Convergence of Adaptive Finite Element Methods for Elliptic Eigenvalue Problems with Applications to Photonic Crystals

submitted by

Stefano Giani

for the degree of Doctor of Philosophy

of the

University of Bath

May, 2008

SUMMARY

In this thesis we consider a convergent adaptive finite element method for elliptic eigenvalue problems on two/three-dimensional domains with applications in photonic crystal fibres (PCFs). We prove the convergence of the adaptive method for simple eigenvalues using linear finite elements. Each step of the adaptive procedure refines elements in which an a posteriori error estimator is large and also refines elements in which the computed eigenfunction has high oscillation. In order to treat PCF problems, we derive an explicit a posteriori error estimator based on residuals for such problems. We prove that the error estimator for the PCF case is reliable and efficient.

The error analysis extends the theory of convergence of adaptive methods for linear elliptic source problems to elliptic eigenvalue problems, and in particular it deals with various complications which arise essentially from the non-linearity of eigenvalue problems. Because of the non-linearity, the convergence result holds under the assumption that the initial finite element mesh is sufficiently fine.

We have collected a rich set of numerical experiments showing the advantages of using h-adaptivity and the convergence of our method. We have also developed two new strategies to improve numerical efficiency. The purpose of the first strategy is to approximate more than one eigenvalue of a generic elliptic eigenvalue problem on a single sequence of adapted meshes. Instead, the second strategy has been designed to solve just PCF problems more efficiently. This second strategy takes advantage of continuity of the bands in the spectra of PCF problems.

ACKNOWLEDGEMENTS

This thesis arose out of three years of research that has been done in Bath where I have worked with a great number of people whose contribution in assorted ways to the research and the making of the thesis deserved special mention.

In the first place I would like to record my gratitude to Ivan Graham for his supervision, advice, and guidance from the very early stage of this research. He also contributed by far the most to this thesis. I am very grateful for his patience, motivation and enthusiasm. I could not have imagined having a better advisor and mentor for my PhD.

I wish to express my warm and sincere thanks to Ilia Kamotski, who introduced me to the field of spectral theory. Furthermore, I would like to thank the following people for kind support and useful discussions: Rob Scheichl, Valery Smyshlyaev, Alastair Spence and in particular Vladimir Kamotski, who answered to so many questions of mine and helped me with the proof of Theorem 2.1.12.

I would like to say a big 'thank-you' to Richard Norton who shared a genuine interest and passion for the subject with me.

Special thanks are due to all my office mates during these years and especially to Simone Mandica, Patrick Lechner, Melina Freitag, Dave Simpson, Laura Hewitt.

Special thanks are due to all the PhD students of the Department of Mathematical Sciences and in particular to Bruce Boutelje.

I would like to particularly thank Alice and my family for all the support and encouragement throughout my postgraduate years. Without their belief in me, none of this would have been possible.

Contents

Chapter 1

Introduction

1.1 The subject of the thesis

The subject of this thesis is a convergent adaptive finite element method (AFEM) for eigenvalue problems. Eigenvalue problems arise naturally in many physical processes and they have a lot of applications in physics and engineering. Example of applications are in structural engineering, weather forecasting and in quantum physics.
We will consider two types of elliptic eigenvalue problems. The first type will be called generic elliptic eigenvalue problem (with eigenpair (λ, u), where $u \neq 0$) and it is defined as follows:

$$-\nabla \cdot \big(\mathcal{A}(x)\ \nabla u(x)\big)\ =\ \lambda\ \mathcal{B}(x)\ u(x), \quad \text{in } \Omega \tag{1.1.1}$$

where Ω is a bounded polygonal or polyhedral region in $\mathbb{R}^d$, with $d = 2, 3$ and subject to homogeneous Dirichlet boundary conditions. Moreover, $\mathcal{A}(x)$ is assumed to be a real piecewise constant valued matrix and uniformly positive definite and bounded above and below by positive numbers. Similarly $\mathcal{B}(x)$ is a real piecewise constant function, which is bounded above and below by positive constants for all $x \in \Omega$.
The second kind of problem, which is considered in this work, is a more complicated elliptic eigenvalue problem, which arises from wave guide applications. We are particularly interested in a new kind of wave guides called photonic crystal fibers (PCFs), which are an evolution of standard fiber optics. In order to understand how light propagates inside PCFs, it is necessary to solve an eigenvalue problem based on Maxwell's equations. This eigenvalue problem is hard to solve, so the standard way to treat such a problem is to use the Floquet transform, which is also widely used in crystallography. The action of the Floquet transform splits the PCF eigenvalue problem into a family of easier eigenvalue problems parameterized by the value of the quasimomentum $\vec{\kappa}$, which is a real vector of dimension 2 and which is defined below. The form of each problem

in the family (with eigenpair (λ, u)) is

$$-(\nabla + i\vec{\kappa}) \cdot \left(\mathcal{A}(x)\ (\nabla + i\vec{\kappa})\ u(x)\right) \ = \ \lambda\ \mathcal{B}(x)\ u(x), \quad \text{in } \Omega, \tag{1.1.2}$$

subject to periodic boundary conditions. The domain of problem (1.1.2) is the unit "cell" of the underlying periodic problem, e.g. (in 2D) a square or, more generally, a polygon with an even number of sides and with opposite sides of the same length and with the same orientation.

Many papers have been published on problems related to PCFs, and in the last years the field has been very active. In this work we are mainly interested in computing eigenvalues of problems like (1.1.2) for a given geometry of the fiber, but many authors have considered different aspects of PCF related problems. For example, in [18, 15] the problem of optimizing the internal structure of the fiber, in order to maximize its efficiency, has been addressed using different methods. In [15] a method based on finite differences has been used, instead [18] it prefers finite elements. Furthermore, in [50] non-linear eigenvalue techniques have been used on eigenvalue problems like (1.1.2). Even if we restrict our attention only to the papers regarding the problems considered in this work, we found that different methods have been proposed. We have methods based on expansions of eigenfunctions, a good example of which is [22], where the localized modes of a PCF have been approximated using expansions in Bloch eigenfunctions. Other authors preferred analytical methods. Such a method has been used in [23] for fibers with simple geometry, in fact analytical methods impose considerable limitations on the geometry of the fiber that they can analysed. There are even papers in which plane-wave expansion methods have been used, like [44, 10]. Despite all the other possibilities, we chose to use FEMs because they are already very widely used to solve many different classes of linear and non-linear problems, and also because they are very flexible methods. There are already some works about PCFs based on finite element methods [8, 16, 17, 29, 33], however, until now no one has used adaptivity on these problems.

Adaptivity is a key factor of the success of FEMs for PDE problems, because it improves the accuracy of computations with, on the other hand, very reasonable computational costs. In this work we implemented h-adaptivity in our methods, which consists in subdivide or "refining" only those elements in a mesh on which some error indicator is sufficiently large. For linear PDEs, there is a vast literature on h-adaptivity and a posteriori error estimators [52, 2, 7, 11, 45]. However, for eigenvalue problems there are only few works [21, 37, 53, 28].

Another kind of adaptivity that could be very useful as well for eigenvalue problems is the hp-adaptivity. In this case, not only the size of the elements are adjusted to improve the accuracy of the simulations, but also the order of the polynomials on each element is tuned appropriately. The exploitation of this kind of adaptivity could be

the topic of further research.
In the last years, it has been possible to prove convergence for adaptive finite element methods (AFEMs) for linear problems [20, 42, 40, 43, 14, 13, 41] and for some examples of non-linear problems [19]. But, for eigenvalue problems, as far as we know, the question of convergence of AFEMs is still open and this is the first result about convergence AFEM for problems (1.1.1) and (1.1.2). More recently, another work about convergence AFEM for eigenvalue problems has appeared [12]. This work is newer than ours and the authors were able to remove the dependence on the oscillations in the convergence proof.

1.2 The aims of the thesis

The main aim of the thesis is to prove an efficient and convergent adaptive finite element method for eigenvalue problems arising from PCF applications. Secondly, we have extended the proof of convergence of our AFEM to generic elliptic eigenvalue problems in 2D and 3D.
In order to obtain such a method we need firstly a good understanding of numerical analysis for elliptic eigenvalue problems. Secondly, we need an error estimator, suitable for problems (1.1.1) and (1.1.2), to drive the mesh adaptivity and for which it is possible to prove the convergence of the method.
We paid much attention to the aspect of computational cost, too. In Chapter 5, we present a new method to compute efficiently the solutions of a family of problems (1.1.2) and also a method to compute many eigenvalues on the same sequence of adapted meshes.

1.3 The main achievements of the thesis

The main achievements of this thesis can be summarized as follows.

(i) Numerical analysis for elliptic eigenvalue problems for PCFs. This analysis is an extension of the standard analysis for elliptic eigenvalue problems [51, 6].

(ii) Explicit a posteriori error estimators based on residuals for general elliptic eigenvalue problems and for problems from PCF applications. In particular, we proved that the error estimator for the PCF case is reliable and efficient. Then we extended the results also to the general elliptic case.

(iii) A convergent adaptive finite element method for general elliptic eigenvalue problems and for PCF eigenvalue problems.

(iv) A code to compute solutions of problems (1.1.1) and (1.1.2), which takes advantage of techniques like Arnoldi's method ARPACK [38] and the fast direct sparse solver for linear problems ME27 [47] contained in the HSL archive.

(v) A rich set of numerical experiments showing the advantages of using h-adaptivity and the convergence of our method.

1.3.1 Definition of the problems and notation

In this section we are going to define rigorously the problems analysed in this work. But before that, we introduce all the necessary notation.

1.3.2 Functional spaces and norms

We are going to use mainly six different Sobolev spaces. Firstly, we are going to use the standard $L^2(\Omega)$, which is a bounded polygonal or polyhedral region in $\mathbb{R}^d$, with $d = 2, 3$. While, all the other functional spaces are defined below:

Definition 1.3.1 (Weighted L^2 spaces)**.** *Let $\mathcal{B}$ be a positive and bounded function on Ω, which is a bounded polygonal or polyhedral region in $\mathbb{R}^d$, with $d = 2, 3$. The $L^2_{\mathcal{B}}$ space on Ω is defined as the set*

$$L^2_{\mathcal{B}}(\Omega) = \{f : \Omega \to \mathbb{C} : \|f\|_{0,\mathcal{B},\Omega} < +\infty\},$$

in which the norm $\|\cdot\|_{0,\mathcal{B},\Omega}$ is defined as follows:

$$\|f\|_{0,\mathcal{B},\Omega} := \left(\int_\Omega \mathcal{B}(x)|f(x)|^2 \, dx \right)^{1/2},$$

where the integral to be understood in the Lebesgue sense.

Definition 1.3.2 (Sobolev space H^1)**.** *Let Ω be a bounded polygonal or polyhedral region in $\mathbb{R}^d$, with $d = 2, 3$. Then, the Sobolev space H^1 on Ω is defined as*

$$H^1(\Omega) = \{f : \Omega \to \mathbb{C},\ f \in L^2(\Omega) : \|f\|_{1,\Omega} < \infty\},$$

where the norm is defined by

$$\|f\|_{1,\Omega} := \left(\sum_{|\alpha| \leq 1} \left\| \frac{\partial^\alpha f}{\partial x^\alpha} \right\|^2_{0,\Omega} \right)^{1/2},$$

with α a multi-index.

Definition 1.3.3 (Sobolev space H_0^1)**.** *Let Ω be a bounded polygonal or polyhedral region in $\mathbb{R}^d$, with $d = 2, 3$. Then, the Sobolev space H_0^1 on Ω is the subspace of $H^1(\Omega)$ containing only the elements with trace equal to 0 on the boundary of Ω.*

Definition 1.3.4 (Sobolev space H_π^1)**.** *Let $\Omega \subset \mathbb{R}^2$ be a polygon with an even number of sides and with opposite sides of the same length and with the same orientation. Then, the Sobolev space H_π^1 on Ω is the subset of $H^1(\Omega)$ containing only the elements satisfying periodic boundary conditions on Ω.*

Definition 1.3.5 (Sobolev space H^t, with $t \in \mathbb{R}$)**.** *Let Ω be a bounded polygonal or polyhedral region in $\mathbb{R}^d$, with $d = 2, 3$. Then, the Sobolev space H^t, with $t \in \mathbb{R}$, is defined by interpolation as shown in [1, Chap. 7]. Thanks to this method, we can define any Sobolev space H^t as an intermediate space between two Sobolev spaces $H^{\underline{t}}$ and $H^{\overline{t}}$, with $\underline{t}$ and $\overline{t}$ integer and with $\underline{t} < t < \overline{t}$.*

1.3.3 Discontinuous coefficients

We define the matrix function $\mathcal{A}$ to be piecewise constant and uniformly positive definite, i.e.

$$\underline{a} \leq \xi^T \mathcal{A}(x)\xi \leq \overline{a} \quad \text{for all} \quad \xi \in \mathbb{R}^2 \text{ with } |\xi| = 1 \text{ and for all} \quad x \in \Omega, \tag{1.3.1}$$

which is also bounded above and below by real numbers $\underline{a}$ and $\overline{a}$ greater than 0. Similarly, we define a piecewise constant function $\mathcal{B}$ in such a way that it is bounded from above and from below by positive constants $\underline{b}$ and $\overline{b}$ for all $x \in \Omega$, i.e.

$$\underline{b} \leq \mathcal{B}(x) \leq \overline{b} \quad \text{for all} \quad x \in \Omega. \tag{1.3.2}$$

1.3.4 Sesquilinear and bilinear forms

We are going to use the following bilinear and sesquilinear forms defined on Ω:

(i) For any u and v in $H_0^1(\Omega)$:

$$a(u, v) := \int_\Omega \mathcal{A}\nabla u \cdot \nabla v; \tag{1.3.3}$$

(ii) For any u and v in $H_\pi^1(\Omega)$ and for any value of the quasimomentum $\vec{\kappa}$, which is a real vector of dimension 2, we have:

$$a_\kappa(u, v) := \int_\Omega \big(\mathcal{A}(\nabla + i\vec{\kappa})u\big) \cdot (\nabla - i\vec{\kappa})\bar{v}; \tag{1.3.4}$$

(iii) For any u and v in $L^2(\Omega)$ or in $L^2_\pi(\Omega)$:

$$(u,v)_{0,\mathcal{B},\Omega} := \int_\Omega (\mathcal{B}u)\cdot\bar{v}; \tag{1.3.5}$$

(iv) Let S be a constant greater than 0. For any u and v in $H^1_\pi(\Omega)$ and for any value of the quasimomentum $\vec{\kappa}$:

$$a_{\kappa,S}(u,v) := a_\kappa(u,v) + S(u,v)_{0,\mathcal{B},\Omega}. \tag{1.3.6}$$

The introduction of the positive constant S has been necessary, since the sesquilinear form (1.3.4) may not be coercive for all values of $\vec{\kappa}$. Instead, in Chapter 2 we prove that (1.3.6) is coercive for any $S > 0$.

1.3.5 Definitions of the problems

In order to simplify the analysis for PCF problems, we consider only square cell crystals, which implies that for those problems the domain Ω is just a square. The analysis holds also for crystals with more general cells.
In this work, we are going to analyse the following problems in variational form. In problem (i) below, we suppose that Ω is a polygonal or polyhedral domain with Dirichlet boundary conditions, and in problems (ii) and (iii) we suppose that Ω is square:

(i) The general elliptic eigenvalue problems is: *seek eigenpairs of the form* $(\lambda_j, u_j) \in \mathbb{R}\times H^1_0(\Omega)$*, with* $\|u_j\|_{0,\mathcal{B},\Omega} = 1$ *such that*

$$a(u_j,v) = \lambda_j (u_j,v)_{0,\mathcal{B},\Omega}\ , \quad \text{for all } v \in H^1_0(\Omega); \tag{1.3.7}$$

(ii) The model problem for PCFs is: *seek eigenpairs of the form* $(\lambda_j, u_j) \in \mathbb{R}\times H^1_\pi(\Omega)$*, with* $\|u_j\|_{0,\mathcal{B},\Omega} = 1$ *such that*

$$a_\kappa(u_j,v) = \lambda_j (u_j,v)_{0,\mathcal{B},\Omega}\ , \quad \text{for all } v \in H^1_\pi(\Omega). \tag{1.3.8}$$

(iii) The shifted version (with $S > 0$) of model problem for PCFs is: *seek eigenpairs of the form* $(\zeta_j, u_j) \in \mathbb{R}\times H^1_\pi(\Omega)$*, with* $\|u_j\|_{0,\mathcal{B},\Omega} = 1$ *such that*

$$a_{\kappa,S}(u_j,v) = \zeta_j (u_j,v)_{0,\mathcal{B},\Omega}\ , \quad \text{for all } v \in H^1_\pi(\Omega). \tag{1.3.9}$$

Note that the shift S defines the relation $(\zeta_j, u_j) = (\lambda_j + S, u_j)$, which is a one-one relation between the spectra of problems (ii) and (iii).

In (i), eigenfunctions u_j are real valued because the bilinear form $a(\cdot,\cdot)$ is real symmetric. In (ii) and (iii), eigenfunctions u_j are in general complex valued. In all cases

eigenvalues ζ_j, λ_j are real, because $a(\cdot,\cdot)$, $a_\kappa(\cdot,\cdot)$ and $a_{\kappa,S}(\cdot,\cdot)$ are sesquilinear forms.

1.4 Photonic Crystal Fibers (PCFs)

Photonic crystals are constructed by assembling portions of periodic media composed of dielectric materials and they are designed to exhibit interesting properties in the propagation of electromagnetic waves, such as spectral band gaps. In other words, monochromatic electromagnetic waves of certain frequencies do not exist in these structures.

Media with band gaps have many potential applications, for example, in optical communications, filters, lasers, switchers, optical transistors; see [32, 31, 46, 35, 3] for an introduction to photonic crystals. But, for all these applications, the employment of materials with band gaps is not enough. It is also necessary to create eigenvalues inside the gaps in the spectra of the media. The common way to create such eigenvalues is by introducing a localized defect in the periodic structures of media [25]. The importance of these eigenvalues is due to the fact that electromagnetic waves, which have frequencies corresponding to these eigenvalues in the gaps, may remain trapped inside the defects [23, 25] and they decay exponentially away from the defects.

PCFs are of special interest. Such structures are much easier to fabricate than general 3D photonic crystals, while they still allow for many important applications. Theoretical analysis for PCFs is significantly simpler than for 3D photonic crystals because a PCF dielectric system has two fundamental types of modes, E polarized (TM mode) and H polarized (TE mode). In each mode, the PCF problem reduces to a one-component wave equation for the E field or H field, respectively, as we shall show in the next subsection.

1.4.1 The physics

The propagation of light inside dielectric materials, which constitute photonic crystals, is governed by Maxwell's equations (in the absence of free charges and currents)

$$\begin{cases} \nabla \times \mathbf{E}(\mathbf{x},t) = -\dfrac{1}{c}\dfrac{\partial \mathbf{B}(\mathbf{x},t)}{\partial t}, \\ \nabla \times \mathbf{H}(\mathbf{x},t) = \dfrac{1}{c}\dfrac{\partial \mathbf{D}(\mathbf{x},t)}{\partial t}, \\ \nabla \cdot \mathbf{B}(\mathbf{x},t) = 0, \\ \nabla \cdot \mathbf{D}(\mathbf{x},t) = 0, \end{cases} \tag{1.4.1}$$

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 speed of light in a vacuum. All vector fields are functions from $\mathbb{R}^3 \times \mathbb{R}$ to $\mathbb{R}^3$. This system is incomplete without the constitutive relations that describe how the fields $\mathbf{D}$ and $\mathbf{B}$ depend on $\mathbf{E}$ and $\mathbf{H}$. Here we assume the linear constitutive relations:

$$\begin{cases} \mathbf{D}(\mathbf{x},t) = \varepsilon(\mathbf{x})\mathbf{E}(\mathbf{x},t), \\ \\ \mathbf{B}(\mathbf{x},t) = \mu(\mathbf{x})\mathbf{H}(\mathbf{x},t), \end{cases} \tag{1.4.2}$$

where ε and μ are the dielectric and magnetic permeability tensors. Inserting relations (1.4.2) into (1.4.1) we obtain:

$$\begin{cases} \nabla \times \mathbf{E}(\mathbf{x},t) = -\dfrac{1}{c}\mu(\mathbf{x})\dfrac{\partial \mathbf{H}(\mathbf{x},t)}{\partial t}, \\ \\ \nabla \times \mathbf{H}(\mathbf{x},t) = \dfrac{1}{c}\varepsilon(\mathbf{x})\dfrac{\partial \mathbf{E}(\mathbf{x},t)}{\partial t}, \\ \\ \nabla \cdot \mu(\mathbf{x})\mathbf{H}(\mathbf{x},t) = 0, \\ \\ \nabla \cdot \varepsilon(\mathbf{x})\mathbf{E}(\mathbf{x},t) = 0. \end{cases} \tag{1.4.3}$$

In order to understand the behavior of light inside these materials, we have to analyse each frequency separately. Monochromatic light of frequency ω can be modeled by

$$\begin{aligned} \mathbf{E}(\mathbf{x},t) &= e^{i\omega t}\tilde{\mathbf{E}}(\mathbf{x}), \\ \\ \mathbf{H}(\mathbf{x},t) &= e^{i\omega t}\tilde{\mathbf{H}}(\mathbf{x}), \end{aligned} \tag{1.4.4}$$

where $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{H}}$ are the modes of the analysed monochromatic light.
So, substituting (1.4.4) into (1.4.3) we obtain a system of differential equations describing the propagation of light of frequency ω in a photonic crystal:

$$\begin{cases} \nabla \times \tilde{\mathbf{E}}(\mathbf{x}) = -\dfrac{i\omega}{c}\mu(\mathbf{x})\tilde{\mathbf{H}}(\mathbf{x}), \\ \\ \nabla \times \tilde{\mathbf{H}}(\mathbf{x}) = \dfrac{i\omega}{c}\varepsilon(\mathbf{x})\tilde{\mathbf{E}}(\mathbf{x}), \\ \\ \nabla \cdot \mu(\mathbf{x})\tilde{\mathbf{H}}(\mathbf{x}) = 0, \\ \\ \nabla \cdot \varepsilon(\mathbf{x})\tilde{\mathbf{E}}(\mathbf{x}) = 0. \end{cases} \tag{1.4.5}$$

The system of equations (1.4.5) is time-independent. Each point in the spectrum of (1.4.5) corresponds to a frequency of light which is allowed to travel through the crystal. On the other hand, any point not in the spectrum of (1.4.5) corresponds to a frequency of light which is not allowed to travel through the crystal.

1.4.2 Periodic media and polarized modes

Photonic crystal fibers (PCFs) are one of the most important applications of photonic crystals. PCFs are a new type of optic fibers, in which, along the axis in the center of the fiber, is embedded a photonic crystal (commonly with defect). Figure 1-1 shows an example of the structure in a section of a PCF. In the structure of a PCF, it is commonly possible to distinguish between two regions: a portion of periodic structure - see the right picture in Figure 1-1- surrounding a "defect" and a "defect" in which the periodicity of the structure is broken - see the center of the left picture in Figure 1-1. The periodic structures used in PCFs have the particular characteristic that they do not allow to all light frequencies to travel within it. So PCFs trap beams of light of characteristic frequencies inside the defect region.

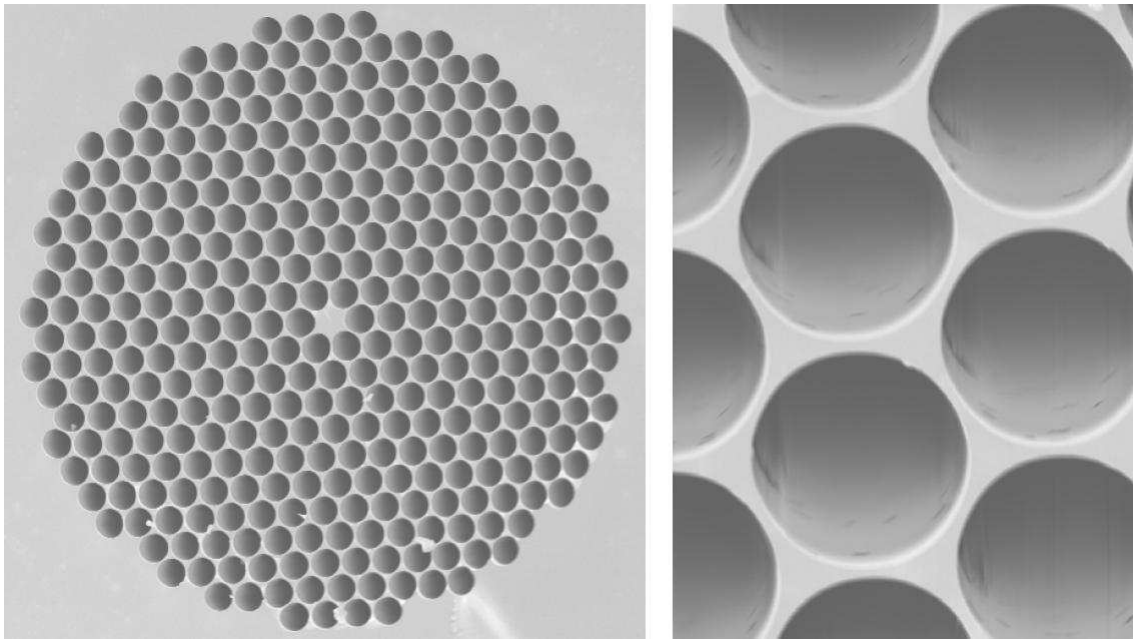

Figure 1-1: An example of micro-structure in the section of a PCF. This picture can be found at the address: http : //en.wikipedia.org/wiki/Photonic_crystal_fibers

The first task, in order to analyse a PCF, is to determine what light frequencies are not allowed to travel across the periodic structure. To simplify the analysis we can take a periodic dielectric medium filling all the real space, instead of considering just the portion of the periodic structure embedded in the PCF.
We are going to consider only "orthotropic" media or in other words, media with a

periodic structure invariant along one axis. This is because the micro-structure in PCFs are invariant along the axis of the fiber. So, in the PCF case, we assume that the tensor ε appearing in (1.4.5) is "orthotropic", i.e. it satisfies

$$\varepsilon = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & 0 \\ \varepsilon_{21} & \varepsilon_{22} & 0 \\ 0 & 0 & \varepsilon_{33} \end{pmatrix}, \tag{1.4.6}$$

with $\varepsilon_{12} = \varepsilon_{21}$ and where each ε_{ij} is a function of x,y only and also we assume that the tensor ε is positive definite and invertible for any value of x and y in the domain of the problem.

Since the structures of orthotropic media are invariant along one axis, that we suppose to be the z-axis, it is straightforward to conclude that also the modes $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{H}}$ in (1.4.5) are invariant along the same axis. For orthotropic media, the system of equations (1.4.5) becomes

$$\begin{cases} \nabla \times \tilde{\mathbf{E}}(x,y) = -\dfrac{i\omega}{c}\tilde{\mathbf{H}}(x,y), \\ \nabla \times \tilde{\mathbf{H}}(x,y) = \dfrac{i\omega}{c}\varepsilon(x,y)\tilde{\mathbf{E}}(x,y), \\ \nabla \cdot \tilde{\mathbf{H}}(x,y) = 0, \\ \nabla \cdot \varepsilon(x,y)\tilde{\mathbf{E}}(x,y) = 0. \end{cases} \tag{1.4.7}$$

where we have assumed that $\mu = 1$, since the common choice of materials for PCFs exhibit values of μ very close to the value for air, which fills the holes of the structures. So, without losing generality we can choose $\mu = 1$.

Now, we want to show that the system of equations (1.4.7) splits naturally in 2 disjoint subproblems: called TE and TM modes.

TM mode

Substituting in (1.4.7) the first equation into the second one we obtain:

$$\nabla \times \left(\nabla \times \tilde{\mathbf{E}}(x,y)\right) \;=\; \frac{\omega^2}{c^2}\varepsilon(x,y)\tilde{\mathbf{E}}(x,y)\;, \tag{1.4.8}$$

such vectorial equation can be written, denoting the components of $\tilde{\mathbf{E}} = (E_1, E_2, E_3)$, as a system of three equations:

$$\begin{cases} E_{2yx} - E_{1yy} - E_{1zz} + E_{3xz} &= \dfrac{\omega^2}{c^2}(\varepsilon_{11}E_1 + \varepsilon_{12}E_2) \ , \\ \\ E_{3yz} - E_{2zz} - E_{2xx} + E_{1xy} &= \dfrac{\omega^2}{c^2}(\varepsilon_{21}E_1 + \varepsilon_{22}E_2) \ , \\ \\ E_{1xz} - E_{3xx} - E_{3yy} + E_{2yz} &= \dfrac{\omega^2}{c^2}\varepsilon_{33}E_3 \ , \end{cases} \tag{1.4.9}$$

where the notation subscribe x, y and z means derivatives along the directions of each axis.

Since the electric field depends only on x and y, we have that all the terms in (1.4.9) involving differentiation along z are zero:

$$\begin{cases} E_{2yx} - E_{1yy} &= \dfrac{\omega^2}{c^2}(\varepsilon_{11}E_1 + \varepsilon_{12}E_2) \ , \\ \\ -E_{2xx} + E_{1xy} &= \dfrac{\omega^2}{c^2}(\varepsilon_{21}E_1 + \varepsilon_{22}E_2) \ , \\ \\ -E_{3xx} - E_{3yy} &= \dfrac{\omega^2}{c^2}\varepsilon_{33}E_3 \ . \end{cases} \tag{1.4.10}$$

Now it is straightforward to see that the first two equations of (1.4.10) form a problem

$$\begin{cases} E_{2yx} - E_{1yy} &= \dfrac{\omega^2}{c^2}(\varepsilon_{11}E_1 + \varepsilon_{12}E_2) \ , \\ \\ -E_{2xx} + E_{1xy} &= \dfrac{\omega^2}{c^2}(\varepsilon_{21}E_1 + \varepsilon_{22}E_2) \ , \end{cases} \tag{1.4.11}$$

and the third equation of (1.4.10) forms another problem independent from the first one

$$-E_{3xx} - E_{3yy} = \frac{\omega^2}{c^2}\varepsilon_{33}E_3 \ , \tag{1.4.12}$$

since the third equation involves only the component E_3, which is absent in the first two equations.

We are going to call (1.4.12) TM mode and, denoting E_3 by a complex valued function $U(x, y)$, the problem (1.4.12) can be written in the simpler form:

$$-\triangle U = \lambda \, \mathcal{B} \, U, \tag{1.4.13}$$

with $\lambda = \omega^2/c^2$ and with $\mathcal{B} = \varepsilon_{33}$. It is clear that the electric field of all the solutions of the TM mode has the form $\tilde{\mathbf{E}} = (0, 0, U)$. Plugging into (1.4.7) the electric field

$\tilde{\mathbf{E}} = (0, 0, U)$, we obtain that the correspondent magnetic field satisfies

$$-\frac{i\omega}{c}\tilde{\mathbf{H}}(x, y) \;=\; (U_y(x, y), -U_x(x, y), 0) \;.$$

TE mode

To obtain a simple formulation of the TE mode, it is necessary to start again from (1.4.7) and then substituting the second equation into the first one to obtain

$$\nabla \times \left(\varepsilon^{-1}(x, y)\nabla \times \tilde{\mathbf{H}}(x, y)\right) \;=\; \frac{\omega^2}{c^2}\tilde{\mathbf{H}}(x, y) \;, \tag{1.4.14}$$

where ε^{-1} is the inverse of ε and which is equal to:

$$\varepsilon^{-1} \;=\; \frac{1}{\varepsilon_{11}\varepsilon_{22} - \varepsilon_{12}\varepsilon_{21}} \begin{pmatrix} \varepsilon_{22} & -\varepsilon_{12} & 0 \\ -\varepsilon_{21} & \varepsilon_{11} & 0 \\ 0 & 0 & \frac{\varepsilon_{11}\varepsilon_{22} - \varepsilon_{12}\varepsilon_{21}}{\varepsilon_{33}} \end{pmatrix}.$$

The vectorial equation (1.4.14) is a set of three scalar equations in the component of the magnetic field $\tilde{\mathbf{H}} = (H_1, H_2, H_3)$. In (1.4.15) below we have reported the third equation of the system, which involves only the component H_3 and it is disjoint from the other two equations:

$$\left(\frac{-\varepsilon_{12}H_{3y} - \varepsilon_{11}H_{3x}}{\varepsilon_{11}\varepsilon_{22} - \varepsilon_{12}\varepsilon_{21}}\right)_x - \left(\frac{\varepsilon_{22}H_{3y} + \varepsilon_{21}H_{3x}}{\varepsilon_{11}\varepsilon_{22} - \varepsilon_{12}\varepsilon_{21}}\right)_y \;=\; \frac{\omega^2}{c^2}H_3 \;. \tag{1.4.15}$$

Denoting the component H_3 of the magnetic field by a complex valued function $U(x, y)$ we have that (1.4.15) can be written in a simpler form:

$$-\nabla \cdot (\mathcal{A}\, \nabla U) \;=\; \lambda\, U, \tag{1.4.16}$$

where $\lambda = \omega^2/c^2$ and where

$$\mathcal{A} = \frac{1}{\varepsilon_{11}\varepsilon_{22} - \varepsilon_{12}\varepsilon_{21}} \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \end{pmatrix}.$$

We use (1.4.16) as the definition of the TE mode. So the magnetic field of the solutions for the TE mode is $\tilde{\mathbf{H}} = (0, 0, U)$ and plugging into (1.4.7) such magnetic field, we obtain that the correspondent electric field satisfies

$$\frac{i\omega}{c}\varepsilon(x, y)\; \tilde{\mathbf{E}}(x, y) \;=\; (U_y(x, y), -U_x(x, y), 0) \;. \tag{1.4.17}$$

1.4.3 Floquet transform

The domain of problems (1.4.13) and (1.4.16) is the whole $\mathbb{R}^2$ filled with a periodic structure. Moreover, we have from the theory [36] that the spectra of periodic problems with smooth coefficients are formed by bands of essential spectrum. Unfortunately, there is not a similar proof for periodic problems with discontinuous coefficients, but it is widely accepted the conjecture that also the spectra of these problems are formed by bands of essential spectrum.
The unboundness of the domain and the nature of their spectra, make problems (1.4.13) and (1.4.16) very difficult to be treated numerically in their stated form.

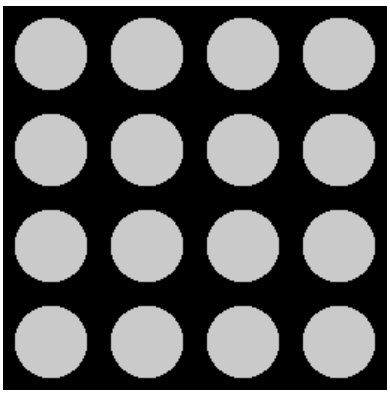

Figure 1-2: A portion of a possible periodic structure medium.

In order to improve our chances to solve efficiently these problems, we are going to use the Floquet transform [34, 35], which is a standard tool of analysis for handling PDEs with periodic coefficients. This transform has been borrowed from crystallography as well as most of the terminology. So, we define the "primitive cell" Ω of a periodic medium to be the smallest portion of the structure of the medium, which if periodically repeated will recover the structure of the whole medium.
A fundamental concept in the description of any crystal structure is the "lattice", which specifies the periodic array in which the repeated primitive cells of the crystal are arranged. A 2D lattice is defined as the linear span of two vectors v_1 and v_2. For any 2D lattice that exists a "reciprocal lattice", which is another 2D lattice generated by vectors w_1 and w_2 such that $v_i \cdot w_j = 2\pi\delta_{ij}$. We define the "first Brillouin zone" $\mathcal{K}$ as the primitive cell of the reciprocal lattice. For example, if the periodic cell of a medium is the unit square, as for the structure in Figure 1-2, the first Brillouin zone $\mathcal{K}$ is the set $[-\pi, +\pi]^2$. In general both the primitive cell and the first Brillouin zone $\mathcal{K}$ are polygonal sets.

The Floquet transform is defined for any function $g \in L^2(\mathbb{R}^2)$ as

$$(\mathcal{F}g)(\vec{\kappa}, \mathbf{x}) = e^{-i\vec{\kappa}\cdot\mathbf{x}} \sum_{\mathbf{n}\in\mathbb{Z}^2} g(\mathbf{x}-\mathbf{n}) e^{i\vec{\kappa}\cdot\mathbf{n}}, \tag{1.4.18}$$

where the "quasimomentum" $\vec{\kappa} \in \mathcal{K}$ acts as a parameter. The main effect of the application of the Floquet transform on an operator with periodic coefficients, is the decomposition of the operator into the direct integral of a family of operators on the primitive cell. Each operator in the family is characterized by a different value of the quasimomentum.
Applying the Floquet transform to problem (1.4.13), and denoting by $u = \mathcal{F}U$, we get

$$-(\nabla + i\vec{\kappa}) \cdot (\nabla + i\vec{\kappa})\ u \ = \ \lambda\ \mathcal{B}\ u,$$

then, multiplying by a test function v and integrating by parts we have:

$$\int_\Omega (\nabla + i\vec{\kappa})\ u \cdot (\nabla - i\vec{\kappa})\ \overline{v} \ = \ \lambda \int_\Omega u\ \mathcal{B}\ \overline{v}, \quad \text{for all } v \in H^1_\pi(\Omega), \tag{1.4.19}$$

which is a special case of problem (1.3.8) with $\mathcal{A} = 1$.
Similarly, applying the Floquet transform to problem (1.4.16), and denoting by $u = \mathcal{F}U$, we get

$$-(\nabla + i\vec{\kappa}) \cdot \mathcal{A}\ (\nabla + i\vec{\kappa})\ u \ = \ \lambda\ u,$$

again, multiplying by a test function v and integrating by parts follows:

$$\int_\Omega \mathcal{A}\ (\nabla + i\vec{\kappa})\ u \cdot (\nabla - i\vec{\kappa})\ \overline{v} \ = \ \lambda \int_\Omega u\ \overline{v},, \quad \text{for all } v \in H^1_\pi(\Omega), \tag{1.4.20}$$

which is another special case of problem (1.3.8) with this time $\mathcal{B} = 1$.
A consequence of the application of the Floquet transform is that the spectra of the TE and TM modes have been decomposed into the spectra of the corresponding problems forming the two families. In order to see that, we can suppose that (λ, U) is an eigenvalue of (1.4.16), then applying the Floquet transform to U we obtain a function u_κ for each value of $\vec{\kappa}$. So, for each value of $\vec{\kappa}$, if we apply the Floquet transform to (1.4.16):

$$\mathcal{F}\big(-\nabla\cdot(\mathcal{A}\ \nabla U)\big)(\vec{\kappa}, \mathbf{x}) \ = \ \mathcal{F}\big(\lambda\ U\big)(\vec{\kappa}, \mathbf{x}),$$

we obtain

$$-(\nabla + i\vec{\kappa}) \cdot \mathcal{A}\ (\nabla + i\vec{\kappa})\ \mathcal{F}(U)(\vec{\kappa}, \mathbf{x}) \ = \ \lambda\ \mathcal{F}(U)(\vec{\kappa}, \mathbf{x}),$$

proving that (λ, u_κ) is an eigenpair of problem (1.4.20). Similarly, we can argue for the TH case mode.
We will see in Chapter 2 that the spectra of problems (1.4.19) and (1.4.20) are discrete

for all values of $\vec{\kappa} \in \mathcal{K}$. To regain the spectrum of problem (1.4.5), it is necessary to take the union of all discrete spectra for all values of $\vec{\kappa}$ and for both problems (1.4.19) and (1.4.20).

1.4.4 Defects and trapped mode

At the beginning of this chapter we described how a light wave could be trapped in the defect of a PCF. So, the topic of this subsection is to explain how it is possible to compute numerically the frequencies (i.e. the eigenvalues) and the shape of the light wave (i.e. eigenfunctions) trapped in the defect.

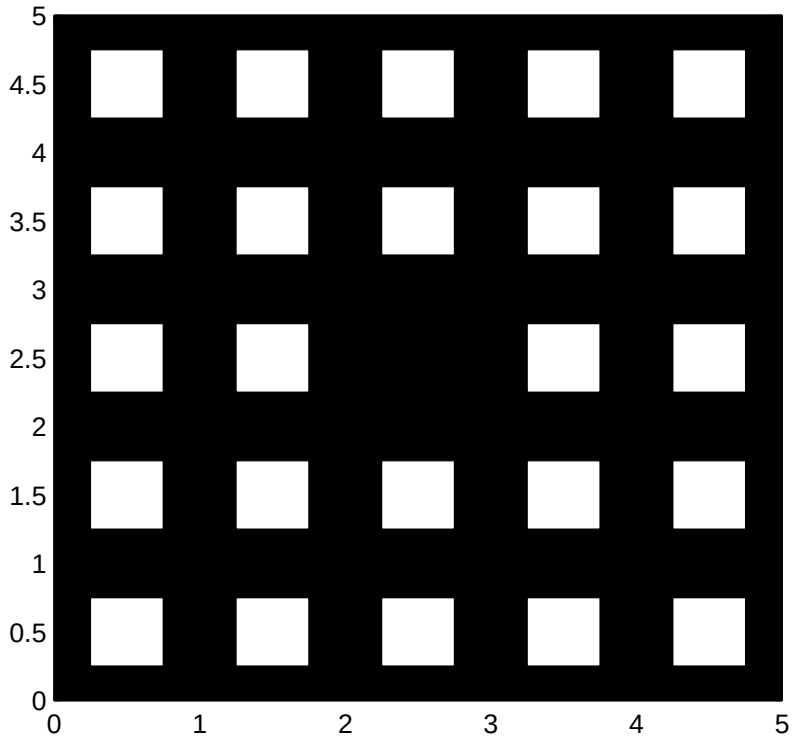

Figure 1-3: An example of supercell composed by five cells per side and with a missing inclusion in the center as a defect.

We already said that the spectrum of a periodic medium is formed by bands of essential spectrum. Then, creating a localized defect somewhere in the medium, we will not change the bands of the essential spectrum [24, Theorem 1], however it would be possible that eigenvalues may appear in the gaps between the bands [24, Theorem 2]. Since we have perturbed the periodic structure of the medium, it is not anymore so simple to apply the Floquet transform. In order to retake the possibility to use the same analysis, that we have used for the pure periodic medium case, we used the "supercell"

framework discussed in [49], where also the convergence proof of this framework is presented. In essence, this framework consists of considering a periodic medium with primitive cell containing the defect of the PCF surrounded by many layers of the periodic structure - in Figure 1-3 it is shown a supercell with two layers of square inclusions surrounding the center of the cell, these square inclusions form two layers of periodic structure. The defect, in this case, is the missing square inclusion in the center of the cell, which would be necessary to complete the symmetry of the cell. Because in the supercell framework there is a defect in each primitive cell, the resulting medium is not any more a compact perturbation of a periodic medium, so the defects create new bands in the spectrum. However, as proved in [49], enlarging the primitive supercell by increasing the number of layers of periodic structure, these new bands will shrink to eigenvalues and we will eventually recover the spectrum of the periodic medium with just one localized defect.

1.5 The structure of the thesis

This thesis is divided into five chapters. Each chapter, except the introduction, treats one main issue of our research. The material in each chapter is linked back to all previous chapters and the layout of the work is constructed in such a way that the reader moves from the abstract theory behind the problems to the numerical results in the last chapter.

In Chapter 2, we illustrate the theory behind elliptic eigenvalue problems and we show how to characterize the spectra of problems (1.3.7), (1.3.8) and (1.3.9). The main results in this chapter are the a priori upper bounds for the energy norm of the error for eigenfunctions and for the absolute value of the error for eigenvalues.

In Chapter 3, we introduce the a posteriori error estimator used to drive the mesh adaptivity. We will introduce an explicit error estimator based on residuals. The main results of this chapter are the proof of reliability and efficiency of our a posteriori error estimator for the PCF case.

In Chapter 4, we present the adaptive FEM for which we can prove convergence. This method embeds two marking schemes: the first one based on the a posteriori error estimator defined in Chapter 3, and the second one based on a different quantity called "oscillations", which is also defined in Chapter 4. We split Chapter 4 into two sections: one devoted to the general elliptic case and the other devoted to the PCF case.

In Chapter 5, we have collected a number of numerical results computed using our convergent adaptive scheme. In particular we present a number of results concerning problems arising from PCF applications such as band gaps and trapped defect modes.

Chapter 2

A priori analysis

In this chapter we characterize the spectra of the three problems analysed in this work, namely: the general elliptic eigenvalue problem (1.3.7), the model problem for PCFs (1.3.8) and its shifted version (1.3.9). We will show that all these problems have some characteristics in common. Moreover, the spectra of all these problems will be shown to be discrete and non-negative.
The analysis presented in this chapter, along with all the results, is not new. In fact it is possible to find similar material in many books. We suggest [6], [51] and [27]. In particular we like how the argument has been treated in [51]. In [51], only the class of regular and elliptic eigenvalue problems has been analysed. So here we have extended the analysis to more general problems.
The structure of this chapter follows. In Section 2.1 we prove the discreteness of all the spectra of the considered problems. We have collected the results of each problem in a different subsection. Then, in Section 2.2 we prove a priori convergence estimates for eigenvalues and eigenfunctions for each problem. Again, for sake of clarity, we assigned a different subsection to each problem.
Before starting with the analysis, we need to define self-adjoint operators. Let us denote by L^* the adjoint of the operator L, then:

Definition 2.0.1 (Self-adjoint operator)**.** *An operator L is self-adjoint if $L = L^*$, which implies:*

1. *L is Hermitian,*
2. *$D(L) = D(L^*)$,*

where $D(L)$ and $D(L^*)$ are the domains of the operator L and of its adjoint.

