dauge *Singularities of electromagnetic fields in polyhedral domains*, Archives for Rational Mechanics and Analysis, 151(3):221-276, 2000.

[15] M. Costabel, M. Dauge *Computation of resonance frequencies for Maxwell equations in non smooth domains*, http:/www.math.univ-rennes1.fr/˜dauge

[16] M. Costabel, M. Dauge *Weighted Regularization of Maxwell Equations in Polyhedral domains. A rehabilitation of Nodal finite elements*, Numer. Math., Vol. 93, 239-277 (2002)

[17] X. Dai, J. Xu, A. Zhou *Convergence and optimal complexity of adaptive finite element eigenvalue computations*, Numer. Math. (2008) 110:313-355

[18] W. Dörfler *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal. 33, 1106-1124 (1996)

[19] W. Dörfler, V. Heuveline *Convergence of an adaptive hp finite element strategy in one space dimension*, Applied Numerical Mathematics 57 (2007) 1108-1124

[20] R. G. Durán, C. Padra, R. Rodríguez *A posteriori error estimates for the finite element approximation of eigenvalue problems*, Mathematical Models and Methods in Applied Sciences Vol. 13, No. 8 (2003) 1219-1229

[21] S. Giani *Convergence of adaptive finite element mehtods with applications to photonic crystals*, Ph.D. thesis

[22] S. Giani, I. G. Graham *A convergent adaptive method for eigenvalue problems*, SIAM J. Numer. Anal., Vol. 47, No. 2, pp. 1067-1091.

[23] V. Heuveline, R. Rannacher *A posteriori error control for finite element approximations of elliptic eigenvalue problems*, Advances in Computational Mathematics 15: 107-138, 2001.

[24] R. Hiptmair *Finite Elements in Computational Electromagnetism*, Acta Numerica (2002), pp. 237-339, Cambridge University Press, 2002 .

[25] J. D. Jackson *Classical electrodynamics*, John Wiley and Sons, 1975.

[26] P. Kuchment *Floquet Theory for Partial Differential Equations*, Birkhäuser Verlag, 1993.

[27] P. Kuchment *The mathematics of photonic crystals*, Mathematical modeling in optical science, pp. 207-272, 2001.

[28] M. G. Larson *A posteriori and a priori error analysis for finite element approximations of self-adjoint elliptic eigenvalue problems*, SIAM J. Numer. Anal. Vol. 38, No. 2, pp. 608-625

[29] P. Morin, R. H. Nochetto, K. Siebert *Convergence of adaptive finite element methods*, SIAM Rev. 44, 631-658 (2002)

[30] J. C. Nédélec *Mixed finite elements in $\mathbb{R}^3$*, Numer. Math. 35, (1980) 315-341

[31] J. C. Nédélec *A new family of mixed finite elements in $\mathbb{R}^3$*, Numer. Math. 50, (1986) 57-81

[32] C. Schwab *p- and hp-Finite Element Methods. Theory and Applications in Solid and Fluid Mechanics*, Clarendon Press, Oxford, 1998.

[33] R. L. Scott, S. Zhang *Finite element interpolation of nonsmooth functios satisfying boundary conditions*, Math. Comp. 54, 483-493.

[34] S. Shu, G. Wittum, J. Xu, L. Zhong *Optimal error estimates for Nédélec edge elements for time-harmonic Maxwell's equations*, Journal of Computational Mathematics, Vol. 27, No. 5, 2009, 563-572.

[35] T. Sorokina, A. J. Worsey *A multi-variate Powell-Sabin interpolant*, Advances in Computational Mathematics, 29(1):71-89, 2008.