Self-adjoint operators have the nice characteristic that they have real spectra; moreover, this property holds for the bigger class of Hermitian operators, as proved in the next theorem.

Theorem 2.0.2. *The spectrum of a Hermitian operator L is real.*

Proof. Since the form $a(\cdot,\cdot)$, which is associated to the operator L, is sesquilinear we have:

$$a(u,v) \;=\; (\lambda u, v),$$

and

$$a(v,u) \;=\; (v, \lambda u),$$

where (λ, u) is an eigenpair of L. Choosing $v \equiv u$ we have

$$(\lambda u, u) \;=\; a(u,u) \;=\; (u, \lambda u),$$

which implies that λ is real, i.e. $\lambda \;=\; \bar{\lambda}$. □

2.1 Characterization of spectra

The purpose of this section is to characterize the spectra of problems (1.3.7), (1.3.8) and the spectrum of the shifted version of the PCF model problem, problem (1.3.9). We start by analysing in the first subsection the problem (1.3.7). Since all three problems are similar in many aspects, we shall modify the framework used for the problem (1.3.7), to analyse also the remaining two problems. This will be done in the following two subsections.

2.1.1 Generalized elliptic problem

We start showing at first that the sesquilinear forms $a(\cdot,\cdot)$ and $(\cdot,\cdot)_{0,\mathcal{B},\Omega}$ of (1.3.7) are continuous and that $a(\cdot,\cdot)$ is also coercive. We prove in Theorem 2.1.2 the coercivity of $a(\cdot,\cdot)$ using the equivalence between the energy norm and the standard norm of the Sobolev space $H_0^1(\Omega)$. The equivalence between the two norms is proven in the next lemma.

Lemma 2.1.1. *The energy norm related to problem (1.3.7) and the standard norm of $H_0^1(\Omega)$ are equivalent:*

$$C' \;\|u\|_{1,\Omega} \;\leq\; a(u,u)^{1/2} \;\leq\; C'' \;\|u\|_{1,\Omega}\;, \quad \textit{for all } u \in H_0^1(\Omega),$$

where the constants C' and C'' are independent of u.

Proof. Using (1.3.1) and the definition of $a(\cdot,\cdot)$, we conclude that

$$a(u,u)^{1/2} \;\leq\; \overline{a}^{1/2}\;|u|_{1,\Omega} \;\leq\; \overline{a}^{1/2}\;\|u\|_{1,\Omega}\;, \quad \text{for all } u \in H_0^1(\Omega). \tag{2.1.1}$$

In order to prove the lower bound for the energy norm, which would complete the proof, we apply the Poincaré inequality

$$\|u\|_{1,\Omega} \;\leq\; C_p\, |u|_{1,\Omega}\;, \quad \text{for all } u \in H^1_0(\Omega),$$

where C_p is a constant depending on the shape of the domain Ω. The application of the Poincaré inequality leads us to the sought lower bound,

$$\|u\|_{1,\Omega} \;\leq\; C_p\, |u|_{1,\Omega} \;\leq\; C_p\, \underline{a}^{-1/2}\, a(u,u)^{1/2}, \quad \text{for all } u \in H^1_0(\Omega). \tag{2.1.2}$$

The results (2.1.1) and (2.1.2) complete the proof. □

The coercivity of the sesquilinear form $a(\cdot,\cdot)$ comes, as a corollary of the Lemma 2.1.1.

Theorem 2.1.2. *The sesquilinear form $a(\cdot,\cdot)$ is coercive with coercivity constant $c_a > 0$, i.e.:*

$$a(u,u) \;\geq\; c_a\, \|u\|^2_{1,\Omega}\;, \quad \textit{for all } u \in H^1_0(\Omega). \tag{2.1.3}$$

Proof. The coercivity is proved just reformulating (2.1.2) as

$$a(u,u) \;\geq\; c_a\, \|u\|^2_{1,\Omega}\;, \quad \text{for all } u \in H^1_0(\Omega), \tag{2.1.4}$$

with constant $c_a \;=\; C_p^{-2}\underline{a}$, which is always greater than 0. □

Remark 2.1.3. *The coercivity of the bilinear form $a(\cdot,\cdot)$ implies that the spectrum is positive, because for any eigenpair (λ, u), with $\|u\|_{0,\mathcal{B},\Omega} = 1$, we have:*

$$0 \;<\; c_a\, \|u\|^2_{1,\mathcal{B},\Omega} \;\leq\; a(u,u) \;=\; \lambda(u,u)_{0,\mathcal{B},\Omega} \;=\; \lambda.$$

Another easy-to-prove property for both the sesquilinear forms $a(\cdot,\cdot)$ and $(\cdot,\cdot)_{0,\mathcal{B},\Omega}$ is continuity.

Theorem 2.1.4. *The sesquilinear form $a(\cdot,\cdot)$ is continuous in $H^1_0(\Omega)$ with continuity constant $C_a \;=\; \overline{a}$:*

$$a(u,v) \;\leq\; C_a\, \|u\|_{1,\Omega}\, \|v\|_{1,\Omega}, \quad \textit{for all } u,v \in H^1_0(\Omega). \tag{2.1.5}$$

Theorem 2.1.5. *The sesquilinear form $(\cdot,\cdot)_{0,\mathcal{B},\Omega}$ is continuous in $L^2_{\mathcal{B}}(\Omega)$, with continuity constant $C_b \;=\; 1$:*

$$(u,v)_{0,\mathcal{B},\Omega} \;\leq\; C_b\, \|u\|_{0,\mathcal{B},\Omega}\, \|v\|_{0,\mathcal{B},\Omega}, \quad \textit{for all } u,v \in L^2_{\mathcal{B}}(\Omega). \tag{2.1.6}$$

The first step in order to prove the discreteness of the spectrum of problem (1.3.7) consists in proving the existence and the uniqueness of the solution for the linear

problem

$$a(u,v) \;=\; (f,v)_{0,\mathcal{B},\Omega}\;, \quad \text{for all } v \in H_0^1(\Omega),$$

for any $f \in L^2_{\mathcal{B}}(\Omega)$. In order to do so we can use Lax-Milgram theorem (see for details [9]) which implies the uniqueness of the solution u. We know that the Lax-Milgram theorem holds in this case, since we have already proved continuity and coercivity for $a(\cdot,\cdot)$ and since the continuity for the linear functional $(f,\cdot)_{0,\mathcal{B},\Omega}$ is straightforward. By the Lax-Milgram theorem, there is a uniquely defined solution operator, $T : L^2_{\mathcal{B}}(\Omega) \longrightarrow H_0^1(\Omega)$ such that

$$\forall f \in L^2_{\mathcal{B}}(\Omega), \quad a(Tf,v) \;=\; (f,v)_{0,\mathcal{B},\Omega}, \quad \text{for all } v \in H_0^1(\Omega).$$

The second necessary step to prove the discreteness of the spectrum of (1.3.7) consists in applying the spectral theorem for compact operators (quoted below as Lemma 2.1.7) to the solution operator T. Let's define what a compact operator is first.

Definition 2.1.6 (Compact operator)**.** *An operator $L : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ on a Hilbert space $\mathcal{H}_1$ is compact if for any bounded sequence $\{v_m\} \in \mathcal{H}_1$ of functions, the resulting sequence $\{Lv_m\} \in \mathcal{H}_2$ has a converging subsequence.*

Lemma 2.1.7 (Spectral theorem for compact self-adjoint operators)**.** *The spectrum of a compact operator consists of eigenvalues of finite multiplicity with the only possible accumulation point at zero, and, possibly, the point zero (which may have infinite multiplicity). Furthermore, eigenfunctions corresponding to distinct eigenvalues are orthogonal to each other, and it is possible to construct an orthogonal basis of eigenfunctions (for details see [30]).*

Now, we are ready to prove in Theorem 2.1.9 the discreteness of the spectrum of (1.3.7).

Lemma 2.1.8. *The solution operator T is compact in $H_0^1(\Omega)$, i.e. $T : H_0^1(\Omega) \longrightarrow H_0^1(\Omega)$ is compact, and its spectrum is discrete.*

Proof. The fact that the solution operator T is bounded comes straightforwardly from the coercivity of $a(\cdot,\cdot)$ and the continuity of $(\cdot,\cdot)_{0,\mathcal{B},\Omega}$. In fact for all $f \in L^2_{\mathcal{B}}(\Omega)$ we have:

$$\|Tf\|^2_{1,\Omega} \;\le\; c_a^{-1}\, a(Tf,Tf) \;=\; c_a^{-1}(f,Tf)_{0,\mathcal{B},\Omega} \;\le\; c_a^{-1}C_b\, \|f\|_{0,\mathcal{B},\Omega}\, \|Tf\|_{0,\mathcal{B},\Omega},$$

which implies that T is a bounded operator for $L^2_{\mathcal{B}}(\Omega)$ to $H_0^1(\Omega)$, i.e.

$$\|Tf\|_{1,\Omega} \;\le\; c_a^{-1}C_b\, \|f\|_{0,\mathcal{B},\Omega}, \quad \text{for all } f \in L^2_{\mathcal{B}}(\Omega). \tag{2.1.7}$$

Then we have that $T : H_0^1(\Omega) \longrightarrow H_0^1(\Omega)$ is compact due to the compactness of embedding $H_0^1(\Omega) \subset L^2_{\mathcal{B}}(\Omega)$ (for the proof see e.g. [1, Theorem 6.3]).

To see that the spectrum of T is discrete we need to use the spectral theorem for compact operators (Lemma 2.1.7). □

Theorem 2.1.9. *The spectrum of problem (1.3.7) is discrete.*

Proof. In Lemma 2.1.8 we have proved that T has discrete spectrum. So, denoting by $(\mu, u) \in \mathbb{R} \times H_0^1(\Omega)$ an eigenpair of T, we have by the definition of the solution operator that

$$a(\mu u, v) \;=\; (u, v)_{0,\mathcal{B},\Omega}, \quad \text{for all } v \in H_0^1(\Omega). \tag{2.1.8}$$

Thanks to the linearity of $a(\cdot,\cdot)$ we have that (2.1.8) is equivalent to

$$a(u, v) \;=\; \mu^{-1}\,(u, v)_{0,\mathcal{B},\Omega}, \quad \text{for all } v \in H_0^1(\Omega), \tag{2.1.9}$$

which shows that for any eigenpair (μ, u) of T, with $\mu \neq 0$, corresponds an eigenpair (λ, u) of the problem (1.3.7), with $\lambda \;=\; \mu^{-1}$. This argument holds also in the other way round, since for any eigenpair (λ, u) of the problem (1.3.7), with $\lambda \neq 0$, we have that, by definition of the solution operator, (λ^{-1}, u) is an eigenpair of T.
In conclusion the spectrum of (1.3.7) is just a transformation of the spectrum of T, where the eigenfunctions remain unchanged and the eigenvalues are transformed as just shown. This prove the discreteness of the spectrum of (1.3.7). □

2.1.2 PCF model problem

In this subsection we are going to show, using the framework of Subsection 2.1.1, that the spectrum of the PCF model problem (1.3.8) is discrete. The analysis for this problem is more complicated because the problem may not be coercive. We show in the next lemma that the sesquilinear form $a_\kappa(\cdot,\cdot)$ is only non-negative definite, which does not imply coercivity.

Lemma 2.1.10. *The sesquilinear form $a_\kappa(\cdot,\cdot)$ of problem (1.3.8) is non-negative definite for any $\vec{\kappa} \in \mathcal{K}$.*

Proof. By direct calculation we have that, for any complex function $u \in H_\pi^1(\Omega)$, which we expand in its real and imaginary parts, i.e. $u = u_r \;+\; i\; u_i$:

$$\begin{aligned}(\nabla + i\vec{\kappa})u \;\cdot\; (\nabla - i\vec{\kappa})\bar{u} \;\;=\;\; \big[\underbrace{(\nabla u_r - \vec{\kappa} u_i)}_{a} \;+\; i\,\underbrace{(\nabla u_i + \vec{\kappa} u_r)}_{b}\big] \\ \cdot\; \big[\underbrace{(\nabla u_r - \vec{\kappa} u_i)}_{a} \;-\; i\,\underbrace{(\nabla u_i + \vec{\kappa} u_r)}_{b}\big],\end{aligned} \tag{2.1.10}$$

where a and b are, by construction, real vector-valued functions. Hence

$$(\nabla + i\vec{\alpha})u \;\cdot\; (\nabla - i\vec{\alpha})\bar{u} \;=\; [a + ib] \;\cdot\; [a - ib] \;=\; a^2 + b^2 \;\geq\; 0,$$

which implies the non-negativeness of the sesquilinear form $a_\kappa(\cdot,\cdot)$. □

Remark 2.1.11. *Because the sesquilinear form $a_\kappa(\cdot,\cdot)$ is Hermitian, we have from Theorem 2.0.2 that the spectrum of the problem is real. Moreover, Lemma 2.1.10 implies that the spectrum of (1.3.8), for any value of $\vec{\kappa} \in \mathcal{K}$, is non-negative:*

$$0 \leq a_\kappa(u,u) = \lambda(u,u)_{0,\mathcal{B},\Omega} = \lambda,$$

for any eigenpair (λ, u), with $\|u\|_{0,\mathcal{B},\Omega} = 1$.

To make problem (1.3.8) coercive we have to introduce a shift in the spectrum. This is the reason why we introduced problem (1.3.9), where S is any constant greater than 0. To simplify the notation we denote by

$$a_{\kappa,S}(u,v) \;:=\; a_\kappa(u,v) \;+\; S(u,v)_{0,\mathcal{B},\Omega}. \tag{2.1.11}$$

Note that trivially any eigenpair (ζ, u) of (1.3.9) corresponds to an eigenpair (λ, u) of (1.3.8), with $\lambda = \zeta - S$. Since the spectrum of (1.3.8) is real and non-negative, we have that the spectrum of (1.3.9) is real and positive, because $S > 0$.
In the next theorem we prove that for any value of $S > 0$, $a_{\kappa,S}(\cdot,\cdot)$ is coercive:

Theorem 2.1.12. *For any $S > 0$ and for any value of the quasimomentum $\vec{\kappa} \in \mathcal{K}$, the sesquilinear form $a_{\kappa,S}(\cdot,\cdot)$ is coercive with coercivity constant $c^{\mathrm{PCF}}_{a,S} \geq \min\{\underline{a}, S\underline{b}\}$, i.e.*

$$a_{\kappa,S}(u,u) \geq c^{\mathrm{PCF}}_{a,S}\|u\|^2_{1,\Omega}, \quad \textit{for all } u \in H^1_\pi(\Omega). \tag{2.1.12}$$

Proof. We want to prove that the sesquilinear form $a_{\kappa,S}(\cdot,\cdot)$ is coercive in the space $H^1_\pi(\Omega)$. Unfortunately, we do not have the Poincaré inequality, since constant functions lie in the space $H^1_\pi(\Omega)$. So, applying the definition of the sesquilinear form $a_{\kappa,S}(\cdot,\cdot)$, we have:

$$\begin{aligned} a_{\kappa,S}(u,u) \;=\; a_\kappa(u,u) \;+\; S\|u\|^2_{0,\mathcal{B},\Omega} \;&=\; \int_\Omega \mathcal{A}\nabla u\cdot\nabla\bar{u} \;-\; \mathcal{A}\nabla u\cdot i\vec{\kappa}\bar{u} \;+\; \mathcal{A}i\vec{\kappa}u\cdot\nabla\bar{u} \\ &+\; \int_\Omega (\mathcal{A}\vec{\kappa})\cdot\vec{\kappa}u\bar{u} \;+\; S\|u\|^2_{0,\mathcal{B},\Omega} \\ &=\; |u|^2_{1,\mathcal{A},\Omega} \;+\; 2i\left(\int_\Omega \mathrm{Im}\big(\mathcal{A}(\vec{\kappa}\cdot\nabla\bar{u})u\big)\right) \\ &+\; \int_\Omega (\mathcal{A}\vec{\kappa})\cdot\vec{\kappa}u\bar{u} \;+\; S\|u\|^2_{0,\mathcal{B},\Omega}. \end{aligned} \tag{2.1.13}$$

In Lemma 2.1.10 we have already proved that, for any $u \in H^1_\pi(\Omega)$, $a_\kappa(u,u)$ is real and non-negative. In view of this, we can conclude that the imaginary term in (2.1.13)

vanishes. Then, what remains from (2.1.13) is

$$a_{\kappa,S}(u,u) \;=\; |u|^2_{1,\mathcal{A},\Omega} \;+\; \int_\Omega (\mathcal{A}\vec{\kappa})\cdot\vec{\kappa}u\bar{u} \;+\; S\|u\|^2_{0,\mathcal{B},\Omega}\;. \tag{2.1.14}$$

Then, manipulating a bit more (2.1.14), we have

$$\begin{aligned} a_{\kappa,S}(u,u) \;&\geq\; \underline{a}|u|^2_{1,\Omega} \;+\; \underline{a}\int_\Omega |\vec{\kappa}|^2 u\bar{u} \;+\; S\underline{b}\|u\|^2_{0,\Omega} \\ &=\; \underline{a}|u|^2_{1,\Omega} \;+\; \underline{a}|\vec{\kappa}|^2\|u\|^2_{0,\Omega} \;+\; S\underline{b}\|u\|^2_{0,\Omega} \\ &\geq\; \underline{a}|u|^2_{1,\Omega} \;+\; S\underline{b}\|u\|^2_{0,\Omega}\;, \end{aligned}$$

which implies that $a_{\kappa,S}(u,u) \geq c^{\mathrm{PCF}}_{a,S}\|u\|^2_{1,\Omega}$, with $c^{\mathrm{PCF}}_{a,S} \geq \min\{\underline{a}, S\underline{b}\}$.

□

In order to show that the spectrum of (1.3.8) is discrete, it is enough to prove that the spectrum of (1.3.9) is discrete, because the spectrum of (1.3.9) is a shifted version of the spectrum of problem (1.3.8). Then, to prove that the spectrum of (1.3.9) is discrete, we are going to argue similarly as in Subsection 2.1.1. The first step is to prove that the sesquilinear form of (1.3.9) is continuous.

Theorem 2.1.13. *For any value of the quasimomentum $\vec{\kappa} \in \mathcal{K}$, the sesquilinear form $a_{\kappa,S}(\cdot,\cdot)$ is continuous with continuity constant $C^{\mathrm{PCF}}_{a,S}$, which depends on $\overline{b}$, $\overline{a}$, S and on the diameter of $\mathcal{K}$:*

$$a_{\kappa,S}(u,v) \leq C^{\mathrm{PCF}}_{a,S}\|u\|_{1,\Omega}\|v\|_{1,\Omega}, \quad \textit{for all } u,v \in H^1_\pi(\Omega). \tag{2.1.15}$$

Proof. The proof is straightforward, it is just necessary to use the Cauchy-Swharz inequality:

$$\begin{aligned} a_{\kappa,S}(u,v) \;&\leq\; \overline{a}\,\big(\,|u|_{1,\Omega}\,|v|_{1,\Omega} \;+\; |\vec{\kappa}|\;\cdot|u|_{1,\Omega}\,\|v\|_{0,\Omega} \\ &\quad + |\vec{\kappa}|\;\cdot|v|_{1,\Omega}\,\|u\|_{0,\Omega} \;+\; (|\vec{\kappa}|^2 \;+ S\;\overline{b}\;\overline{a}^{-1})\;\|u\|_{0,\Omega}\,\|v\|_{0,\Omega}\,\big) \\ &\leq\; \overline{a}\;\max_{\vec{\kappa}\in\mathcal{K}}\{1,|\vec{\kappa}|,|\vec{\kappa}|^2 + S\;\overline{b}\;\overline{a}^{-1}\}\;(\|u\|_{0,\Omega} + |u|_{1,\Omega})\;(\|v\|_{0,\Omega} + |v|_{1,\Omega}) \\ &\leq\; C^{\mathrm{PCF}}_{a,S}\|u\|_{1,\Omega}\;\|v\|_{1,\Omega}\;, \end{aligned}$$

with $C^{\mathrm{PCF}}_{a,S} := 2\overline{a}\;\max_{\vec{\kappa}\in\mathcal{K}}\{1,|\vec{\kappa}|,|\vec{\kappa}|^2 + S\;\overline{b}\;\overline{a}^{-1}\}$. □

Corollary 2.1.14. *For any value of the quasimomentum $\vec{\kappa} \in \mathcal{K}$, the sesquilinear form $a_\kappa(\cdot,\cdot)$ is continuous with continuity constant C^{PCF}_a, which depends on $\overline{a}$ and on the*

diameter of $\mathcal{K}$:

$$a_\kappa(u,v) \leq C_a^{\mathrm{PCF}} \|u\|_{1,\Omega} \|v\|_{1,\Omega}, \quad \textit{for all } u,v \in H^1_\pi(\Omega). \tag{2.1.16}$$

The next step is to prove that the solution operator T^{PCF} of problem (1.3.9) is compact in $H^1_\pi(\Omega)$. We can define a solution operator $T^{\mathrm{PCF}} : L^2_{\mathcal{B}}(\Omega) \longrightarrow H^1_\pi(\Omega)$ as:

$$\forall f \in L^2_{\mathcal{B}}(\Omega), \quad a_{\kappa,S}(Tf,v) \;=\; (f,v)_{0,\mathcal{B},\Omega}, \quad \text{for all } v \in H^1_\pi(\Omega).$$

Lemma 2.1.15. *The solution operator T^{PCF} is compact and its spectrum is discrete.*

Proof. The proof is analogous to the proof of Lemma 2.1.8, since $a_{\kappa,S}(\cdot,\cdot)$ is coercive from Theorem 2.1.12 and the imbedding $H^1_\pi(\Omega) \subset L^2_{\mathcal{B}}(\Omega)$ is compact. □

Theorem 2.1.16. *The spectrum of (1.3.9) is discrete for any $\vec{\kappa} \in \mathcal{K}$.*

Proof. The spectrum of problem (1.3.9) is a transformation of the spectrum of T^{PCF}. For the details see the proof of Theorem 2.1.9, since the transformation is the same. □

We would like to conclude this section remarking that, because the spectrum of (1.3.9) is a shifted version of the spectrum of (1.3.8), Theorem 2.1.16 also implies that the spectrum of (1.3.8) is discrete.

2.2 Convergence estimates

In this section we prove a priori convergence estimates for finite element approximation of both eigenvalues and eigenfunctions. We shall also introduce the FEM that we are going to use. We start with problem (1.3.7), then we will adapt the theory to cope with the PCF model problem in the following sections.

2.2.1 Finite element approximation for general elliptic eigenvalue problems

Now we introduce the definition of the discrete version of problem (1.3.7). Accordingly, let $\mathcal{T}_n$, $n = 1, 2, \ldots$ denote a family of conforming triangular ($d = 2$) or tetrahedral ($d = 3$) meshes on Ω. Each mesh consists of elements denoted by $\tau \in \mathcal{T}_n$. We assume that for each n, $\mathcal{T}_{n+1}$ is a refinement of $\mathcal{T}_n$. For a typical element τ of any mesh $\mathcal{T}_n$, its diameter is denoted H_τ and the diameter of its largest inscribed ball is denoted ρ_τ. Moreover all the meshes are to be considered conforming (the definition can be found for example in [9]) and we use only shape regular meshes, i.e. there exists a constant C_{reg} independent of n such that

$$H_\tau \leq C_{\mathrm{reg}}\ \rho_\tau, \quad \text{for all} \quad \tau \in \mathcal{T}_n. \tag{2.2.1}$$

We denote with $\mathcal{F}_n$ the set of all the edges (faces) of the elements of the mesh $\mathcal{T}_n$, and we assume to have already chosen an ordering and a preorientated unit normal vector $\vec{n}_f$ for each $f \in \mathcal{F}_n$. Furthermore, we denote by $\tau_1(f)$ and $\tau_2(f)$ the elements sharing $f \in \mathcal{F}_n$. Finally we define

$$H_n^{\max} := \max_{\tau \in \mathcal{T}_n} \{H_\tau\}.$$

We assume that the meshes $\mathcal{T}_n$ form a sequence $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$, on which the quantity $H_n^{\max}$ goes to 0 when n goes to infinity.

Our problems may have discontinuous coefficients, but we assume that in the interior of each element τ of any mesh $\mathcal{T}_n$ the values of $\mathcal{A}$ and $\mathcal{B}$ are constants. To enforce this requirement we only consider problems with piecewise constant coefficients where discontinuities are resolved on the coarsest mesh.

On any mesh $\mathcal{T}_n$ we denote by $V_n \subset C^0(\Omega)$ the finite dimensional space, of dimension N, of linear polynomials on each element τ of the mesh.

The discrete formulation of problem (1.3.7) is:

seek eigenpairs of the form $(\lambda_n, u_n) \in \mathbb{R} \times V_n$, *with* $\|u_n\|_{0,\mathcal{B},\Omega} = 1$ *such that*

$$a(u_n, v_n) = \lambda_n (u_n, v_n)_{0,\mathcal{B},\Omega} \ , \quad \text{for all } v_n \in V_n. \tag{2.2.2}$$

For any n, the spectrum of problem (2.2.2) is discrete due to the fact that the space V_n is finite dimensional.

In order to carry out the analysis in the rest of the section, we assume that the eigenfunctions of the problem (1.3.7) are contained in the Sobolev space $H^{1+s}(\Omega)$ for some $s > 0$. We make the following regularity assumption for the elliptic problem (1.3.7):

Assumption 2.2.1. *We assume that there exists a constant* $C_{\text{ell}} > 0$ *and* $s \in [0, 1]$ *with the following property. For* $f \in L^2(\Omega)$, *if* $v \in H_0^1(\Omega)$ *solves the problem* $a(v, w) = (f, w)_{0,\Omega}$, *for all* $w \in H_0^1(\Omega)$, *then*

$$\|v\|_{1+s,\Omega} \leq C_{\text{ell}} \|f\|_{0,\Omega} \ . \tag{2.2.3}$$

Assumption 2.2.1 is satisfied with $s = 1$ when $\mathcal{A}$ is constant (or smooth) and Ω is convex. In a range of other practical cases $s \in (0, 1)$, for example Ω non-convex (see [39]), or $\mathcal{A}$ having a discontinuity across an interior interface (see [5]).

Assumption 2.2.1 is stated for the linear problem $a(v, w) = (f, w)_{0,\Omega}$, so in order to apply Assumption 2.2.1 to the eigenvalue problem (1.3.7), i.e. $a(u_j, v) = \lambda_j (u_j, v)_{0,\mathcal{B},\Omega}$, we need to substitute the data f with the eigenpair (λ_j, u_j), where $\|u_j\|_{0,\mathcal{B},\Omega} = 1$, and also it is necessary to take in account the fact that the inner product of (1.3.7) is weighted by $\mathcal{B}$, so (2.2.3) becomes:

$$\|u_j\|_{1+s,\Omega} \ \leq \ C_{\text{ell}} \lambda_j \bar{b} \ .$$

The next preliminary result comes as a standard result from approximation theory:

Lemma 2.2.2. *For any function $u \in H^{1+s}(\Omega) \cap H_0^1(\Omega)$ we have that*

$$\inf_{v_n \in V_n} \|u - v_n\|_{1,\Omega} \ \leq \ C_{\text{app}} (H_n^{\max})^s |u|_{1+s,\Omega},$$

Proof. For a proof see e.g. [48]. □

A consequence of Lemma 2.2.2 is that the space V_n becomes dense in $H^{1+s}(\Omega) \cap H_0^1(\Omega)$, when n goes to infinity due to the assumptions on the sequence $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$, i.e.

$$H^{1+s}(\Omega) \cap H_0^1(\Omega) = \overline{\lim_{n\to\infty} V_n}. \tag{2.2.4}$$

The next theorem comes from [6] and it is fundamental for the a priori analysis of elliptic eigenvalue problems.

Theorem 2.2.3. *The sequence $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$ converges in norm to the solution operator $\mathcal{T}$ when n goes to infinity. This implies that also the spectrum of problem* (2.2.2) *converges to the spectrum of the continuous problem* (1.3.7) *when n goes to infinity.*

Remark 2.2.4. *From Theorem 2.2.3 we have that for each eigenvalue λ_j of multiplicity $R+1$, it is possible to construct $R+1$ sequences of computed eigenpairs $(\lambda_{l+r,n}, u_{l+r,n})$, with $r = 0, \ldots, R$, such that $\lambda_{l+r,n}$ converges to λ_j when n goes to infinity, for all $r = 0, \ldots, R$. Moreover, for any n all the eigenfunctions $u_{l,n}, \ldots, u_{l+r,n}$ are orthogonal to each other.*

2.2.2 Convergence estimates for the general elliptic eigenvalue case

In Section 2.1.1, we have already proved that the spectrum of the problem (1.3.7) is positive and discrete. But we have not yet defined a way to actually determine the eigenvalues of such problem. Now, it is time turn our attention to this particular aspect. In Definition 2.2.7 the Rayleigh quotient is introduced and the following theorem uses this functional to characterize the eigenvalues of problem (1.3.7).

Notation 2.2.5. *In this subsection, we write $A \lesssim B$ with $A,\ B \in \mathbb{R}$ when A/B is bounded by a constant which may depend on the functions $\mathcal{A}$ and $\mathcal{B}$, on c_a in (2.1.3), on C_a in (2.1.5), on C_b in (2.1.6), on C_{reg} in (2.2.1), on C_{ell} in Assumption 2.2.1 or on C_{app} in Lemma 2.2.2, but not on n. The notation $A \cong B$ means $A \lesssim B$ and $A \gtrsim B$.*

Since we know that the spectrum of (1.3.7) is positive and discrete, we are able to sort the eigenvalues in increasing order:

$$0 < \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ldots$$

Let E_j be the eigenspace of problem (1.3.7) corresponding to λ_j.

Definition 2.2.6. *For the first j eigenvalues, we define the space*

$$\mathcal{E}^j = \operatorname{span}\{E_i : i = 1, \dots, j\} \ .$$

Moreover, we also define the space

$$\mathcal{E}_1^j = \{u \in \mathcal{E}^j : \|u\|_{0,\mathcal{B},\Omega} = 1\} \ .$$

Definition 2.2.7 (Rayleigh quotient for general elliptic eigenvalue problems)**.** *We define the Rayleigh quotient as*

$$\mathcal{R}(v) \ = \ \frac{a(v,v)}{(v,v)_{0,\mathcal{B},\Omega}},$$

where $v \in H_0^1(\Omega)$.

Theorem 2.2.8. *Any eigenvalue λ_j, with $j \geq 1$, of problem* (1.3.7) *can be characterized in the following way using the Rayleigh quotient (with $v \neq 0$):*

$$\lambda_j = \min_{\substack{v \in H_0^1(\Omega) \\ \|v\|_{0,\mathcal{B},\Omega}=1 \\ v \perp \mathcal{E}_1^{j-1}}} \mathcal{R}(v) \ ,$$

where $\mathcal{E}_1^0$ is to be interpreted as the empty set (see [51, Chapter 6, page 220] for the proof).

An equivalent way to characterize these eigenvalues is using the minimum-maximum principle explained in [51, Chapter 6, page 221]. If $\mathcal{R}(v)$ is maximized over an j-dimensional subspace $V_j \subset H_0^1(\Omega)$, then we have:

$$\lambda_j = \min_{V_j \subset H_0^1(\Omega)} \max_{\substack{v \in V_j \\ \|v\|_{0,\mathcal{B},\Omega}=1}} \mathcal{R}(v) \ , \tag{2.2.5}$$

where the minimum is taken over all j-dimensional subspaces of $H_0^1(\Omega)$.
The characterization of the spectrum of (2.2.2) follows. Let $E_{j,n}$ denote the discrete eigenspace of problem (2.2.2) corresponding to the eigenvalue λ_j in view of Remark 2.2.4 and let also

$$\mathcal{E}_{1,n}^{j-1} \ = \ \left\{v \in \operatorname{span}\{E_{1,n}, \dots, E_{j-1,n}\} : \|v\|_{0,\mathcal{B},\Omega} = 1\right\} \ ,$$

where $\mathcal{E}_{1,n}^0$ is to be interpreted as the empty set.

Theorem 2.2.9. *Any eigenvalue $\lambda_{j,n}$, with $j \leq N = \dim V_n$, of problem* (2.2.2) *can*

be characterized in the following way using the Rayleigh quotient (with $v \neq 0$):

$$\lambda_{j,n} = \min_{\substack{v \in V_n \\ \|v\|_{0,\mathcal{B},\Omega}=1 \\ v \perp \mathcal{E}_{1,n}^{j-1}}} \mathcal{R}(v),$$

(see [6, page 699].)

Also for the discrete problem there is an equivalent way to characterize the spectrum based on the minimum-maximum principle, which is explained in [51, Chapter 6, page 223]. This time the minimum is over all j dimensional subspaces $V_{j,n}$ contained in V_n:

$$\lambda_{j,n} = \min_{V_{j,n} \subset V_n} \max_{\substack{v \in V_{j,n} \\ \|v\|_{0,\mathcal{B},\Omega}=1}} \mathcal{R}(v). \tag{2.2.6}$$

Since V_n, for all n, is contained in $H_0^1(\Omega)$ by construction, we have that, for the same value j, the minimum (2.2.5) is always smaller than the minimum (2.2.6). So it follows directly that $\lambda_j \leq \lambda_{j,n}$ for problem (1.3.7).
In the rest of this section we will primarily consider an eigenvalue λ_l of problem (1.3.7) with multiplicity $R+1$, where $R \geq 0$. So, from the positiveness of the spectrum of (1.3.7) we have:

$$0 < \lambda_l = \lambda_{l+1} = \cdots = \lambda_{l+R}.$$

The remainder of this section is devoted to the proof of convergence of approximate eigenvalues and eigenfunctions of problem (1.3.7). The main results are in Theorem 2.2.10, where we also illustrate how the convergence depends on $H_n^{\max}$. The treatment below is an extension of the theory in [51], however, we covered the multiple eigenvalue case, too.

Theorem 2.2.10. *Let s be as given in Assumption 2.2.1 and suppose that $H_n^{\max}$ is small enough. Then considering the eigenvalue λ_l, we have that the following statements hold:*

(i) In view of Remark 2.2.4, let λ_l be an eigenvalue of (1.3.7) and let $(\lambda_{l,n}, u_{l,n})$ be a computed eigenpair of problem (2.2.2), with $\lambda_{l,n}$ converging to λ_l when n goes to infinity, then

$$0 \leq \lambda_{l,n} - \lambda_l \lesssim (H_n^{\max})^{2s} \ . \tag{2.2.7}$$

(ii) Let λ_l be an eigenvalue of problem (1.3.7) with multiplicity $R+1$, with $R \geq 0$ and let u_l be any eigenfunction of λ_l with $\|u_l\|_{0,\mathcal{B},\Omega} = 1$, then there exists a sequence $\{w_{l,n}\}_{n \in \mathbb{N}}$ with $w_{l,n} \in E_{l,n}$ for all n and with $\|w_{l,n}\|_{0,\mathcal{B},\Omega} = 1$ such that

$$\|u_l - w_{l,n}\|_{0,\mathcal{B},\Omega} \ \lesssim \ C_{\text{spec1}} (H_n^{\max})^{2s} \ , \tag{2.2.8}$$

$$a(u_l - w_{l,n}, u_l - w_{l,n})^{1/2} \lesssim C_{\text{spec2}}(H_n^{\max})^s \ . \tag{2.2.9}$$

Where the constants C_{spec1} and C_{spec2} depends on the spectral information λ_j, u_j, $j = 1, \ldots, l + R$.

The proof of this theorem above is postponed to the end of the section. Let us start with a lemma that should clarify our strategy to prove Theorem 2.2.10:

Lemma 2.2.11. *Let (λ_l, u_l) be a true eigenpair of problem (1.3.7) with $\|u_l\|_{0,\mathcal{B},\Omega} = 1$ and let $(\lambda_{j,n}, u_{j,n})$ be a computed eigenpair of problem (2.2.2) with $\|u_{j,n}\|_{0,\mathcal{B},\Omega} = 1$. Then we have:*

$$a(u_l - u_{j,n}, u_l - u_{j,n}) \ = \ \lambda_l \|u_l - u_{j,n}\|^2_{0,\mathcal{B},\Omega} \ + \ \lambda_{j,n} - \lambda_l.$$

Proof. Using the linearity of the bilinear form $a(\cdot, \cdot)$ and using (1.3.7), (2.2.2); we have

$$a(u_l - u_{j,n}, u_l - u_{j,n}) \ = \ \lambda_l \ + \ \lambda_{j,n} \ - \ 2\lambda_l(u_l, u_{j,n})_{0,\mathcal{B},\Omega}. \tag{2.2.10}$$

Furthermore, by analogous arguments we obtain

$$\|u_l - u_{j,n}\|^2_{0,\mathcal{B},\Omega} \ = \ 2 \ - \ 2(u_l, u_{j,n})_{0,\mathcal{B},\Omega}. \tag{2.2.11}$$

Substituting (2.2.11) into (2.2.10) we obtain the sought result. □

Corollary 2.2.12. *Let (λ_l, u_l) be a true eigenpair of problem (1.3.7) with $\|u_l\|_{0,\mathcal{B},\Omega} = 1$ and let $(\lambda_{j,n}, u_{j,n})$ be a computed eigenpair of problem (2.2.2) with $\|u_{j,n}\|_{0,\mathcal{B},\Omega} = 1$. Then we have:*

$$\lambda_{j,n} - \lambda_l \ \leq \ a(u_l - u_{j,n}, u_l - u_{j,n}) \ .$$

Proof. The proof is straightforward from Lemma 2.2.11 since the quantity $\lambda_l \|u_l - u_{j,n}\|^2_{0,\mathcal{B},\Omega}$ is always greater than 0. □

In the proof of Theorem 2.2.10 below we first prove (2.2.7), and then (2.2.8). Afterward, thanks to Lemma 2.2.11, (2.2.9) follows easily.
Now we start to prove (2.2.7). In the next definition we introduce the projection operator Q_n which for a given $u \in H_0^1(\Omega)$, it returns the best approximation in the energy norm of u in the finite space V_n.

Definition 2.2.13 (Rayleigh-Ritz projection operator for general elliptic problems)**.** *We define the projection operator $Q_n : H_0^1(\Omega) \longrightarrow V_n$ as the operator that, for any given function $u \in H_0^1(\Omega)$, it returns the function $Q_n u \in V_n$ which satisfies:*

$$a(u, v_n) \ = \ a(Q_n u, v_n), \quad \textit{for all } v_n \in V_n.$$

From the definition of Q_n it is straightforward to see the orthogonality of the projection, i.e.

$$a(u - Q_n u, v_n) \;=\; 0, \quad \textit{for all } v_n \in V_n.$$

In other words, if u is the solution to the problem $-\triangle u = f$, $Q_n u$ would be exactly its Ritz approximation u_n. This guarantees that:

$$\|u - Q_n u\|_{1,\Omega} \lesssim (H_n^{\max})^s \|u\|_{1+s,\Omega} \;, \tag{2.2.12}$$

which comes from Lemma 2.2.2 and Cèa's lemma. See [27, Theorem 8.4.14].
In the next lemma, we prove an upper bound for the computed eigenvalues using the true ones. This result, together with the fact that computed eigenvalues are always greater than the true ones, thanks to the minimum-maximum principle, is the pivot to prove (2.2.7).

Lemma 2.2.14. *Let us define the quantity σ_{l+R}^n as*

$$\sigma_{l+R}^n := \max_{u \in \mathcal{E}_1^{l+R}} \Big| 2(u, u - Q_n u)_{0,\mathcal{B},\Omega} \;-\; (u - Q_n u, u - Q_n u)_{0,\mathcal{B},\Omega} \Big|. \tag{2.2.13}$$

Provided that $H_n^{\max}$ is small enough such that $\sigma_{l+R}^n < 1$, then the computed eigenvalue $\lambda_{l,n}$, with $l \leq N$ where $N = \dim V_n$, is bounded above and below by:

$$\lambda_l \;\leq\; \lambda_{l,n} \;\leq\; \frac{\lambda_l}{1 - \sigma_{l+R}^n}. \tag{2.2.14}$$

Remark 2.2.15. *The quantity σ_{l+R}^n has a geometrical interpretation:*

$$\begin{aligned} 2(u, u - Q_n u)_{0,\mathcal{B},\Omega} \;-\; (u - Q_n u, u - Q_n u)_{0,\mathcal{B},\Omega} \;&=\; (u + Q_n u, u - Q_n u)_{0,\mathcal{B},\Omega} \\ &=\; (u, u)_{0,\mathcal{B},\Omega} \;-\; (Q_n u, Q_n u)_{0,\mathcal{B},\Omega} \;. \end{aligned}$$

As can be seen, the quantity σ_{l+R}^n is related to the difference between the norm of true eigenfunction and the norm of the projection of the eigenfunction on the finite element space.

Proof. Since $\|u - Q_n u\|_{0,\mathcal{B},\Omega} \to 0$ as $H_n^{\max} \to 0$, so $\sigma_{l+R}^n < 1$ when $H_n^{\max}$ is small enough.
Now, we can turn our attention to (2.2.14). From the minimum-maximum principle (2.2.6), we have for the space $\mathcal{E}_1^{l+R}$, which is defined in Definition 2.2.6, that

$$\lambda_{l,n} \;\leq\; \max_{v_n \in Q_n \mathcal{E}_1^{l+R}} \mathcal{R}(v_n) = \max_{u \in \mathcal{E}_1^{l+R}} \frac{a(Q_n u, Q_n u)}{(Q_n u, Q_n u)_{0,\mathcal{B},\Omega}}, \tag{2.2.15}$$

where $v_n = Q_n u$. The numerator of (2.2.15) is bounded from above by:

$$a(Q_n u, Q_n u) \;\leq\; a(u,u), \tag{2.2.16}$$

since Q_n by definition is a projection in the energy norm. Furthermore for any $u \in \mathcal{E}_1^{l+R}$, the denominator of (2.2.15) is bounded from below by

$$\begin{aligned}
(Q_n u, Q_n u)_{0,\mathcal{B},\Omega} &= -(u - Q_n u, Q_n u)_{0,\mathcal{B},\Omega} \;+\; (u, Q_n u)_{0,\mathcal{B},\Omega} \\
&= (u - Q_n u, u - Q_n u)_{0,\mathcal{B},\Omega} \;-\; (u - Q_n u, u)_{0,\mathcal{B},\Omega} \;+\; (u, Q_n u)_{0,\mathcal{B},\Omega} \\
&= (u,u)_{0,\mathcal{B},\Omega} \;-\; 2(u, u - Q_n u)_{0,\mathcal{B},\Omega} \\
&+ (u - Q_n u, u - Q_n u)_{0,\mathcal{B},\Omega} \;\geq\; 1 - \sigma_{l+R}^n \;.
\end{aligned} \tag{2.2.17}$$

To conclude the proof, we substitute (2.2.16) and (2.2.17) into (2.2.15):

$$\lambda_{l,n} \;\leq\; \max_{u \in \mathcal{E}_1^{l+R}} \frac{a(u,u)}{1 - \sigma_{l+R}^n} \;=\; \frac{\lambda_l}{1 - \sigma_{l+R}^n}.$$

□

The last result that we need in order to prove Theorem 2.2.10(i) is the next lemma.

Lemma 2.2.16. *Let u be a function in $\mathcal{E}_1^{l+R}$, then the following equality holds*

$$(u, u - Q_n u)_{0,\mathcal{B},\Omega} \;=\; \sum_{i=1}^{l+R} c_i \lambda_i^{-1} a(u_i - Q_n u_i, u - Q_n u). \tag{2.2.18}$$

Proof. By definition $u = \sum_1^{l+R} c_i \; u_i$, where u_i are eigenfunctions of (1.3.7) and c_i are real values. Applying the decomposition for u yields:

$$(u, u - Q_n u)_{0,\mathcal{B},\Omega} \;=\; \sum_{i=1}^{l+R} c_i \; (u_i, u - Q_n u)_{0,\mathcal{B},\Omega}. \tag{2.2.19}$$

Since all u_i are true eigenfunctions with corresponding eigenvalue λ_i, we have:

$$(u_i, u - Q_n u)_{0,\mathcal{B},\Omega} \;=\; \lambda_i^{-1} a(u_i, u - Q_n u). \tag{2.2.20}$$

Furthermore, from the orthogonality of the projection operator Q_n we have:

$$a(Q_n u_i, u - Q_n u) \;=\; 0 \;. \tag{2.2.21}$$

Now, subtracting (2.2.21) from (2.2.20) we have

$$(u_i, u - Q_n u)_{0,\mathcal{B},\Omega} \;=\; \lambda_i^{-1} a(u_i - Q_n u_i, u - Q_n u) \quad \text{for } i = 1, \ldots, l+R\,. \tag{2.2.22}$$

To complete the proof, we substitute (2.2.22) into (2.2.19). □

Now we return to the proof of part (i) of Theorem 2.2.10.

Proof of Theorem 2.2.10(i). From (2.2.14) we have that if $H_n^{\max}$ is small enough so that $\sigma_{l+R}^n \leq 1/2$, then:

$$\lambda_{l,n} \;\leq\; \frac{\lambda_l}{1 - \sigma_{l+R}^n} \;\leq\; \lambda_l\,(1 + 2\sigma_{l+R}^n). \tag{2.2.23}$$

So, the only missing piece, in order to prove (2.2.7), is an estimate for σ_{l+R}^n in terms of $H_n^{\max}$. We are going to estimate the two terms in σ_{l+R}^n separately. The first term can be estimated using Lemma 2.2.16 for any function $u = \sum_1^{l+R} c_i u_i$ in $\mathcal{E}_1^{l+R}$ and also using (2.1.5):

$$\begin{aligned} 2\big|(u, u - Q_n u)_{0,\mathcal{B},\Omega}\big| \;&=\; 2\Big|\sum_{i=1}^{l+R} c_i \lambda_i^{-1}\; a(u_i - Q_n u_i, u - Q_n u)\Big| \\ &\lesssim\; \Big\|(I - Q_n)\sum_{i=1}^{l+R} c_i \lambda_i^{-1} u_i\Big\|_{1,\Omega}\; \big\|(I - Q_n)u\big\|_{1,\Omega}. \end{aligned}$$

Then, applying (2.2.12), we obtain:

$$2|(u, u - Q_n u)_{0,\mathcal{B},\Omega}| \;\lesssim\; (H_n^{\max})^{2s}\; \Big\|\sum_{i=1}^{l+R} c_i \lambda_i^{-1} u_i\Big\|_{1+s,\Omega}\; \big\|u\big\|_{1+s,\Omega}. \tag{2.2.24}$$

To treat the second term of σ_{l+R}^n, we can use the usual Aubin-Nitsche duality argument. Let us denote $e_n := u - Q_n u$ and let us define φ to be the solution of the linear problem

$$a(v, \varphi) \;=\; (v, e_n)_{0,\mathcal{B},\Omega} \quad \text{for all} \quad v \in H_0^1(\Omega). \tag{2.2.25}$$

We have from the orthogonality of Q_n, i.e. $a(e_n, v_n) = 0$ for all $v_n \in V_n$, that:

$$\|e_n\|_{0,\mathcal{B},\Omega}^2 = a(e_n, \varphi) = a(e_n, \varphi - v_n) \quad \text{for all} \quad v_n \in V_n\,.$$

Then applying Cauchy-Schwarz we obtain

$$\|e_n\|_{0,\mathcal{B},\Omega}^2 \lesssim\; |\varphi - v_n|_{1,\Omega}\; |e_n|_{1,\Omega}, \quad \text{for all} \quad v_n \in V_n. \tag{2.2.26}$$

Using Lemma 2.2.2 (together with Assumption 2.2.1) in (2.2.26) we get

$$\begin{aligned} \|e_n\|^2_{0,\mathcal{B},\Omega} &\lesssim (H_n^{\max})^s |\varphi|_{1+s,\Omega} |e_n|_{1,\Omega} \\ &\lesssim (H_n^{\max})^s \|e_n\|_{0,\mathcal{B},\Omega} |e_n|_{1,\Omega}. \end{aligned} \tag{2.2.27}$$

The last step of the argument consists of dividing both sides of (2.2.27) by $\|e_n\|_{0,\mathcal{B},\Omega}$ and applying the regularity result (2.2.12)

$$\|e_n\|_{0,\mathcal{B},\Omega} \lesssim (H_n^{\max})^{2s} |u|_{1+s,\Omega}. \tag{2.2.28}$$

So, applying (2.2.28) to the second term of σ^n_{l+R} we obtain:

$$(u - Q_n u, u - Q_n u)_{0,\mathcal{B},\Omega} \lesssim (H_n^{\max})^{4s} |u|^2_{1+s,\Omega}. \tag{2.2.29}$$

Now, substituting (2.2.24) and (2.2.29) into (2.2.23), we have:

$$\begin{aligned} \lambda_{l,n} \lesssim \lambda_l + 2\lambda_l \Bigg(& (H_n^{\max})^{2s} \max_{\substack{c_1,\ldots,c_{l+R} \\ \sum |c_i|^2 = 1}} \Big\| \sum_{i=1}^{l+R} c_i \lambda_i^{-1} u_i \Big\|_{1+s,\Omega} \max_{u \in \mathcal{E}_1^{l+R}} \|u\|_{1+s,\Omega} \\ & + (H_n^{\max})^{4s} \max_{u \in \mathcal{E}_1^{l+R}} |u|^2_{1+s,\Omega} \Bigg). \end{aligned}$$

Yields:

$$\lambda_{l,n} \lesssim \lambda_l + \lambda_l (H_n^{\max})^{2s}.$$

□

In order to prove (2.2.8), we use the following argument:

$$\|u_j - w_{j,n}\|_{0,\mathcal{B},\Omega} \leq \|u_j - \beta_j w_{j,n}\|_{0,\mathcal{B},\Omega} + \|(\beta_j - 1) w_{j,n}\|_{0,\mathcal{B},\Omega}, \tag{2.2.30}$$

for any scalar β_j and where $w_{j,n} \in E_{j,n}$. Then we make the choice $\beta_j = (Q_n u_j, w_{j,n})_{0,\mathcal{B},\Omega}$. The proof of (2.2.8) consists of proving the convergence of the two terms on the right hand side of (2.2.30). The first term is treated in Lemma 2.2.18 and in Lemma 2.2.19. We need both lemmas because the analysis is different for either simple or multiple eigenvalues. After those lemmas we give the proof of Theorem 2.2.10(ii) where we treat the second term. First we prove a preliminary lemma.

Lemma 2.2.17. *Let (λ_l, u_l) be a true eigenpair of problem* (1.3.7) *and let $(\lambda_{j,n}, u_{j,n})$ be a computed eigenpair. Then we have:*

$$(\lambda_{j,n} - \lambda_l)(Q_n u_l, u_{j,n})_{0,\mathcal{B},\Omega} = \lambda_l (u_l - Q_n u_l, u_{j,n})_{0,\mathcal{B},\Omega}. \tag{2.2.31}$$

Proof. By Definition 2.2.13 of Q_n we have

$$a(Q_n u_l, u_{j,n}) \;=\; a(u_l, u_{j,n}), \tag{2.2.32}$$

Since $u_{j,n}$ and u_l are eigenfunctions with corresponding eigenvalues $\lambda_{j,n}$ and λ_l, (2.2.32) yields to

$$\lambda_{j,n}(Q_n u_l, u_{j,n})_{0,\mathcal{B},\Omega} \;=\; \lambda_l(u_l, u_{j,n})_{0,\mathcal{B},\Omega}, \tag{2.2.33}$$

which is equivalent to (2.2.31). □

Lemma 2.2.18 (For simple eigenvalues)**.** *Let s be as given in Assumption 2.2.1 and let λ_l be an eigenvalue of* (1.3.7) *with multiplicity $R+1 = 1$, i.e. λ_l is a simple eigenvalue. In view of Remark 2.2.4, let $(\lambda_{l,n}, u_{l,n})$ be the computed eigenpair, whose eigenvalue converges to λ_l. Moreover, let u_l be any eigenfunction of λ_l with $\|u_l\|_{0,\mathcal{B},\Omega} = 1$. Then, there exists a function $w_{l,n} \in E_{l,n}$, with $\|w_{l,n}\|_{0,\mathcal{B},\Omega} = 1$ such that:*

$$\|u_l - \beta_l w_{l,n}\|_{0,\mathcal{B},\Omega} \lesssim C_{\text{spec1}}(H_n^{\max})^{2s} \;, \tag{2.2.34}$$

where $\beta_l = (Q_n u_l, w_{l,n})_{0,\mathcal{B},\Omega}$.

Proof. Let $\{w_{1,n}, w_{2,n}, \ldots, w_{N,n}\}$ be a orthonormal basis in the $L^2_{\mathcal{B}}$ norm for the space V_n constituted by eigenfunctions of the discrete problem and containing $w_{l,n} \in E_{l,n}$. For $u_l \in E_l$ we have

$$Q_n u_l \;=\; \sum_{i=1}^{N} (Q_n u_l, w_{i,n})_{0,\mathcal{B},\Omega}\; w_{i,n}. \tag{2.2.35}$$

Since we have supposed that λ_j is a simple eigenvalue, we define ρ_l as

$$\rho_l = \max_{\substack{i \le N \\ i \neq l}} \frac{\lambda_l}{|\lambda_{i,n} - \lambda_l|} \;, \tag{2.2.36}$$

where N is the dimension of V_n. The quantity ρ_l is well defined for $H_n^{\max}$ small enough (by Theorem 2.2.10(i) which we already proved). In order to prove (2.2.34) we can use the triangle inequality:

$$\|u_l - \beta_l w_{l,n}\|_{0,\mathcal{B},\Omega} \;\le\; \|u_l - Q_n u_l\|_{0,\mathcal{B},\Omega} \;+\; \|Q_n u_l - \beta_l w_{l,n}\|_{0,\mathcal{B},\Omega}. \tag{2.2.37}$$

Then, we estimate the second term on the right hand side of (2.2.37) by:

$$\begin{aligned}
\|Q_n u_l - \beta_l w_{l,n}\|^2_{0,\mathcal{B},\Omega} &= \|Q_n u_l - (Q_n u_l, w_{l,n})_{0,\mathcal{B},\Omega}\, w_{l,n}\|^2_{0,\mathcal{B},\Omega} \\
&= \left\| \sum_{i=1}^{N} (Q_n u_l, w_{i,n})_{0,\mathcal{B},\Omega}\, w_{i,n} - (Q_n u_l, w_{l,n})_{0,\mathcal{B},\Omega}\, w_{l,n} \right\|^2_{0,\mathcal{B},\Omega} \\
&= \left\| \sum_{\substack{i=1 \\ i\neq l}}^{N} (Q_n u_l, w_{i,n})_{0,\mathcal{B},\Omega}\, w_{i,n} \right\|^2_{0,\mathcal{B},\Omega} \\
&= \sum_{\substack{i=1 \\ i\neq l}}^{N} (Q_n u_l, w_{i,n})^2_{0,\mathcal{B},\Omega} \|w_{i,n}\|^2_{0,\mathcal{B},\Omega}. \qquad (2.2.38)
\end{aligned}$$

Applying Lemma 2.2.17 to (2.2.38), for each i, and using (2.2.36), we obtain

$$\begin{aligned}
\|Q_n u_l - \beta_l w_{l,n}\|^2_{0,\mathcal{B},\Omega} &= \sum_{\substack{i=1 \\ i\neq l}}^{N} \left(\frac{\lambda_l}{\lambda_{i,n} - \lambda_l} \right)^2 (u_l - Q_n u_l, w_{i,n})^2_{0,\mathcal{B},\Omega} \\
&\leq \sum_{\substack{i=1 \\ i\neq l}}^{N} \rho_l^2 (u_l - Q_n u_l, w_{i,n})^2_{0,\mathcal{B},\Omega} \\
&\leq \rho_l^2 \|u_l - Q_n u_l\|^2_{0,\mathcal{B},\Omega}\ , \qquad (2.2.39)
\end{aligned}$$

where in the last step we used the fact that all $w_{i,n}$ are normalized in $L^2_{\mathcal{B}}$. So from (2.2.37), (2.2.39) and (2.2.29), we have that

$$\begin{aligned}
\|u_l - \beta_l w_{l,n}\|_{0,\mathcal{B},\Omega} &\leq \|u_l - Q_n u_l\|_{0,\mathcal{B},\Omega} + \|Q_n u_l - \beta_l w_{l,n}\|_{0,\mathcal{B},\Omega} \\
&\leq (1+\rho_l)\|u_l - Q_n u_l\|_{0,\mathcal{B},\Omega} \\
&\lesssim (1+\rho_l)\ (H_n^{\max})^{2s} |u_l|_{1+s,\Omega}.
\end{aligned}$$

□

Lemma 2.2.19 (For multiple eigenvalues)**.** *Let s be as given in Assumption 2.2.1 and let λ_l be an eigenvalue of* (1.3.7) *with multiplicity $R+1$, with $R+1>1$. In view of Remark 2.2.4, let $(\lambda_{l+i,n}, u_{l+i,n})$, with $0 \leq i \leq R$, be the $R+1$ computed eigenpairs, whose eigenvalues converge to λ_l. Moreover, let u_l be any eigenfunction of λ_l with*

$\|u_l\|_{0,\mathcal{B},\Omega} = 1$. *Then defining* $\tilde{\beta}_i = (Q_n u_l, u_{l+i,n})_{0,\mathcal{B},\Omega}$, *for* $0 \leq i \leq R$, *then we have*

$$\left\| u_l - \sum_{i=0}^{R} \tilde{\beta}_i u_{l+i,n} \right\|_{0,\mathcal{B},\Omega} \lesssim C_{\text{spec1}} (H_n^{\max})^{2s}. \tag{2.2.40}$$

Proof. Let $\{u_{1,n}, u_{2,n}, \ldots, u_{N,n}\}$ be an orthonormal basis with respect to $(\cdot,\cdot)_{0,\mathcal{B}\Omega}$ for the space V_n constituted by eigenfunctions of the discrete problem. For $u_l \in E_l$ we have

$$Q_n u_l \;=\; \sum_{i=1}^{N} (Q_n u_l, u_{i,n})_{0,\mathcal{B},\Omega}\; u_{i,n}. \tag{2.2.41}$$

Since we have supposed that λ_l is a multiple eigenvalue, we define ρ_l as

$$\rho_l \;=\; \max_{\substack{i \leq N \\ i \neq l, l+1, \ldots, l+R}} \frac{\lambda_l}{|\lambda_{i,n} - \lambda_l|}\;, \tag{2.2.42}$$

where $N = \dim(V_n)$. In order to prove (2.2.40) we can use the triangle inequality:

$$\left\| u_l - \sum_{i=0}^{R} \tilde{\beta}_i u_{l+i,n} \right\|_{0,\mathcal{B},\Omega} \leq \|u_l - Q_n u_l\|_{0,\mathcal{B},\Omega} \;+\; \left\| Q_n u_l \;-\; \sum_{i=0}^{R} \tilde{\beta}_i u_{l+i,n} \right\|_{0,\mathcal{B},\Omega}. \tag{2.2.43}$$

Then we estimate the second term on the right hand side of (2.2.43) by:

$$\begin{aligned}
\left\| Q_n u_l - \sum_{i=0}^{R} \tilde{\beta}_i u_{l+i,n} \right\|_{0,\mathcal{B},\Omega}^2 &= \left\| \sum_{i=1}^{N} (Q_n u_l, u_{i,n})_{0,\mathcal{B},\Omega}\; u_{i,n} \right. \\
&\quad \left. - \sum_{i=0}^{R} (Q_n u_l, u_{l+i,n})_{0,\mathcal{B},\Omega}\; u_{l+i,n} \right\|_{0,\mathcal{B},\Omega}^2 \\
&= \left\| \sum_{\substack{i=1 \\ i \neq l, \ldots, l+R}}^{N} (Q_n u_l, u_{i,n})_{0,\mathcal{B},\Omega}\; u_{i,n} \right\|_{0,\mathcal{B},\Omega}^2 \\
&= \sum_{\substack{i=1 \\ i \neq l, \ldots, l+R}}^{N} (Q_n u_j, u_{i,n})_{0,\mathcal{B},\Omega}^2.
\end{aligned} \tag{2.2.44}$$

Then, applying Lemma 2.2.17 to (2.2.44), for each i, and using (2.2.42), we obtain

$$\begin{aligned}
\left\| Q_n u_l - \sum_{i=0}^{R} \tilde{\beta}_i u_{l+i,n} \right\|_{0,\mathcal{B},\Omega}^2 &\leq \rho_l^2 \sum_{i=1, i \neq l, \ldots, l+R}^{N} (u_l - Q_n u_l, u_{i,n})_{0,\mathcal{B},\Omega}^2 \\
&\leq \rho_l^2 \|u_l - Q_n u_l\|_{0,\mathcal{B},\Omega}^2.
\end{aligned} \tag{2.2.45}$$

So from (2.2.43), (2.2.45) and (2.2.29), we have that

$$\begin{aligned}\left\| u_l - \sum_{i=0}^{R} \tilde{\beta}_i u_{l+i,n} \right\|_{0,\mathcal{B},\Omega} &\leq \|u_l - Q_n u_l\|_{0,\mathcal{B},\Omega} + \left\| Q_n u_l - \sum_{i=0}^{R} \tilde{\beta}_i u_{l+i,n} \right\|_{0,\mathcal{B},\Omega} \\ &\leq (1+\rho_l)\|u_l - Q_n u_l\|_{0,\mathcal{B},\Omega} \\ &\lesssim (1+\rho_l)\ (H_n^{\max})^{2s} |u_l|_{1+s,\Omega}\end{aligned}$$

□

Finally we prove part (ii) of Theorem 2.2.10.

Proof of Theorem 2.2.10(ii). Let us consider (ii) for simple eigenvalues at first. Since we are supposing that λ_l is simple, we have that $E_{l,n} = \text{span}\{u_{l,n}\}$, where $u_{l,n}$ is a computed eigenvalue. So, in this case the only two possibilities for $w_{l,n}$ are plus or minus $u_{l,n}$. Let choose $w_{l,n}$ in such a way that $\beta_l = (Q_n u_l, w_{l,n})_{0,\mathcal{B},\Omega} \geq 0$.
Since, we have already proved that the first term of (2.2.30) is $\mathcal{O}(H_n^{\max})^{2s}$ - see Lemma 2.2.18. What remains is to prove that also the second term on the right hand side of (2.2.30) is converging with $\mathcal{O}(H_n^{\max})^{2s}$. To do this we write

$$\begin{aligned}|\beta_l - 1|\ \|w_{l,n}\|_{0,\mathcal{B},\Omega} &= \left|(\beta_l - 1)\ \|w_{l,n}\|_{0,\mathcal{B},\Omega}\right| = \left|\beta_l \|w_{l,n}\|_{0,\mathcal{B},\Omega} - \|u_l\|_{0,\mathcal{B},\Omega}\right| \\ &= \left|\|\beta_l w_{l,n}\|_{0,\mathcal{B},\Omega} - \|u_l\|_{0,\mathcal{B},\Omega}\right| \leq \|\beta_l w_{l,n} - u_l\|_{0,\mathcal{B},\Omega}.\end{aligned} \tag{2.2.46}$$

Putting (2.2.46) into (2.2.30) and using Lemma 2.2.18 we have

$$\|u_l - w_{l,n}\|_{0,\mathcal{B},\Omega} \leq 2\|u_l - \beta_l w_{l,n}\|_{0,\mathcal{B},\Omega} \lesssim C_{\text{spec1}}(H_n^{\max})^{2s}.$$

To prove the statement (2.2.9) we start from Lemma 2.2.11 and using (2.2.8) together with (2.2.7) we have

$$\begin{aligned}a(u_l - w_{l,n}, u_l - w_{l,n}) &= \lambda_l \|u_l - w_{l,n}\|^2_{0,\mathcal{B},\Omega} + |\lambda_{l,n} - \lambda_l| \\ &\lesssim \lambda_l\ C^2_{\text{spec1}}(H_n^{\max})^{4s} + (H_n^{\max})^{2s}.\end{aligned}$$

The proof for multiple eigenvalues is a bit more complicated:
We chose

$$w_{l,n} = \frac{\sum_{i=0}^{R} \tilde{\beta}_i\ u_{l+i,n}}{\|\sum_{i=0}^{R} \tilde{\beta}_i\ u_{l+i,n}\|_{0,\mathcal{B},\Omega}},$$

where $\tilde{\beta}_i = (Q_n u_l, u_{l+i,n})_{0,\mathcal{B},\Omega}$. We also set $\beta_l = (Q_n u_l, w_{l,n})_{0,\mathcal{B},\Omega}$. Again we choose the

sign of $w_{l,n}$ in such a way that $\beta_l > 0$. It comes straightforwardly that

$$\beta_l \; w_{l,n} = \sum_{i=0}^{R} \tilde{\beta}_i \; u_{l+i,n} \; .$$

In view of (2.2.8), we can use the triangular inequality:

$$\begin{aligned} \|u_l - w_{l,n}\|_{0,\mathcal{B},\Omega} \quad &\leq \|u_l - \beta_l \; w_{l,n}\|_{0,\mathcal{B},\Omega} \\ &+ \|\beta_l \; w_{l,n} - w_{l,n}\|_{0,\mathcal{B},\Omega}, \end{aligned} \tag{2.2.47}$$

where the first term on the right hand side has already been analysed in Lemma 2.2.19. So the proof of statement (2.2.8) would be complete if we found an upper bound for $\|\beta_l \; w_{l,n} - w_{l,n}\|_{0,\mathcal{B},\Omega}$. This could be done in the same way as for the case of simple eigenvalue.

The statement (2.2.9) for multiple eigenvalues can be proved in a similar way as in the case for simple eigenvalues. From Lemma 2.2.11 and using (2.2.8) together with (2.2.7) we have

$$\begin{aligned} a(u_l - w_{l,n}, u_l - w_{l,n}) \quad &\leq \lambda_l \|u_l - w_{l,n}\|^2_{0,\mathcal{B},\Omega} + \max_{i=0\ldots,R} |\lambda_{l+i,n} - \lambda_l| \\ &\lesssim \lambda_l \; C^2_{\text{spec1}} (H_n^{\max})^{4s} + (H_n^{\max})^{2s}. \end{aligned}$$

□

2.2.3 Finite element approximation for PCF model problems

Now we introduce the definition of the discrete versions of problems (1.3.8) and (1.3.9). Since the FEMs for these problems are very similar to the FEM for generic elliptic eigenvalue problems, we are going to discuss only the differences between these methods. Again, let $\mathcal{T}_n$, $n = 1, 2, \ldots$ denote a family of conforming and periodic triangular meshes on Ω where Ω is a square.

On any mesh $\mathcal{T}_n$ we denote by $V_n \subset C^0(\Omega)$ the finite dimensional space of linear polynomials on each element τ of the mesh, let the dimension of this space be N. For problem (1.3.8) the space $V_n \subset H^1_\pi(\Omega)$, since the problem has periodic boundary conditions.

The discrete formulation of problem (1.3.8) is:

seek eigenpairs of the form $(\lambda_{i,n}, u_{i,n}) \in \mathbb{R} \times V_n$*, with* $\|u_{i,n}\|_{0,\mathcal{B},\Omega} = 1$ *such that*

$$a_\kappa(u_{i,n}, v_n) = \lambda_{i,n}(u_{i,n}, v_n)_{0,\mathcal{B},\Omega} \; , \quad \text{for all } v_n \in V_n. \tag{2.2.48}$$

Furthermore, the discrete formulation of problem (1.3.9) is:

seek eigenpairs of the form $(\zeta_{i,n}, u_{i,n}) \in \mathbb{R} \times V_n$*, with* $\|u_{i,n}\|_{0,\mathcal{B},\Omega} = 1$ *such that*

$$a_{\kappa,S}(u_{i,n}, v_n) = \zeta_{i,n}(u_{i,n}, v_n)_{0,\mathcal{B},\Omega} \ , \quad \text{for all } v_n \in V_n. \tag{2.2.49}$$

Assumption 2.2.20. *We assume that there exists a constant* $C_{\text{ell}}^{\text{PCF}} > 0$ *and* $s \in [0,1]$ *with the following property. For* $f \in L^2(\Omega)$*, if* $v \in H^1_\pi(\Omega)$ *solves the problem* $a_{\kappa,S}(v,w) = (f,w)_{0,\Omega}$ *for all* $w \in H^1_\pi(\Omega)$*, then*

$$\|v\|_{1+s,\Omega} \leq C_{\text{ell}}^{\text{PCF}} \|f\|_{0,\Omega} \ . \tag{2.2.50}$$

The result above comes from the standard theory used in Assumption 2.2.1. In fact, for any couple of f and v satisfying the shifted problem with periodic boundary conditions, we have that the same couple of functions satisfy the problem $a_{\kappa,S}(v,w) = (f,w)_{0,\Omega}$ with Dirichlet boundary conditions matching the function v on the border of the domain Ω. Under Assumption 2.2.20 it follows that for any eigenpair (λ_j, u_j) with $\|u_j\|_{0,\mathcal{B},\Omega} = 1$ of the problem (1.3.9), i.e. $a_{\kappa,S}(u_j, v) = \lambda_j (u_j, v)_{0,\mathcal{B},\Omega}$, we have that inequality (2.2.50) becomes $\|u_j\|_{1+s,\Omega} \leq C_{\text{ell}}^{\text{PCF}} \lambda_j \overline{b}$, where we have substituted f with $\lambda_j u_j \mathcal{B}$.
Also for PCF problems, we have a result similar to Lemma 2.2.2:

Lemma 2.2.21. *Let the finite dimensional space* V_n *be constructed on a mesh* $\mathcal{T}_n$*, with mesh size* $H_n^{\max}$*. For any function* $u \in H^{1+s}(\Omega) \cap H^1_\pi(\Omega)$ *we have that*

$$\inf_{v_n \in V_n} \|u - v_n\|_{1,\Omega} \ \leq \ C_{\text{app}}^{\text{PCF}} (H_n^{\max})^s |u|_{1+s,\Omega} \ .$$

Proof. The proof is based on the material in [48], which is easy to extend to the periodic case, since the definition of the Scott-Zhang quasi-interpolation operator $I_n : H^1(\Omega) \longrightarrow V_n$ is elementwise. So we can keep the same definition on each element, but, since our problem has periodic boundary conditions, summing the contribution from all elements we end up with the definition $I_n : H^1_\pi(\Omega) \longrightarrow V_n$. Moreover, in [48] it is proved the following result for any element τ in a shape-regular mesh:

$$\|u - I_n u\|_{1,\tau} \ \leq \ C h_\tau^s \|u\|_{1+s,\omega_\tau} \ , \tag{2.2.51}$$

where ω_τ is the union of all the elements which are neighbours of τ and where the constant C is not depending on the size of the element τ. Summing (2.2.51) on all the elements in the mesh $\mathcal{T}_n$ we obtain:

$$\|u - I_n u\|^2_{1,\Omega} \ = \ \sum_{\tau \in V_n} \|u - I_n u\|^2_{1,\tau} \ \leq \ C^2 \sum_{\tau \in V_n} h_\tau^{2s} \|u\|^2_{1+s,\omega_\tau} \ \leq \ C' C^2 (H_n^{\max})^{2s} \|u\|^2_{1+s,\Omega} \ ,$$

where the constant C' depends on the overlapping of the patches ω_τ. We conclude the proof denoting by $C_{\text{app}}^{\text{PCF}} = C'^{1/2} C$ and taking the infimum over all the functions in V_n,

i.e.

$$\inf_{v_n \in V_n} \|u - v_n\|_{1,\Omega} \ \leq \ \|u - I_n u\|_{1,\Omega} \ \leq \ C_{\text{app}}^{\text{PCF}} (H_n^{\max})^s \|u\|_{1+s,\Omega} \ .$$

□

A consequence of Lemma 2.2.21 is that the space V_n becomes dense in $H^{1+s}(\Omega) \cap H^1_\pi(\Omega)$, when n goes to infinity due to the assumptions on the sequence $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$, i.e.

$$H^{1+s}(\Omega) \cap H^1_\pi(\Omega) = \overline{\lim_{n\to\infty} V_n}. \tag{2.2.52}$$

2.2.4 Convergence estimates for the PCF case

In this section we apply the framework in Section 2.2.2 to PCF problems (1.3.8) and (1.3.9). For these problems we have already proved the discreteness and non-negativity of the spectrum in Section 2.1.2.
The framework in Section 2.2.2 can be easily adapted for problem (1.3.9), since this problem is coercive. In view of this, we are able to state for (1.3.9) results analogous to Theorem 2.2.10. Then, the convergence estimates for problem (1.3.8) will come at once from the relation between the spectra of the two problems, which has been analysed in Section 2.1.2.

Notation 2.2.22. *In this subsection, we write $A \lesssim B$ when A/B is bounded by a constant which may depend on the functions $\mathcal{A}$ and $\mathcal{B}$, on $c_{a,S}^{\text{PCF}}$ in (2.1.12), on $C_{a,S}^{\text{PCF}}$ in (2.1.15), on C_b in (2.1.6), on C_{reg} in (2.2.1), on $C_{\text{ell}}^{\text{PCF}}$, or on $C_{\text{app}}^{\text{PCF}}$ in Lemma 2.2.21, but not on n. The notation $A \cong B$ means $A \lesssim B$ and $A \gtrsim B$.*

Remark 2.2.23. *Similarly to what we have already done for general elliptic eigenvalue problems, we have from Theorem 2.2.3 that the sequence $\{\mathcal{T}_n\}_{n\in\mathbb{N}}$ converges in norm to the solution operator $\mathcal{T}$ when n goes to infinity. This implies that also the spectrum of problem* (2.2.49) *converges to the spectrum of the continuous problem* (1.3.9) *when n goes to infinity. So, for each eigenvalue ζ_j of multiplicity $R+1$, it is possible to construct $R+1$ sequences of computed eigenpairs $(\zeta_{l+r,n}, u_{l+r,n})$, with $r = 0, \ldots, R$, such that $\zeta_{l+r,n}$ converges to ζ_j when n goes to infinity, for all $r = 0, \ldots, R$. Moreover, for any n all the eigenfunctions $u_{l,n}, \ldots, u_{l+r,n}$ are orthogonal to each other.*

From now on we will consider an eigenvalue ζ_l of problem (1.3.9) with multiplicity $R+1$, where $R \geq 0$. Moreover, let $E_{l,n}^{\text{PCF}}$ be the computed eigenspace corresponding to the true eigenvalue ζ_l in view of Remark 2.2.23. The application of the general framework to the PCF problem leads us to the following result..

Theorem 2.2.24. *Let s be as given in Assumption 2.2.20 and suppose that $H_n^{\max}$ is small enough. Then considering the eigenvalue λ_l, we have that the following statements hold:*

(i) In view of Remark 2.2.23, let ζ_l be an eigenvalue of (1.3.9) and let $(\zeta_{l,n}, u_{l,n})$ be a computed eigenpair of problem (2.2.49), with $\zeta_{l,n}$ converging to ζ_l when n goes to infinity, then

$$0 \leq \zeta_{l,n} - \zeta_l \lesssim (H_n^{\max})^{2s} \ . \tag{2.2.53}$$

(ii) Let ζ_l be an eigenvalue of problem (1.3.9) with multiplicity $R+1$, with $R \geq 0$ and let u_l be any eigenfunction of ζ_l with $\|u_l\|_{0,\mathcal{B},\Omega} = 1$, then there exists a sequence $\{w_{l,n}\}_{n\in\mathbb{N}}$ with $w_{l,n} \in E_{l,n}^{\mathrm{PCF}}$ for all n and with $\|w_{l,n}\|_{0,\mathcal{B},\Omega} = 1$ such that

$$\|u_l - w_{l,n}\|_{0,\mathcal{B},\Omega} \ \lesssim \ C_{\mathrm{spec1}}^{\mathrm{PCF}} (H_n^{\max})^{2s} \ , \tag{2.2.54}$$

$$a_{\kappa,S}(u_l - w_{l,n}, u_l - w_{l,n})^{1/2} \ \lesssim \ C_{\mathrm{spec2}}^{\mathrm{PCF}} (H_n^{\max})^{s} \ . \tag{2.2.55}$$

Where the constants $C_{\mathrm{spec1}}^{\mathrm{PCF}}$ and $C_{\mathrm{spec2}}^{\mathrm{PCF}}$ depends on the spectral information ζ_i, u_i, $i = 1, \ldots, l$.

The structure of the proof of Theorem 2.2.24 is very similar to the proof of Theorem 2.2.10. So we are not going to rewrite it. Instead we state some of the intermediate results used to prove the theorem. We start defining the Rayleigh-Ritz projection operator for this problem.

Definition 2.2.25 (Rayleigh-Ritz projection operator for the PCF case)**.** *We define the projection operator $Q_n^{\mathrm{PCF}} : H_\pi^1(\Omega) \longrightarrow V_n$ as the operator that for a given function $u \in H_\pi^1(\Omega)$ returns the function $Q_n^{\mathrm{PCF}} u \in V_n$:*

$$a_{\kappa,S}(u - Q_n^{\mathrm{PCF}} u, v_n) = 0 \quad \textit{for all } v_n \in V_n.$$

To prove the estimates for eigenfunctions we have to adapt Lemma 2.2.11 and Lemma 2.2.18 to this problem. To modifications are very simple since we need just to change the sesquilinear form.

Lemma 2.2.26. *Let (ζ_l, u_l) be a true eigenpair of problem* (1.3.9) *with $\|u_l\|_{0,\mathcal{B},\Omega} = 1$and let $(\zeta_{j,n}, u_{j,n})$ be a computed eigenpair of problem* (2.2.49) *with $\|u_{j,n}\|_{0,\mathcal{B},\Omega} = 1$. Then we have:*

$$a_{\kappa,S}(u_l - u_{j,n}, u_l - u_{j,n}) = \zeta_l \|u_l - u_{j,n}\|_{0,\mathcal{B},\Omega}^2 + |\zeta_{j,n} - \zeta_l|.$$

Corollary 2.2.27. *Let (ζ_l, u_l) be a true eigenpair of problem* (1.3.9) *and let $(\zeta_{j,n}, u_{j,n})$ be a computed eigenpair of problem* (2.2.49)*. Then we have:*

$$|\zeta_{j,n} - \zeta_l| \ \leq \ a_{\kappa,S}(u_l - u_{j,n}, u_l - u_{j,n}) \ .$$

Lemma 2.2.28. *Let define the quantity* σ_{l+R}^n *as*

$$\sigma_{l+R}^n := \max_{u \in \mathcal{E}_1^{l+R}} \left| (u, u - Q_n^{\mathrm{PCF}} u)_{0,\mathcal{B},\Omega} + (u - Q_n^{\mathrm{PCF}} u, u)_{0,\mathcal{B},\Omega} - (u - Q_n^{\mathrm{PCF}} u, u - Q_n^{\mathrm{PCF}} u)_{0,\mathcal{B},\Omega} \right|. \tag{2.2.56}$$

Provided that $H_n^{\max}$ *is small enough so that* $\sigma_{l+R}^n < 1$*, then the computed eigenvalue* $\zeta_{l,n}$*, with* $l \leq N$ *where* $N = \dim V_n$*, is bounded above and below by:*

$$\zeta_l \;\leq\; \zeta_{l,n} \;\leq\; \frac{\zeta_l}{1 - \sigma_{l+R}^n}. \tag{2.2.57}$$

Lemma 2.2.29 (For simple eigenvalues)**.** *Let* s *be as given in Assumption 2.2.20 and let* ζ_l *be an eigenvalue of* (1.3.9) *with multiplicity* $R+1 = 1$*, i.e.* ζ_l *is a simple eigenvalue. In view of Remark 2.2.23, let* $(\zeta_{l,n}, u_{l,n})$ *be the computed eigenpair, whose eigenvalue converges to* ζ_l*. Moreover, let* u_l *be any eigenfunction of* ζ_l *with* $\|u_l\|_{0,\mathcal{B},\Omega} = 1$*. Then, there exists a function* $w_{l,n} \in E_{l,n}^{\mathrm{PCF}}$*, with* $\|w_{l,n}\|_{0,\mathcal{B},\Omega} = 1$ *such that:*

$$\|u_l - \beta_l w_{l,n}\|_{0,\mathcal{B},\Omega} \lesssim C_{\mathrm{spec1}}^{\mathrm{PCF}} (H_n^{\max})^{2s} \;, \tag{2.2.58}$$

where $\beta_l = (Q_n^{\mathrm{PCF}} u_l, w_{l,n})_{0,\mathcal{B},\Omega}$*.*

Lemma 2.2.29 needs some modification to be suitable for multiple eigenvalues.

Lemma 2.2.30 (For multiple eigenvalues)**.** *Let* ζ_l *be an eigenvalue of* (1.3.9) *with multiplicity* $R+1$*, with* $R+1 > 1$*. In view of Remark 2.2.23, let* $(\zeta_{l+i,n}, u_{l+i,n})$*, with* $0 \leq i \leq R$*, be the* $R+1$ *computed eigenpairs, whose eigenvalues converge to* ζ_l*. Moreover, let* u_j *be any eigenfunction of* ζ_l*. Then defining* $\tilde{\beta}_i = (Q_n^{\mathrm{PCF}} u_l, u_{l+i,n})_{0,\mathcal{B},\Omega}$*, for* $0 \leq i \leq R$*, then we have*

$$\left\| u_l - \sum_{i=0}^{R} \tilde{\beta}_i u_{l+i,n} \right\|_{0,\mathcal{B},\Omega} \;\lesssim\; C_{\mathrm{spec1}}^{\mathrm{PCF}} (H_n^{\max})^{2s}. \tag{2.2.59}$$

We conclude this chapter stating the converging estimates for both eigenvalues and eigenvectors for problem (1.3.8). These results comes easily from Theorem 2.2.24 undoing the effect of the shift on the spectrum.

Lemma 2.2.31. *Let* (λ_l, u_l) *be a true eigenpair of problem* (1.3.8) *with* $\|u_l\|_{0,\mathcal{B},\Omega} = 1$*and let* $(\lambda_{j,n}, u_{j,n})$ *be a computed eigenpair of problem* (2.2.48) *with* $\|u_{j,n}\|_{0,\mathcal{B},\Omega} = 1$*. Then we have:*

$$a_\kappa(u_l - u_{j,n}, u_l - u_{j,n}) = \lambda_l \|u_l - u_{j,n}\|_{0,\mathcal{B},\Omega}^2 + |\lambda_{j,n} - \lambda_l|.$$

Corollary 2.2.32. *Let* (λ_l, u_l) *be a true eigenpair of problem* (1.3.8) *and let* $(\lambda_{j,n}, u_{j,n})$ *be a computed eigenpair of problem* (2.2.48)*. Then we have:*

$$|\lambda_{j,n} - \lambda_l| \ \leq \ a_\kappa(u_l - u_{j,n}, u_l - u_{j,n}) \ .$$

Theorem 2.2.33. *Let s be as given in Assumption 2.2.20 and suppose that $H_n^{\max}$ is small enough. Then considering the eigenvalue λ_l of problem (1.3.8) with multiplicity $R+1>0$, we have that the following statements hold:*

(i) In view of Remark 2.2.23, let λ_l be an eigenvalue of (1.3.8) and let $(\lambda_{l,n}, u_{l,n})$ be a computed eigenpair of problem (2.2.48), with $\lambda_{l,n}$ converging to λ_l when n goes to infinity, then

$$\lambda_l \leq \lambda_{l,n} \lesssim \lambda_l + (H_n^{\max})^{2s} \ . \tag{2.2.60}$$

(ii) Let λ_l be an eigenvalue of problem (1.3.8) with multiplicity $R+1$, with $R \geq 0$ and let u_l be any eigenfunction of λ_l with $\|u_l\|_{0,\mathcal{B},\Omega} = 1$, then there exists a sequence $\{w_{j,n}\}_{n\in\mathbb{N}}$ with $w_{j,n} \in E_{j,n}^{\mathrm{PCF}}$ for all n and with $\|w_{j,n}\|_{0,\mathcal{B},\Omega} = 1$ such that

$$\|u_l - w_{l,n}\|_{0,\mathcal{B},\Omega} \ \lesssim \ C_{\mathrm{spec1}}^{\mathrm{PCF}}(H_n^{\max})^{2s} \ , \tag{2.2.61}$$

$$a_\kappa(u_l - w_{l,n}, u_l - w_{l,n})^{1/2} \ \lesssim \ C_{\mathrm{spec2}}^{\mathrm{PCF}}(H_n^{\max})^{s} \ . \tag{2.2.62}$$

Where the constants $C_{\mathrm{spec1}}^{\mathrm{PCF}}$ and $C_{\mathrm{spec2}}^{\mathrm{PCF}}$ depends on the spectral information λ_i, u_i, $i = 1, \ldots, l$.

Chapter 3

A posteriori error estimator

In the last decades, a posteriori error estimates have become essential tools in engineering and physics to improve accuracy of numerical solutions. A comprehensive survey on the topic is in [52]. However, an a posteriori error estimate for eigenvalue problems is still quite a new piece of technology. There are only a few works on the topic: [37, 53, 21, 52, 28, 12]. The approach presented in [52] and [28] is different because in these works eigenvalue problems are treated as particular cases of general non linear problems. As far as we are aware there is no a posteriori error estimate used together with mesh adaptivity for photonic crystal eigenvalue problems.
The a posteriori error estimator we present is based on residuals (defined in Section 3.2). Its most important characteristics are *reliability* and *efficiency*: the first ensures that the actual error is always smaller than the residual multiplied by a constant (ignoring higher order terms). The latter ensures that the residual is proportional to the actual error (plus higher order terms). We will state all the result for linear elements, but the same analysis holds also for any higher order. Since the presence of higher order terms in such results, we will refer to them as asymptotic reliability and asymptotic efficiency.
In Section 3.1 we prove some preliminary results - Theorem 3.1.4, Theorem 3.1.7 and Theorem 3.1.8 - which will be useful in order to prove reliability and efficiency for our a posteriori error estimator. In Theorem 3.1.4, Theorem 3.1.7 and Theorem 3.1.8 we rework the a priori convergence estimates of Theorem 2.2.10(ii), Theorem 2.2.24(ii) and Theorem 2.2.33(ii) in Chapter 2. Such results in Chapter 2 estimate in different norms the quantity $u_l - w_{l,n}$, where u_l is a true eigenfunction and where $w_{l,n}$ is a linear combination of computed eigenfunctions. So, this quantity describes how well a true eigenfunction is approximated by the computed ones. But, for the a posteriori analysis, especially in the context of adaptive methods, it would be more useful to estimate how good a computed eigenfunction $u_{l,n}$ is an approximation of a true eigenfunction U_l. In particular, U_l is the true eigenfunction with minimum distance from $u_{l,n}$ in the $L^2_{\mathcal{B}}$ norm and, since $u_{l,n}$ depends on n, consequently also U_l depends on n. The quantities

$u_l - w_{l,n}$ and $U_l - u_{l,n}$ are not equivalent from a practical point of view, because $u_{l,n}$ is an eigenfunction of the discrete problem and it is a known quantity coming out from the computations, instead $w_{l,n}$ in general is not an eigenfunction of the discrete problem and moreover it is unknown, because without knowing u_l, it is not possible to construct the linear combination to obtain $w_{l,n}$. So, in Theorem 3.1.4, Theorem 3.1.7 and Theorem 3.1.8 we estimate the quantity $U_l - u_{l,n}$ both in the $L^2_{\mathcal{B}}$ norm and in the energy norm.
The outline of this chapter is as follows: in Section 3.1 we prove Theorem 3.1.4 and Theorem 3.1.7, then in Section 3.2 we define residuals. Further, in Section 3.3 we give the proof of asymptotic reliability for the PCF case and in the following section, Section 3.4, we adapt the reliability results to the TE and TM mode problems and to the general elliptic eigenvalue problem (1.3.7). Then, Section 3.5 contains the proof of asymptotic efficiency of our a posteriori error estimator for the PCF case.

Notation 3.0.34. *In this chapter, we write $A \lesssim B$ when A/B is bounded by a constant which may depend on the functions $\mathcal{A}$ and $\mathcal{B}$, on c_a in (2.1.3), on $c_{a,S}^{\mathrm{PCF}}$ in (2.1.12), on C_a in (2.1.5), on C_a^{PCF} in (2.1.16), on $C_{a,S}^{\mathrm{PCF}}$ in (2.1.15), on C_b in (2.1.6), on C_{reg} in (2.2.1) and on the multiplicity R of eigenvalues, but* **not on the mesh parameters**. *The notation $A \cong B$ means $A \lesssim B$ and $A \gtrsim B$.*

3.1 Further a priori convergence results

This section is split into two subsections one devoted to the general elliptic case and the other to the PCF case. The subdivision has been done for sake of clarity.

3.1.1 The general elliptic case

Let us use the same notation as in Chapter 2: λ_l is an eigenvalue of multiplicity $R+1$ and E_l and $E_{l,n}$ are the true and computed eigenspaces corresponding to λ_l, in the sense of Remark 2.2.4. We denote by $\{u_{l+r}\}_{r=0}^R$ a orthonormal basis for E_l with respect to the inner product $(\cdot,\cdot)_{0,\mathcal{B},\Omega}$ and from Theorem 2.2.10(ii) we have that for each $r = 0, \dots, R$ there is a sequence $\{w_{l+r,n}\}_{n\in\mathbb{N}}$, with $w_{l+r,n} \in E_{l,n}$, that converges to u_{l+r} in both the L^2 and the energy norms.
We can define the $R+1 \times R+1$ matrix Ψ_n, whose entries are

$$[\Psi_n]_{r,i} \;:=\; (Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega} \Big/ \Big\| \sum_{m=0}^{R} (Q_n u_{l+r}, u_{l+m,n})_{0,\mathcal{B},\Omega}\; u_{l+m,n} \Big\|_{0,\mathcal{B},\Omega}, \qquad (3.1.1)$$

where the projection operator Q_n is defined in Definition 2.2.13. We would like to show that the definition of Ψ_n is well posed for $H_n^{\max}$ small enough, since in such case the quantities $\Big\| \sum_{m=0}^{R} (Q_n u_{l+r}, u_{l+m,n})_{0,\mathcal{B},\Omega}\; u_{l+m,n} \Big\|_{0,\mathcal{B},\Omega}$ are different from 0 for all r.

Looking for a contradiction, we suppose that exists an r such that for any value $H_{n'}^{\max}$ in a subsequence of $H_n^{\max}$ we have that

$$\Big\| \sum_{m=0}^{R} (Q_{n'} u_{l+r}, u_{l+m,n'})_{0,\mathcal{B},\Omega} \; u_{l+m,n'} \Big\|_{0,\mathcal{B},\Omega} \;=\; 0 \; . \tag{3.1.2}$$

Since the set of vectors $\{u_{l+m,n'}\}$ is an orthonormal basis for $E_{l,n'}$, we have that (3.1.2) is equivalent to

$$\forall m, \quad (Q_{n'} u_{l+r}, u_{l+m,n'})_{0,\mathcal{B},\Omega} \;=\; 0 \; . \tag{3.1.3}$$

Using the linearity of the inner product we obtain

$$\forall m, \quad (Q_{n'} u_{l+r} - w_{l+r,n'}, u_{l+m,n'})_{0,\mathcal{B},\Omega} \;+\; (w_{l+r,n'}, u_{l+m,n'})_{0,\mathcal{B},\Omega} \;=\; 0 \; . \tag{3.1.4}$$

Let's start analysing the quantity $Q_{n'} u_{l+r} - w_{l+r,n'}$, using the fact that $w_{l+r,n'}$ converges to u_{l+r} and also using the properties of $Q_{n'}$ we have

$$\begin{aligned} \lim_{H_{n'}^{\max} \to 0} \|Q_{n'} u_{l+r} - w_{l+r,n'}\|_{0,\mathcal{B},\Omega} \;\leq\; & \lim_{H_{n'}^{\max} \to 0} \|Q_{n'} u_{l+r} - u_{l+r}\|_{0,\mathcal{B},\Omega} \\ & + \lim_{H_{n'}^{\max} \to 0} \|u_{l+r} - w_{l+r,n'}\|_{0,\mathcal{B},\Omega} \;=\; 0 \; . \end{aligned} \tag{3.1.5}$$

So, when $H_{n'}^{\max} \to 0$ the first inner product in (3.1.4) goes to 0 for all m. Then, the contradiction we are looking for should raise from the second inner product in (3.1.4), i.e. $(w_{l+r,n'}, u_{l+m,n'})_{0,\mathcal{B},\Omega}$. We know that $w_{l+r,n'}$ is an unit vector in $E_{l,n'}$, then

$$w_{l+r,n'} \;=\; \sum_{m=0}^{R} (w_{l+r,n'}, u_{l+m,n'})_{0,\mathcal{B},\Omega} \; u_{l+m,n'} \; ,$$

since $w_{l+r,n'}$ is an unit vector, we have that it is not possible that all $(w_{l+r,n'}, u_{l+m,n'})_{0,\mathcal{B},\Omega}$ are 0 at the same time for any value of $H_{n'}^{\max}$. This is the contradiction we were looking for.

To have more insights on the definition of Ψ_n, we can also analyse the quantity $(Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}$. Using the definition of problem (1.3.7) and the properties of

Q_n we have that:

$$\begin{aligned}(Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega} &= \frac{1}{\lambda_{l+i,n}}\lambda_{l+i,n}(Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega} \\ &= \frac{1}{\lambda_{l+i,n}}a(Q_n u_{l+r}, u_{l+i,n}) \\ &= \frac{1}{\lambda_{l+i,n}}a(u_{l+r}, u_{l+i,n}) \;=\; \frac{1}{\lambda_{l+i,n}}\lambda_{l+r}(u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}\ .\end{aligned} \tag{3.1.6}$$

So, the quantities $(Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}$ are proportional to the simpler quantities $(u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}$. In the next Lemma we prove that for $H_n^{\max}$ small enough the infinity norm - defined below - of the matrix Ψ_n is bounded from above by 1.

Lemma 3.1.1. *For $H_n^{\max}$ small enough, there is a constant C_Ψ independent of $H_n^{\max}$ such that*

$$\|\Psi_n\|_\infty \;\leq\; C_\Psi,$$

where the infinity norm of the matrix Ψ_n is defined as $\|\Psi_n\|_\infty := \max_r\{\sum_{i=0}^R |[\Psi_n]_{r,i}|\}$.

Proof. From the definition of the infinity norm for matrices and from (3.1.1) we have:

$$\begin{aligned}\|\Psi_n\|_\infty &= \max_r\Big\{\sum_{i=0}^R \Big|[\Psi_n]_{r,i}\Big|\Big\} \\ &= \max_r\left\{\frac{\sum_{i=0}^R \Big|(Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}\Big|}{\Big\|\sum_{m=0}^R (Q_n u_{l+r}, u_{l+m,n})\ u_{l+m,n}\Big\|_{0,\mathcal{B},\Omega}}\right\}.\end{aligned} \tag{3.1.7}$$

The quantities $\Big|(Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}\Big|$ in (3.1.7) are all bounded by 1 since u_{l+r} and $u_{l+i,n}$, for all r and i, are unit vectors in $\|\cdot\|_{0,\mathcal{B},\Omega}$, so from (3.1.7) we obtain:

$$\begin{aligned}\|\Psi_n\|_\infty &\leq \max_r\left\{\frac{R+1}{\Big\|\sum_{m=0}^R (Q_n u_{l+r}, u_{l+m,n})\ u_{l+m,n}\Big\|_{0,\mathcal{B},\Omega}}\right\} \\ &= \frac{R+1}{\min_r\Big\|\sum_{m=0}^R (Q_n u_{l+r}, u_{l+m,n})\ u_{l+m,n}\Big\|_{0,\mathcal{B},\Omega}}\ .\end{aligned} \tag{3.1.8}$$

In order to conclude the proof, we need to find a lower bound of $\min_r \Big\|\sum_{m=0}^R (Q_n u_{l+r}, u_{l+m,n})\ u_{l+m,n}\Big\|_{0,\mathcal{B},\Omega}$ which is independent of $H_n^{\max}$. We have already proved above that, for $H_n^{\max}$ small

enough and for all r, the quantities $\left\| \sum_{m=0}^{R} (Q_n u_{l+r}, u_{l+m,n})\ u_{l+m,n} \right\|_{0,\mathcal{B},\Omega}$ are different from 0 and now we want to prove that the limit of the quantities $\left\| \sum_{m=0}^{R} (Q_n u_{l+r}, u_{l+m,n})\ u_{l+m,n} \right\|_{0,\mathcal{B},\Omega}$, for all r, is 1. This will imply that for $H_n^{\max}$ small enough there exists a constant $C > 0$, which is independent of $H_n^{\max}$, bounding from below all those quantities and that

$$\min_r \left\| \sum_{m=0}^{R} (Q_n u_{l+r}, u_{l+m,n})\ u_{l+m,n} \right\|_{0,\mathcal{B},\Omega} \ > \ C \ .$$

Let's start manipulating the quantity $\left\| \sum_{i=0}^{R} (Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}\ u_{l+i,n} \right\|_{0,\mathcal{B},\Omega}$: since the eigenvectors $u_{l+i,n}$ are orthonormal to each other with respect to the inner product $(\cdot,\cdot)_{0,\mathcal{B},\Omega}$, we obtain that

$$\begin{aligned} \left\| \sum_{i=0}^{R} (Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}\ u_{l+i,n} \right\|_{0,\mathcal{B},\Omega} &= \left\{ \sum_{i=0}^{R} |(Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}|^2 \right\}^{1/2} \\ &= \left\{ \sum_{i=0}^{R} |(Q_n u_{l+r} - u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega} \right. \\ &\quad \left. + (u_{l+r} - w_{l+r,n}, u_{l+i,n})_{0,\mathcal{B},\Omega} + (w_{l+r,n}, u_{l+i,n})_{0,\mathcal{B},\Omega}|^2 \right\}^{1/2} . \end{aligned} \tag{3.1.9}$$

In view of (3.1.9) we have that for all r:

$$\begin{aligned} \lim_{n\to\infty} \left\| \sum_{i=0}^{R} (Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}\ u_{l+i,n} \right\|_{0,\mathcal{B},\Omega} &= \left\{ \sum_{i=0}^{R} \Big(\lim_{n\to\infty} (Q_n u_{l+r} - u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega} \right. \\ &\quad + \lim_{n\to\infty} (u_{l+r} - w_{l+r,n}, u_{l+i,n})_{0,\mathcal{B},\Omega} \\ &\quad \left. + \lim_{n\to\infty} (w_{l+r,n}, u_{l+i,n})_{0,\mathcal{B},\Omega} \Big)^2 \right\}^{1/2} . \end{aligned} \tag{3.1.10}$$

From the properties of the projection operator Q_n we have that

$$\lim_{n\to\infty} (Q_n u_{l+r} - u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega} \ = \ 0 \ . \tag{3.1.11}$$

Moreover, from Theorem 2.2.10(ii) we have that

$$\lim_{n\to\infty} (u_{l+r} - w_{l+r,n}, u_{l+i,n})_{0,\mathcal{B},\Omega} \ = \ 0 \ . \tag{3.1.12}$$

Then, substituting (3.1.11) and (3.1.12) into (3.1.10), we obtain:

$$\begin{aligned}
\lim_{n\to\infty} \Big\| \sum_{i=0}^{R} (Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}\ u_{l+i,n} \Big\|_{0,\mathcal{B},\Omega} &= \left\{ \sum_{i=0}^{R} \Big(\lim_{n\to\infty} (w_{l+r,n}, u_{l+i,n})_{0,\mathcal{B},\Omega} \Big)^2 \right\}^{1/2} \\
&= \lim_{n\to\infty} \left\{ \sum_{i=0}^{R} \Big((w_{l+r,n}, u_{l+i,n})_{0,\mathcal{B},\Omega} \Big)^2 \right\}^{1/2} \\
&= \lim_{n\to\infty} \Big\| \sum_{i=0}^{R} (w_{l+r,n}, u_{l+i,n})_{0,\mathcal{B},\Omega}\ u_{l+i,n} \Big\|_{0,\mathcal{B},\Omega} \\
&= \lim_{n\to\infty} \|w_{l+r,n}\|_{0,\mathcal{B},\Omega} \ = \ 1 \ .
\end{aligned}$$

□

Lemma 3.1.2. *For $H_n^{\max}$ small enough, the infinity norm of the matrix Ψ_n, i.e. $\|\Psi_n\|_\infty := \max_r\{\sum_{i=0}^{R} |[\Psi_n]_{r,i}|\}$, is bounded from below by* 1.

Proof. From the definition of the infinity norm for matrices and from (3.1.1) we have:

$$\begin{aligned}
\|\Psi_n\|_\infty &= \max_r \left\{ \sum_{i=0}^{R} \Big| [\Psi_n]_{r,i} \Big| \right\} \\
&= \max_r \left\{ \frac{\sum_{i=0}^{R} \Big| (Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega} \Big|}{\Big\| \sum_{m=0}^{R} (Q_n u_{l+r}, u_{l+m,n})\ u_{l+m,n} \Big\|_{0,\mathcal{B},\Omega}} \right\} .
\end{aligned} \tag{3.1.13}$$

Now, since the eigenvectors $u_{l+i,n}$ are orthonormal with respect to the inner product $(\cdot,\cdot)_{0,\mathcal{B},\Omega}$ to each other we obtain that

$$\begin{aligned}
\Big\| \sum_{i=0}^{R} (Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}\ u_{l+i,n} \Big\|_{0,\mathcal{B},\Omega} &= \left\{ \sum_{i=0}^{R} |(Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}|^2 \right\}^{1/2} \\
&\le \sum_{i=0}^{R} |(Q_n u_{l+r}, u_{l+i,n})_{0,\mathcal{B},\Omega}| \ .
\end{aligned} \tag{3.1.14}$$

The result follows directly by inserting estimates (3.1.14) into (3.1.13).

□

We have already implicitly used the matrix Ψ_n in the proof of Theorem 2.2.10(ii) in

Chapter 2, in fact the vectors $w_{l+r,n}$ can be equivalently defined as

$$w_{l+r,n} = \sum_{i=0}^{R} [\Psi_n]_{r,i} u_{l+i,n} \ . \tag{3.1.15}$$

In the next Lemma we prove that also the infinity norm of the inverse of Ψ_n is bounded for $H_n^{\max}$ small enough.

Lemma 3.1.3. *For $H_n^{\max}$ small enough, the matrix Ψ_n^{-1}, which is the inverse of Ψ_n, exists and we have also*

$$\|\Psi_n^{-1}\|_\infty \ \leq \ C_{\Psi^{-1}} \ ,$$

where the constant $C_{\Psi^{-1}}$ is independent of $H_n^{\max}$.

Proof. By contradiction suppose that is not true that for $H_n^{\max}$ small enough the matrix Ψ_n^{-1} exists, so we should have a subsequence $\{H_m^{\max}\}_{m=1}^{\infty}$ of $\{H_n^{\max}\}_{n=1}^{\infty}$ such that for each m the matrix Ψ_m is not invertible, since its kernel is not trivial and its image has dimension less than $R+1$. Equivalently using (3.1.15), there are unit vectors $\vec{\mathbf{x}}_m \in \mathbb{R}^{R+1}$ different from 0 for each m such that

$$\sum_{r=0}^{R} x_{m,r} \ \sum_{i=0}^{R} [\Psi_n]_{r,i} u_{l+i,n} \ = \ \sum_{r=0}^{R} x_{m,r} \ w_{l+r,m} \ = \ 0 \ , \tag{3.1.16}$$

where $x_{m,r}$ is the r-component of the vector $\vec{\mathbf{x}}_m$.
Denote with $\{\vec{x}_{m'}\}_{n=1}^{\infty}$ a subsequence of unit vectors of the sequence $\{\vec{x}_m\}_{n=1}^{\infty}$ that converges to a unit vector called $\vec{\mathbf{x}}'$, then rewriting (3.1.16) for the subsequence $\{\vec{x}_{m'}\}_{n=1}^{\infty}$ we have

$$\sum_{r=0}^{R} x_{m',r} \ w_{l+r,m'} \ = \ 0 \ . \tag{3.1.17}$$

Taking the limit of (3.1.17) we obtain

$$0 \ = \ \lim_{m' \to \infty} \sum_{r=0}^{R} x_{m',r} \ w_{l+r,m'} \ = \ \sum_{r=0}^{R} x'_r \ u_{l+r} \ , \tag{3.1.18}$$

that is the contradiction we were looking for since all the vectors $\{u_{l+r}\}_{r=0}^{R}$ are orthogonal to each other, so the only vector $\vec{\mathbf{x}}'$ that should satisfies (3.1.18) is the 0 vector, which is not a unit vector.
Since we have already proved above the existence of the inverse of Ψ for $H_n^{\max}$ small enough, what remains to prove is the existence of a constant $C_{\Psi^{-1}}$ such that for $H_n^{\max}$ small enough

$$\|\Psi_n^{-1}\|_\infty \ \leq \ C_{\Psi^{-1}} \ .$$

Suppose, seeking a contradiction, that there is a subsequence $\{H_m^{\max}\}$ of $\{H_n^{\max}\}$ such

that $\|\Psi_m^{-1}\|_\infty \to \infty$ as $m \to \infty$. This is equivalent to $\|\Psi_m^{-T}\|_1 \to \infty$ and by equivalence of norms on finite dimensional spaces (here the space of $(R+1)\times(R+1)$ matrices), it is in turn equivalent to $\|\Psi_m^{-T}\|_\infty \to \infty$. Thus there exists a sequence of vectors $\vec{\mathbf{v}}_m \in \mathbb{R}^{R+1}$ such that $\|\vec{\mathbf{v}}_m\|_\infty = 1$ for all m but

$$\lim_{m\to\infty} \|\vec{\mathbf{v}'}_m\|_\infty \;=\; \infty, \quad \text{where} \quad \vec{\mathbf{v}'}_m \;=\; \Psi_m^{-T}\vec{\mathbf{v}}_m \;.$$

Hence,

$$\lim_{m\to\infty} \left(\frac{\vec{\mathbf{v}}'_m}{\|\vec{\mathbf{v}}'_m\|_\infty}\right)^T \Psi_m \;=\; \lim_{m\to\infty} \frac{\vec{\mathbf{v}}_m}{\|\vec{\mathbf{v}}'_m\|_\infty} \;=\; \mathbf{0} \;. \tag{3.1.19}$$

Equation (3.1.19) also implies that

$$0 \;=\; \lim_{m\to\infty} \sum_{r=0}^{R} \frac{v'_{m,r}}{\|\vec{\mathbf{v}}'_m\|_\infty} \sum_{i=0}^{R} [\Psi_m]_{r,i}\; u_{l+i,m} \;=\; \lim_{m\to\infty} \sum_{r=0}^{R} \frac{v'_{m,r}}{\|\vec{\mathbf{v}}'_m\|_\infty}\; w_{l+r,m} \;, \tag{3.1.20}$$

where we denoted by $v'_{m,r}$ the r-component of the vector $\vec{\mathbf{v}}'_m$. Thanks to the properties of the limits and using the fact that for all i, $w_{l+i,m}$ converges to u_{l+i}, we obtain from (3.1.20):

$$0 \;=\; \sum_{r=0}^{R} \left(\lim_{m\to\infty} \frac{v'_{m,r}}{\|\vec{\mathbf{v}}'_m\|_\infty}\right) \left(\lim_{m\to\infty} w_{l+r,m}\right) \;=\; \sum_{r=0}^{R} \left(\lim_{m\to\infty} \frac{v'_{m,r}}{\|\vec{\mathbf{v}}'_m\|_\infty}\right) u_{l+r} \;. \tag{3.1.21}$$

Since all vectors $\{u_{l+r}\}_{r=0}^{R}$ are orthogonal to each other, (3.1.21) implies that for all r

$$\lim_{m\to\infty} \frac{v'_{m,r}}{\|\vec{\mathbf{v}}'_m\|_\infty} \;=\; 0,$$

which means that

$$\lim_{m\to\infty} \frac{\vec{\mathbf{v}}'_m}{\|\vec{\mathbf{v}}'_m\|_\infty} \;=\; 0 \;, \tag{3.1.22}$$

which is in contradiction with the fact that all vectors $\vec{\mathbf{v}}'_m/\|\vec{\mathbf{v}}'_m\|_\infty$ are constructed to be unit vectors in the infinity norm.

□

Now, it is time to introduce the main results of this section. The point of the next theorem is to show that for each n the computed eigenfunction $u_{l+i,n}$ is an approximation of a true eigenfunction of the continuous problem. Next theorem is an extension of the results in [51], since it holds also in the multiple eigenvalue case.

Theorem 3.1.4. *Let s be as given in Assumption 2.2.1, and let λ_l be an eigenvalue of multiplicity $R+1$ and let $(\lambda_{l+i,n}, u_{l+i,n})$ be computed eigenpairs spanning the computed eigenspace $E_{l,n}$, in the sense of Remark 2.2.4. Then, there exist true eigenfunctions U_{l+i} such that:*

$$\|U_{l+i} - u_{l+i,n}\|_{0,\mathcal{B},\Omega} \lesssim C_{\text{spec1}}(H_n^{\max})^{2s} , \tag{3.1.23}$$

and

$$a(U_{l+i} - u_{l+i,n}, U_{l+i} - u_{l+i,n})^{1/2} \lesssim C_{\text{spec2}}(H_n^{\max})^{s} , \tag{3.1.24}$$

where C_{spec1} and C_{spec2} are defined in Theorem 2.2.10.

Proof. In order to prove (3.1.23), we define $U_{l+i} = \sum_{r=0}^{R} [\Psi_n^{-1}]_{i,r} u_{l+r}$ and then we make use of Lemma 3.1.3 and Theorem 2.2.10(ii):

$$\begin{aligned}
\|U_{l+i} - u_{l+i,n}\|_{0,\mathcal{B},\Omega} &= \Big\| \sum_{r=0}^{R} [\Psi_n^{-1}]_{i,r} (u_{l+r} - w_{l+r,n}) \Big\|_{0,\mathcal{B},\Omega} \\
&\lesssim \left\| \Psi_n^{-1} \right\|_{\infty} \sum_{r=0}^{R} \|u_{l+r} - w_{l+r,n}\|_{0,\mathcal{B},\Omega} \\
&\lesssim C_{\Psi^{-1}}(R+1)\, C_{\text{spec1}}(H_n^{\max})^{2s} \lesssim C_{\text{spec1}}(H_n^{\max})^{2s} .
\end{aligned}$$

The result (3.1.24) is just a simple application of Lemma 2.2.11 and Theorem 2.2.10(i). □

Remark 3.1.5. *Note that each U_{l+i} in general depends on n.*

The next theorem extends a standard result for linear problems to eigenvalue problems:

Theorem 3.1.6. *Let s be as given in Assumption 2.2.1, and let λ_j be an eigenvalue of multiplicity $R+1$ and let $(\lambda_{j+i,n}, u_{j+i,n})$ be computed eigenpairs spanning the computed eigenspace $E_{j,n}$, in the sense of Remark 2.2.4. Then, there is a constant $C_{\text{adj}} > 0$ depending on the spectral information λ_l, E_l, $l = 1, \ldots, j$ such that:*

(i) let, for each $0 \le i \le R$, $w_{j+i,n}$ be as in Theorem 2.2.10, then we have:

$$\|u_{j+i} - w_{j+i,n}\|_{0,\mathcal{B},\Omega} \lesssim C_{\text{adj}}(H_n^{\max})^{s} a(u_{j+i} - w_{j+i,n}, u_{j+i} - w_{j+i,n})^{1/2}, \tag{3.1.25}$$

(ii) let U_{j+i} be as in Theorem 3.1.4 for $0 \le i \le R$, then we have:

$$\sum_{i=0}^{R} \|U_{j+i} - u_{j+i,n}\|_{0,\mathcal{B},\Omega} \lesssim C_{\text{adj}}(H_n^{\max})^{s} \sum_{i=0}^{R} a(U_{j+i} - u_{j+i,n}, U_{j+i} - u_{j+i,n})^{1/2}. \tag{3.1.26}$$

Proof. The proof of (3.1.25) is obtained by reworking the results in Chapter 2. Using the triangle inequality we have:

$$\|u_{j+i} - w_{j+i,n}\|_{0,\mathcal{B},\Omega} \le \|u_{j+i} - \beta_{j+i} w_{j+i,n}\|_{0,\mathcal{B},\Omega} + \|\beta_{j+i} w_{j+i,n} - w_{j+i,n}\|_{0,\mathcal{B},\Omega} , \tag{3.1.27}$$

where the value of the constant β_{j+i} is defined in the proof of Theorem 2.2.10. The second term on the right hand side of (3.1.27) can be treated as in the proof of Theorem 2.2.10 in order to obtain:

$$\|\beta_{j+i}w_{j+i,n} - w_{j+i,n}\|_{0,\mathcal{B},\Omega} \;\leq\; \|u_{j+i} - \beta_{j+i}w_{j+i,n}\|_{0,\mathcal{B},\Omega}\;. \tag{3.1.28}$$

Then, on the quantity $\|u_{j+i} - \beta_{j+i}w_{j+i,n}\|_{0,\mathcal{B},\Omega}$ appearing in both (3.1.27) and (3.1.28) it can be applied the same arguments as in Lemma 2.2.19 to have:

$$\|u_{j+i} - \beta_{j+i}w_{j+i,n}\|_{0,\mathcal{B},\Omega} \;\leq\; (1+\rho_{j+i})\|u_{j+i} - Q_n u_{j+i}\|_{0,\mathcal{B},\Omega}, \tag{3.1.29}$$

where ρ_{j+i} is defined within the proof of Lemma 2.2.19. Substituting (3.1.28) and (3.1.29) in (3.1.27) we get:

$$\|u_{j+i} - w_{j+i,n}\|_{0,\mathcal{B},\Omega} \;\lesssim\; (1+\rho_{j+i})\|u_{j+i} - Q_n u_{j+i}\|_{0,\mathcal{B},\Omega}\;. \tag{3.1.30}$$

The usual Aubin-Nitsche duality argument can be applied to obtain the L^2 convergence for $u_{j+i} - Q_n u_{j+i}$. Let us denote $e_{j+i,n} := u_{j+i} - Q_n u_{j+i}$ and let us define φ to be the solution of the linear problem

$$a(\varphi, w) \;=\; (e_{n,j+i}, w)_{0,\mathcal{B},\Omega}\;, \quad \text{for all} \quad w \in H_0^1(\Omega). \tag{3.1.31}$$

We have

$$\|e_{j+i,n}\|^2_{0,\mathcal{B},\Omega} = a(\varphi, e_{j+i,n}) = a(\varphi - v_n, e_{j+i,n})\;, \quad \text{for all} \quad v_n \in V_n,$$

where in the last step we used the orthogonality of $e_{j+i,n}$ to the space V_n. Then applying Cauchy-Schwarz we obtain

$$\|e_{j+i,n}\|^2_{0,\mathcal{B},\Omega} \lesssim\; |\varphi - v_n|_{1,\Omega}\; |e_{j+i,n}|_{1,\Omega}, \quad \text{for all} \quad v_n \in V_n. \tag{3.1.32}$$

Using Lemma 2.2.2 (together with the Assumption 2.2.1) in (3.1.32) we get

$$\begin{aligned}
\|e_{j+i,n}\|^2_{0,\mathcal{B},\Omega} &\lesssim\; C_{\text{app}}\;(H_n^{\max})^s |\varphi|_{1+s,\Omega} |e_{j+i,n}|_{1,\Omega} \\
&\leq\; C_{\text{app}}\; C_{\text{ell}} (H_n^{\max})^s \|\mathcal{B}e_{j+i,n}\|_{0,\Omega} |e_{j+i,n}|_{1,\Omega} \\
&\lesssim\; C_{\text{app}}\; C_{\text{ell}} (H_n^{\max})^s \|e_{j+i,n}\|_{0,\mathcal{B},\Omega} |e_{j+i,n}|_{1,\Omega}.
\end{aligned} \tag{3.1.33}$$

The last step of the argument consists of dividing both sides of (3.1.33) by $\|e_{j+i,n}\|_{0,\mathcal{B},\Omega}$ and applying the coercivity of the bilinear form $a(\cdot,\cdot)$

$$\|e_{j+i,n}\|_{0,\mathcal{B},\Omega} \lesssim\; C_{\text{app}} C_{\text{ell}} (H_n^{\max})^s a(e_{j+i,n}, e_{j+i,n})^{1/2}\;. \tag{3.1.34}$$

Combining (3.1.30) and (3.1.34) we obtain

$$\|u_{j+i} - w_{j+i,n}\|_{0,\mathcal{B},\Omega} \lesssim C_{\text{adj}}(H_n^{\max})^s a(u_{j+i} - Q_n u_{j+i}, u_{j+i} - Q_n u_{j+i})^{1/2} . \quad (3.1.35)$$

The result (3.1.25) comes from (3.1.35) noticing that $Q_n u_{j+i}$ is the best approximation of u_{j+i} in the energy norm, so $w_{j+i,n}$ should not be a better approximation than $Q_n u_{j+i}$. Now, we start to prove (3.1.26). Using properties of the matrix Ψ_n^{-1} as well as Lemma 3.1.3 we have:

$$\begin{aligned} \sum_{i=0}^{R} \|U_{j+i} - u_{j+i,n}\|_{0,\mathcal{B},\Omega} &\lesssim \sum_{i=0}^{R} \left\|\Psi_n^{-1}\right\|_\infty \sum_{r=0}^{R} \|u_{j+r} - w_{j+r,n}\|_{0,\mathcal{B},\Omega} \\ &\leq (R+1) C_{\Psi^{-1}} \sum_{r=0}^{R} \|u_{j+r} - w_{j+r,n}\|_{0,\mathcal{B},\Omega} . \end{aligned} \quad (3.1.36)$$

Then, using (3.1.25) on (3.1.36), we obtain:

$$\sum_{i=0}^{R} \|U_{j+i} - u_{j+i,n}\|_{0,\mathcal{B},\Omega} \lesssim \sum_{r=0}^{R} C_{\text{adj}}(H_n^{\max})^s a(u_{j+r} - w_{j+r,n}, u_{j+r} - w_{j+r,n})^{1/2} . \quad (3.1.37)$$

To conclude the proof of (3.1.26), it is just necessary to use the properties of the matrix Ψ_n and Lemma 3.1.1:

$$\begin{aligned} \sum_{i=0}^{R} \|U_{j+i} - u_{j+i,n}\|_{0,\mathcal{B},\Omega} &\lesssim (R+1) C_{\text{adj}}(H_n^{\max})^s \sum_{r=0}^{R} \left\|\Psi_n\right\|_\infty a(U_{j+r} - u_{j+r,n}, U_{j+r} - u_{j+r,n})^{1/2} \\ &\lesssim C_{\text{adj}}(H_n^{\max})^s \sum_{r=0}^{R} a(U_{j+r} - u_{j+r,n}, U_{j+r} - u_{j+r,n})^{1/2} . \end{aligned} \quad (3.1.38)$$

□

3.1.2 The PCF case

In analogy to what we have done above in Theorem 3.1.4, we have that also for problems (1.3.8) and (1.3.9) it is possible to prove that for each n the computed eigenfunction $u_{l+i,n}$ is an approximation to a true eigenfunction of the continuous problem.

Theorem 3.1.7. *Let s be as given in Assumption 2.2.20, and let λ_l be an eigenvalue of problem (1.3.8) with multiplicity $R+1$ and let $(\lambda_{l+i,n}, u_{l+i,n})$ be computed eigenpairs of problem* (2.2.48) *spanning the computed eigenspace $E_{l,n}^{\text{PCF}}$, in the sense of Remark 2.2.23. Then, there exist true eigenfunctions U_{l+i} of problem* (1.3.8) *such that:*

$$\|U_{j+i} - u_{j+i,n}\|_{0,\mathcal{B},\Omega} \lesssim C_{\text{spec1}}^{\text{PCF}}(H_n^{\max})^{2s} , \tag{3.1.39}$$

and

$$a_\kappa(U_{j+i} - u_{j+i,n}, U_{j+i} - u_{j+i,n})^{1/2} \lesssim C_{\text{spec2}}^{\text{PCF}}(H_n^{\max})^{s} . \tag{3.1.40}$$

where $C_{\text{spec1}}^{\text{PCF}}$ and $C_{\text{spec2}}^{\text{PCF}}$ are defined in Theorem 2.2.24.

Theorem 3.1.8. *Let s be as given in Assumption 2.2.20, and let ζ_l be an eigenvalue of problem* (1.3.9) *with multiplicity $R+1$ and let $(\zeta_{l+i,n}, u_{l+i,n})$ be computed eigenpairs of problem* (2.2.49) *spanning the computed eigenspace $E_{l,n}^{\text{PCF}}$, in the sense of Remark 2.2.23. Then, there exist true eigenfunctions U_{l+i} of problem* (1.3.9) *such that:*

$$\|U_{j+i} - u_{j+i,n}\|_{0,\mathcal{B},\Omega} \lesssim C_{\text{spec1}}^{\text{PCF}}(H_n^{\max})^{2s} , \tag{3.1.41}$$

and

$$a_{\kappa,S}(U_{j+i} - u_{j+i,n}, U_{j+i} - u_{j+i,n})^{1/2} \lesssim C_{\text{spec2}}^{\text{PCF}}(H_n^{\max})^{s} . \tag{3.1.42}$$

where $C_{\text{spec1}}^{\text{PCF}}$ and $C_{\text{spec2}}^{\text{PCF}}$ are defined in Theorem 2.2.24.

Theorem 3.1.9. *Let s be as given in Assumption 2.2.20, and let ζ_j be an eigenvalue of problem* (1.3.9) *with multiplicity $R+1$ and let $(\zeta_{j+i,n}, u_{j+i,n})$ be computed eigenpairs spanning the computed eigenspace $E_{j,n}^{\text{PCF}}$, in the sense of Remark 2.2.23. Then, there is a constant $C_{\text{adj}}^{\text{PCF}} > 0$ depending on the spectral information ζ_l, E_l^{PCF}, $l = 1, \ldots, j$ such that:*

(i) let, for each $0 \le i \le R$, $w_{j+i,n}$ be as in Theorem 2.2.24, then we have:

$$\|u_{j+i} - w_{j+i,n}\|_{0,\mathcal{B},\Omega} \lesssim C_{\text{adj}}^{\text{PCF}}(H_n^{\max})^{s} a_{\kappa,S}(u_{j+i} - w_{j+i,n}, u_{j+i} - w_{j+i,n})^{1/2}, \tag{3.1.43}$$

(ii) let U_{j+i} be as in Theorem 3.1.8 for $0 \le i \le R$, then we have:

$$\sum_{i=0}^{R} \|U_{j+i} - u_{j+i,n}\|_{0,\mathcal{B},\Omega} \lesssim C_{\text{adj}}^{\text{PCF}}(H_n^{\max})^{s} \sum_{i=0}^{R} a_{\kappa,S}(U_{j+i} - u_{j+i,n}, U_{j+i} - u_{j+i,n})^{1/2}. \tag{3.1.44}$$

3.2 Residual error estimators - the PCF case

In this section we define the "residual estimator" $\eta_{j,n}$ for the computed eigenpair $(\zeta_{j,n}, u_{j,n})$, which is computed on the mesh $\mathcal{T}_n$, for the shifted problem (1.3.9). We decided to start with the definition of residuals for the problem (1.3.9), because the residuals for all the other problems treated in this work are just particular cases of

the residuals for (1.3.9). In Section 3.4 we derive from $\eta_{j,n}$ other residual estimators suitable for other problems, namely: the unshifted problem (1.3.8), the TE and TM mode problems and for the general elliptic eigenvalue problem (1.3.7).
The residual estimator $\eta_{j,n}$ is defined as a sum of element residuals and edge (face) residuals, which are all computable quantities. To simplify the notation, we define the functional $[\cdot]_f$ as follow

Definition 3.2.1. *We can define for any function $g : \Omega \to \mathbb{C}$ and for any $f \in \mathcal{F}_n$*

$$[g]_f(x) := \left(\lim_{\substack{\tilde{x}\in\tau_1(f)\\ \tilde{x}\to x}} g(\tilde{x}) \;-\; \lim_{\substack{\tilde{x}\in\tau_2(f)\\ \tilde{x}\to x}} g(\tilde{x}) \right), \quad \textit{with } x \in f.$$

Definition 3.2.2 (Residual)**.** *The definition of the residual estimator $\eta_{j,n}$ involves two functionals: the functional $R_I(\cdot,\cdot)$, which expresses the contributions of the elements in the mesh:*

$$R_I(u,\zeta)(x) := \big((\nabla + i\vec{\kappa}) \cdot \mathcal{A}(\nabla + i\vec{\kappa})u \;+\; \zeta\mathcal{B}u\big)(x), \quad \textit{with } x \in \textit{int}(\tau), \quad \tau \in \mathcal{T}_n,$$

and the functional $R_F(\cdot)$, which expresses the contributions from the edges (faces) of the elements

$$R_F(u)(x) := \big[\vec{n}_f \cdot \mathcal{A}(\nabla + i\vec{\kappa})u\big]_f(x), \quad \textit{with } x \in \textit{int}(f), \quad f \in \mathcal{F}_n.$$

Then the residual estimator $\eta_{j,n}$ for the computed eigenpair $(\zeta_{j,n}, u_{j,n})$ is defined as:

$$\eta_{j,n} := \left\{ \sum_{\tau\in\mathcal{T}_n} H_\tau^2 \|R_I(u_{j,n}, \zeta_{j,n} - S)\|_{0,\tau}^2 \;+\; \sum_{f\in\mathcal{F}_n} H_f \|R_F(u_{j,n})\|_{0,f}^2 \right\}^{1/2}, \qquad (3.2.1)$$

where H_τ is the diameter of the element τ and H_f is the diameter of the edge (face) f.

3.3 Asymptotic reliability - the PCF case

In this section, we are going to prove asymptotic reliability of our error estimator for problem (1.3.9). So, in this section ζ_j is an eigenvalue of multiplicity $R+1$ of problem (1.3.9) for some value of $\vec{\kappa}$ and we denote by $(\zeta_{j+i,n}, u_{j+i,n})$ the computed eigenpairs for the same value of $\vec{\kappa}$ spanning the computed eigenspace $E_{j,n}^{\mathrm{PCF}}$ in the sense of Remark 2.2.23.
In Theorem 3.3.5 and Theorem 3.3.7 we prove the reliability of our error estimator for eigenfunctions and eigenvalues of problem (1.3.9). The main difference between the two results is the presence of $\sum_{i=0}^{R} \eta_{j+i,n}^2$ - in Theorem 3.3.7 - in the bound for the error for eigenvalues, instead of just $\sum_{i=0}^{R} \eta_{j+i,n}$, which appears in the bound for the error for eigenfunctions - in Theorem 3.3.5. This difference reflects the different rate

of convergence for eigenvalues and eigenfunctions that we have already encountered in the a priori analysis. Unsurprisingly we have recovered the same discrepancy in the rates of convergence also in the a posteriori analysis.
Furthermore, the terms $\sum_{i=0}^{R} G_{j+i,n}$ and $\sum_{i=0}^{R} G'_{j+i,n}$ in Theorem 3.3.5 and Theorem 3.3.7 should not go unnoticed. These terms, which do not appear in reliability results for linear problems, come from the non-linearity of the problem. In Section 3.4 we will show that these are asymptotically higher order terms and there is nothing to worry about them.
In order to prove reliability in Theorem 3.3.5 and Theorem 3.3.7, we need some preliminary lemmas:

Lemma 3.3.1. *Let $(\zeta_{j,n}, u_{j,n})$ be a calculated eigenpair of the discrete problem (2.2.49) for some value of the parameter $\vec{\kappa}$ and (ζ_j, u_j) be an eigenpair of the continuous problem (1.3.9) for the same value of $\vec{\kappa}$. Then denoting by $e_{j,n} := u_j - u_{j,n}$, we have*

$$(\zeta_j u_j - \zeta_{j,n} u_{j,n}, e_{j,n})_{0,\mathcal{B},\Omega} = \frac{1}{2}(\zeta_j + \zeta_{j,n})(e_{j,n}, e_{j,n})_{0,\mathcal{B},\Omega} \; + \; \mathrm{i}(\zeta_{j,n} - \zeta_j)\mathrm{Im}(u_j, u_{j,n})_{0,\mathcal{B},\Omega}. \tag{3.3.1}$$

Remark 3.3.2. *The result in this lemma holds even if the computed eigenpair $(\zeta_{j,n}, u_{j,n})$ does not converge to (ζ_j, u_j).*

Proof. Using the sesquilinearity of $(\cdot,\cdot)_{0,\mathcal{B},\Omega}$ and exploiting the fact that $(\zeta_{j,n}, u_{j,n})$ and (ζ_j, u_j) are respectively two normalized eigenpairs of (2.2.49) and of (1.3.9), we have:

$$\begin{aligned}
(\zeta_j u_j - \zeta_{j,n} u_{j,n}, e_{j,n})_{0,\mathcal{B},\Omega} &= (\zeta_j u_j - \zeta_{j,n} u_{j,n}, u_j)_{0,\mathcal{B},\Omega} \; - \; (\zeta_j u_j - \zeta_{j,n} u_{j,n}, u_{j,n})_{0,\mathcal{B},\Omega} \\
&= \zeta_j + \zeta_{j,n} \; - \; \zeta_{j,n}\overline{(u_j, u_{j,n})}_{0,\mathcal{B},\Omega} \; - \; \zeta_j (u_j, u_{j,n})_{0,\mathcal{B},\Omega} \\
&= (\zeta_j + \zeta_{j,n})(1 - \mathrm{Re}(u_j, u_{j,n})_{0,\mathcal{B},\Omega}) \\
&- \; i(\zeta_j - \zeta_{j,n})\mathrm{Im}(u_j, u_{j,n})_{0,\mathcal{B},\Omega}
\end{aligned} \tag{3.3.2}$$

Another use of sesquilinearity gives us:

$$\begin{aligned}
(e_{j,n}, e_{j,n})_{0,\mathcal{B},\Omega} &= (u_j, u_j)_{0,\mathcal{B},\Omega} + (u_{j,n}, u_{j,n})_{0,\mathcal{B},\Omega} \; - (u_j, u_{j,n})_{0,\mathcal{B},\Omega} - \overline{(u_j, u_{j,n})}_{0,\mathcal{B},\Omega} \\
&= 2 - 2\mathrm{Re}(u_j, u_{j,n})_{0,\mathcal{B},\Omega}.
\end{aligned} \tag{3.3.3}$$

The insertion of (3.3.3) into (3.3.2) concludes the proof.

□

Lemma 3.3.3. *Let $(\zeta_{j,n}, u_{j,n})$ be a computed eigenpair of problem (2.2.49) for some value of the parameter $\vec{\kappa}$ and (ζ_j, u_j) an eigenpair of problem (1.3.9) for the same value of $\vec{\kappa}$. Then, for any $v \in H^1_\pi(\Omega)$,*

$$\begin{aligned} a_{\kappa,S}(u_j - u_{j,n}, v) \;&= \sum_{\tau\in\mathcal{T}_n}\int_\tau R_I(u_{j,n}, \zeta_{j,n} - S)\overline{v} \;-\; \sum_{f\in\mathcal{F}_n}\int_f R_F(u_{j,n})\overline{v} \\ &\quad +(\zeta_j u_j - \zeta_{j,n}u_{j,n}, v)_{0,\mathcal{B},\Omega}. \end{aligned} \tag{3.3.4}$$

Remark 3.3.4. *Again, the result in this lemma holds even if the computed eigenpair $(\zeta_{j,n}, u_{j,n})$ does not converge to (ζ_j, u_j) in the sense of Remark 2.2.23.*

Proof. The equation (3.3.4) results from integration by parts. We start from the term on the left hand side of (3.3.4): using the fact that (ζ_j, u_j) is an eigenpair of (1.3.9) yields

$$\begin{aligned} a_{\kappa,S}(u_j - u_{j,n}, v) \;&=\; a_{\kappa,S}(u_j, v) \;-\; a_{\kappa,S}(u_{j,n}, v) \\ &=\; \zeta_j(u_j, v)_{0,\mathcal{B},\Omega} \;-\; a_{\kappa,S}(u_{j,n}, v). \end{aligned} \tag{3.3.5}$$

The first step in order to derive the right hand side of (3.3.4) is to apply element wise integration by parts to $a_\kappa(u_{j,n}, v)$, yielding:

$$\begin{aligned} a_\kappa(u_{j,n}, v) \;&=\; \sum_{\tau\in\mathcal{T}_n}\int_\tau \mathcal{A}(\nabla + i\vec{\kappa})u_{j,n}\cdot(\nabla - i\vec{\kappa})\overline{v} \\ &=\; -\sum_{\tau\in\mathcal{T}_n}\int_\tau \Big((\nabla + i\vec{\kappa})\cdot\mathcal{A}(\nabla + i\vec{\kappa})u_{j,n}\Big)\overline{v} \\ &+\; \sum_{f\in\mathcal{F}_n}\int_f [\vec{n}_f\cdot\mathcal{A}(\nabla + i\vec{\kappa})u_{j,n}]_f\;\overline{v}. \end{aligned} \tag{3.3.6}$$

The domain Ω, of problem (1.3.9), is a closed surface, i.e. it has no boundaries. So, in this case all the faces $f \in \mathcal{F}_n$ are within the domain.

Using the fact that $a_{\kappa,S}(\cdot,\cdot) := a_\kappa(\cdot,\cdot) + S(\cdot,\cdot)_{0,\mathcal{B},\Omega}$, then (3.3.6) and (3.3.5) yield

$$\begin{aligned} a_{\kappa,S}(u_j - u_{j,n}, v) &= -a_\kappa(u_{j,n}, v) \; - \; S(u_{j,n}, v)_{0,\mathcal{B},\Omega} \; + \; \zeta_j (u_j, v)_{0,\mathcal{B},\Omega} \\ &= \sum_{\tau \in \mathcal{T}_n} \int_\tau \Big((\nabla + i\vec{\kappa}) \cdot \mathcal{A}(\nabla + i\vec{\kappa}) u_{j,n} \Big) \overline{v} \\ &\quad - \sum_{f \in \mathcal{F}_n} \int_f [\vec{n}_f \cdot \mathcal{A}(\nabla + i\vec{\kappa}) u_{j,n}]_f \; \overline{v} \\ &\quad - S(u_{j,n}, v)_{0,\mathcal{B},\Omega} \; + \; \zeta_j (u_j, v)_{0,\mathcal{B},\Omega}. \end{aligned} \tag{3.3.7}$$

Finally we obtain (3.3.4) from (3.3.7) by noticing that $\zeta_j (u_j, v)_{0,\mathcal{B},\Omega} = \zeta_{j,n}(u_{j,n}, v)_{0,\mathcal{B},\Omega} + (\zeta_j u_j - \zeta_{j,n} u_{j,n}, v)_{0,\mathcal{B},\Omega}$ and then, splitting elementwise the two last linear terms on the right hand side of (3.3.7):

$$\begin{aligned} a_{\kappa,S}(u_j - u_{j,n}, v) &= \sum_{\tau \in \mathcal{T}_n} \left(\int_\tau (\nabla + i\vec{\kappa}) \cdot \mathcal{A}(\nabla + i\vec{\kappa}) u_{j,n} \; - \; S\mathcal{B} u_{j,n} \; + \; \zeta_{j,n} \mathcal{B} u_{j,n} \right) \overline{v} \\ &\quad - \sum_{F \in \mathcal{F}_n} \int_f n_f \cdot [\mathcal{A}(\nabla + i\vec{\kappa}) u_{j,n}]_f \; \overline{v} \\ &\quad + (\zeta_j u_j - \zeta_{j,n} u_{j,n}, v)_{0,\mathcal{B},\Omega}. \end{aligned}$$

□

The proof of reliability for eigenfunctions comes as an application of the previous lemmas. But before that, let us introduce the Scott-Zhang quasi-interpolation operator (see [48] for details). An important role in the next proof is played by this operator $I_n : H^1(\Omega) \to V_n$, which satisfies for any $v \in H^1(\Omega)$:

$$\|v - I_n v\|_{0,\tau} \lesssim H_\tau |v|_{1,\omega_\tau}, \tag{3.3.8}$$

$$\|v - I_n v\|_{0,f} \lesssim H_f^{\frac{1}{2}} |v|_{1,\omega_f}, \tag{3.3.9}$$

where ω_τ is the union of all the elements sharing at least a point with τ and where ω_f is the union of all the elements sharing at least a point with f. Since the nature of our problems, we restrict the use of the operator I_n to functions $v \in H^1_\pi(\Omega)$.

Theorem 3.3.5 (Asymptotic reliability for eigenfunctions)**.** *Let ζ_j be an eigenvalue of (1.3.9) of multiplicity $R+1$ and let $(\zeta_{j+i,n}, u_{j+i,n})$ be computed eigenpairs for the same value of $\vec{\kappa}$ spanning the computed eigenspace $E^{\mathrm{PCF}}_{j,n}$, in the sense of Remark 2.2.23.*

Let also the true eigenfunctions $U_{j+i} \in E_j^{\mathrm{PCF}}$, for $i = 0, \ldots, R$, be defined as in Theorem 3.1.8. Then we have for $e_{j+i,n} = U_{j+i} - u_{j+i,n}$, for $i = 0, \ldots, R$, that

$$\sum_{i=0}^{R} a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})^{1/2} \lesssim \sum_{i=0}^{R} \eta_{j+i,n} \; + \; \sum_{i=0}^{R} G_{j+i,n}, \tag{3.3.10}$$

where

$$G_{j+i,n} = \frac{1}{2}(\zeta_j + \zeta_{j+i,n}) \frac{(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}}{a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})^{1/2}}. \tag{3.3.11}$$

Remark 3.3.6. *In Theorem 3.4.1 in Section 3.4 we will prove that the terms $G_{j+i,n}$ are "higher order" (in a sense which will be made precise below).*

Proof. We are going to prove firstly that for all $i = 0, \ldots, R$:

$$a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})^{1/2} \lesssim \eta_{j+i,n} \; + \; G_{j+i,n}, \tag{3.3.12}$$

then in order to prove (3.3.10) it is just necessary to sum (3.3.12) over i.
Note first that, since (ζ_j, U_{j+i}) and $(\zeta_{j+i,n}, u_{j+i,n})$ respectively solve the eigenvalue problems (1.3.9) and (2.2.49), we have, for all $w_n \in V_n$,

$$\begin{aligned}
a_{\kappa,S}(e_{j+i,n}, e_{j+i,n}) &= a_{\kappa,S}(e_{j+i,n}, e_{j+i,n} - w_n) \; + \; a_{\kappa,S}(e_{j+i,n}, w_n) \\
&= a_{\kappa,S}(e_{j+i,n}, e_{j+i,n} - w_n) \; + \; a_{\kappa,S}(U_{j+i}, w_n) \; - \; a_{\kappa,S}(u_{j+i,n}, w_n) \\
&= a_{\kappa,S}(e_{j+i,n}, e_{j+i,n} - w_n) \; + \; (\zeta_j U_{j+i} - \zeta_{j+i,n} u_{j+i,n}, w_n)_{0,\mathcal{B},\Omega} \\
&= a_{\kappa,S}(e_{j+i,n}, e_{j+i,n} - w_n) \; - \; (\zeta_j U_{j+i} - \zeta_{j+i,n} u_{j+i,n}, e_{j+i,n} - w_n)_{0,\mathcal{B},\Omega} \\
&+ \; (\zeta_j U_{j+i} - \zeta_{j+i,n} u_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega} \; .
\end{aligned} \tag{3.3.13}$$

We will expand the first and the third terms on the right-hand side of (3.3.13) using Lemma 3.3.1 and Lemma 3.3.3, then the middle term will be cancelled out.
Using Lemma 3.3.3, we have for all $v \in H^1_\pi(\Omega)$,

$$\begin{aligned}
a_{\kappa,S}(e_{j+i,n}, v) &= \sum_{\tau \in \mathcal{T}_n} \int_\tau R_I(u_{j+i,n}, \zeta_{j+i,n} - S)\overline{v} \; - \; \sum_{f \in \mathcal{F}_n} \int_f R_F(u_{j+i,n})\overline{v} \\
&+ \; (\zeta_j U_{j+i} - \zeta_{j+i,n} u_{j+i,n}, v)_{0,\mathcal{B},\Omega} \; .
\end{aligned} \tag{3.3.14}$$

Hence for all $w_n \in V_n$,

$$\begin{aligned}
a_{\kappa,S}(e_{j+i,n}, e_{j+i,n} - w_n) &= \sum_{\tau \in \mathcal{T}_n} \int_\tau R_I(u_{j+i,n}, \zeta_{j+i,n} - S)\overline{(e_{j+i,n} - w_n)} \\
&- \; \sum_{f \in \mathcal{F}_n} \int_f R_F(u_{j+i,n})\overline{(e_{j+i,n} - w_n)} \\
&+ \; (\zeta_j U_{j+i} - \zeta_{j+i,n} u_{j+i,n}, e_{j+i,n} - w_n)_{0,\mathcal{B},\Omega}.
\end{aligned} \tag{3.3.15}$$

Moreover, from Lemma 3.3.1 we have

$$\begin{aligned}(\zeta_j U_{j+i} - \zeta_{j+i,n} u_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega} &= \frac{1}{2}(\zeta_j + \zeta_{j+i,n})(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega} \\ &\quad + \mathrm{i}(\zeta_{j+i,n} - \zeta_j)\mathrm{Im}(U_{j+i}, u_{j+i,n})_{0,\mathcal{B},\Omega}.\end{aligned} \tag{3.3.16}$$

Substituting (3.3.15) and (3.3.16) into (3.3.13), we obtain:

$$\begin{aligned}a_{\kappa,S}(e_{j+i,n}, e_{j+i,n}) &= \sum_{\tau\in\mathcal{T}_n}\int_\tau R_I(u_{j+i,n}, \zeta_{j+i,n} - S)\overline{(e_{j+i,n} - w_n)} \\ &- \sum_{f\in\mathcal{F}_n}\int_f R_F(u_{j+i,n})\overline{(e_{j+i,n} - w_n)} \\ &+ \frac{1}{2}(\zeta_j + \zeta_{j+i,n})(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega} \\ &+ \mathrm{i}(\zeta_{j+i,n} - \zeta_j)\mathrm{Im}(U_{j+i}, u_{j+i,n})_{0,\mathcal{B},\Omega}.\end{aligned} \tag{3.3.17}$$

Noticing that $a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})$, $\zeta_{j+i,n}$ and ζ_j are all real, we have $a_{\kappa,S}(e_{j+i,n}, e_{j+i,n}) \leq |\mathrm{Re}\ a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})|$ and applying the triangle inequality, yields

$$\begin{aligned}a_{\kappa,S}(e_{j+i,n}, e_{j+i,n}) &\leq \left|\sum_{\tau\in\mathcal{T}_n}\int_\tau R_I(u_{j+i,n}, \zeta_{j+i,n} - S)\overline{(e_{j+i,n} - w_n)}\right| \\ &+ \left|\sum_{f\in\mathcal{F}_n}\int_f R_F(u_{j+i,n})\overline{(e_{j+i,n} - w_n)}\right| \\ &+ \frac{1}{2}(\zeta_j + \zeta_{j+i,n})(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}.\end{aligned} \tag{3.3.18}$$

In particular we are allowed to choose $w_n = I_n e_{j+i,n}$ where I_n is the Scott-Zhang interpolation operator, defined above in (3.3.8) and (3.3.9).
Now substituting $w_n = I_n e_{j+i,n}$ in (3.3.18) and using Cauchy-Schwarz, together with

the inequalities (3.3.8) and (3.3.9), we obtain:

$$\begin{aligned} a_{\kappa,S}(e_{j+i,n},e_{j+i,n}) &\leq \sum_{\tau\in\mathcal{T}_n}\|R_I(u_{j+i,n},\zeta_{j+i,n}-S)\|_{0,\tau}\|e_{j+i,n}-I_ne_{j+i,n}\|_{0,\tau} \\ &+ \sum_{f\in\mathcal{F}_n}\|R_F(u_{j+i,n})\|_{0,f}\|e_{j+i,n}-I_ne_{j+i,n}\|_{0,f} \\ &+ \frac{1}{2}(\zeta_j+\zeta_{j+i,n})(e_{j+i,n},e_{j+i,n})_{0,\mathcal{B},\Omega} \\ &\lesssim \sum_{\tau\in\mathcal{T}_n}H_\tau\|R_I(u_{j+i,n},\zeta_{j+i,n}-S)\|_{0,\tau}|e_{j+i,n}|_{1,\omega_\tau} \\ &+ \sum_{f\in\mathcal{F}_n}H_f^{1/2}\|R_F(u_{j+i,n})\|_{0,f}|e_{j+i,n}|_{1,\omega_f} \\ &+ \frac{1}{2}(\zeta_j+\zeta_{j+i,n})(e_{j+i,n},e_{j+i,n})_{0,\mathcal{B},\Omega}. \end{aligned} \tag{3.3.19}$$

Furthermore, manipulating the weights of the 1-seminorm in (3.3.19) we obtain:

$$\begin{aligned} a_{\kappa,S}(e_{j+i,n},e_{j+i,n}) &\lesssim \sum_{\tau\in\mathcal{T}_n}H_\tau\|R_I(u_{j+i,n},\zeta_{j+i,n}-S)\|_{0,\tau}|e_{j+i,n}|_{1,\mathcal{A},\omega_\tau} \\ &+ \sum_{f\in\mathcal{F}_n}H_f^{1/2}\|R_F(u_{j+i,n})\|_{0,f}|e_{j+i,n}|_{1,\mathcal{A},\omega_f} \\ &+ \frac{1}{2}(\zeta_j+\zeta_{j+i,n})(e_{j+i,n},e_{j+i,n})_{0,\mathcal{B},\Omega} \end{aligned} \tag{3.3.20}$$

Another easy application of the discrete version of the Cauchy-Schwarz inequality yields

$$\begin{aligned} a_{\kappa,S}(e_{j+i,n},e_{j+i,n}) &\lesssim \eta_{j+i,n}\Big\{\sum_{\tau\in\mathcal{T}_n}|e_{j+i,n}|^2_{1,\mathcal{A},\omega_\tau} + \sum_{f\in\mathcal{F}_n}|e_{j+i,n}|^2_{1,\mathcal{A},\omega_f}\Big\}^{1/2} \\ &+\frac{1}{2}(\zeta_j+\zeta_{j+i,n})(e_{j+i,n},e_{j+i,n})_{0,\mathcal{B},\Omega} \\ &\lesssim \eta_{j+i,n}|e_{j+i,n}|_{1,\mathcal{A},\Omega} + \frac{1}{2}(\zeta_j+\zeta_{j+i,n})(e_{j+i,n},e_{j+i,n})_{0,\mathcal{B},\Omega}\ . \end{aligned} \tag{3.3.21}$$

Now to complete the treatment of the terms in (3.3.21), we can use Theorem 2.1.12 to get:

$$a_{\kappa,S}(e_{j+i,n},e_{j+i,n}) \lesssim \eta_{j+i,n}a_{\kappa,S}(e_{j+i,n},e_{j+i,n})^{1/2} + \frac{1}{2}(\zeta_j+\zeta_{j+i,n})(e_{j+i,n},e_{j+i,n})_{0,\mathcal{B},\Omega}. \tag{3.3.22}$$

Finally, in order to conclude the proof we have just to divide both sides of (3.3.22) by $a_{\kappa,S}(e_{j+i,n},e_{j+i,n})^{1/2}$ and sum over i.

□

The last result of this section is the asymptotic reliability for eigenvalues.

Theorem 3.3.7 (Asymptotic reliability for eigenvalues)**.** *Under the same assumptions as in Theorem 3.3.5 and denoting by $e_{j+i,n} = U_{j+i} - u_{j+i,n}$, we have:*

$$\sum_{i=0}^{R} |\zeta_{j+i,n} - \zeta_j| \lesssim \sum_{i=0}^{R} \eta_{j+i,n}^2 \ + \ \sum_{i=0}^{R} G'_{j+i,n},$$

where

$$G'_{j+i,n} = \eta_{j+i,n} \frac{1}{2}(\zeta_j + \zeta_{j+i,n}) \frac{(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}}{a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})^{1/2}} \ + \ \frac{1}{2}(\zeta_j - \zeta_{j+i,n})(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}.$$

Remark 3.3.8. *In Theorem 3.4.2 in Section 3.4 we will prove that the terms $G'_{j+i,n}$ are also "higher order".*

Proof. In Lemma 2.2.26 we have shown that

$$|\zeta_{j+i,n} - \zeta_j| = a_{\kappa,S}(e_{j+i,n}, e_{j+i,n}) \ - \ \zeta_j (e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}. \tag{3.3.23}$$

Hence, for any $i = 0, \ldots, R$, substituting (3.3.12) twice in (3.3.23) leads to the result:

$$\begin{aligned}
|\zeta_{j+i,n} - \zeta_j| \ &\lesssim \eta_{j+i,n} a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})^{1/2} \ + \ \frac{1}{2}(\zeta_{j+i,n} + \zeta_j)(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega} \\
&\quad - \zeta_j (e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega} \\
&\lesssim \eta_{j+i,n} a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})^{1/2} \ + \ \frac{1}{2}(\zeta_{j+i,n} - \zeta_j)(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega} \\
&\lesssim \eta_{j+i,n}^2 \ + \ \eta_{j+i,n} \frac{1}{2}(\zeta_{j+i,n} + \zeta_j) \frac{(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}}{a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})^{1/2}} \\
&\quad + \frac{1}{2}(\zeta_{j+i,n} - \zeta_j)(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}.
\end{aligned}$$

Then the proof is concluded summing over i. □

3.4 Further asymptotic reliability results

In this section we have collected other asymptotic reliability results. Some of them are related to the TE and TM mode problems, while others are related to the general elliptic eigenvalue problem (1.3.7). The first two theorems show that the terms $G_{j+i,n}$

in Theorem 3.3.5 and the terms $G'_{j+i,n}$ in Theorem 3.3.7 are asymptotically higher order terms.
In this section we assume that the a priori upper bounds proved in Theorem 2.2.33 and in Theorem 2.2.10 are sharp. With $e_{j+i,n} = U_{j+i} - u_{j+i,n}$, we see from (3.1.42) that $a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})^{1/2} = \mathcal{O}\big((H_n^{\max})^s\big)$, where $0 < s \leq 1$. What we want to prove now is that the asymptotic order of $G_{j+i,n}$ is greater than s for all $i = 0, \ldots, R$, i.e. $G_{j+i,n}$ is a higher order term. Moreover, if, for all $i = 0, \ldots, R$, $G_{j+i,n}$ is a higher order term, from the inequality (3.3.10) it is possible to conclude that each $\eta_{j+i,n}$ should have at least the same asymptotic order as the energy norm of the error $a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})^{1/2}$.

Theorem 3.4.1. *Let $(\zeta_{j+i,n}, u_{j+i,n})$ be a calculated eigenpair of the discrete problem (2.2.49) for some value of $\vec{\kappa}$ and let (ζ_j, U_{j+i}) be the corresponding true eigenpair of the problem (1.3.9). Then we have that the term $G_{j+i,n}$ in Theorem 3.3.5 has higher order with respect to the energy norm of the error:*

$$G_{j+i,n} = \mathcal{O}\big((H_n^{\max})^{2s}\big).$$

Proof. We start from the definition of $G_{j+i,n}$ given in Theorem 3.3.5, then using Theorem 2.1.12, we have

$$\begin{aligned} G_{j+i,n} &= \frac{1}{2}(\zeta_j + \zeta_{j+i,n}) \frac{(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}}{a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})^{1/2}} \\ &\lesssim \frac{1}{2}(\zeta_j + \zeta_{j+i,n})(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}^{1/2}. \end{aligned}$$

Since, from (3.1.39), we have that $(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}^{1/2} = \mathcal{O}(H_n^{\max})^{2s}$, then

$$G_{j+i,n} \lesssim C_{\text{spec1}}^{\text{PCF}} \frac{1}{2}(\zeta_j + \zeta_{j+i,n}) \, (H_n^{\max})^{2s}.$$

□

Form (2.2.53) we know that $|\zeta_{j+i,n} - \zeta_j| = \mathcal{O}(H_n^{\max})^{2s}$, where $0 < s \leq 1$. What we want to prove is that the term $G'_{j+i,n}$ appearing in Theorem 3.3.7 is $\mathcal{O}\big((H_n^{\max})^{2s}\big)$. In the following theorem we do even better.

Theorem 3.4.2. *Let $(\zeta_{j+i,n}, u_{j+i,n})$ be a calculated eigenpair of the discrete problems (2.2.49) for some value of $\vec{\kappa}$ and let (ζ_j, U_{j+i}) be the corresponding true eigenpair for the same value of $\vec{\kappa}$. Then we have that the term $G'_{j+i,n}$ in Theorem 3.3.7 has higher order than the error of the eigenvalues:*

$$G'_{j+i,n} = \mathcal{O}(H_n^{\max})^{3s}.$$

Proof. We start from the definition of $G'_{j+i,n}$ and using Theorem 2.1.12, we have

$$\begin{aligned} G'_{j+i,n} &= \eta_{j+i,n}\frac{1}{2}(\zeta_j+\zeta_{j+i,n})\frac{(e_{j+i,n},e_{j+i,n})_{0,\mathcal{B},\Omega}}{a_{\kappa,S}(e_{j+i,n},e_{j+i,n})^{1/2}} \;+\; \frac{1}{2}(\zeta_j-\zeta_{j+i,n})(e_{j+i,n},e_{j+i,n})_{0,\mathcal{B},\Omega} \\ &\lesssim \eta_{j+i,n}\frac{1}{2}(\zeta_j+\zeta_{j+i,n})(e_{j+i,n},e_{j+i,n})^{1/2}_{0,\mathcal{B},\Omega} \;+\; \frac{1}{2}(\zeta_j-\zeta_{j+i,n})(e_{j+i,n},e_{j+i,n})_{0,\mathcal{B},\Omega}. \end{aligned} \tag{3.4.1}$$

Using (2.2.53) and (3.1.39) on the right hand side of (3.4.1) we obtain

$$G'_{j+i,n} \;\lesssim\; \eta_{j+i,n}\frac{1}{2}C^{\mathrm{PCF}}_{\mathrm{spec1}}(\zeta_j+\zeta_{j+i,n})(H_n^{\max})^{2s} \;+\; \frac{1}{2}(C^{\mathrm{PCF}}_{\mathrm{spec1}})^2\;(H_n^{\max})^{6s},$$

since from Theorem 3.4.1 and from (2.2.55) we know that $\eta_{j+i,n}$ has at least order $(H_n^{\max})^s$, we conclude that $G'_{j+i,n}$ has at least order $3s$. □

Now, we move to prove the asymptotic reliability result for the un-shifted problem (1.3.8). The difference between the problem (1.3.8) and the problem (1.3.9) is the linear term $S(u,v)_{0,\mathcal{B},\Omega}$. This term introduces a shift in the spectrum of the problem, but it has no effect on the eigenfunctions. So, for any eigenvalue ζ_j of (1.3.9), there is a corresponding eigenvalue $\lambda_j = \zeta_j - S$ of (1.3.8). The same happens to the eigenvalues of (2.2.48) and (2.2.49), i.e. $\lambda_{j,n} = \zeta_{j,n} - S$. Moreover, for any function $u \in H^1_\pi(\Omega)$ and for some value of $S > 0$, the bilinear form $a_{\kappa,S}(u,u) \geq a_\kappa(u,u)$.
Theorem 3.3.5 and Theorem 3.3.7 can be easily adapted as follows to the un-shifted problem:

Theorem 3.4.3 (Asymptotic reliability for eigenfunctions)**.** *Let λ_j be an eigenvalue of (1.3.8) of multiplicity $R+1$ and let $(\lambda_{j+i,n}, u_{j+i,n})$ be computed eigenpairs for the same value of $\vec{\kappa}$ forming the computed eigenspace $E^{\mathrm{PCF}}_{j,n}$, in the sense of Remark 2.2.23. Let also the true eigenfunctions $U_{j+i} \in E^{\mathrm{PCF}}_j$, for $i = 0,\dots,R$, be defined in Theorem 3.1.7. Then we have for $e_{j+i,n} = U_{j+i} - u_{j+i,n}$, for $i = 0,\dots,R$, that*

$$\sum_{i=0}^{R} a_\kappa(e_{j+i,n},e_{j+i,n})^{1/2} \;\lesssim\; \sum_{i=0}^{R}\eta_{j+i,n} \;+\; \sum_{i=0}^{R} D_{j+i,n}, \tag{3.4.2}$$

where

$$D_{j+i,n} \;=\; \frac{1}{2}(\lambda_j+\lambda_{j+i,n}+2S)(e_{j+i,n},e_{j+i,n})^{1/2}_{0,\mathcal{B},\Omega}.$$

Proof. For any value of $S > 0$ we have that $a_\kappa(e_{j+i,n},e_{j+i,n}) \leq a_{\kappa,S}(e_{j+i,n},e_{j+i,n})$. So applying (3.3.10) we obtain

$$\sum_{i=0}^{R} a_\kappa(e_{j+i,n},e_{j+i,n})^{1/2} \;\leq\; \sum_{i=0}^{R} a_{\kappa,S}(e_{j+i,n},e_{j+I,n})^{1/2} \;\lesssim\; \sum_{i=0}^{R}\eta_{j+i,n} \;+\; \sum_{i=0}^{R} G_{j+i,n}. \tag{3.4.3}$$

Moreover, the computed value of the residual R_I is not changed by the shift because:

$$\begin{aligned} R_I(u_{j+i,n}, \zeta_{j+i,n} - S)(x) &:= ((\nabla + i\vec{\kappa}) \cdot \mathcal{A}(\nabla + i\vec{\kappa})u_{j+i,n} \ - \ S\mathcal{B}u_{j+i,n} \ + \ \zeta_{j+i,n}\mathcal{B}u_{j+i,n})(x) \\ &= ((\nabla + i\vec{\kappa}) \cdot \mathcal{A}(\nabla + i\vec{\kappa})u_{j+i,n} \ + \ \lambda_{j+i,n}\mathcal{B}u_{j+i,n})(x) \\ &=: R_I(u_{j+i,n}, \lambda_{j+i,n})(x). \end{aligned}$$

The residual R_F is also unaffected by the shift because, in its case, its value does not depend on the computed eigenvalue. So, we can conclude that the computed value of the residual estimator $\eta_{j+i,n}$ is unaffected by the value of the shift $S > 0$.
The term $D_{j+i,n}$ comes from the term $G_{j+i,n}$, to which we apply Theorem 2.1.12 and we undo the shift:

$$\begin{aligned} G_{j+i,n} &:= \frac{1}{2}(\zeta_j + \zeta_{j+i,n})\frac{(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}}{a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})^{1/2}} \\ &= \frac{1}{2}(\lambda_j + \lambda_{j+i,n} + 2S)\frac{(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}}{a_{\kappa,S}(e_{j+i,n}, e_{j+i,n})^{1/2}} \\ &\lesssim \frac{1}{2}(\lambda_j + \lambda_{j+i,n} + 2S)(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}^{1/2} \ =: \ D_{j+i,n}. \end{aligned}$$

□

Theorem 3.4.4 (Asymptotic reliability for eigenvalues)**.** *Under the same assumptions as Theorem 3.4.3 we have:*

$$\sum_{i=0}^{R} |\lambda_{j+i,n} - \lambda_j| \ \lesssim \ \sum_{i=0}^{R} \eta_{j+i,n}^2 \ + \ \sum_{i=0}^{R} D'_{j+i,n},$$

where we have denoting by $e_{j+i,n} = U_{j+i} - u_{j+i,n}$ that:

$$D'_{j+i,n} \ = \ \eta_{j+i,n} \ \frac{1}{2}(\lambda_j + \lambda_{j+i,n} + 2S)(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}^{1/2} \ + \ \frac{1}{2}(\lambda_j - \lambda_{j+i,n})(e_{j+i,n}, e_{j+i,n})_{0,\mathcal{B},\Omega}.$$

Proof. Applying Theorem 3.3.7 and noticing that $\lambda_{j+i,n} - \lambda_j = \zeta_{j+i,n} - \zeta_j$, we have:

$$\sum_{i=0}^{R} |\lambda_{j+i,n} - \lambda_j| \ \lesssim \ \sum_{i=0}^{R} \eta_{j+i,n}^2 + \sum_{i=0}^{R} G'_{j+i,n}.$$

We have already seen in Theorem 3.4.3 that the residual estimator $\eta_{j+i,n}$ is unaffected by the shift. What remains to show is what happens to the term $G'_{j+i,n}$ shifting back

the problem:

$$
\begin{aligned}
G'_{j+i,n} &:= \eta_{j+i,n}\,\frac{1}{2}(\zeta_j+\zeta_{j+i,n})\frac{(e_{j+i,n},e_{j+i,n})_{0,\mathcal{B},\Omega}}{a_{\kappa,S}(e_{j+i,n},e_{j+i,n})^{1/2}} \;+\; \frac{1}{2}(\zeta_j-\zeta_{j+i,n})(e_{j+i,n},e_{j+i,n})_{0,\mathcal{B},\Omega} \\
&\lesssim \eta_{j+i,n}\,\frac{1}{2}(\lambda_j+\lambda_{j+i,n}+2S)(e_{j+i,n},e_{j+i,n})^{1/2}_{0,\mathcal{B},\Omega} \;+\; \frac{1}{2}(\lambda_j-\lambda_{j+i,n})(e_{j+i,n},e_{j+i,n})_{0,\mathcal{B},\Omega} \\
&=: D'_{j+i,n},
\end{aligned}
$$

where we have made use of Theorem 2.1.12. □

Remark 3.4.5. *The terms $D_{j+i,n}$ and $D'_{j+i,n}$ are higher order terms from the same arguments used for $G_{j+i,n}$ and $G'_{j+i,n}$ - Theorem 3.4.1 and Theorem 3.4.2.*

Remark 3.4.6. *The TE and TM mode problems are particular cases of problem (1.3.8): in the TE case we have that $\mathcal{B}=1$, instead in the TM case $\mathcal{A}=1$. So the asymptotic reliability result is applicable to the TE and to the TM mode problems, too.*

Remark 3.4.7. *The proof of asymptotic reliability for the general elliptic problem (1.3.7) is not more involved. This problem has Dirichlet boundary conditions, so the bilinear form $a(\cdot,\cdot)$ is already coercive. Then we do not need to introduce a shift. This implies that the reliability result for (1.3.7) comes from Theorem 3.3.5 and Theorem 3.3.7 (with $\vec{\kappa}=(0,0)$), as before, but this time we are allowed to choose $S=0$.*

3.5 Asymptotic efficiency - the PCF case

This section contains the proof of asymptotic efficiency for our residual estimator applied to the unshifted problem (1.3.8) (the same proof holds also for the general elliptic problem (1.3.7)). We are not going to prove asymptotic efficiency for the shifted problem (1.3.9) because it does not come from a physical model. It was introduced in the first place just to let us prove easily reliability.
The asymptotic efficiency guarantees that the residual estimator is not asymptotically unreasonably greater than the actual error. In order to prove the efficiency, we need first a weaker result called "local efficiency". Then the asymptotic efficiency will be proved in Theorem 3.5.6. The same approach has been used in [52] and in [53].

Notation 3.5.1. *In this section we extend the Notation 3.0.34 in such a way that $\lesssim$ and $\gtrsim$ will hide constants depending also on H_τ and H_f only under the condition that such constants will remain bounded above and below when H_τ and H_f go to 0. So, we have e.g. $1+H_\tau \lesssim 1$.*

In this section we are going to use bubble functions, which are in general smooth and positive real valued functions with compact supports and bounded by 1 in the L^∞ norm. The proof of efficiency for a posteriori error estimators is normally carried out with bubble functions, which have many useful characteristics. Firstly, these functions have local support, so it is possible to define a bubble function on each element and on each edge in the mesh. This will reduce the proof of efficiency from the whole mesh to a local result. Furthermore, it is possible to prove inverse estimates for bubble functions of standard results involving norms, thanks to their regularity. These estimates are collected in the next proposition. We define for any edge (face) f the set Δ_f, which is the union of the two elements sharing f. In particular we need for any element τ a real-valued bubble function ψ_τ with support in τ which vanishes on the edge of τ and for any edge f, and we need a real-valued bubble function ψ_f that vanishes outside the closure of Δ_f. In [52, Lemma 3.3], such bubble functions ψ_τ, ψ_f are constructed using polynomials. Moreover, it is proven that ψ_τ, ψ_f satisfy the following properties:

Proposition 3.5.2. *There are constants, which only depend on the regularity of the mesh $\mathcal{T}_n$, such that the inequalities on an element τ*

$$\|v\|_{0,\tau} \lesssim \|\psi_\tau^{1/2} v\|_{0,\tau}, \tag{3.5.1}$$

$$|\psi_\tau v|_{1,\tau} \lesssim H_\tau^{-1} \|v\|_{0,\tau}, \tag{3.5.2}$$

and on a edge (face) f

$$\|\omega\|_{0,f} \lesssim \|\psi_f^{1/2}\, \omega\|_{0,f}, \tag{3.5.3}$$

$$|\psi_f\, \omega|_{1,\Delta_f} \lesssim H_f^{-1/2} \|\omega\|_{0,f}, \tag{3.5.4}$$

$$\|\psi_f\, \omega\|_{0,\Delta_f} \lesssim H_f^{1/2} \|\omega\|_{0,f}, \tag{3.5.5}$$

hold for all $\tau \in \mathcal{T}_n$, all $f \in \mathcal{F}_n$, for all polynomials v and for all polynomials ω.

Proof. See [52, Lemma 3.3]. □

In the next two lemmas we bound the residuals R_I and R_F (defined in Definition 3.2.2 above) in terms of the energy norm of the error.

Lemma 3.5.3. *Let $(\lambda_{j,n}, u_{j,n})$ be a computed eigenpair on $\mathcal{T}_n$ of (2.2.48) for some value of $\vec{\kappa}$ and (λ_j, u_j) be a true eigenpair of (1.3.8) for the same value of $\vec{\kappa}$, then for any element $\tau \in \mathcal{T}_n$ we have*

$$\begin{aligned} H_\tau \|R_I(u_{j,n}, \lambda_{j,n})\|_{0,\tau} \lesssim\ & \|\mathcal{A}^{1/2}(\nabla + i\vec{\kappa})(u_j - u_{j,n})\|_{0,\tau} \\ & + H_\tau \|\lambda_{j,n} u_{j,n} - \lambda_j u_j\|_{0,\mathcal{B},\tau}. \end{aligned} \tag{3.5.6}$$

Proof. Let ψ_τ be the real-valued bubble function introduced above and set

$$w_\tau \;=\; \psi_\tau \; R_I(u_{j,n}, \lambda_{j,n}).$$

Because we are using P_1 elements and since $\mathcal{A}$, $\mathcal{B}$ are assumed constant in the interior of each element, the residual R_I is a polynomial function on τ. This fact together with (3.5.1) leads to

$$\|R_I(u_{j,n}, \lambda_{j,n})\|^2_{0,\tau} \;\lesssim\; \|\psi_\tau^{1/2} R_I(u_{j,n}, \lambda_{j,n})\|^2_{0,\tau},$$

hence by the positivity of ψ_τ:

$$\|R_I(u_{j,n}, \lambda_{j,n})\|^2_{0,\tau} \lesssim \int_\tau \psi_\tau |R_I(u_{j,n}, \lambda_{j,n})|^2 = \int_\tau R_I(u_{j,n}, \lambda_{j,n}) \overline{w}_\tau \tag{3.5.7}$$

Since supp $\psi_\tau = \overline{\tau}$, we can integrate by parts the right hand side of (3.5.7), using the fact that ψ_τ vanishes on $\partial\tau$, to get

$$\begin{aligned} \|R_I(u_{j,n}, \lambda_{j,n})\|^2_{0,\tau} \;&\lesssim\; \int_\tau \big((\nabla + i\vec{\kappa}) \cdot \mathcal{A}(\nabla + i\vec{\kappa}) u_{j,n} \;+\; \lambda_{j,n}\, \mathcal{B}\, u_{j,n}\big)\, \overline{w}_\tau \\ &=\; \big(- a_\kappa(u_{j,n}, w_\tau) \;+\; \lambda_{j,n}(u_{j,n}, w_\tau)_{0,\mathcal{B},\tau}\big). \end{aligned} \tag{3.5.8}$$

Because we have supposed that (λ_j, u_j) is an eigenpair of the continuous problem (1.3.8), it satisfies:

$$a_\kappa(u_j, w_\tau) \;=\; \lambda_j (u_j, w_\tau)_{0,\mathcal{B},\Omega}. \tag{3.5.9}$$

Then adding (3.5.9) to (3.5.8) and noticing that supp $\psi_\tau = \overline{\tau}$ we have

$$\|R_I(u_{j,n}, \lambda_{j,n})\|^2_{0,\tau} \;\lesssim\; \left[-\int_\tau \big((\nabla + i\vec{\kappa})(u_j - u_{j,n}) \cdot \mathcal{A}(\nabla - i\vec{\kappa})\overline{w_\tau}\big) + (\lambda_{j,n} u_{j,n} - \lambda_j u_j, w_\tau)_{0,\mathcal{B},\tau} \right].$$

Hence by the Cauchy-Schwarz inequality and applying (2.1.16) yields:

$$\begin{aligned} \|R_I(u_{j,n}, \lambda_{j,n})\|^2_{0,\tau} \;&\lesssim\; \big\|\mathcal{A}^{1/2}(\nabla + i\vec{\kappa})(u_j - u_{j,n})\big\|_{0,\tau} \, \big\|\mathcal{A}^{1/2}(\nabla - i\vec{\kappa})\overline{w}_\tau\big\|_{0,\tau} \\ &+\; \|\lambda_{j,n} u_{j,n} - \lambda_j u_j\|_{0,\mathcal{B},\tau} \, \|w_\tau\|_{0,\mathcal{B},\tau} \\ &\lesssim\; \big\|\mathcal{A}^{1/2}(\nabla + i\vec{\kappa})(u_j - u_{j,n})\big\|_{0,\tau} \, \|w_\tau\|_{1,\tau} \\ &+\; \|\lambda_{j,n} u_{j,n} - \lambda_j u_j\|_{0,\mathcal{B},\tau} \, \|w_\tau\|_{0,\mathcal{B},\tau}. \end{aligned} \tag{3.5.10}$$

The last step of the proof is quite straightforward: using the definition of w_τ and using

(3.5.2), then we obtain from (3.5.10):

$$\|R_I(u_{j,n},\lambda_{j,n})\|_{0,\tau}^2 \lesssim \Big[(1+H_\tau^{-1})\ \big\|\mathcal{A}^{1/2}(\nabla+i\vec{\kappa})(u_j-u_{j,n})\big\|_{0,\tau} + \|\lambda_{j,n}u_{j,n}-\lambda_j u_j\|_{0,\mathcal{B},\tau}\Big]\|R_I(u_{j,n},\lambda_{j,n})\|_{0,\tau},$$

then multiplying each side by $H_\tau\|R_I(u_{j,n},\lambda_{j,n})\|_{0,\tau}^{-1}$ yields the result.

□

Lemma 3.5.4. *Let $(\lambda_{j,n},u_{j,n})$ be a computed eigenpair on $\mathcal{T}_n$ of (2.2.48) for some value of $\vec{\kappa}$ and (λ_j,u_j) be a true eigenpair of (1.3.8) for the same value of $\vec{\kappa}$, then we have for any face f in $\mathcal{F}_n$*

$$\begin{aligned} H_f^{1/2}\|R_F(u_{j,n})\|_{0,f} &\lesssim \textstyle\sum_{\tau\in\Delta_f}\ \|\mathcal{A}^{1/2}(\nabla+i\vec{\kappa})(u_j-u_{j,n})\|_{0,\tau} \\ &+ \textstyle\sum_{\tau\in\Delta_f} H_f\ \|\lambda_{j,n}u_{j,n}-\lambda_j u_j\|_{0,\mathcal{B},\tau}. \end{aligned} \tag{3.5.11}$$

Proof. Let ψ_f be the real-valued bubble function introduced above and set

$$w_f := \psi_f\ R_F(u_{j,n}).$$

Applying Lemma 3.3.3 to problem (1.3.8), i.e. choosing $S=0$ in Lemma 3.3.3, and also exploiting the fact that supp $\psi_f = \overline{\Delta}_f$, we obtain

$$\begin{aligned} \int_f R_F(u_{j,n})\overline{w}_f &= \sum_{\tau\in\Delta_f}\int_\tau R_I(u_{j,n},\lambda_{j,n})\overline{w}_f \ -\ a_\kappa(u_j-u_{j,n},w_f) \\ &+ (\lambda_j u_j-\lambda_{j,n}u_{j,n},w_f)_{0,\mathcal{B},\Omega}\ . \end{aligned} \tag{3.5.12}$$

Then using the Cauchy-Schwarz inequality and (3.5.3) on (3.5.12), we get:

$$\begin{aligned} \|R_F(u_{j,n})\|_{0,f}^2 &\lesssim \sum_{\tau\in\Delta_f}\|R_I(u_{j,n},\lambda_{j,n})\|_{0,\tau}\ \|w_f\|_{0,\tau} \\ &+ \|\mathcal{A}^{1/2}(\nabla+i\vec{\kappa})(u_j-u_{j,n})\|_{0,\Delta_f}\ \|\mathcal{A}^{1/2}(\nabla+i\vec{\kappa})w_f\|_{0,\Delta_f} \\ &+ \|\lambda_{j,n}u_{j,n}-\lambda_j u_j\|_{0,\mathcal{B},\Delta_f}\ \|w_f\|_{0,\mathcal{B},\Delta_f}\ . \end{aligned} \tag{3.5.13}$$

Now, we have to estimate each of the three terms on the right-hand side of (3.5.13). We start from the sum at the beginning of the right hand side of (3.5.13): this sum

can be treated using (3.5.5) and (3.5.6)

$$\begin{aligned}\sum_{\tau\in\Delta_f}\|R_I(u_{j,n},\lambda_{j,n})\|_{0,\tau}\ \|w_f\|_{0,\tau} \ &\lesssim\ H_f^{1/2}\ \sum_{\tau\in\Delta_f}\|R_I(u_{j,n},\lambda_{j,n})\|_{0,\tau}\ \|R_F(u_{j,n})\|_{0,f}\\ &\lesssim\ H_f^{1/2}\ \|R_F(u_{j,n})\|_{0,f}\ \sum_{\tau\in\Delta_f}H_\tau^{-1}\Big(\|\mathcal{A}^{1/2}(\nabla+i\vec{\kappa})(u_j-u_{j,n})\|_{0,\tau}\\ &\quad+\ H_\tau\ \|\lambda_{j,n}u_{j,n}-\lambda_j u_j\|_{0,\mathcal{B},\tau}\Big).\end{aligned}\tag{3.5.14}$$

Now it is the turn for the second term on the right hand side of (3.5.13). We are interested just in the component $\|\mathcal{A}^{1/2}(\nabla+i\vec{\kappa})w_f\|_{0,\Delta_f}$ of this term. On this component we can use (2.1.16) to obtain:

$$\begin{aligned}\|\mathcal{A}^{1/2}(\nabla+i\vec{\kappa})w_f\|_{0,\Delta_f}\ &=\ a_\kappa(w_f,w_f)^{1/2}\ \lesssim\ \|w_f\|_{1,\Delta_f}\\ &\leq\ \big(\|w_f\|_{0,\Delta_f}\ +\ |w_f|_{1,\Delta_f}\big).\end{aligned}$$

Then using (3.5.4) and (3.5.5) we get:

$$\|\mathcal{A}^{1/2}(\nabla+i\vec{\kappa})w_f\|_{0,\Delta_f}\ \lesssim\big(H_f^{1/2}\ +\ H_f^{-1/2}\big)\ \|R_F(u_{j,n})\|_{0,f}.\tag{3.5.15}$$

The remaining term to treat is the last term on the right hand side of (3.5.13). Again we are just interested in $\|w_f\|_{0,\mathcal{B},\Delta_f}$ and not in the whole term. We can use (3.5.5) in order to obtain:

$$\|w_f\|_{0,\mathcal{B},\Delta_f}\ \lesssim\ \|w_f\|_{0,\Delta_f}\ \lesssim\ H_f^{1/2}\ \|R_F(u_{j,n})\|_{0,f}\tag{3.5.16}$$

Now substituting (3.5.14), (3.5.15) and (3.5.16) in (3.5.13) we get:

$$\begin{aligned}\|R_F(u_{j,n})\|_{0,f}^2\ &\lesssim\ \|R_F(u_{j,n})\|_{0,f}\ \sum_{\tau\in\Delta_f}(H_f^{1/2}+H_f^{-1/2})\ \|\mathcal{A}^{1/2}(\nabla+i\vec{\kappa})(u_j-u_{j,n})\|_{0,\tau}\\ &\quad+H_f^{1/2}\ \|\lambda_{j,n}u_{j,n}-\lambda_j u_j\|_{0,\mathcal{B},\tau}\ .\end{aligned}$$

To conclude the proof we have to multiply both sides by $H_f^{1/2}\|R_F(u_{j,n})\|_{0,f}^{-1}$:

$$\begin{aligned}H_f^{1/2}\|R_F(u_{j,n})\|_{0,f}\ &\lesssim\ \sum_{\tau\in\Delta_f}\|\mathcal{A}^{1/2}(\nabla+i\vec{\kappa})(u_j-u_{j,n})\|_{0,\tau}\\ &\quad+\ H_f\ \|\lambda_{j,n}u_{j,n}-\lambda_j u_j\|_{0,\mathcal{B},\tau}\ .\end{aligned}$$

□

In Lemma 3.5.5 we prove a local version of the efficiency, this result is extended to whole domain Ω in Theorem 3.5.6.

Lemma 3.5.5 (Local asymptotic efficiency)**.** *Let λ_j be an eigenvalue of (1.3.8) of multiplicity $R+1$ and let $(\lambda_{j+i,n}, u_{j+i,n})$ be computed eigenpairs for the same value of $\vec{\kappa}$ forming the computed eigenspace $E_{j,n}^{\mathrm{PCF}}$, in the sense of Remark 2.2.23. Let also the true eigenfunctions $U_{j+i} \in E_j^{\mathrm{PCF}}$, for $i = 0, \ldots, R$, be defined in Theorem 3.1.7. Then for each $i = 0, \ldots, R$ we have*

$$\begin{aligned}\eta_{j+i,n,\Delta_f}^2 \;&:=\; \left(\sum_{\tau \in \Delta_f} \left(H_\tau^2 \; \|R_I(u_{j+i,n}, \lambda_{j+i,n})\|_{0,\tau}^2 \right) \;+\; H_f \; \|R_F(u_{j+i,n})\|_{0,f}^2 \right) \\ &\lesssim \sum_{\tau \in \Delta_f} \left(\|\mathcal{A}^{1/2}(\nabla + i\vec{\kappa})(U_{j+i} - u_{j+i,n})\|_{0,\tau}^2 \;+\; H_\tau^2 \|\lambda_{j+i,n} u_{j+i,n} - \lambda_j U_{j+i}\|_{0,\mathcal{B},\tau}^2 \right).\end{aligned} \tag{3.5.17}$$

Proof. The local efficiency result (3.5.17) for the convex hull Δ_f comes as an application of Lemma 3.5.3 to the two element $\tau_1(f)$ and $\tau_2(f)$ and an application of Lemma 3.5.4 to f. □

Theorem 3.5.6 (Asymptotic efficiency)**.** *Let λ_j be an eigenvalue of (1.3.8) of multiplicity $R+1$ and let $(\lambda_{j+i,n}, u_{j+i,n})$ be computed eigenpairs for the same value of $\vec{\kappa}$ forming the computed eigenspace $E_{j,n}^{\mathrm{PCF}}$, in the sense of Remark 2.2.23. Let also the true eigenfunctions $U_{j+i} \in E_j^{\mathrm{PCF}}$, for $i = 0, \ldots, R$, be defined in Theorem 3.1.7. Then we have that the global residual estimator is bounded by the energy norm of the error:*

$$\sum_{i=0}^{R} \eta_{j+i,n} \lesssim \sum_{i=0}^{R} a_\kappa(U_{j+i} - u_{j+i,n}, U_{j+i} - u_{j+i,n})^{1/2} + \sum_{i=0}^{R} \|H_\tau(\lambda_{j+i,n} u_{j+i,n} - \lambda_j U_{j+i})\|_{0,\mathcal{B},\Omega}. \tag{3.5.18}$$

Proof. To prove the global efficiency we have to sum (3.5.17) for all edge (face) f and then for all i. So, summing (3.5.17) for all f, we have:

$$\begin{aligned}\eta_{j+i,n}^2 \;\lesssim \sum_{f \in \mathcal{F}_n} \Bigg\{ \sum_{\tau \in \Delta_f} \Bigg(&\|\mathcal{A}^{1/2}(\nabla + i\vec{\kappa})(U_{j+i} - u_{j+i,n})\|_{0,\tau}^2 \\ &+ H_\tau^2 \|\lambda_{j+i,n} u_{j+i,n} - \lambda_j U_{j+i}\|_{0,\mathcal{B},\tau}^2 \Bigg) \Bigg\}.\end{aligned} \tag{3.5.19}$$

The subsets Δ_f, for each value of f, are not all disjoint. Because we are using triangle elements, the maximum number of overlapping subdomains Δ_f at any point in the

interior of an element is 3. So we can put an upper-bound to (3.5.19) as

$$\begin{aligned} \eta^2_{j+i,n} \quad &\lesssim \; a_\kappa(U_{j+i} - u_{j+i,n}, U_{j+i} - u_{j+i,n}) \\ &\quad + \; \|H_\tau(\lambda_{j+i,n} u_{j+i,n} - \lambda_j U_{j+i})\|^2_{0,\mathcal{B},\Omega} \; . \end{aligned} \tag{3.5.20}$$

Then summing (3.5.20) for all $i = 0, \ldots, R$, we get the global efficiency result.

□

Remark 3.5.7. *Using Theorem 2.2.33(i) and (ii) on the term* $\|H_\tau(\lambda_{j+i,n} u_{j+i,n} - \lambda_j U_{j+i})\|_{0,\mathcal{B},\Omega}$ *in (3.5.18), we have that it is a higher order term respect to the energy norm of the error:*

$$\begin{aligned} \|H_\tau(\lambda_{j+i,n} u_{j+i,n} - \lambda_j U_{j+i})\|_{0,\mathcal{B},\Omega} \quad &\lesssim \; H_n^{\max}\big(|\lambda_{j+i,n} - \lambda_j| \; \|u_{j+i,n}\|_{0,\mathcal{B},\Omega} \\ &\quad + \; \lambda_j \; \|u_{j+i,n} - U_{j+i}\|_{0,\mathcal{B},\Omega}\big) \; = \; \mathcal{O}(H_n^{\max})^{2s+1}. \end{aligned}$$

This concludes the proof of the global efficiency for the model problem (1.3.8). This result and the local version of it holds also for the TE and TM mode problems and for the general elliptic eigenvalue problem (1.3.7), since they are particular cases of that problem. In particular for (1.3.7) you have to repeat the proof with $\vec{\kappa} = (0, 0)$ and you have to take account of the different boundary conditions.

Chapter 4

Convergent AFEM for eigenvalue problems

In the last decades, mesh adaptivity has been widely used to improve the accuracy of numerical solutions of many scientific problems. The basic idea is to refine the mesh only where the error is supposed to be large, together with the aim of achieving an accurate solution using an optimal number of degrees of freedom. There is a large numerical analysis literature on adaptivity, in particular on reliable and efficient a posteriori error estimates (e.g. [2]). Recently the question of convergence for adaptive methods has produced a great amount of interest and a number of convergence results for boundary value problems have appeared (e.g. [20, 42, 14, 13]). The only other work about convergence for eigenvalue problems, that we are aware of, is [12], which is actually more recent the ours.

The main result of this section is the proof of convergence for our adaptive FEM for elliptic eigenvalue problems, however the result presented in this work holds only for simple eigenvalues. We are going to use linear conforming finite elements on triangles. The domains of the considered problems would be bounded polygonals or polyhedrals and the problems would be subject rather to Dirichlet boundary conditions or to periodic boundary conditions. In particular, we are going to discuss the convergence of the method applied to problems (1.3.7), (1.3.8) and (1.3.9).

The outline of this chapter is as follows. The first Section 4.1 is devoted to the proof of convergence for the general elliptic eigenvalue problem (1.3.7). The same results have been submitted for publication in [26]. In the second section, Section 4.2, the convergence proof for problems arising from PCF applications, and in particular for the model problem (1.3.8), is exposed.

4.1 Convergent AFEM for generic elliptic eigenvalue problems

The outline of this section is as follows. In Subsection 4.1.2, the convergence result for problem (1.3.7), which is the main result of this section, is presented. Meanwhile, in Subsection 4.1.1 we prove that mesh refining ensures error reduction (up to oscillation of the computed eigenfunction).
Our refinement procedure is based on two elementwise defined quantities, firstly the a posteriori error estimator coming from Definition 3.2.2 and secondly a measure of the variability (or "oscillation") of the computed eigenfunction. Measures of "data oscillation" appear in other convergence results for linear boundary value problems (e.g. [42]). The definition of the error estimator η_n, when adapted to problem (1.3.7), becomes:

$$\eta_n := \Big\{ \sum_{\tau\in\mathcal{T}_n} H_\tau^2 \|R_I(u_n,\lambda_n)\|_{0,\tau}^2 \;+\; \sum_{f\in\mathcal{F}_n} H_f \|R_F(u_n)\|_{0,f}^2 \Big\}^{1/2}, \tag{4.1.1}$$

where

$$R_F(u_n)(x) := \big[\vec{n}_f \cdot \mathcal{A}\nabla u_n\big]_f(x), \quad \text{with } x \in int(f), \quad f \in \mathcal{F}_n.$$

and

$$R_I(u_n,\lambda_n)(x) := \big(\nabla\cdot\mathcal{A}\nabla u_n \;+\; \lambda_n \mathcal{B} u_n\big)(x) = \big(\lambda_n \mathcal{B} u_n\big)(x), \quad \text{with } x \in int(\tau), \quad \tau \in \mathcal{T}_n,$$

where in the last equality we exploited the fact that we use linear elements on triangles. Our algorithm performs local refinement on all elements on which at least one of these two local quantities is sufficiently large. We prove that the adaptive method converges provided the initial mesh is sufficiently fine. The latter condition, which is absent in adaptive methods for linear symmetric elliptic boundary value problems, commonly appears for nonlinear problems and it can be thought of as a manifestation of the nonlinearity of the problem.
The mesh refinement that we adopted is the same already used in [20], [42]. The idea is to refine a subset of the elements of the mesh $\mathcal{T}_n$ whose side residuals sum up to a fixed proportion of the total residual η_n.

Definition 4.1.1 (Marking Strategy 1)**.** *Given a parameter $0 < \theta < 1$, the procedure is: mark the sides in a minimal subset $\hat{\mathcal{F}}_n$ of $\mathcal{F}_n$ such that*

$$\Big(\sum_{f\in\hat{\mathcal{F}}_n} \eta_{f,n}^2 \Big)^{1/2} \;\geq\; \theta\; \eta_n \;, \tag{4.1.2}$$

where $\eta_{f,n}$ is:

$$\eta_{f,n}^2 := \frac{1}{3}\|H_\tau \; R_I(u_n,\lambda_n)\|_{0,\Delta_f}^2 \; + \; \|H_f^{1/2} R_F(u_n)\|_{0,f}^2 \; , \tag{4.1.3}$$

where we denoted by Δ_f the union of the two elements $\tau_1(f)$ and $\tau_2(f)$ sharing f.

To satisfy the condition (4.1.2), we need first of all to compute all the "local residuals" $\eta_{f,n}$ and sort them according their values. Then the edges (faces) f are inserted into $\hat{\mathcal{F}}_n$ in decreasing order of $\eta_{f,n}$, starting from the edge (face) with the biggest local residual, until the condition (4.1.2) is satisfied. Note that a minimal subset $\hat{\mathcal{F}}_n$ may not be unique. Then, we construct another set $\hat{\mathcal{T}}_n$, containing all the elements of $\mathcal{T}_n$ which share at least one edge (face) $f \in \hat{\mathcal{F}}_n$.
In order to prove the convergence of the adaptive method, we require an additional marking strategy, which will be defined in Definition 4.1.4 below. The latter marking strategy is driven by oscillations. The same argument has been already used in some papers about convergence for source problems (see [42] and [40]), but to our knowledge has not yet been used for analysing convergent algorithms for eigenvalue problems.
The concept of "oscillations" is just a measure of how well a function may be approximated by piecewise constant elements on a particular mesh. For any function $v \in L^2(\Omega)$, and any mesh $\mathcal{T}_n$, we introduce its orthogonal projection $P_n v$ onto piecewise constants defined by:

$$(P_n v)|_\tau = \frac{1}{|\tau|}\int_\tau v, \quad \text{for all} \quad \tau \in \mathcal{T}_n. \tag{4.1.4}$$

Notation 4.1.2. *In this chapter we define H_n to be a piecewise constant function which assumes in the interior of each element τ of the mesh $\mathcal{T}_n$ the size of the element, i.e.*

$$\forall \tau \in \mathcal{T}_n, \quad H_n|_\tau \; = \; H_\tau \; .$$

In the next definition we make use of the projection operator P_n:

Definition 4.1.3 (Oscillations)**.** *On a mesh $\mathcal{T}_n$, we define*

$$\mathrm{osc}(v,\mathcal{T}_n) := \|H_n(v - P_n v)\|_{0,\mathcal{B},\Omega}. \tag{4.1.5}$$

Note that

$$\mathrm{osc}(v,\mathcal{T}_n) = \Big(\sum_{\tau\in\mathcal{T}_n} H_\tau^2 \|v - P_n v\|_{0,\mathcal{B},\tau}^2\Big)^{1/2}.$$

and that (by standard approximation theory and the coercivity of $a(\cdot,\cdot)$),

$$\mathrm{osc}(v,\mathcal{T}_n) \; \lesssim \; (H_n^{\max})^2 a(v,v)^{1/2} \; , \quad \text{for all} \quad v \in H_0^1(\Omega) \; . \tag{4.1.6}$$

The second marking strategy (introduced below) aims to reduce the quantity osc corresponding to a particular approximate eigenfunction u_n.

Definition 4.1.4 (Marking Strategy 2)**.** *Given a parameter $0 < \tilde{\theta} < 1$: mark the elements in a minimal subset $\tilde{\mathcal{T}}_n$ of $\mathcal{T}_n$ such that*

$$\operatorname{osc}(u_n, \tilde{\mathcal{T}}_n) \;\geq\; \tilde{\theta}\, \operatorname{osc}(u_n, \mathcal{T}_n)\ . \tag{4.1.7}$$

Note that a minimal subset $\tilde{\mathcal{T}}_n$ may not be unique. To satisfy the condition (4.1.7), we need first of all to compute all the local terms $H_\tau^2 \|(u_n - P_n u_n)\|_{0,\mathcal{B},\tau}^2$ forming $\operatorname{osc}(u_n, \mathcal{T}_n)$ and sort them according their values. Then the elements τ are inserted into $\tilde{\mathcal{T}}_n$ in decreasing order of the size of those local terms, until the condition (4.1.7) is satisfied. Our adaptive algorithm can then be stated:

Algorithm 1 Converging algorithm

Require: $0 < \theta < 1$
Require: $0 < \tilde{\theta} < 1$
loop
Compute the approximated eigenpair on the mesh $\mathcal{T}_n$
Mark the elements using the first marking strategy (Definition 4.1.1)
Mark any additional unmarked elements using the second marking strategy (Definition 4.1.4)
Construct the mesh $\mathcal{T}_{n+1}$ refining the elements in $\hat{\mathcal{T}}_n \cup \tilde{\mathcal{T}}_n$ using the bisection5 scheme in Figure 4-1.
end loop

Remark 4.1.5. *From now on we fix the value of j because we restrict our analysis to the true eigenpair (λ_j, u_j) and to the computed eigenpair on the mesh $\mathcal{T}_n$ $(\lambda_{j,n}, u_{j,n})$ converging to (λ_j, u_j) in the sense described in Theorem 2.2.10. So we can drop the subscript j and we simply write (λ, u) for the eigenpair of* (1.3.7) *and (λ_n, u_n) for the eigenpair of* (2.2.2)*.*

Remark 4.1.6. *In this chapter we suppose that λ is a simple eigenvalue. This implies that the corresponding eigenspace has dimension 1 and it is possible to find two unit eigenvectors corresponding to λ, namely u or $-u$. In other words, there is not a unique eigenvector corresponding to λ, but two. The same ambiguity holds also for all the eigenvalues λ_n computed in Algorithm 1, which approximate λ. In fact, for each n, both (λ_n, u_n) and $(\lambda_n, -u_n)$ are acceptable eigenpairs for the discrete problem. To make the arguments in this chapter not ambiguous, we assume that u_0 is the eigenfunction actually computed in the first iteration of Algorithm 1. Then we suppose that the true eigenfunction $u := U$, where U is constructed as in the proof of Theorem 3.1.4. Then, we set for each $n > 0$ the eigenfunction $u_n := w_n$, where w_n comes from Theorem 2.2.10. So, denoting by u_n^* the eigenfunction actually computed*

in the n-th iteration of Algorithm 1, we have that either $w_n = u_n^*$ *or* $w_n = -u_n^*$. *In general not all the eigenfunctions* u_n *appearing in the results below coincide with the computed ones, i.e.* $u_n = u_n^*$, *for some n it could be possible that* $u_n = -u_n^*$. *Anyway, from a computational point of view the signs are not important, since the error estimator used in Algorithm 1 is independent of the signs of the eigenfunctions. Moreover, Algorithm 1 generates a sequence of eigenvalues* λ_n *converging to* λ *and a sequence of computed eigenfunctions* u_n^* *converging into the true eigenspace of* λ. *But, without taking control of the signs of the computed eigenfunctions, what could happen is that a subsequence of computed eigenfunctions would converge to the true eigenfunction u and another subsequence would converge to the true eigenfunction* $-u$.

In the 2D-case, at the $n-th$ iteration of Algorithm 1, each element in the set $\hat{\mathcal{T}}_n \cup \tilde{\mathcal{T}}_n$ is refined using the "bisection5" procedure (which has been used also in [42]), which is illustrated in Figure 4-1c. An advantage of this technique is the creation of a new node in the middle of each marked side in $\hat{\mathcal{F}}_n$ and also a new node in the interior of each marked element.

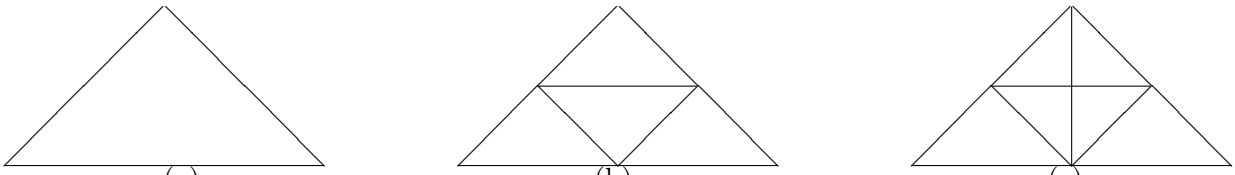

Figure 4-1: The refinement procedure applied to an element of the mesh. In (a) the element before the refinement, in (b) after the three sides as been refined and in (c) after the bisection of one of the three new segments.

In the 3D-case, we use a suitable refinement that creates a new node on each marked face in $\hat{\mathcal{F}}_n$ and a node in the interior of each marked element. These requirements are analogous to the requirements satisfied by bisection5 in 2D-case.
In [42] and [40] it has been shown (for linear source problems) that the reduction of the error is trigged by the decay of the quantity osc on the sequence of constructed meshes. This is only for problems like $a(u,v) = (f,v)_{0,\Omega}$, where f is a given function. For the eigenvalue problem (2.2.2) the quantity $\lambda_n u_n$ plays the role of data and in principle we have to ensure that the value of osc for this quantity, is sufficiently small. However $\lambda_n u_n$ may change if the mesh changes and so the proof of error reduction for eigenvalue problems is not as simple as it is for linear source problems. This is the essence of the theoretical problems solved in this paper.

Notation 4.1.7. *We write* $A \lesssim B$ *when* A/B *is bounded by a constant which may depend on the functions* $\mathcal{A}$ *and* $\mathcal{B}$ *in* (1.3.1) *and* (1.3.2), *on* C_{ell} *in Assumption 2.2.1 and* C_{reg} *in* (2.2.1). *The notation* $A \cong B$ *means* $A \lesssim B$ *and* $A \gtrsim B$.

All the constants depending on the spectrum, namely C_{adj} in (3.1.25) and C_{spec1} and C_{spec2} in Theorem 2.2.10, are handled explicitly. Similarly all mesh size dependencies are explicit. Note that all eigenvalues of (2.2.2) satisfy $\lambda_n \gtrsim 1$, since $\lambda_n \geq \lambda_1 = a(u_1, u_1) \gtrsim |u_1|^2_{1,\Omega} \gtrsim \|u_1\|^2_{0,\Omega} \gtrsim \|u_1\|^2_{0,\mathcal{B},\Omega} = 1$.

4.1.1 Error Reduction

In this subsection we give the proof of error reduction for Algorithm 1. The proof has been inspired by the corresponding theory for source problems in [42]. However the nonlinearity of the eigenvalue problem introduces new complications, so there are several lemmas before the main theorem (Theorem 4.1.15).
In Lemma 4.1.14 below, we are going to use the reliability result for general elliptic eigenvalue problems, which is Theorem 3.3.5 modified as prescribed by Remark 3.4.7. To improve the readability of this section, the reliability for general elliptic eigenvalue problems used below is stated here:

Theorem 4.1.8 (Reliability for eigenfunctions)**.** *Let λ be a simple eigenvalue of (1.3.7) and let (λ_n, u_n) be computed eigenpairs, in the sense of Remark 2.2.4. Let also the true eigenfunction u and the approximated one u_n be defined in the sense of Remark 4.1.6. Then we have for $e_n = u - u_n$ that*

$$a(e_n, e_n)^{1/2} \;\lesssim\; \eta_n \;+\; G_n, \tag{4.1.8}$$

where the quantity η_n is defined in 4.1.1 and where

$$G_n \;=\; \frac{1}{2}(\lambda + \lambda_n)\frac{(e_n, e_n)_{0,\mathcal{B},\Omega}}{a(e_n, e_n)^{1/2}}. \tag{4.1.9}$$

Notation 4.1.9. *In this chapter we denote by $|||u|||_\Omega$ the norm $a(u,u)^{1/2}$.*

The next theorem is a generalization to eigenvalue problems of the standard monotone convergence property for linear symmetric elliptic PDEs, namely that if you enrich the finite dimensional space, then the error is bound to decrease. This result fails to hold for eigenvalue problems (even for symmetric elliptic partial differential operators), because of the nonlinearity of such problems. The best that we can do is to show that if the finite dimensional space is enriched, then the error will not increase very much. This is the subject of Theorem 4.1.10.

Theorem 4.1.10. *Let λ be a simple eigenvalue of (1.3.7) and let (λ_n, u_n) and (λ_m, u_m) be computed eigenpairs, in the sense of Remark 2.2.4. Let also the true eigenfunction u and the approximated ones u_n and u_m be defined in the sense of Remark 4.1.6. Then there exists a constant $q > 1$ such that, for all $m \geq n$, the corresponding computed*

eigenpair (λ_m, u_m) *satisfies:*

$$|||u - u_m|||_\Omega \leq q\, |||u - u_n|||_\Omega \ . \tag{4.1.10}$$

Proof. From Theorem 3.1.6, we obtain

$$\|u - u_m\|_{0,\mathcal{B},\Omega} \lesssim C_{\text{adj}} (H_m^{\max})^s\, |||u - Q_m u|||_\Omega \tag{4.1.11}$$

Since $\mathcal{T}_m$ is a refinement of $\mathcal{T}_n$, it follows that $V_n \subset V_m$ and so the best approximation property of Q_m ensures that

$$|||u - Q_m u|||_\Omega \ \leq |||u - Q_n u|||_\Omega \ .$$

Hence from (4.1.11) and using the fact that $H_m^{\max} \leq H_n^{\max}$, we have

$$\|u - u_m\|_{0,\mathcal{B},\Omega} \ \leq \ C_{\text{adj}} (H_n^{\max})^s\, |||u - Q_n u|||_\Omega. \tag{4.1.12}$$

Now, using Lemma 2.2.11 we get:

$$|||u - u_m|||_\Omega^2 \ = \ |\lambda - \lambda_m| \ + \ \lambda \|u - u_m\|_{0,\mathcal{B},\Omega}^2 \ . \tag{4.1.13}$$

Then, combining (4.1.12) with (4.1.13) and using the minimum-maximum principle, we obtain

$$\begin{aligned} |||u - u_m|||_\Omega^2 \ &\leq \ |\lambda - \lambda_m| \ + \ \lambda C_{\text{adj}}^2 (H_n^{\max})^{2s}\, |||u - Q_n u|||_\Omega^2 \\ &\leq \ |\lambda - \lambda_n| \ + \ \lambda C_{\text{adj}}^2 (H_n^{\max})^{2s}\, |||u - Q_n u|||_\Omega^2. \end{aligned} \tag{4.1.14}$$

Hence, using Corollary 2.2.12

$$|||u - u_m|||_\Omega^2 \ \leq \ |||u - u_n|||_\Omega^2 \ + \ \lambda C_{\text{adj}}^2 (H_n^{\max})^{2s}\, |||u - Q_n u|||_\Omega^2. \tag{4.1.15}$$

But since Q_n yields the best approximation in the energy norm, we have

$$|||u - u_m|||_\Omega^2 \ \leq \ (1 + \lambda C_{\text{adj}}^2 (H_0^{\max})^{2s})\, |||u - u_n|||_\Omega^2 \ , \tag{4.1.16}$$

which is in the required form. □

The next lemma is similar to [42, Lemma 4.2] for the 2D-case. But we are going to extend the result to the 3D-case, too.

Lemma 4.1.11. *Let* (λ_n, u_n) *be an approximated eigenpair on the mesh* $\mathcal{T}_n$ *and let* $\hat{\mathcal{F}}_n$ *be as defined in Definition 4.1.1 and let* P_n *be as defined in (4.1.4). For any* $f \in \hat{\mathcal{F}}_n$,

there exists a function $\Phi_f \in V_{n+1}$ *such that* $\mathrm{supp}(\Phi_f) = \Delta_f$*, and also*

$$\int_{\Delta_f} \lambda_n \ \mathcal{B}(P_n u_n)\Phi_f - \int_f R_F(u_n)\Phi_f \ = \ \|H_n \ \lambda_n \mathcal{B} P_n u_n\|^2_{0,\Delta_f} + \|H_f^{1/2} R_F(u_n)\|^2_{0,f}, \tag{4.1.17}$$

and

$$|||\Phi_f|||^2_{\Delta_f} \ \lesssim \ \|H_n \ \lambda_n \mathcal{B} P_n u_n\|^2_{0,\Delta_f} + \|H_f^{1/2} R_F(u_n)\|^2_{0,f}. \tag{4.1.18}$$

Figure 4-2: Two cases of refined couples of elements .

Proof. Figure 4-2 illustrates two possible configurations of the domain Δ_f (in the 2D case): in Figure 4-2a we have that both corners opposite to the common edge have been bisected, while Figure 4-2b shows a different choice of bisected corners. The point x_f is the node created on the shared edge f by the refinement while the points x_1 and x_2 are the nodes created in the interior of the refined elements $\tau_1(f)$ and $\tau_2(f)$ respectively.
The two situations in Figure 4-2 do not exhaust all the possible configurations for couples of adjacent refined elements. There could be other possible configurations different from Figure 4-2b, in which the green refinements are applied to different edges. However, the way in which the green-refinements split the elements is irrelevant for the proof, since the only important thing is the existence of a new node on the shared edge and two nodes in the interior of the elements.
In the 3D case we denote by $\tau_1(f)$ and $\tau_2(f)$ the elements sharing the face f and, similarly to the 2D case, we denote by x_f the node created in the middle of the shared face f while the points x_1 and x_2 are the nodes created in the interior of the refined elements $\tau_1(f)$ and $\tau_2(f)$ respectively. We have not included a picture of the refinement for the 3D case, since it would be very difficult to draw.
We start proving (4.1.17). The proof of this result is not affected by the number of dimensions of the domain, instead the proof of (4.1.18) slightly differs according to the number of dimensions.
We then define

$$\Phi_f := \alpha_f \varphi_f + \beta_1 \varphi_1 + \beta_2 \varphi_2, \tag{4.1.19}$$

where φ_f and φ_i are the nodal basis functions associated with the points x_f and x_i in

$\mathcal{T}_{n+1}$, and α_f, β_i are defined by

$$\alpha_f = \begin{cases} -\dfrac{\|H_f^{1/2} R_F(u_n)\|_{0,f}^2}{\int_f R_F(u_n)\varphi_f} & \text{if } R_F(u_n) \neq 0, \\ \\ 0 & \text{otherwise,} \end{cases} \tag{4.1.20}$$

and

$$\beta_i = \begin{cases} \dfrac{\|H_n\ \lambda_n \mathcal{B} P_n u_n\|_{0,\tau_i(f)}^2 - \alpha_f \int_{\tau_i(f)} \lambda_n \mathcal{B} P_n u_n\ \varphi_f}{\int_{\tau_i(f)} \lambda_n \mathcal{B} P_n u_n\ \varphi_i} & \text{if } P_n u_n|_{\tau_i(f)} \neq 0, \\ \\ 0 & \text{otherwise,} \end{cases} \tag{4.1.21}$$

for $i = 1, 2$.
Using the fact that $\text{supp}(\varphi_i) = \tau_i(f)$, for $i = 1, 2$ we can easily see that the above formulae imply

$$\alpha_f \int_f R_F(u_n)\varphi_f \quad = \quad -\|H_f^{1/2} R_F(u_n)\|_{0,f}^2, \tag{4.1.22}$$

$$\int_{\Delta_f} \lambda_n \mathcal{B} P_n u_n\ (\alpha_f \varphi_f + \beta_1 \varphi_1 + \beta_2 \varphi_2) \quad = \quad \|H_n\ \lambda_n \mathcal{B} P_n u_n\|_{0,\Delta_f}^2, \tag{4.1.23}$$

(these formulae remain true even if $R_F(u_n)$ or $P_n u_n|_{\tau_i(f)}$ vanish). Hence

$$\int_{\Delta_f} \lambda_n \mathcal{B} P_n u_n\ \Phi_f - \int_f R_F(u_n)\Phi_f \ = \ \int_{\Delta_f} \lambda_n \mathcal{B}\ P_n u_n(\alpha_f \varphi_f + \beta_1 \varphi_1 + \beta_2 \varphi_2) - \int_f R_F(u_n)\alpha_f \varphi_f$$

and (4.1.17) follows immediately on using (4.1.22) and (4.1.23).
To prove (4.1.18) in the 2D case, we use (4.1.19) and the facts that $|\varphi_f|_{1,\Delta_f} \lesssim 1$ and $|\varphi_i|_{1,\Delta_f} \lesssim 1$ to obtain

$$|||\Phi_f|||_{\Delta_f}^2 \ \lesssim \ |\alpha_f|^2 + |\beta_1|^2 + |\beta_2|^2 \ . \tag{4.1.24}$$

Now, since $R_F(u_n)$ is constant on f and $\int_f \varphi_f \sim H_f$, we have

$$|\alpha_f| \ \lesssim \ \frac{|R_F(u_n)| \|H_f^{1/2}\|_{0,f}^2}{H_f} \ \lesssim \ |R_F(u_n)| H_f \ \sim \ \|H_f^{1/2} R_F(u_n)\|_{0,f} \ . \tag{4.1.25}$$

Also since $P_n u_n$ is constant on each $\tau_i(f)$ and since $\int_{\tau_i(f)} \varphi_i \sim H^2_{\tau_i(f)}$, we have

$$\begin{aligned} |\beta_i| &\lesssim \frac{|\lambda_n \mathcal{B} P_n u_n|_{\tau_i(f)}| \, \|H_n\|^2_{0,\tau_i(f)} \; + \; |\alpha_f| H^2_{\tau_i(f)}}{H^2_{\tau_i(f)}} \\ &\lesssim |\lambda_n \mathcal{B} P_n u_n|_{\tau_i(f)}| \; H^2_{\tau_i(f)} + |\alpha_f| \;\sim\; \|H_n \lambda_n \mathcal{B} P_n u_n\|_{0,\tau_i(f)} + |\alpha_f| \end{aligned}$$

This implies

$$\begin{aligned} |\beta_i|^2 &\lesssim \|H_n \lambda_n \mathcal{B} P_n u_n\|^2_{0,\tau_i(f)} + |\alpha_f|^2 \\ &\lesssim \|H_n \lambda_n \mathcal{B} P_n u_n\|^2_{0\tau_i(f)} + \|H_f^{1/2} R_F(u_n)\|^2_{0,f} \;, \end{aligned} \tag{4.1.26}$$

and the proof is completed by combining (4.1.24) with (4.1.25) and (4.1.26).
To prove (4.1.18) in the 3D case, we use (4.1.19), and the facts that $|\varphi_f|_{1,\Delta_f} \lesssim H_f^{1/2}$ and $|\varphi_i|_{1,\Delta_f} \lesssim H^{1/2}_{\tau_i(f)}$ to obtain

$$|||\Phi_f|||^2_{\Delta_f} \;\lesssim\; H_f |\alpha_f|^2 + H_{\tau_1(f)} |\beta_1|^2 + H_{\tau_2(f)} |\beta_2|^2 \;. \tag{4.1.27}$$

Now, since $R_F(u_n)$ is constant on S and $\int_f \varphi_f \sim H_f^2$, we have

$$|\alpha_f| \;\lesssim\; \frac{|R_F(u_n)| \|H_f^{1/2}\|^2_{0,f}}{H_f^2} \;\lesssim\; |R_F(u_n)| H_f \;\sim\; H_f^{-1/2} \|H_f^{1/2} R_F(u_n)\|_{0,f} \;. \tag{4.1.28}$$

Also since $P_n u_n$ is constant on each $\tau_i(f)$ and since $\int_{\tau_i(f)} \varphi_i \sim H^3_{\tau_i(f)}$, we have

$$\begin{aligned} |\beta_i| &\lesssim \frac{|\lambda_n \mathcal{B} P_n u_n|_{\tau_i(f)}| \, \|H_n\|^2_{0,\tau_i(f)} \; + \; |\alpha_f| H^3_{\tau_i(f)}}{H^3_{\tau_i(f)}} \\ &\lesssim |\lambda_n \mathcal{B} P_n u_n|_{\tau_i(f)}| \; H^2_{\tau_i(f)} + |\alpha_f| \;\sim\; H^{-1/2}_{\tau_i(f)} \|H_n \lambda_n \mathcal{B} P_n u_n\|_{0,\tau_i(f)} + |\alpha_f| \end{aligned}$$

This implies

$$\begin{aligned} |\beta_i|^2 &\lesssim H^{-1}_{\tau_i(f)} \|H_n \lambda_n \mathcal{B} P_n u_n\|^2_{0,\tau_i(f)} + |\alpha_f|^2 \\ &\lesssim H^{-1}_{\tau_i(f)} \|H_n \lambda_n \mathcal{B} P_n u_n\|^2_{0,\tau_i(f)} + H_f^{-1} \|H_f^{1/2} R_F(u_n)\|^2_{0,f} \;, \end{aligned} \tag{4.1.29}$$

and the proof is completed by combining (4.1.27) with (4.1.28) and (4.1.29).

□

Remark 4.1.12. *The reason why in this chapter we present convergence results for linear elements only, is that we have not found a way to extend Lemma 4.1.11 to higher order elements yet. This could be the subject of further investigations.*

In the next lemma we bound the local error estimator from above by the local difference

of two discrete solutions coming from consecutive meshes, plus higher order terms. This kind of result is called "discrete local efficiency" by many authors.
Recall that $\mathcal{T}_{n+1}$ is the refinement of $\mathcal{T}_n$ obtained by applying Algorithm 1.

Lemma 4.1.13. *Let (λ_n, u_n) be an approximate eigenpair on a mesh $\mathcal{T}_n$, let $\mathcal{T}_{n+1}$ be the mesh obtained by one iteration of Algorithm 1 and let (λ_{n+1}, u_{n+1}) be an approximate eigenpair on a mesh $\mathcal{T}_{n+1}$. Let the eigenfunctions u, u_n and u_{n+1} be defined in the sense of Remark 4.1.5. Then, for any $f \in \hat{\mathcal{F}}_n$, we have*

$$\begin{aligned} \eta_{f,n}^2 \quad \lesssim |||u_{n+1} - u_n|||_{\Delta_f}^2 \quad + \quad \|H_n(\lambda_{n+1}u_{n+1} - \lambda_n P_n u_n)\|_{0,\mathcal{B},\Delta_f}^2 \\ + \|H_n \mathcal{B} \lambda_n (u_n - P_n u_n)\|_{0,\Delta_f}^2 \ , \end{aligned} \tag{4.1.30}$$

where $\eta_{f,n}$ is defined in 4.1.3.

Proof. Since the function Φ_f defined in Lemma 4.1.11 is in V_{n+1} and $\mathrm{supp}(\Phi_f) = \Delta_f$, we have

$$\begin{aligned} a(u_{n+1} - u_n, \Phi_f) &= a(u_{n+1}, \Phi_f) - a(u_n, \Phi_f) \\ &= \lambda_{n+1} \int_{\Delta_f} \mathcal{B} u_{n+1} \Phi_f \ - \ a(u_n, \Phi_f). \end{aligned} \tag{4.1.31}$$

Now applying integration by parts to the last term on the right-hand side of (4.1.31), we obtain

$$a(u_{n+1} - u_n, \Phi_f) \quad = \quad \lambda_{n+1} \int_{\Delta_f} \mathcal{B} u_{n+1} \Phi_f \ - \ \int_f R_F(u_n) \Phi_f. \tag{4.1.32}$$

Combining (4.1.32) with (4.1.17), we obtain

$$\begin{aligned} a(u_{n+1} - u_n, \Phi_f) - \int_{\Delta_f} \mathcal{B}(\lambda_{n+1}u_{n+1} - \lambda_n P_n u_n)\Phi_f \\ &= \sum_{\tau \in \Delta_f} \int_\tau \lambda_n \mathcal{B} P_n u_n \Phi_f - \int_f R_F(u_n)\Phi_f \\ &= \|H_n \ \lambda_n \mathcal{B} P_n u_n\|_{0,\Delta_f}^2 + \|H_f^{1/2} R_F(u_n)\|_{0,f}^2. \end{aligned} \tag{4.1.33}$$

Rearranging (4.1.33), and then applying the triangle and Cauchy-Schwarz inequalities, we obtain:

$$
\begin{aligned}
&\|H_n\ \lambda_n \mathcal{B} P_n u_n\|^2_{0,\Delta_f} + \|H_f^{1/2} R_F(u_n)\|^2_{0,f} \\
&\qquad \leq\ |a(u_{n+1}-u_n, \Phi_f)| + \left| \int_{\Delta_f} \mathcal{B}(\lambda_{n+1}u_{n+1} - \lambda_n P_n u_n)\Phi_f \right| \\
&\qquad \leq\ |||u_{n+1}-u_n|||_{\Delta_f} |||\Phi_f|||_{\Delta_f} + \|\lambda_{n+1}u_{n+1} - \lambda_n P_n u_n\|_{0,\mathcal{B},\Delta_f} \|\Phi_f\|_{0,\mathcal{B},\Delta_f} \\
&\qquad \lesssim\ \Big(|||u_{n+1}-u_n|||_{\Delta_f} + \|H_n(\lambda_{n+1}u_{n+1} - \lambda_n P_n u_n)\|_{0,\mathcal{B},\Delta_f} \Big) |||\Phi_f|||_{\Delta_f}\ ,
\end{aligned}
\tag{4.1.34}
$$

where in the final step of (4.1.34) we made use of the Poincaré inequality

$$
\|\Phi_f\|_{0,\mathcal{B},\Delta_f} \lesssim H_f |\Phi_f|_{1,\Delta_f}\ ,
$$

the coercivity $|\Phi_f|_{1,\Delta_f} \lesssim |||\Phi_f|||_{\Delta_f}$ and also the shape-regularity of the meshes. In view of (4.1.18), yields

$$
\begin{aligned}
&\|H_n\ \lambda_n \mathcal{B} P_n u_n\|^2_{0,\Delta_f} + \|H_f^{1/2} R_F(u_n)\|^2_{0,f} \\
&\qquad \lesssim\ |||u_{n+1}-u_n|||^2_{\Delta_f} + \|H_n(\lambda_{n+1}u_{n+1} - \lambda_n P_n u_n)\|^2_{0,\mathcal{B},\Delta_f}\ .
\end{aligned}
\tag{4.1.35}
$$

From the definition of $\eta_{f,n}$ in (4.1.3), and the triangle inequality, we have

$$
\begin{aligned}
\eta^2_{f,n} \quad &\lesssim \quad \|H_n\ \lambda_n \mathcal{B} P_n u_n\|^2_{0,\Delta_f} \\
&+ \quad \|H_f^{1/2} R_F(u_n)\|^2_{0,f} + \|H_n \mathcal{B} \lambda_n (u_n - P_n u_n)\|^2_{0,\Delta_f}\ .
\end{aligned}
\tag{4.1.36}
$$

The required inequality (4.1.30) now follows from (4.1.35) and (4.1.36). □

In the main result of this section, Theorem 4.1.15 below, we achieve error reduction of the form $|||u-u_{n+1}|||_\Omega \leq \alpha |||u-u_n|||_\Omega$, for some $\alpha < 1$. In the case of source problems (see [42]) this is approached by writing

$$
\begin{aligned}
|||u-u_n\ |||^2_\Omega &= |||u-u_{n+1}+u_{n+1}-u_n\ |||^2_\Omega \\
&= |||u-u_{n+1}\ |||^2_\Omega + |||u_{n+1}-u_n\ |||^2_\Omega \\
&\quad + 2a(u-u_{n+1}, u_{n+1}-u_n).
\end{aligned}
\tag{4.1.37}
$$

and making use of the fact that the last term on the right-hand side vanishes due to Galerkin orthogonality. However this approach is not available in the eigenvalue problem context. Therefore a more technical approach is needed to bound the two terms on the right-hand side of (4.1.37) from below. The main technical result is in the following lemma. Recall the convention in Notation 4.1.7.

Lemma 4.1.14. *Under the same assumptions of Lemma 4.1.13 we have:*

$$|||u_{n+1}-u_n|||_\Omega^2 \;\gtrsim\; \theta^2\,|||u-u_n|||_\Omega^2 \;-\; \mathrm{osc}(\lambda_n u_n,\mathcal{T}_n)^2 - L_n^2\;, \tag{4.1.38}$$

where θ is defined in the marking strategy in Definition 4.1.1 and L_n satisfies the estimate:

$$L_n \;\lesssim\; \hat{C}\;(H_n^{\max})^s|||u-u_n|||_\Omega\;, \tag{4.1.39}$$

where $\hat{C}$ depends on θ, λ_n, C_{spec}, C_{adj} and q.

Proof. By Lemma 4.1.13 and Definition 4.1.1 we have

$$\begin{aligned}\theta^2\eta_n^2 \;&\leq\; \textstyle\sum_{f\in\hat{\mathcal{F}}_n}\eta_{f,n}^2\\ &\lesssim |||u_{n+1}-u_n\;|||_\Omega^2 + \|H_n(\lambda_{n+1}u_{n+1}-\lambda_n P_n u_n)\|_{0,\mathcal{B},\Omega}^2\\ &+\;\mathrm{osc}(\lambda_n u_n,\mathcal{T}_n)^2\;.\end{aligned}$$

Hence, rearranging and making use of Theorem 4.1.8, we have

$$\begin{aligned}|||u_{n+1}-u_n|||_\Omega^2 \;&\gtrsim\; \theta^2\;\eta_n^2 - \|H_n(\lambda_{n+1}u_{n+1}-\lambda_n P_n u_n)\|_{0,\mathcal{B},\Omega}^2 \;-\; \mathrm{osc}(\lambda_n u_n,\mathcal{T}_n)^2\\ &\gtrsim\; \theta^2\;|||u-u_n|||_\Omega^2 \;-\; \mathrm{osc}(\lambda_n u_n,\mathcal{T}_n)^2 - \theta^2\;G_n^2\\ &\qquad -\|H_n(\lambda_{n+1}u_{n+1}-\lambda_n P_n u_n)\|_{0,\mathcal{B},\Omega}^2\;.\end{aligned} \tag{4.1.40}$$

We now estimate the last two terms in (4.1.40) separately.
To estimate G_n, we use (4.1.9), combined with the Poincaré inequality (and the H^1 - ellipticity of $a(\cdot,\cdot)$) and then Theorem 3.1.6 to obtain

$$G_n \;\lesssim\; \frac{1}{2}(\lambda+\lambda_n)\|u-u_n\|_{0,\mathcal{B},\Omega} \;\lesssim\; \frac{1}{2}(\lambda+\lambda_n)C_{\mathrm{adj}}(H_n^{\max})^s|||u-u_n|||_\Omega. \tag{4.1.41}$$

To estimate the last term in (4.1.40), we first use the triangle inequality to obtain

$$\|H_n(\lambda_{n+1}u_{n+1}-\lambda_n P_n u_n)\|_{0,\mathcal{B},\Omega} \;\leq\; \|H_n(\lambda_{n+1}u_{n+1}-\lambda_n u_n)\|_{0,\mathcal{B},\Omega} \;+\; \lambda_n\mathrm{osc}(u_n,\mathcal{T}_n). \tag{4.1.42}$$

For the first term on the right-hand side of (4.1.42), we have

$$\|H_n(\lambda_{n+1}u_{n+1}-\lambda_n u_n)\|_{0,\mathcal{B},\Omega} \leq H_n^{\max}\big(\|\lambda u-\lambda_{n+1}u_{n+1}\|_{0,\mathcal{B},\Omega}+\|\lambda u-\lambda_n u_n\|_{0,\mathcal{B},\Omega}\big). \tag{4.1.43}$$

From Corollary 2.2.12 we have that

$$|\lambda-\lambda_{n+1}| \;\leq\; |||u-u_{n+1}|||_\Omega^2\;,$$

then using this result and Theorem 3.1.6, we obtain

$$\begin{aligned}\|\lambda u - \lambda_{n+1}u_{n+1}\|_{0,\mathcal{B},\Omega} &\leq |\lambda - \lambda_{n+1}|\|u\|_{0,\mathcal{B},\Omega} + \lambda_{n+1}\|u - u_{n+1}\|_{0,\mathcal{B},\Omega} \\ &\lesssim |||u - u_{n+1}|||_\Omega^2 \\ &\quad + \lambda_{n+1}C_{\text{adj}}(H_n^{\max})^s|||u - u_{n+1}|||_\Omega \ . \end{aligned} \tag{4.1.44}$$

Using Theorem 3.1.4 and using the fact that $\lambda_{n+1} \leq \lambda_n$ from the minimum-maximum principle we have

$$\|\lambda u - \lambda_{n+1}u_{n+1}\|_{0,\mathcal{B},\Omega} \lesssim (C_{\text{spec2}} + \lambda_n C_{\text{adj}})(H_n^{\max})^s|||u - u_{n+1}|||_\Omega \ . \tag{4.1.45}$$

Finally, using Theorem 4.1.10 we obtain

$$\|\lambda u - \lambda_{n+1}u_{n+1}\|_{0,\mathcal{B},\Omega} \lesssim q(C_{\text{spec2}} + \lambda_n C_{\text{adj}})(H_n^{\max})^s|||u - u_n|||_\Omega \ . \tag{4.1.46}$$

An identical argument shows

$$\|\lambda u - \lambda_n u_n\|_{0,\mathcal{B},\Omega} \lesssim (C_{\text{spec2}} + \lambda_n C_{\text{adj}})(H_n^{\max})^s|||u - u_n|||_\Omega \ . \tag{4.1.47}$$

Combining (4.1.46) and (4.1.47) with (4.1.43), we obtain

$$\|H_n(\lambda_{n+1}u_{n+1} - \lambda_n u_n)\|_{0,\mathcal{B},\Omega} \lesssim (1+q)(C_{\text{spec2}} + \lambda_n C_{\text{adj}})(H_n^{\max})^{s+1}|||u - u_n|||_\Omega \ . \tag{4.1.48}$$

Now combining (4.1.40) with (4.1.48), (4.1.41) and (4.1.42) we obtain the result.

□

The next theorem contains the main result of this section. It shows that provided we start with a "fine enough" mesh $\mathcal{T}_n$, the mesh adaptivity algorithm will reduce the error in the energy norm.

Theorem 4.1.15 (Error reduction)**.** *For each $\theta \in (0,1)$, there exists a sufficiently fine mesh threshold $H_n^{\max}$ and constants $\mu > 0$ (all of which may depend on θ and on the eigenvalue λ) and $\alpha \in (0,1)$, with the following property. For any $\varepsilon > 0$ the inequality*

$$\text{osc}(\lambda_n u_n, \mathcal{T}_n) \leq \mu\varepsilon, \tag{4.1.49}$$

implies either $|||u - u_n|||_\Omega \leq \varepsilon$ or

$$|||u - u_{n+1}|||_\Omega \leq \alpha|||u - u_n|||_\Omega \ ,$$

where the constant α may depend also on the parameter θ and on the considered eigenvalue.

Proof. In view of the equation (4.1.37) and remembering that $u_{n+1} - u_n \in V_{n+1}$ we have

$$\begin{aligned} |||u - u_n|||_\Omega^2 - |||u - u_{n+1}|||_\Omega^2 &= |||u_{n+1} - u_n|||_\Omega^2 + 2a(u - u_{n+1}, u_{n+1} - u_n) \\ &= |||u_{n+1} - u_n|||_\Omega^2 + 2(\lambda u - \lambda_{n+1}u_{n+1}, u_{n+1} - u_n)_{0,\mathcal{B},\Omega} . \end{aligned}$$

Now using on the second term on the right hand side the Cauchy-Schwarz and the Young inequality $2ab \leq \frac{1}{4C_{\mathrm{PF}}^2}a^2 + 4C_{\mathrm{PF}}^2 b^2$, where C_{PF} is the constant of the Poincaré inequality, we get

$$\begin{aligned} |||u - u_n|||_\Omega^2 - |||u - u_{n+1}|||_\Omega^2 &\geq |||u_{n+1} - u_n|||_\Omega^2 \\ &\quad - 2\|\lambda u - \lambda_{n+1}u_{n+1}\|_{0,\mathcal{B},\Omega}\|u_{n+1} - u_n\|_{0,\mathcal{B},\Omega} \\ &\geq |||u_{n+1} - u_n|||_\Omega^2 - \frac{1}{4C_{\mathrm{PF}}^2}\|u_{n+1} - u_n\|_{0,\mathcal{B},\Omega}^2 \\ &\quad - 4C_{\mathrm{PF}}^2\|\lambda u - \lambda_{n+1}u_{n+1}\|_{0,\mathcal{B},\Omega}^2 \\ &\geq \frac{3}{4}|||u_{n+1} - u_n|||_\Omega^2 - 4C_{\mathrm{PF}}^2\|\lambda u - \lambda_{n+1}u_{n+1}\|_{0,\mathcal{B},\Omega}^2. \end{aligned} \tag{4.1.50}$$

Hence

$$|||u - u_{n+1}|||_\Omega^2 \leq |||u - u_n|||_\Omega^2 - \frac{3}{4}|||u_{n+1} - u_n|||_\Omega^2 + 4C_{\mathrm{PF}}^2\|\lambda u - \lambda_{n+1}u_{n+1}\|_{0,\mathcal{B},\Omega}^2 .$$

Applying Lemma 4.1.14 we obtain

$$\begin{aligned} |||u - u_{n+1}|||_\Omega^2 \leq &\left(1 - \frac{3}{4}\theta^2 + \hat{C}^2\,(H_n^{\max})^{2s}\right)|||u - u_n|||_\Omega^2 \\ &+ 4C_{\mathrm{PF}}^2\|\lambda u - \lambda_{n+1}u_{n+1}\|_{0,\mathcal{B},\Omega}^2 + \mathrm{osc}(\lambda_n u_n, \mathcal{T}_n)^2 \end{aligned}$$

Then making use of (4.1.46) we have

$$|||u - u_{n+1}|||_\Omega^2 \leq \beta_n\,|||u - u_n|||_\Omega^2 + \mathrm{osc}(\lambda_n u_n, \mathcal{T}_n)^2. \tag{4.1.51}$$

with

$$\beta_n := \left[1 - \frac{3}{4}\theta^2 + \big((C')^2 C_{\mathrm{PF}}^2 q^2 (C_{\mathrm{spec2}} + \lambda_n C_{\mathrm{adj}})^2 + \hat{C}^2\big)(H_n^{\max})^{2s}\right], \tag{4.1.52}$$

where C' is the constant hidden in (4.1.46).
Note that $H_n^{\max}$ can be chosen sufficiently small so that $\beta_m \leq \beta < 1$ for all $m \geq n$.
Consider now the consequences of the inequality (4.1.49). If $|||u - u_n|||_\Omega > \varepsilon$ then (4.1.51) implies

$$|||u - u_{n+1}|||_\Omega^2 \leq (\beta + \mu^2)\,|||u - u_n|||_\Omega^2 .$$

Now choose μ small enough so that

$$\alpha := (\beta + \mu^2)^{1/2} < 1 \ , \tag{4.1.53}$$

to complete the proof. □

4.1.2 Proof of convergence

The main result of this chapter is Theorem 4.1.17 below which proves convergence of the adaptive method and also demonstrates the decay of the quantity osc on the sequence of approximate eigenfunctions. Before proving the convergence result we need a final lemma.

Lemma 4.1.16. *There exists a constant $\tilde{\alpha} \in (0,1)$ such that*

$$\operatorname{osc}(u_{n+1}, \mathcal{T}_{n+1}) \ \leq \ \tilde{\alpha} \operatorname{osc}(u_n, \mathcal{T}_n) \ + \ (1+q)(H_n^{\max})^2 \, |||u - u_n|||_\Omega. \tag{4.1.54}$$

Proof. First recall that one of the key results in [42, Lemma 3.8] is the proof that the value of osc of any fixed function $v \in H_0^1(\Omega)$ is reduced by applying one refinement based on Marking Strategy 2 (Definition 4.1.4). Thus we have (in view of Algorithm 1):

$$\operatorname{osc}(u_n, \mathcal{T}_{n+1}) \ \leq \ \tilde{\alpha} \operatorname{osc}(u_n, \mathcal{T}_n), \tag{4.1.55}$$

where $0 < \tilde{\alpha} < 1$ is independent of u_n. Thus, a simple application of the triangle inequality combined with (4.1.55) yields

$$\begin{aligned} \operatorname{osc}(u_{n+1}, \mathcal{T}_{n+1}) \quad &\leq \quad \operatorname{osc}(u_n, \mathcal{T}_{n+1}) \ + \ \operatorname{osc}(u_{n+1} - u_n, \mathcal{T}_{n+1}) \\ &\leq \quad \tilde{\alpha} \operatorname{osc}(u_n, \mathcal{T}_n) \ + \ \operatorname{osc}(u_{n+1} - u_n, \mathcal{T}_{n+1}) \end{aligned} \tag{4.1.56}$$

A further application of the triangle inequality and then (4.1.6) yields

$$\begin{aligned} \operatorname{osc}(u_{n+1} - u_n, \mathcal{T}_{n+1}) \quad &\leq \quad \operatorname{osc}(u - u_{n+1}, \mathcal{T}_{n+1}) \ + \ \operatorname{osc}(u - u_n, \mathcal{T}_{n+1}) \\ &\lesssim \quad (H_{n+1}^{\max})^2 \, (|||u - u_{n+1}|||_\Omega + |||u - u_n|||_\Omega) \end{aligned} \tag{4.1.57}$$

and then combining (4.1.56) and (4.1.57) and applying Theorem 4.1.10 completes the proof. □

Theorem 4.1.17 (Convergence)**.** *Let (λ, u) be a simple eigenvalue of the continuous problem, then provided that the initial mesh $\mathcal{T}_0$ is chosen in such a way that $H_0^{\max}$ is small enough, there exists a constant $p \in (0,1)$ such that the recursive application of Algorithm 1 to solve problem* (1.3.7) *yields a convergent sequence of approximate*

eigenvectors, with the properties:

$$|||u - u_n|||_\Omega \ \leq \ C_0 q p^n, \tag{4.1.58}$$

and

$$\lambda_n \ \mathrm{osc}(u_n, \mathcal{T}_n) \ \leq \ C_1 p^n, \tag{4.1.59}$$

where C_0 and C_1 are constants and q is the constant defined in Theorem 4.1.10.

Remark 4.1.18. *The initial mesh convergence threshold and the constants C_1 and C_2 may depend on θ, $\tilde{\theta}$ and λ.*

Proof. The proof of this theorem is by induction and the induction step contains an application of Theorem 4.1.15. In order to ensure the reduction of the error, we have to assume that the starting mesh $\mathcal{T}_0$ is fine enough and that μ, which is defined in Theorem 4.1.15, is small enough such that for the chosen value of θ, the quantity α in (4.1.53) satisfies $\alpha < 1$.
Then with $\tilde{\alpha}$ as in Lemma 4.1.16, we set

$$\max\{\alpha, \tilde{\alpha}\} \ < \ p \ < \ 1 \ .$$

We also set

$$C_1 \ = \ \mathrm{osc}(\lambda_0 u_0, \mathcal{T}_0) \quad \text{and} \quad C_0 \ = \ \max\{\mu^{-1} p^{-1} C_1, \ |||u - u_0|||_\Omega\}.$$

First note that by the definition of C_0 and Theorem 4.1.10,

$$|||u - u_0|||_\Omega \leq C_0 \leq C_0 q,$$

since $q > 1$. Combined with the definition of C_1, it proves the result for $n = 0$.
Now, suppose that for some $n > 0$ the inequalities (4.1.58) and (4.1.59) hold.
Then let us consider the outcomes, depending on whether the inequality

$$|||u - u_n|||_\Omega \leq C_0 p^{n+1}, \tag{4.1.60}$$

holds or not. If (4.1.60) holds then we can apply Theorem 4.1.10 to conclude that

$$|||u - u_{n+1}|||_\Omega \leq q \ |||u - u_n|||_\Omega \leq q C_0 p^{n+1},$$

which proves (4.1.58) for $n + 1$.
On the other hand, if (4.1.60) does not hold then, by definition of C_0,

$$|||u - u_n|||_\Omega \ > \ C_0 p^{n+1} \ \geq \ \mu^{-1} C_1 p^n. \tag{4.1.61}$$

Also, since we have assumed (4.1.59) for n, we have

$$\lambda_n \operatorname{osc}(u_n, \mathcal{T}_n) \leq \mu\varepsilon \quad \text{with} \quad \varepsilon := \mu^{-1} C_1 p^n \ . \tag{4.1.62}$$

Then (4.1.61) and (4.1.62) combined with Theorem 4.1.15 yields

$$|||u - u_{n+1}|||_\Omega \ \leq \ \alpha |||u - u_n|||_\Omega$$

and so using the inductive hypothesis (4.1.58) combined with the definition of p, we have

$$|||u - u_{n+1}|||_\Omega \ \leq \ \alpha C_0 q p^n \leq q C_0 p^{n+1},$$

which again proves (4.1.58) for $n+1$.
To conclude the proof, we have to show that also (4.1.59) holds for $n+1$. Using Lemma 4.1.16 and the inductive hypothesis, we have

$$\begin{aligned} \lambda_{n+1} \operatorname{osc}(u_{n+1}, \mathcal{T}_{n+1}) \ &\leq \ \tilde{\alpha} C_1 p^n + (1+q)(H_n^{\max})^2 \lambda_n C_0 q p^n \\ &\leq \ (\tilde{\alpha} C_1 + (1+q)(H_0^{\max})^2 \lambda_0 C_0 q) p^n. \end{aligned} \tag{4.1.63}$$

Now, (recalling that $\tilde{\alpha} < p$), in addition to the condition already imposed on $H_0^{\max}$ we can further require that

$$\tilde{\alpha} C_1 + (1+q)(H_0^{\max})^2 |\lambda_0| C_0 q \ \leq \ p C_1.$$

This ensures that

$$\lambda_n \operatorname{osc}(u_{n+1}, \mathcal{T}_{n+1}) \ \leq \ C_1 p^{n+1},$$

thus concluding the proof. □

Corollary 4.1.19 (Convergence)**.** *Provided the initial mesh $\mathcal{T}_0$ is chosen so that $H_0^{\max}$ is small enough, there exists a constant $p \in (0,1)$ such that the recursive application of Algorithm 1 to solve problem* (1.3.7) *yields a convergent sequence of approximate eigenvalues, with the property:*

$$|\lambda - \lambda_n| \ \leq \ C_0^2 q^2 p^{2n}. \tag{4.1.64}$$

Proof. The proof is a straightforward application of Corollary 2.2.12 to (4.1.58). □

4.2 Convergent AFEM for PCF eigenvalue problems

The outline of this section is as follows. In Subsection 4.2.2 the convergence result for problem (1.3.9), which is the main result of this section, is presented. Meanwhile, in Subsection 4.2.1, we prove that mesh refining ensures error reduction (up to oscillation of the computed eigenfunction). Moreover, in Subsection 4.2.3, we present the convergence result for problem (1.3.8).

Assumption 4.2.1. *In Theorem 2.1.12 in Chapter 2 we proved that $a_{\kappa,S}(\cdot,\cdot)$ is coercive form any $S > 0$. But, in order to simplify the arguments below, we are going to assume in this section that $S \geq \overline{a}\underline{b}^{-1} \max_{\vec{\kappa}\in\mathcal{K}} |\vec{\kappa}|^2$. We would like to remark that all the results below holds also without this assumption, but in such case the proof is more complicated.*

We are going to use the same algorithm, Algorithm 1, which has been already used in the previous section. So, we are again going to use the error estimator η_n, defined in 4.1.1, and the quantity osc to drive the adaptivity. We recall from Chapter 3 that for PCF problems the error estimator η_n is defined as:

$$\eta_n := \Big\{ \sum_{\tau\in\mathcal{T}_n} H_\tau^2 \|R_I(u_n,\zeta_n)\|_{0,\tau}^2 \ + \ \sum_{f\in\mathcal{F}_n} H_f \|R_F(u_n)\|_{0,f}^2 \Big\}^{1/2},$$

where

$$R_I(u_n,\zeta_n)(x) := \big((\nabla + i\vec{\kappa})\cdot\mathcal{A}(\nabla + i\vec{\kappa})u_n \ + \ \zeta_n \mathcal{B} u_n\big)(x), \quad \text{with } x \in int(\tau), \quad \tau \in \mathcal{T}_n,$$

and

$$R_F(u_n)(x) := \big[\vec{n}_f \cdot \mathcal{A}(\nabla + i\vec{\kappa})u_n\big]_f(x), \quad \text{with } x \in int(f), \quad f \in \mathcal{F}_n.$$

Definition 4.2.2. *We define $\eta_{f,n}$ as:*

$$\eta_{f,n}^2 := \frac{1}{3}\|H_\tau \ R_I(u_n,\zeta_n - S)\|_{0,\Delta_f}^2 \ + \ \|H_f^{1/2} R_F(u_n)\|_{0,f}^2 \ , \tag{4.2.1}$$

where we denoted by Δ_f the union of the two elements $\tau_1(f)$ and $\tau_2(f)$ sharing f.

Since we are going to reuse Algorithm 1, we invite the reader to refer to the definitions of the two marking strategies contained in Section 4.1. The only remark that we would like to make about the marking strategies is that in the PCF context, (4.1.6) becomes

$$\mathrm{osc}(v,\mathcal{T}_n) \ \lesssim \ (H_n^{\max})^2 a_{\kappa,S}(v,v)^{1/2} \ , \quad \text{for all} \quad v \in H_\pi^1(\Omega) \ . \tag{4.2.2}$$

To simplify the notation in this section, we are going to embrace the same notation used in Section 4.1. So, from now on we fix the value of j because we restrict our analysis to the true eigenpair (ζ_j, u_j) and to the computed eigenpair on the mesh $\mathcal{T}_n$ $(\zeta_{j,n}, u_{j,n})$. So

we can drop the subscript j and we simply write (ζ, u) for the eigenpair of (1.3.9) and (ζ_n, u_n) for the eigenpair of (2.2.49). Moreover, we introduce the following notations:

Notation 4.2.3. *We write $A \lesssim B$ when A/B is bounded by a constant which may depend on the functions $\mathcal{A}$ and $\mathcal{B}$ in* (1.3.1) *and* (1.3.2)*, on S in* (1.3.9)*, on $C_{\text{ell}}^{\text{PCF}}$ in Assumption 2.2.20 and C_{reg} in* (2.2.1)*. The notation $A \cong B$ means $A \lesssim B$ and $A \gtrsim B$. All the constants depending on the spectrum, namely $C_{\text{adj}}^{\text{PCF}}$ in (3.1.43) and $C_{\text{spec1}}^{\text{PCF}}$ and $C_{\text{spec2}}^{\text{PCF}}$ in Theorem 2.2.24, are handled explicitly. Similarly all mesh size dependencies are explicit. Note that all eigenvalues of (1.3.9) satisfy $\zeta_n \gtrsim 1$, since $\zeta_n \geq \zeta_1 = a_{\kappa,S}(u_1, u_1) \gtrsim \|u_1\|_{1,\Omega}^2 \gtrsim \|u_1\|_{0,\mathcal{B},\Omega}^2 = 1$.*

Notation 4.2.4. *In this section we denote by $|||u|||_{\kappa,S,\Omega}$ the norm $a_{\kappa,S}(u,u)^{1/2}$, which is related to the problem* (1.3.9)*. Moreover, we are going to apply the same notation for H_n explained in Notation 4.1.2.*

Remark 4.2.5. *We assume in this chapter that ζ is a simple eigenvalue. This implies that the corresponding eigenspace has dimension 1 and that it is possible to find two unit eigenvectors corresponding to ζ, namely u or $-u$. In other words, there is not a unique eigenvector corresponding to ζ, but two. The same is true for all the eigenvalues ζ_n computed in Algorithm 1, which approximate ζ. In fact, for each n, both (ζ_n, u_n) and $(\zeta_n, -u_n)$ are acceptable eigenpairs for the discrete problem. Similarly to what we have done in Remark 4.1.6 for generic elliptic eigenvalue problems, we assume that u_0 is the eigenfunction actually computed in the first iteration of Algorithm 1, then we set $u := U$, where U is constructed as in the proof of Theorem 3.1.8. Then, we set for each $n > 0$ the eigenfunction $u_n := w_n$, where w_n comes from Theorem 2.2.24.*

The next theorem extends the result of Theorem 4.1.10 to the PCF case. The proof of this theorem follows by the same arguments used in the proof of Theorem 4.1.10.

Theorem 4.2.6. *Let ζ be a simple eigenvalue of (1.3.9) and let (ζ_n, u_n) and (ζ_m, u_m) be computed eigenpairs, in the sense of Remark 2.2.23. Let also the true eigenfunction u and the approximated ones u_n and u_m be defined in the sense of Remark 4.2.5. Then there exists a constant $q^{\text{PCF}} > 1$ such that, for all $m \geq n$, the corresponding computed eigenpair (ζ_m, u_m) satisfies:*

$$|||u - u_m|||_{\kappa,S,\Omega} \leq q^{\text{PCF}} \, |||u - u_n|||_{\kappa,S,\Omega} \, . \tag{4.2.3}$$

4.2.1 Error Reduction

In this section we give the proof of error reduction for Algorithm 1 for problem (1.3.9). The proof has been inspired by the corresponding theory for source problems in [42]. However the nonlinearity of the eigenvalue problem introduces new complications and there are several lemmas before the main theorem (Theorem 4.2.11).

The first lemma is similar to Lemma 4.1.11, but in this case we are going to treat only the 2D case, since in this work we analyse only PCF problems, which are in the end 2D problems.

Lemma 4.2.7. *Let $\hat{\mathcal{F}}_n$ be as defined in Definition 4.1.1 and let P_n be as defined in (4.1.4). For any $f \in \hat{\mathcal{F}}_n$, there exists a function $\Phi_f \in V_{n+1}$ such that* $\mathrm{supp}(\Phi_f) = \Delta_f$, *where Δ_f is the union of the two elements $\tau_1(f)$ and $\tau_2(f)$ sharing f, and also*

$$\int_{\Delta_f} P_n R_I(u_n, \zeta_n - S)\overline{\Phi_f} - \int_f R_F(u_n)\overline{\Phi_f} \;=\; \|H_n\, P_n R_I(u_n, \zeta_n - S)\|^2_{0,\Delta_f} + \|H_f^{1/2} R_F(u_n)\|^2_{0,f}, \tag{4.2.4}$$

and

$$|||\Phi_f|||^2_{\Delta_f} \;\lesssim\; \big(1 + H_f^2\big)\Big(\|H_n\; P_n R_I(u_n, \zeta_n - S)\|^2_{0,\Delta_f} + \|H_f^{1/2} R_F(u_n)\|^2_{0,f}\Big)\,. \tag{4.2.5}$$

Remark 4.2.8. *The function $P_n R_I(u_n, \zeta_n - S)$ in Lemma 4.2.7 is the projection of the elementwise linear functional $R_I(u_n, \zeta_n - S)$ on the set of elementwise constant functions. Using the linearity of the projection operator P_n we have:*

$$\begin{aligned} P_n R_I(u_n, \zeta_n - S) \;&=\; P_n\big((\nabla + i\vec{\kappa}) \cdot \mathcal{A}(\nabla + i\vec{\kappa})u_n \;+\; (\zeta_n \;-\; S)\mathcal{B}u_n\big) \\ &= \nabla \cdot \mathcal{A} i\vec{\kappa} u_n \;+\; i\vec{\kappa} \cdot \mathcal{A}\nabla u_n \;- \vec{\kappa} \cdot \mathcal{A}\vec{\kappa} P_n u_n \;+\; (\zeta_n \;-\; S)\mathcal{B}P_n u_n, \end{aligned}$$

the reason why the term $\nabla \cdot \mathcal{A}\nabla u_n$ disappeared is because we are using linear elements, instead, the reason why the operator P_n does not appear in all terms is because these terms are already elementwise constant.

Proof. We invite the reader to refer to Figure 4-2 in Section 4.1, which illustrates possible configuration for Δ_f. The point x_f is the node created by the red-refinement in the middle of the shared edge f while the points x_1 and x_2 are the nodes created in the interior of the refined elements $\tau_1(f)$ and $\tau_2(f)$ respectively.
The two situations in Figure 4-2 do not exhaust all the possible configurations for couples of adjacent refined elements. There could be other possible configurations different from Figure 4-2b, in which the green-refinements are applied to different edges. However, the way in which the green-refinements split the elements is irrelevant for the proof, since the only important thing is the existence of an new node on the shared edge and two nodes in the interior of the elements.
We denote by $\tau_1(f)$ and $\tau_2(f)$ the elements sharing the edge f and, we denote by x_f the node created in the middle of the shared edge f while the points x_1 and x_2 are the nodes created in the interior of the refined elements $\tau_1(f)$ and $\tau_2(f)$ respectively.
We start proving (4.2.4). We then define

$$\Phi_f := \alpha_f \varphi_f + \beta_1 \varphi_1 + \beta_2 \varphi_2, \tag{4.2.6}$$

where φ_f and φ_i are the nodal basis functions associated with the points x_f and x_i on $\mathcal{T}_{n+1}$, and α_f, β_i are defined by

$$\overline{\alpha_f} = \begin{cases} -\dfrac{\|H_f^{1/2} R_F(u_n)\|_{0,f}^2}{\int_f R_F(u_n)\overline{\varphi_f}} & \text{if } R_F(u_n) \neq 0, \\ \\ 0 & \text{otherwise,} \end{cases} \tag{4.2.7}$$

and

$$\overline{\beta_i} = \begin{cases} \dfrac{\|H_n\ P_n R_I(u_n, \zeta_n - S)\|_{0,\tau_i(f)}^2 - \overline{\alpha_f} \int_{\tau_i(f)} P_n R_I(u_n, \zeta_n - S)\ \overline{\varphi_f}}{\int_{\tau_i(f)} P_n R_I(u_n, \zeta_n - S)\ \overline{\varphi_i}} & \text{if } P_n R_I(u_n, \zeta_n)|_{\tau_i(f)} \neq 0, \\ \\ 0 & \text{otherwise,} \end{cases} \tag{4.2.8}$$

for $i = 1, 2$.
Using the fact that $\mathrm{supp}(\varphi_i) = \tau_i(f)$, for $i = 1, 2$ we can easily see that the above formulae imply

$$\overline{\alpha_f} \int_f R_F(u_n)\overline{\varphi_f} \;=\; -\|H_f^{1/2} R_F(u_n)\|_{0,f}^2, \tag{4.2.9}$$

$$\int_{\Delta_f} P_n R_I(u_n, \zeta_n - S)\overline{(\alpha_f\varphi_f + \beta_1\varphi_1 + \beta_2\varphi_2)} \;=\; \|H_n P_n R_I(u_n, \zeta_n - S)\|_{0,\Delta_f}^2, \tag{4.2.10}$$

(and that these formulae remain true even if $R_F(u_n)$ or $P_n R_I(u_n, \zeta_n)|_{\tau_i(f)}$ vanish). Hence

$$\begin{aligned} \textstyle\int_{\Delta_f} P_n R_I(u_n, \zeta_n - S)\overline{\Phi_f} - \int_f R_F(u_n)\overline{\Phi_f} \;&=\; \textstyle\int_{\Delta_f} P_n R_I(u_n, \zeta_n - S)\overline{(\alpha_f\varphi_f + \beta_1\varphi_1 + \beta_2\varphi_2)} \\ &\quad - \textstyle\int_f R_F(u_n)\overline{\alpha_f\varphi_f} \end{aligned}$$

and (4.2.4) follows immediately on using (4.2.9) and (4.2.10).
To prove (4.2.5), we use (4.2.6), and the facts that $|\varphi_f|_{1,\Delta_f} \lesssim 1$, $|\varphi_i|_{1,\Delta_f} \lesssim 1$, $|\varphi_f|_{0,\Delta_f} \lesssim H_f$, $|\varphi_i|_{0,\Delta_f} \lesssim H_{\tau_i(f)}$ and the shape regularity of the mesh to obtain

$$|||\Phi_f|||_{\kappa,S,\Delta_f}^2 \;\lesssim\; \left(1 + H_f^2\right)\left(|\alpha_f|^2 + |\beta_1|^2 + |\beta_2|^2\right). \tag{4.2.11}$$

Now, since $R_F(u_n)$ is constant on f and $\int_f \varphi_f \sim H_f$, we have

$$|\alpha_f| \;\lesssim\; \frac{|R_F(u_n)|\|H_f^{1/2}\|_{0,f}^2}{H_f} \;\lesssim\; |R_F(u_n)|H_f \;\sim\; \|H_f^{1/2} R_F(u_n)\|_{0,f}. \tag{4.2.12}$$

Also since $P_n R_I(u_n, \zeta_n)$ is constant on each $\tau_i(f)$ and since $\int_{\tau_i(f)} \varphi_i \sim H^2_{\tau_i(f)}$, we have

$$\begin{aligned} |\beta_i| &\lesssim \frac{|P_n R_I(u_n, \zeta_n - S)|_{\tau_i(f)}| \, \|H_n\|^2_{0,\tau_i(f)} \; + \; |\alpha_f| H^2_{\tau_i(f)}}{H^2_{\tau_i(f)}} \\ &\lesssim |P_n R_I(u_n, \zeta_n - S)|_{\tau_i(f)}| \; H^2_{\tau_i(f)} + |\alpha_f| \; \sim \; \|H_n P_n R_I(u_n, \zeta_n - S)\|_{0,\tau_i(f)} + |\alpha_f| \end{aligned}$$

This implies

$$\begin{aligned} |\beta_i|^2 &\lesssim \|H_n P_n R_I(u_n, \zeta_n - S)\|^2_{0,\tau_i(f)} + |\alpha_f|^2 \\ &\lesssim \|H_n P_n R_I(u_n, \zeta_n - S)\|^2_{0,\tau_i(f)} + \|H_f^{1/2} R_F(u_n)\|^2_{0,f} \; , \end{aligned} \tag{4.2.13}$$

and the proof is completed by combining (4.2.11) with (4.2.12) and (4.2.13).

□

In the next lemma we bound the local error estimator from above by the local difference of two discrete solutions coming from consecutive meshes, plus higher order terms. This kind of result is called "discrete local efficiency" by many authors.
Recall that $\mathcal{T}_{n+1}$ is the refinement of $\mathcal{T}_n$ obtained by applying Algorithm 1.

Lemma 4.2.9. *For any $f \in \hat{\mathcal{F}}_n$, we have*

$$\begin{aligned} \eta^2_{f,n} \lesssim \big(1 + H_f^2\big)\Big(& |||u_{n+1} - u_n|||^2_{\kappa,S,\Delta_f} \\ & + \|H_n(\zeta_{n+1} u_{n+1} - \zeta_n P_n u_n)\|^2_{0,\mathcal{B},\Delta_f} + S^2 \|H_n(u_n - P_n u_n)\|^2_{0,\mathcal{B},\Delta_f} \\ & + \big((\zeta_n - S)^2 + S^2\big)\|H_n \mathcal{B}(u_n - P_n u_n)\|^2_{0,\Delta_f}\Big) \; . \end{aligned} \tag{4.2.14}$$

Proof. Since the function Φ_f defined in Lemma 4.2.7 is in V_{n+1} and $\mathrm{supp}(\Phi_f) = \Delta_f$, we have

$$\begin{aligned} a_{\kappa,S}(u_{n+1} - u_n, \Phi_f) &= a_{\kappa,S}(u_{n+1}, \Phi_f) - a_{\kappa,S}(u_n, \Phi_f) \\ &= \zeta_{n+1} \int_{\Delta_f} \mathcal{B} u_{n+1} \overline{\Phi_f} \; - \; a_{\kappa,S}(u_n, \Phi_f). \end{aligned} \tag{4.2.15}$$

Now applying integration by parts to the last term on the right-hand side of (4.2.15), we obtain

$$\begin{aligned} a_{\kappa,S}(u_{n+1} - u_n, \Phi_f) &= \zeta_{n+1} \int_{\Delta_f} \mathcal{B} u_{n+1} \overline{\Phi_f} \\ &+ \sum_{\tau \in \Delta_f} \int_\tau ((\nabla + i\vec{\kappa}) \cdot \mathcal{A}(\nabla + i\vec{\kappa}) u_n - S\mathcal{B} u_n) \overline{\Phi_f} - \int_f R_F(u_n) \overline{\Phi_f} \; . \end{aligned} \tag{4.2.16}$$

Combining (4.2.16) with (4.2.4) and using Remark 4.2.8, we obtain

$$\begin{aligned} a_{\kappa,S}(u_{n+1}-u_n,\Phi_f) &- \int_{\Delta_f}\mathcal{B}(\zeta_{n+1}u_{n+1}-\zeta_nP_nu_n)\overline{\Phi_f} \\ &+ S\int_{\Delta_f}\mathcal{B}(u_n-P_nu_n)\overline{\Phi_f}+\int_{\Delta_f}\vec{\kappa}\cdot\mathcal{A}\vec{\kappa}(u_n-P_nu_n)\overline{\Phi_f} \\ &= \sum_{\tau\in\Delta_f}\int_{\tau}P_nR_I(u_n,\zeta_n-S)\overline{\Phi_f}-\int_f R_F(u_n)\overline{\Phi_f} \\ &= \|H_n\ P_nR_I(u_n,\zeta_n-S)\|^2_{0,\Delta_f}+\|H_f^{1/2}R_F(u_n)\|^2_{0,f}. \quad (4.2.17) \end{aligned}$$

Rearranging (4.2.17) and then applying the triangle inequality, we obtain:

$$\begin{aligned} &\|H_n\ P_nR_I(u_n,\zeta_n-S)\|^2_{0,\Delta_f}+\|H_f^{1/2}R_F(u_n)\|^2_{0,f} \\ &\quad\le\ |a_{\kappa,S}(u_{n+1}-u_n,\Phi_f)|+\left|\int_{\Delta_f}\mathcal{B}(\zeta_{n+1}u_{n+1}-\zeta_nP_nu_n)\overline{\Phi_f}\right| \\ &\qquad+\left|S\int_{\Delta_f}\mathcal{B}(u_n-P_nu_n)\overline{\Phi_f}\right|+\left|\int_{\Delta_f}\vec{\kappa}\cdot\mathcal{A}\vec{\kappa}(u_n-P_nu_n)\overline{\Phi_f}\right| \quad (4.2.18) \end{aligned}$$

The last term of (4.2.18) can be absorbed in the term $S\int_{\Delta_f}\mathcal{B}(u_n-P_nu_n)\overline{\Phi_f}$, since we have assumed in Assumption 4.2.1 that $|\vec{\kappa}|^2\lesssim S$, so

$$\int_{\Delta_f}\vec{\kappa}\cdot\mathcal{A}\vec{\kappa}(u_n-P_nu_n)\overline{\Phi_f}\lesssim S\int_{\Delta_f}\mathcal{B}(u_n-P_nu_n)\overline{\Phi_f},$$

in view of this fact, (4.2.18) becomes:

$$\begin{aligned} &\|H_n\ P_nR_I(u_n,\zeta_n-S)\|^2_{0,\Delta_f}+\|H_f^{1/2}R_F(u_n)\|^2_{0,f} \\ &\quad\lesssim\ |a_{\kappa,S}(u_{n+1}-u_n,\Phi_f)|+\left|\int_{\Delta_f}\mathcal{B}(\zeta_{n+1}u_{n+1}-\zeta_nP_nu_n)\overline{\Phi_f}\right| \\ &\qquad+\left|S\int_{\Delta_f}\mathcal{B}(u_n-P_nu_n)\overline{\Phi_f}\right|. \quad (4.2.19) \end{aligned}$$

Then applying the Cauchy-Schwarz inequalities to (4.2.19), we get:

$$\begin{aligned} &\|H_n\ P_nR_I(u_n,\zeta_n-S)\|^2_{0,\Delta_f}+\|H_f^{1/2}R_F(u_n)\|^2_{0,f} \\ &\quad\lesssim\ |||u_{n+1}-u_n|||_{\kappa,S,\Delta_f}|||\Phi_f|||_{\kappa,S,\Delta_f}+\|\zeta_{n+1}u_{n+1}-\zeta_nP_nu_n\|_{0,\mathcal{B},\Delta_f}\|\Phi_f\|_{0,\mathcal{B},\Delta_f} \\ &\qquad+S\|u_n-P_nu_n\|_{0,\mathcal{B},\Delta_f}\|\Phi_f\|_{0,\mathcal{B},\Delta_f} \\ &\quad\lesssim\ \Big(|||u_{n+1}-u_n|||_{\kappa,S,\Delta_f}+\|H_n(\zeta_{n+1}u_{n+1}-\zeta_nP_nu_n)\|_{0,\mathcal{B},\Delta_f} \\ &\qquad+S\|H_n(u_n-P_nu_n)\|_{0,\mathcal{B},\Delta_f}\Big)|||\Phi_f|||_{\kappa,S,\Delta_f}\ , \quad (4.2.20) \end{aligned}$$

where in the final step of (4.2.20) we made use of the Poincaré inequality

$$\|\Phi_f\|_{0,\mathcal{B},\Delta_f} \lesssim H_f|\Phi_f|_{1,\Delta_f} \ ,$$

the coercivity of the bilinear form $|\Phi_f|_{1,\Delta_f} \lesssim |||\Phi_f|||_{\kappa,S,\Delta_f}$ and also the shape-regularity of the meshes.
In view of (4.2.5), yields

$$\begin{aligned}\|H_n\ P_nR_I(u_n,\zeta_n - S)\|^2_{0,\Delta_f} + \|H_f^{1/2}R_F(u_n)\|^2_{0,f}& \\ \lesssim\ \big(1+H_f^2\big)\Big(|||u_{n+1}-u_n|||^2_{\kappa,S,\Delta_f} + \|H_n(\zeta_{n+1}u_{n+1}-\zeta_nP_nu_n)\|^2_{0,\mathcal{B},\Delta_f}& \\ + S^2\|H_n(u_n-P_nu_n)\|^2_{0,\mathcal{B},\Delta_f}\Big)\ .& \qquad (4.2.21)\end{aligned}$$

From the definition of $\eta_{f,n}$ in (4.2.1), and the triangle inequality, we have

$$\begin{aligned}\eta^2_{f,n} \ \lesssim\ & \|H_n\ P_nR_I(u_n,\zeta_n-S)\|^2_{0,\Delta_f} \\ +\ & \|H_f^{1/2}R_F(u_n)\|^2_{0,f} + \big\|H_n\big((\zeta_n-S)\mathcal{B}-\vec{\kappa}\cdot\mathcal{A}\vec{\kappa}\big)(u_n-P_nu_n)\big\|^2_{0,\Delta_f}\end{aligned} \qquad (4.2.22)$$

where we have used $R_I(u_n,\zeta_n) = P_nR_I(u_n,\zeta_n-S) + \big((\zeta_n-S)\mathcal{B}-\vec{\kappa}\cdot\mathcal{A}\vec{\kappa}\big)(u_n-P_nu_n)$. In order to simplify the result, we can use again the fact $|\vec{\kappa}|^2 \lesssim S$ as follows:

$$\begin{aligned}&\big\|H_n\big((\zeta_n-S)\mathcal{B}-\vec{\kappa}\cdot\mathcal{A}\vec{\kappa}\big)(u_n-P_nu_n)\big\|^2_{0,\Delta_f} \\ &\quad\le\ \|H_n(\zeta_n-S)\mathcal{B}(u_n-P_nu_n)\|^2_{0,\Delta_f} + \|H_n\vec{\kappa}\cdot\mathcal{A}\vec{\kappa}(u_n-P_nu_n)\|^2_{0,\Delta_f} \\ &\quad\lesssim\ \|H_n(\zeta_n-S)\mathcal{B}(u_n-P_nu_n)\|^2_{0,\Delta_f} + \|H_nS\mathcal{B}(u_n-P_nu_n)\|^2_{0,\Delta_f}\ .\end{aligned} \qquad (4.2.23)$$

The required inequality (4.2.14) now follows from (4.2.21), (4.2.22) and (4.2.23). □

In the main result of this section, Theorem 4.2.11 below, we achieve error reduction of the form $|||u-u_{n+1}|||_{\kappa,S,\Omega} \le \alpha|||u-u_n|||_{\kappa,S,\Omega}$ for some $\alpha<1$. In the case of source problems (see [42]) this is approached by writing

$$\begin{aligned}|||u-u_n\ |||^2_{\kappa,S,\Omega} &= |||u-u_{n+1}+u_{n+1}-u_n\ |||^2_{\kappa,S,\Omega} \\ &= |||u-u_{n+1}\ |||^2_{\kappa,S,\Omega} + |||u_{n+1}-u_n\ |||^2_{\kappa,S,\Omega} \\ &\quad + 2a_{\kappa,S}(u-u_{n+1},u_{n+1}-u_n).\end{aligned} \qquad (4.2.24)$$

and making use of the fact that the last term on the right-hand side vanishes due to Galerkin orthogonality. However this approach is not available in the eigenvalue problem context. Therefore a more technical approach is needed to bound the two terms on the right-hand side of (4.2.24) from below. The main technical result is in the following lemma. Recall the convention in Notation 4.2.3.

Lemma 4.2.10.

$$\begin{aligned}|||u_{n+1}-u_n|||^2_{\kappa,S,\Omega} \quad &\gtrsim\ \theta^2\ \big(1+(H_n^{\max})^2\big)^{-1}|||u-u_n|||^2_{\kappa,S,\Omega} \\ &\quad -\big((\zeta_n-S)^2\overline{b}+S^2(1+\overline{b})+\zeta_n^2\big)\mathrm{osc}(u_n,\mathcal{T}_n)^2-L_n^2\ ,\end{aligned} \tag{4.2.25}$$

where θ is defined in the marking strategy in Definition 4.1.1 and L_n satisfies the estimate:

$$L_n\ \lesssim\ \hat{C}\ (H_n^{\max})^s|||u-u_n|||_{\kappa,S,\Omega}\ , \tag{4.2.26}$$

where $\hat{C}$ depends on θ, ζ_n, $C^{\mathrm{PCF}}_{\mathrm{spec2}}$, $C^{\mathrm{PCF}}_{\mathrm{adj}}$ and q^{PCF}.

Proof. By Lemma 4.2.9 and Definition 4.1.1 we have

$$\begin{aligned}\theta^2\eta_n^2 \quad &\leq\ \textstyle\sum_{f\in\hat{\mathcal{F}}_n}\eta_{f,n}^2 \\ &\lesssim\ \big(1+(H_n^{\max})^2\big)\Big(\ |||u_{n+1}-u_n\ |||^2_{\kappa,S,\Omega}+\|H_n(\zeta_{n+1}u_{n+1}-\zeta_nP_nu_n)\|^2_{0,\mathcal{B},\Omega} \\ &+\ \big((\zeta_n-S)^2\overline{b}+S^2(1+\overline{b})\big)\mathrm{osc}(u_n,\mathcal{T}_n)^2\Big)\ .\end{aligned}$$

Hence, rearranging and making use of Theorem 3.3.5, we have

$$\begin{aligned}|||u_{n+1}-u_n|||^2_{\kappa,S,\Omega} \quad &\gtrsim \quad \theta^2\ \big(1+(H_n^{\max})^2\big)^{-1}\eta_n^2-\|H_n(\zeta_{n+1}u_{n+1}-\zeta_nP_nu_n)\|^2_{0,\mathcal{B},\Omega} \\ &\qquad -\big((\zeta_n-S)^2\overline{b}+S^2(1+\overline{b})\big)\mathrm{osc}(u_n,\mathcal{T}_n)^2 \\ &\gtrsim \quad \theta^2\ \big(1+(H_n^{\max})^2\big)^{-1}|||u-u_n|||^2_{\kappa,S,\Omega} \\ &\qquad -\big((\zeta_n-S)^2\overline{b}^2+S^2(1+\overline{b})\big)\mathrm{osc}(u_n,\mathcal{T}_n)^2-\theta^2\ \big(1+(H_n^{\max})^2\big)^{-1}G_n^2 \\ &\qquad -\|H_n(\zeta_{n+1}u_{n+1}-\zeta_nP_nu_n)\|^2_{0,\mathcal{B},\Omega}\ .\end{aligned} \tag{4.2.27}$$

We now estimate the last two terms in (4.2.27) separately.
To estimate G_n, we use (3.3.11), combined with the H^1 - ellipticity of $a_{\kappa,S}(\cdot,\cdot)$ and then Theorem 3.1.9 to obtain

$$G_n\ \lesssim\ \frac{1}{2}(\zeta+\zeta_n)\|u-u_n\|_{0,\mathcal{B},\Omega}\ \lesssim\ \frac{1}{2}(\zeta+\zeta_n)C^{\mathrm{PCF}}_{\mathrm{adj}}(H_n^{\max})^s|||u-u_n|||_{\kappa,S,\Omega}. \tag{4.2.28}$$

To estimate the last term in (4.2.27), we first use the triangle inequality to obtain

$$\|H_n(\zeta_{n+1}u_{n+1}-\zeta_nP_nu_n)\|_{0,\mathcal{B},\Omega}\ \leq\ \|H_n(\zeta_{n+1}u_{n+1}-\zeta_nu_n)\|_{0,\mathcal{B},\Omega}\ +\ \zeta_n\mathrm{osc}(u_n,\mathcal{T}_n). \tag{4.2.29}$$

For the first term on the right-hand side of (4.2.29), we have

$$\|H_n(\zeta_{n+1}u_{n+1}-\zeta_nu_n)\|_{0,\mathcal{B},\Omega}\leq H_n^{\max}\big(\|\zeta u-\zeta_{n+1}u_{n+1}\|_{0,\mathcal{B},\Omega}+\|\zeta u-\zeta_nu_n\|_{0,\mathcal{B},\Omega}\big). \tag{4.2.30}$$

From Corollary 2.2.27 we have that

$$|\zeta - \zeta_{n+1}| \ \leq \ |||u - u_{n+1}|||^2_{\kappa,S,\Omega} \ ,$$

then using this result and Theorem 3.1.9, we obtain

$$\begin{aligned} \|\zeta u - \zeta_{n+1}u_{n+1}\|_{0,\mathcal{B},\Omega} &\leq |\zeta - \zeta_{n+1}|\|u\|_{0,\mathcal{B},\Omega} \ + \ \zeta_{n+1}\|u - u_{n+1}\|_{0,\mathcal{B},\Omega} \\ &\leq |||u - u_{n+1}|||^2_{\kappa,S,\Omega} \\ &\quad + \ \zeta_{n+1}C^{\mathrm{PCF}}_{\mathrm{adj}}(H^{\max}_n)^s|||u - u_{n+1}|||_{\kappa,S,\Omega} \ . \end{aligned} \tag{4.2.31}$$

Using Theorem 3.1.9 again, the minimum-maximum principle and then Theorem 4.2.6, this implies

$$\begin{aligned} \|\zeta u - \zeta_{n+1}u_{n+1}\|_{0,\mathcal{B},\Omega} &\lesssim (C^{\mathrm{PCF}}_{\mathrm{spec2}} + \zeta_{n+1}C^{\mathrm{PCF}}_{\mathrm{adj}})(H^{\max}_n)^s|||u - u_{n+1}|||_{\kappa,S,\Omega} \\ &\leq q^{\mathrm{PCF}}(C^{\mathrm{PCF}}_{\mathrm{spec2}} + \zeta_n C^{\mathrm{PCF}}_{\mathrm{adj}})(H^{\max}_n)^s|||u - u_n|||_{\kappa,S,\Omega}, \end{aligned} \tag{4.2.32}$$

An identical argument shows

$$\|\zeta u - \zeta_n u_n\|_{0,\mathcal{B},\Omega} \ \lesssim \ (C^{\mathrm{PCF}}_{\mathrm{spec2}} + \zeta_n C^{\mathrm{PCF}}_{\mathrm{adj}})(H^{\max}_n)^s|||u - u_n|||_{\kappa,S,\Omega} \ . \tag{4.2.33}$$

Combining (4.2.32) and (4.2.33) with (4.2.30), we obtain

$$\|H_n(\zeta_{n+1}u_{n+1} - \zeta_n u_n)\|_{0,\mathcal{B},\Omega} \ \lesssim \ (1 + q^{\mathrm{PCF}})(C^{\mathrm{PCF}}_{\mathrm{spec2}} + \zeta_n C^{\mathrm{PCF}}_{\mathrm{adj}})(H^{\max}_n)^{s+1}|||u - u_n|||_{\kappa,S,\Omega} \ . \tag{4.2.34}$$

Now combining (4.2.27) with (4.2.34), (4.2.28) and (4.2.29) we obtain the result.

□

The next theorem contains the main result of this section. It shows that provided that we start with a "fine enough" mesh $\mathcal{T}_n$, the mesh adaptivity algorithm will reduce the error in the energy norm.

Theorem 4.2.11 (Error reduction)**.** *For each $\theta \in (0,1)$, exists a sufficiently fine mesh threshold $H^{\max}_n$ and constants $\mu > 0$ (both of them may depend on θ and on the eigenvalue λ) and $\alpha \in (0,1)$, with the following property. For any $\varepsilon > 0$ the inequality*

$$\left((\zeta_n - S)^2\overline{b} + S^2(1 + \overline{b}) + \zeta_n^2\right) \ \mathrm{osc}(u_n, \mathcal{T}_n) \leq \mu\varepsilon, \tag{4.2.35}$$

implies either $|||u - u_n|||_{\kappa,S,\Omega} \leq \varepsilon$ or

$$|||u - u_{n+1}|||_{\kappa,S,\Omega} \ \leq \ \alpha|||u - u_n|||_{\kappa,S,\Omega} \ ,$$

where the constant α may depend also on the parameter θ and on λ.

Proof. In view of the equation (4.2.24) and remembering that $u_{n+1} - u_n \in V_{n+1}$ we have

$$\begin{aligned} |||u - u_n|||^2_{\kappa,S,\Omega} - |||u - u_{n+1}|||^2_{\kappa,S,\Omega} &= |||u_{n+1} - u_n|||^2_{\kappa,S,\Omega} + 2a_{\kappa,S}(u - u_{n+1}, u_{n+1} - u_n) \\ &= |||u_{n+1} - u_n|||^2_{\kappa,S,\Omega} + 2(\zeta u - \zeta_{n+1}u_{n+1}, u_{n+1} - u_n)_{0,\mathcal{B},\Omega} . \end{aligned} \tag{4.2.36}$$

In the next step we will use the following inequality, which easily comes from Theorem 2.1.12:

$$\|u\|^2_{0,\mathcal{B},\Omega} \leq \frac{\overline{b}}{c^{\mathrm{PCF}}_{a,S}} |||u|||^2_{\kappa,S,\Omega} , \quad \text{for all } u \in H^1_\pi(\Omega),$$

in order to simplify the notation, we will denote the constant $C_{\mathcal{B},S} := \frac{\overline{b}}{c^{\mathrm{PCF}}_{a,S}}$.
Now using on the second term on the right hand side of (4.2.36) the Cauchy-Schwarz and the Young inequality $2ab \leq \frac{1}{4C_{\mathcal{B},S}}a^2 + 4C_{\mathcal{B},S}b^2$, where $C_{\mathcal{B},S}$ is the constant of the Poincaré inequality, we get

$$\begin{aligned} |||u - u_n|||^2_{\kappa,S,\Omega} - |||u - u_{n+1}|||^2_{\kappa,S,\Omega} &\geq |||u_{n+1} - u_n|||^2_{\kappa,S,\Omega} \\ &\quad - 2\|\zeta u - \zeta_{n+1}u_{n+1}\|_{0,\mathcal{B},\Omega}\|u_{n+1} - u_n\|_{0,\mathcal{B},\Omega} \\ &\geq |||u_{n+1} - u_n|||^2_{\kappa,S,\Omega} - \frac{1}{4C_{\mathcal{B},S}}\|u_{n+1} - u_n\|^2_{0,\mathcal{B},\Omega} \\ &\quad - 4C_{\mathcal{B},S}\|\zeta u - \zeta_{n+1}u_{n+1}\|^2_{0,\mathcal{B},\Omega} \\ &\geq \frac{3}{4} |||u_{n+1} - u_n|||^2_{\kappa,S,\Omega} - 4C_{\mathcal{B},S}\|\zeta u - \zeta_{n+1}u_{n+1}\|^2_{0,\mathcal{B},\Omega}. \end{aligned} \tag{4.2.37}$$

Hence

$$|||u-u_{n+1}|||^2_{\kappa,S,\Omega} \leq |||u-u_n|||^2_{\kappa,S,\Omega} - \frac{3}{4}|||u_{n+1}-u_n|||^2_{\kappa,S,\Omega} + 4C_{\mathcal{B},S}\|\zeta u-\zeta_{n+1}u_{n+1}\|^2_{0,\mathcal{B},\Omega} .$$

Applying Lemma 4.2.10 we obtain

$$\begin{aligned} |||u - u_{n+1}|||^2_{\kappa,S,\Omega} \leq & \left(1 - \frac{3}{4}\theta^2 \left(1 + (H^{\max}_n)^2\right)^{-1} + \hat{C}^2 (H^{\max}_n)^{2s}\right) |||u - u_n|||^2_{\kappa,S,\Omega} \\ & + 4C_{\mathcal{B},S}\|\zeta u - \zeta_{n+1}u_{n+1}\|^2_{0,\mathcal{B},\Omega} \\ & + \big((\zeta_n - S)^2\overline{b} + S^2(1 + \overline{b}) + \zeta_n^2\big)\mathrm{osc}(u_n, \mathcal{T}_n)^2 \end{aligned}$$

Then making use of (4.2.32) we have

$$\begin{aligned} |||u - u_{n+1}|||^2_{\kappa,S,\Omega} \leq & \ \beta_n |||u - u_n|||^2_{\kappa,S,\Omega} \\ & + \big((\zeta_n - S)^2\overline{b} + S^2(1 + \overline{b}) + \zeta_n^2\big)\mathrm{osc}(u_n, \mathcal{T}_n)^2 . \end{aligned} \tag{4.2.38}$$

with

$$\beta_n := \left[1 - \frac{3}{4}\theta^2 \left(1 + (H_n^{\max})^2\right)^{-1} + \left((C')^2 C_{\mathcal{B},S} (q^{\mathrm{PCF}})^2 (C_{\mathrm{spec2}}^{\mathrm{PCF}} + \zeta_n C_{\mathrm{adj}}^{\mathrm{PCF}})^2 + \hat{C}^2\right)(H_n^{\max})^{2s}\right], \tag{4.2.39}$$

where C' is the constant hidden in (4.2.32).
Note that $H_n^{\max}$ can be chosen sufficiently small so that $\beta_m \leq \beta < 1$ for all $m \geq n$.
Consider now the consequences of the inequality (4.2.35). If $|||u - u_n|||_{\kappa,S,\Omega} > \varepsilon$ then (4.2.38) implies

$$|||u - u_{n+1}|||^2_{\kappa,S,\Omega} \;\leq\; (\beta + \mu^2)\; |||u - u_n|||^2_{\kappa,S,\Omega}\,.$$

Now choose μ small enough so that

$$\alpha := (\beta + \mu^2)^{1/2} < 1 \tag{4.2.40}$$

to complete the proof. □

4.2.2 Proof of convergence

The main result of this section is Theorem 4.2.13 below which proves convergence of the adaptive method and also demonstrates the decay of the quantity osc on the sequence of approximate eigenfunctions. Before proving this result we need a final lemma.

Lemma 4.2.12. *There exists a constant $\tilde{\alpha} \in (0,1)$ such that*

$$\mathrm{osc}(u_{n+1}, \mathcal{T}_{n+1}) \;\leq\; \tilde{\alpha}\, \mathrm{osc}(u_n, \mathcal{T}_n) \;+\; (1 + q^{\mathrm{PCF}})(H_n^{\max})^2\, |||u - u_n|||_{\kappa,S,\Omega}. \tag{4.2.41}$$

Proof. First recall that one of the key results in [42] is the proof that the value of osc of any fixed function $v \in H_0^1(\Omega)$ is reduced by applying one refinement based on Marking Strategy 2 (Definition 4.1.4). Similarly, it is possible to prove the same result for any fixed function $v \in H_\pi^1(\Omega)$. Thus we have (in view of Algorithm 1):

$$\mathrm{osc}(u_n, \mathcal{T}_{n+1}) \;\leq\; \tilde{\alpha}\, \mathrm{osc}(u_n, \mathcal{T}_n), \tag{4.2.42}$$

where $0 < \tilde{\alpha} < 1$ is independent of u_n. Thus, a simple application of the triangle inequality combined with (4.2.42) yields

$$\begin{aligned} \mathrm{osc}(u_{n+1}, \mathcal{T}_{n+1}) &\leq \mathrm{osc}(u_n, \mathcal{T}_{n+1}) \;+\; \mathrm{osc}(u_{n+1} - u_n, \mathcal{T}_{n+1}) \\ &\leq \tilde{\alpha}\, \mathrm{osc}(u_n, \mathcal{T}_n) \;+\; \mathrm{osc}(u_{n+1} - u_n, \mathcal{T}_{n+1}) \end{aligned} \tag{4.2.43}$$

A further application of the triangle inequality and then (4.2.2) yields

$$\begin{aligned}\operatorname{osc}(u_{n+1}-u_n,\mathcal{T}_{n+1}) &\leq \operatorname{osc}(u-u_{n+1},\mathcal{T}_{n+1}) + \operatorname{osc}(u-u_n,\mathcal{T}_{n+1})\\ &\lesssim (H_{n+1}^{\max})^2(|||u-u_{n+1}|||_{\kappa,S,\Omega}\\ &\quad +|||u-u_n|||_{\kappa,S,\Omega}) \end{aligned} \tag{4.2.44}$$

and then combining (4.2.43) and (4.2.44) and applying Theorem 4.2.6 completes the proof. □

Theorem 4.2.13 (Convergence)**.** *Provided the initial mesh $\mathcal{T}_0$ is chosen so that $H_0^{\max}$ is small enough, there exists a constant $p \in (0,1)$ such that the recursive application of Algorithm 1 to solve problem* (1.3.9) *yields a convergent sequence of approximate eigenvectors, with the property:*

$$|||u-u_n|||_{\kappa,S,\Omega} \leq C_0 q^{\mathrm{PCF}} p^n, \tag{4.2.45}$$

and

$$\left((\zeta_n-S)^2\overline{b}+S^2(1+\overline{b})+\zeta_n^2\right)\operatorname{osc}(u_n,\mathcal{T}_n) \leq C_1 p^n, \tag{4.2.46}$$

where C_0 and C_1 are constants and q^{PCF} is the constant defined in Theorem 4.2.6.

Remark 4.2.14. *The initial mesh convergence threshold and the constants C_1 and C_2 may depend on θ, $\tilde{\theta}$ and ζ.*

Proof. The proof of this theorem is by induction and the induction step contains an application of Theorem 4.2.11. In order to ensure the reduction of the error, we have to assume that the starting mesh $\mathcal{T}_0$ is fine enough and μ in Theorem 4.2.11 is small enough such that for the chosen value of θ, the quantity α in (4.2.40) satisfies $\alpha < 1$. Then with $\tilde{\alpha}$ as in Lemma 4.2.12, we set

$$\max\{\alpha,\tilde{\alpha}\} < p < 1 .$$

We also set

$$C_1 = \left((\zeta_0-S)^2\overline{b}+S^2(1+\overline{b})+\zeta_0^2\right)\operatorname{osc}(u_0,\mathcal{T}_0) \quad \text{and} \quad C_0 = \max\{\mu^{-1}p^{-1}C_1,\ |||u-u_0|||_{\kappa,S,\Omega}\}.$$

First note that by the definition of C_0 and Theorem 4.2.6,

$$|||u-u_0|||_{\kappa,S,\Omega} \leq C_0 \leq C_0 q^{\mathrm{PCF}},$$

since $q > 1$. Combined with the definition of C_1, it proves the result for $n = 0$.
Now, suppose that for some $n > 0$ the inequalities (4.2.45) and (4.2.46) hold.

Then let us consider the outcomes, depending on whether the inequality

$$|||u - u_n|||_{\kappa,S,\Omega} \leq C_0 p^{n+1}, \tag{4.2.47}$$

holds or not. If (4.2.47) holds then we can apply Theorem 4.2.6 to conclude that

$$|||u - u_{n+1}|||_{\kappa,S,\Omega} \leq q^{\mathrm{PCF}} \, |||u - u_n|||_{\kappa,S,\Omega} \leq q^{\mathrm{PCF}} C_0 p^{n+1},$$

which proves (4.2.45) for $n+1$.
On the other hand, if (4.2.47) does not hold then, by definition of C_0,

$$|||u - u_n|||_{\kappa,S,\Omega} \;>\; C_0 p^{n+1} \;\geq\; \mu^{-1} C_1 p^n. \tag{4.2.48}$$

Also, since we have assumed (4.2.46) for n, we have

$$\big((\zeta_n - S)^2 \overline{b} + S^2(1+\overline{b}) + \zeta_n^2\big) \, \mathrm{osc}(u_n, \mathcal{T}_n) \leq \mu\varepsilon \quad \text{with} \quad \varepsilon := \mu^{-1} C_1 p^n \; . \tag{4.2.49}$$

Then (4.2.48) and (4.2.49) combined with Theorem 4.2.11 yields

$$|||u - u_{n+1}|||_{\kappa,S,\Omega} \;\leq\; \alpha |||u - u_n|||_{\kappa,S,\Omega}$$

and so using the inductive hypothesis (4.2.45) combined with the definition of p, we have

$$|||u - u_{n+1}|||_{\kappa,S,\Omega} \;\leq\; \alpha C_0 q^{\mathrm{PCF}} p^n \leq q^{\mathrm{PCF}} C_0 p^{n+1},$$

which again proves (4.2.45) for $n+1$.
To conclude the proof, we have to show that also (4.2.46) holds for $n+1$. Using Lemma 4.2.12, the minimum-maximum principle, which we applied to ζ_{n+1} and to $(\zeta_{n+1} - S) = \lambda_{n+1}$, and the inductive hypothesis, we have

$$\begin{aligned}
&\big((\zeta_{n+1} - S)^2 \overline{b} + S^2(1+\overline{b}) + \zeta_{n+1}^2\big) \, \mathrm{osc}(u_{n+1}, \mathcal{T}_{n+1}) \\
&\quad\leq\; \tilde{\alpha} C_1 p^n + (1 + q^{\mathrm{PCF}})(H_n^{\max})^2 \big((\zeta_n - S)^2 \overline{b} + S^2(1+\overline{b}) + \zeta_n^2\big) C_0 q^{\mathrm{PCF}} p^n \\
&\quad\leq\; \Big(\tilde{\alpha} C_1 + (1 + q^{\mathrm{PCF}})(H_0^{\max})^2 \big((\zeta_0 - S)^2 \overline{b} + S^2(1+\overline{b}) + \zeta_0^2\big) C_0 q^{\mathrm{PCF}}\Big) p^n.
\end{aligned} \tag{4.2.50}$$

Now, (recalling that $\tilde{\alpha} < p$), in addition to the condition already imposed on $H_0^{\max}$ we can further require that

$$\tilde{\alpha} C_1 + (1 + q^{\mathrm{PCF}})(H_0^{\max})^2 \big((\zeta_0 - S)^2 \overline{b} + S^2(1+\overline{b}) + \zeta_0^2\big) C_0 q^{\mathrm{PCF}} \;\leq\; p C_1.$$

This ensures that

$$\left((\zeta_{n+1}-S)^2\overline{b}+S^2(1+\overline{b})+\zeta_{n+1}^2\right)\,\mathrm{osc}(u_{n+1},\mathcal{T}_{n+1})\;\leq\;C_1p^{n+1},$$

thus concluding the proof. □

Corollary 4.2.15 (Convergence)**.** *Provided the initial mesh $\mathcal{T}_0$ is chosen so that $H_0^{\max}$ is small enough, there exists a constant $p\in(0,1)$ such that the recursive application of Algorithm 1 to solve problem* (1.3.9) *yields a convergent sequence of approximate eigenvalues, with the property:*

$$|\zeta-\zeta_n|\;\leq\;C_0^2(q^{\mathrm{PCF}})^2p^{2n}. \tag{4.2.51}$$

Proof. The proof is straightforward applying

$$|\zeta-\zeta_n|\;\leq\;|||u-u_n|||_\Omega^2\;,$$

from Lemma 2.2.26, to (4.2.45). □

4.2.3 Other convergence results

In this section we present convergence result for problem (1.3.8). The convergence proof is based on Algorithm 1.
The next theorem is very similar to Theorem 4.2.13. In fact it comes as a consequence of Theorem 4.2.13, since the two problems (1.3.8) and (1.3.9) are very close.

Theorem 4.2.16 (Convergence)**.** *Provided the initial mesh $\mathcal{T}_0$ is chosen so that $H_0^{\max}$ is small enough, there exists a constant $p\in(0,1)$ such that the recursive application of Algorithm 1 to solve problem* (1.3.8) *yields a convergent sequence of approximate eigenvectors, with the properties:*

$$a_\kappa(u-u_n,u-u_n)^{1/2}\;\leq\;C_0q^{\mathrm{PCF}}p^n, \tag{4.2.52}$$

$$|\lambda-\lambda_n|\;\leq\;C_0^2(q^{\mathrm{PCF}})^2p^{2n}, \tag{4.2.53}$$

and

$$(\lambda_n^2(1+\overline{b})+S^2(2+\overline{b})+2S\lambda_n)\,\mathrm{osc}(u_n,\mathcal{T}_n)\;\leq\;C_1p^n, \tag{4.2.54}$$

where C_0 and C_1 are constants and q^{PCF} is the constant defined in Theorem 4.2.6.

Proof. The result (4.2.52) comes straightforwardly from (4.2.45), since the eigenfunctions of problems (1.3.8) and (1.3.9) are the same and since $a_\kappa(u-u_n,u-u_n)^{1/2}\leq a_{\kappa,S}(u-u_n,u-u_n)^{1/2}$.

Using the relation between the spectra of problems (1.3.8) and (1.3.9) is possible to deduce (4.2.53), since $|\lambda-\lambda_n| = |\zeta-\zeta_n|$, where ζ and ζ_n are the eigenvalues corresponding to λ and λ_n. Similarly comes (4.2.54). □

Chapter 5

Numerics

In this chapter we present numerical results illustrating the convergence of our adaptive FEM. We have considered the problems (1.3.7) and (1.3.8). In particular, concerning problem (1.3.8), we solved the TE case mode because we believe that it is more interesting from a mathematical point of view, since it could present localized singularities in the gradient of the solutions. The reason why we haven't done any computation regarding problem (1.3.9) is because this problem has been introduced just to simplify the analysis for problem (1.3.8).

All the numerical results in this chapter have been computed using our on research codes which make use of ARPACK [38] and of the fast direct sparse solver for linear problems ME27 [47] contained in the HSL archive. One of the advantages of ARPACK is the possibility to compute just the approximations of the few eigenpairs of interest. Especially, we used it to compute the smallest part of the spectrum, when we were searching for gaps in periodic media. Then, we used again ARPACK to look for trapped modes in periodic structures with defects just computing the approximations of eigenpairs with eigenvalues inside the gaps. Despite the actual computation of the wanted eigenpairs, which has been done using these free packages, we wrote all the code necessary to do all the other tasks, like: generate the meshes, discretize the problems, compute the error estimations and refine the meshes.

The structure of the chapter is as follows: in Section 5.1 we present the numerical experiments on the general elliptic eigenvalue problem and the TE mode problem. In particular, concerning the latter problem, we have done numerical experiments on both purely periodic media and periodic media with defects. We also would like to bring to the attention of the reader that in Subsection 5.1.5 we present an efficient way to compute a bundle of eigenvalues for the TE case problem using just one sequence of adapted meshes. In Section 5.2 we applied our AFEM, not just to a point in the spectrum of the TE problem, but to entire bands of the spectra. We concentrated our efforts on bands belonging to trapped modes of supercells. Finally, in Section 5.3 we present a more efficient method to compute entire bands of the spectrum.

5.1 Adaptivity and convergence

In this section, a number of results from simulations concerning the convergence of our adaptive method are collected. However, in the first part of this section we shall present some extra results about the error estimator η_n (introduced in (3.2.1) above), which are particularly useful in practice.
In our computations we used Algorithm 2 below, which is very similar to the algorithm presented in Chapter 4. The only difference is the presence of a condition to terminate the execution of the loop. This condition is based on the value of the error estimator and on the number of iterations already done. For this reason, we have introduced in the algorithm the parameter tol, which sets the wanted tolerance for the error estimator η_n, and the parameter $\max_n$, which sets the maximum number of iterations that we are prepared to do.

5.1.1 Preliminary results

The first set of theorems show the conditions under which the high order terms in the results of Theorem 3.3.5, Theorem 3.3.7, Theorem 3.4.3 and Theorem 3.4.4 can be ignored. For sake of clarity we have grouped the results for the PCF case in the first subsection and the results for the general elliptic case in the second one.

PCF case

Theorem 5.1.1. *Let ζ_j be an eigenvalue of (1.3.9) of multiplicity 1 and let $(\zeta_{j,n}, u_{j,n})$ be computed eigenpairs for the same value of $\vec{\kappa}$ spanning the computed eigenspace $E_{j,n}^{\mathrm{PCF}}$, in the sense of Remark 2.2.23. Let also the true eigenfunction $U_j \in E_j^{\mathrm{PCF}}$ be defined as in Theorem 3.1.8. Then we have for $e_{j,n} = U_j - u_{j,n}$ that if $H_n^{\max}$ is small enough:*

$$a_{\kappa,S}(e_{j,n}, e_{j,n})^{1/2} \lesssim \eta_n \ , \tag{5.1.1}$$

where the hidden constant in 5.1.1 is different from the hidden constant in 3.3.10.

Proof. The proof comes applying the results of Chapter 3. From Theorem 3.3.5 we have that:

$$a_{\kappa,S}(e_{j,n}, e_{j,n})^{1/2} \lesssim \eta_n \ + \ G_n, \tag{5.1.2}$$

where $G_n = \frac{1}{2}(\zeta_j + \zeta_{j,n})(e_{j,n}, e_{j,n})_{0,\mathcal{B},\Omega} / a_{\kappa,S}(e_{j,n}, e_{j,n})^{1/2}$ is a higher order term, as proved in Theorem 3.4.1. Now, applying Theorem 3.1.6(ii) to (5.1.2), we have

$$\begin{aligned} a_{\kappa,S}(e_{j,n}, e_{j,n})^{1/2} &\lesssim \eta_n \ + \ \frac{1}{2}(\zeta_j + \zeta_{j,n}) \ \frac{(e_{j,n}, e_{j,n})_{0,\mathcal{B},\Omega}}{a_{\kappa,S}(e_{j,n}, e_{j,n})^{1/2}} \\ &\lesssim \eta_n \ + \ \frac{1}{2}(\zeta_j + \zeta_{j,n}) C_{\mathrm{adj}}^2 (H_n^{max})^{2s} \ a_{\kappa,S}(e_{j,n}, e_{j,n})^{1/2}. \end{aligned} \tag{5.1.3}$$

From the minimum-maximum principle we know that $\zeta_j \leq \zeta_{j,n}$. So supposing that $H_n^{\max}$ is small enough, we obtain

$$\frac{1}{2}(\zeta_j + \zeta_{j,n}) C_{\text{adj}}^2 (H_n^{max})^{2s} \ \leq \ \zeta_{j,n} \ C_{\text{adj}}^2 (H_n^{max})^{2s} \ < \ 1 \ ,$$

and then from (5.1.3) we have that there is a constant C such that

$$a_{\kappa,S}(e_{j,n}, e_{j,n})^{1/2} \ \leq \ C \ \eta_n \ .$$

□

Theorem 5.1.2. *Under the same assumptions as in Theorem 5.1.1 we have:*

$$|\zeta_{j,n} - \zeta_j| \lesssim \eta_n^2 \ .$$

Proof. The proof is straightforward from in view of Corollary 2.2.27 and Theorem 5.1.1. □

Theorem 5.1.3. *Let λ_j be an eigenvalue of (1.3.8) of multiplicity 1 and let $(\lambda_{j,n}, u_{j,n})$ be computed eigenpairs for the same value of $\vec{\kappa}$ spanning the computed eigenspace $E_{j,n}^{\text{PCF}}$, in the sense of Remark 2.2.23. Let also the true eigenfunction $U_j \in E_j^{\text{PCF}}$ be defined as in Theorem 3.1.7. Then we have for $e_{j,n} = U_j - u_{j,n}$ that if $H_n^{\max}$ is small enough:*

$$a_\kappa(e_{j,n}, e_{j,n})^{1/2} \ \lesssim \ \eta_n \ . \tag{5.1.4}$$

Proof. The proof is straightforward in view of Theorem 5.1.1 and since $a_\kappa(e_{j,n}, e_{j,n}) \leq a_{\kappa,S}(e_{j,n}, e_{j,n})$. □

Theorem 5.1.4. *Under the same assumptions as Theorem 5.1.3 we have:*

$$|\lambda_{j,n} - \lambda_j| \ \lesssim \ \eta_n^2 \ .$$

Proof. The proof is straightforward in view of Corollary 2.2.32 and Theorem 5.1.3. □

The next corollary is very important for computations, since it proves that if the error estimator η_n goes to 0, this implies convergence to the exact eigenpair. This justifies our procedure of refining the elements which have big associated residual values.

Corollary 5.1.5. *Let $(\lambda_{j,n}, u_{j,n})$ be a calculated eigenpair of the problem (2.2.48) for some value of $\vec{\kappa}$ and (λ_j, U_j) be the corresponding eigenpair in the sense of Theorem 3.1.7 of the continuous problem (1.3.8) for the same value of $\vec{\kappa}$. Then if the residual error estimator η_n goes to 0, the energy norm of the error $a_\kappa(U_j - u_{j,n}, U_j - u_{j,n})^{1/2}$ and error for eigenvalues $|\lambda_{j,n} - \lambda_j|$ go to 0. Moreover, if the eigenpair $(\lambda_{j,n}, u_{j,n})$ converges to (λ_j, U_j), then the residual error estimator η_n goes to 0.*

Proof. The first statement comes straightforwardly from Theorem 5.1.3 and Theorem 5.1.4.
The second statement comes straightforwardly from Theorem 3.5.6. □

General elliptic case

In this subsection we have collected for the general elliptic case the analogous results proved above.

Theorem 5.1.6. *Let λ_j be an eigenvalue of (2.2.2) of multiplicity 1 and let $(\lambda_{j,n}, u_{j,n})$ be computed eigenpairs spanning the computed eigenspace $E_{j,n}$, in the sense of Remark 2.2.4. Let also the true eigenfunction $U_j \in E_j$ be defined as in Theorem 3.1.4. Then we have for $e_{j,n} = U_j - u_{j,n}$ that if $H_n^{\max}$ is small enough:*

$$a(e_{j,n}, e_{j,n})^{1/2} \lesssim \eta_n \ . \tag{5.1.5}$$

Theorem 5.1.7. *Under the same assumptions as Theorem 5.1.6 we have:*

$$|\lambda_{j,n} - \lambda_j| \lesssim \eta_n^2 \ .$$

Corollary 5.1.8. *Let $(\lambda_{j,n}, u_{j,n})$ be a calculated eigenpair of the problem (2.2.2) and (λ_j, U_j) be the correspondent eigenpair in the sense of Theorem 3.1.4 of the continuous problem (1.3.7). Then if the residual error estimator η_n goes to 0, then the energy norm of the error $a(U_j - u_{j,n}, U_j - u_{j,n})^{1/2}$ and error for eigenvalues $|\lambda_{j,n} - \lambda_j|$ go to 0. Moreover, if the eigenpair $(\lambda_{j,n}, u_{j,n})$ converges to (λ_j, U_j), then the residual error estimator η_n goes to 0.*

Algorithm 2 Converging algorithm

Require: $0 < \theta < 1$
Require: $0 < \tilde{\theta} < 1$
Require: $\text{tol} > 0$
Require: $\max_n > 0$
Require: $\mathcal{T}_0$
  $n = 0$
  repeat
    Compute (λ_n, u_n) on $\mathcal{T}_n$
    Mark the elements using the first marking strategy (Definition 4.1.1)
    Mark any additional unmarked elements using the second marking strategy (Definition 4.1.4)
    Refine the mesh $\mathcal{T}_n$ and construct $\mathcal{T}_{n+1}$
    $n = n + 1$
  until $\eta_n \geq \text{tol}$ AND $n \leq \max_n$

5.1.2 Laplace operator

In the first set of simulations we have solved the Laplace eigenvalue problem on a unit square with Dirichlet boundary conditions.
In Table 5.1, we compare different runs of Algorithm 2 using different values for θ and $\tilde{\theta}$. Since the problem is smooth, it follows from Theorem 2.2.10 that using uniform refinement the rate of convergence for eigenvalues should be $\mathcal{O}(H_n^{\max})^2$, or equivalently the rate of convergence in the number of degrees of freedom (DOFs) N should be $\mathcal{O}(N^{-1})$. We measure the rate of convergence by conjecturing that $|\lambda - \lambda_n| = \mathcal{O}(N^{-\beta})$ and estimating β for each pair of computations by the formula $\beta = -\log(|\lambda - \lambda_n|/|\lambda - \lambda_{n-1}|)/\log(\text{DOFs}_n/\text{DOFs}_{n-1})$. In addition, in Figure 5-1 we plotted the values of β for more iterations for $\theta = \tilde{\theta} = 0.2$ and for $\theta = \tilde{\theta} = 0.5$, since β for those simulations were not yet settle down in the first few iterations in Table 5.1. As can be seen in the graph, the values of β soon starts to oscillates around 1, which is the asymptotic order of convergence for this problem. Similarly Table 5.2 and Figure 5-2 contain the same kind of information relative to the fourth smallest eigenvalue of the problem. As can be seen the rate of convergence is sensitive to the values of θ and $\tilde{\theta}$. Moreover, our results for the adaptive method show a convergence rate close to $\mathcal{O}(N^{-1})$ for θ, $\tilde{\theta}$ sufficiently large.
In the theory presented in [51] it is shown how the error in computed eigenvalues for smooth problems is proportional to the square of the considered eigenvalue, i.e. $|\lambda - \lambda_n| \leq C\ \lambda^2\ (H_n^{\max})^2$. The same result can be deduced from our results in Chapter 2 with the appropriate modifications, since here we are supposing that the problem has better regularity. Since the Laplace problem is very well understood, we know from the theory the values for the first and the fourth eigenvalues, namely: 19.7392089 and 78.9568352. Comparing errors in Tables 5.1 and 5.2, corresponding to similar numbers of degrees of freedom (DOFs), we see that the error grows roughly with the square of the eigenvalue.

5.1.3 Elliptic operator with discontinuous coefficients

In this second example we investigate how our method copes with discontinuous coefficients. In order to do that we modified the smooth problem from the previous example. We inserted a square subdomain of side 0.5 in the center of the unit square domain. We also choose the function $\mathcal{A}$ (introduced in (1.3.1)) to be a scalar piecewise constant and to assume the value 100 inside the subdomain and the value 1 outside it.
The jump in the value of $\mathcal{A}$ could produce a jump in the gradient of the eigenfunctions all along the boundary of the subdomain. So the eigenfunctions now lie in $H^{s+1}(\Omega)$ with $s > 1/2 - \varepsilon$, for all $\varepsilon > 0$ in general. We remark that from [45, Example 2.1] we also know that $u \in H^{s+1}(\Omega_i)$ where $s > 2/3 + \mathcal{O}(1/\overline{a})$ in each subdomain Ω_i on which $\mathcal{A}$ is constant, since singularities in the gradient of the eigenfunctions may arise

	$\theta = \tilde{\theta} = 0.2$			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
n	$\lvert\lambda - \lambda_n\rvert$	N	β	$\lvert\lambda - \lambda_n\rvert$	N	β	$\lvert\lambda - \lambda_n\rvert$	N	β
1	0.1350	400	-	0.1350	400	-	0.1350	400	-
2	0.1327	498	0.0802	0.1177	954	0.1581	0.0529	1989	0.5839
3	0.1293	613	0.1228	0.0779	1564	0.8349	0.0176	5205	1.1407
4	0.1256	731	0.1645	0.0501	1977	1.8788	0.0073	15980	0.7877
5	0.1215	854	0.2138	0.0351	2634	1.2383	0.0024	48434	0.9836
6	0.1165	970	0.3340	0.0176	4004	0.7885	0.0009	122699	1.0673
7	0.1069	1097	0.6962	0.0121	6588	0.7217	0.0003	312591	1.0083

Table 5.1: Comparison of the reduction of the error and DOFs of the adaptive method for the smallest eigenvalue for the Laplace problem on the unit square.

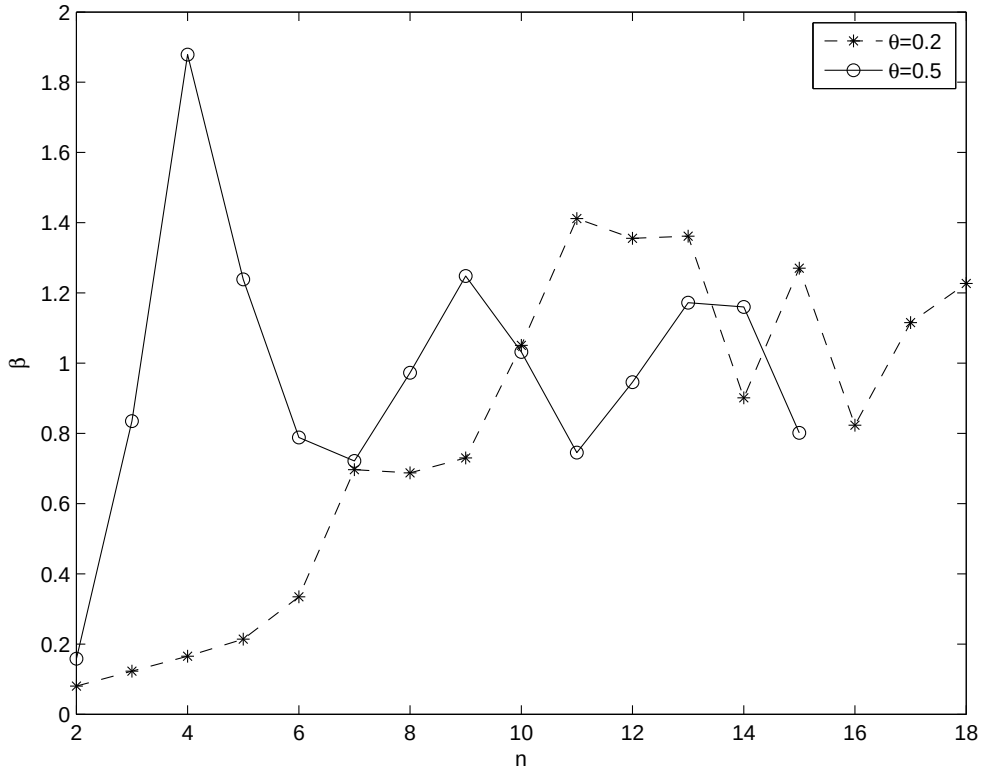

Figure 5-1: The graph contains the values of β for the smallest eigenvalue for the Laplace problem on the unit square for $\theta = \tilde{\theta} = 0.2$ and for $\theta = \tilde{\theta} = 0.5$.

	$\theta = \tilde{\theta} = 0.2$			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
n	$\lvert\lambda - \lambda_n\rvert$	N	β	$\lvert\lambda - \lambda_n\rvert$	N	β	$\lvert\lambda - \lambda_n\rvert$	N	β
1	2.1439	400	-	2.1439	400	-	2.1439	400	-
2	2.0997	505	0.0895	1.8280	1016	0.1658	0.7603	2039	0.6365
3	2.0549	626	0.1004	1.0850	1636	1.1662	0.2439	6793	0.9447
4	1.9945	759	0.1548	0.7792	12254	1.0331	0.0917	18717	0.9652
5	1.9164	883	0.2638	0.4936	3067	1.4826	0.0331	54113	0.9583
6	1.7717	1017	0.5557	0.3484	4681	0.8240	0.0120	146056	1.0181
7	1.6463	1131	0.6911	0.2578	7321	0.6730	0.0046	382024	0.9970

Table 5.2: Comparison of the reduction of the error and DOFs of the adaptive method for the fourth smallest eigenvalue for the Laplace problem on the unit square.

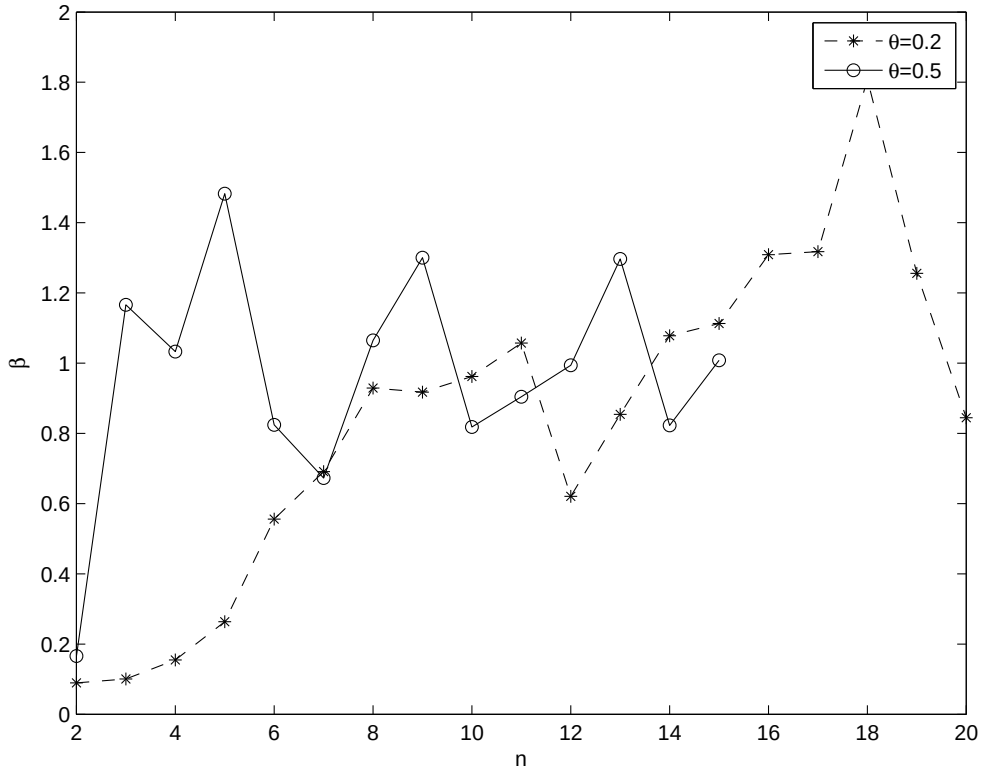

Figure 5-2: The graph contains the values of β for the fourth smallest eigenvalue for the Laplace problem on the unit square for $\theta = \tilde{\theta} = 0.2$ and for $\theta = \tilde{\theta} = 0.5$.

in the corners of the subdomains. Since we resolve exactly the interface between the subdomains, in our numerical results we see a convergence speed coming from just the singularities arising at the corners of the subdomains.

From Theorem 2.2.10 and using uniform refinement, the rate of convergence for eigenvalues should be at least $\mathcal{O}(H_n^{\max})^{2s}$ or equivalently $\mathcal{O}(N^{-s})$, where N is the number of DOFs. In Table 5.3 there are the results of the computations using a sequence of uniform meshes; the value of β is computed as explained before and it could be considered as a rough approximation to s. In this case the exact eigenvalue λ is unknown, but we approximate it by computing the eigenvalue on a very fine mesh involving about half a million of DOFs.

Using our adaptive method we obtain greater orders of convergence for big enough value of θ and $\tilde{\theta}$, as can be seen from Table 5.4. In fact the rate of convergence for $\theta = \tilde{\theta} = 0.5$ or 0.8 is close to the rate of convergence for smooth problems showed in Table 5.1 and Table 5.2. To make the comparison between our method and uniform refinement easier, we summarize the results in Table 5.5. From Table 5.5 it is clear the advantage in using our adaptive method, since the error for eigenvalues is much lower with the same number of DOFs. However, the performance of our adaptive method is sensitive to the value of θ and $\tilde{\theta}$; from our computations, it resulted that for this problem the best value for both θ and $\tilde{\theta}$ is 0.8.

To illustrate Theorem 5.1.7, we have constructed Table 5.6, where in the columns labeled by C_r we have estimated numerically the value of the hidden constant in the result of Theorem 5.1.7. To compute the values of C_r, we have used: $C_r = \sqrt{|\lambda - \lambda_n|/\eta_n^2}$. The fact that the values of C_r are all contained in a small range, is a numerical evidence that the result in Theorem 5.1.7 underlines the behavior of our residual-based error estimator and that the effects of higher order terms are negligible. Moreover, it shows that in this case the hidden constant C_r is of very moderate size. In order to show the quality of our error estimator, we have also compared in Table 5.6 the true errors with the value of the residuals for different choices of θ and $\tilde{\theta}$. From Table 5.6 is clear that the error-residual value η_n^2 is always an upper bound for the true error and, moreover, it is possible to see that η_n strictly mimics the decay of the true error, since, as said above, the values of C_r are in a small range. This latter fact is particularly interesting since it implies that the error-residual can be used as an indicator for the behavior of the true error. Unfortunately, due to the small value of C_r, the quantity η_n can not be used as a good indicator of the value of the true error, at least not for this particular problem.

In Table 5.7, we compare computational estimations of the value of p introduced in Theorem 4.1.17. To compute the values p, we used the formula $p = \sqrt{|\lambda - \lambda_n|/|\lambda - \lambda_{n-1}|}$. It is clear that the values of p, and then the rate of convergence, is sensitive to the values of θ and $\tilde{\theta}$. In particular, greater values of θ and $\tilde{\theta}$ lead to smaller p and conse-

quentially to a faster convergence. Another interesting thing to notice is that the value of p remains almost constant during each run of the algorithm, this is a consequence of the monotone decay of the error that we experienced in our simulations. Such behavior of the error is better than what predicted in Theorem 4.1.17, since that result does not imply a monotone decay of the error, but just the monotone decay of an upper bound of the error. So, according to that result, the error could oscillate.
In Figure 5-3 we depict the mesh coming from the fourth iteration of Algorithm 2 with $\theta = \tilde{\theta} = 0.8$. This mesh is the result of multiple refinements using both marking strategies 1 and 2 each time. As can be seen the corners of the subdomain are much more refined than the rest of the domain. This is clearly the effect of the first marking strategy, since the residual has detected singularities in the corners.
Finally in Figure 5-4 we depict the eigenfunction corresponding to the smallest eigenvalue of the problem with discontinuous coefficients.

5.1.4 TE case problem on periodic medium

Now, we are going to consider an example arising from PCF applications. We will consider the TE case problem for a periodic medium with square inclusions. The unit cell, on which we are going to solve this problem, is the unit square with a square inclusion of side 0.5 which is centered within the unit cell. We choose the function $\mathcal{A}$ to be piecewise constant and to assume the value 10000 inside the subdomain and the value 1 outside it. This is an academic example, since expected jumps in dielectric properties of real PCFs, are much more moderate than this.
As already seen for the general elliptic eigenvalue problem, the jump in the value of $\mathcal{A}$ could produce a jump in the gradient of the eigenfunctions all along the boundary of the subdomain. As above, the eigenfunctions lie in $H^{s+1}(\Omega)$, with $s > 1/2 - \varepsilon$, for all $\varepsilon > 0$ in general. However, since we resolve exactly the interface also in this example, we see a convergence speed coming from the regularity of the eigenfunctions in each subdomain, which is $u \in H^{s+1}(\Omega_i)$ where $s > 2/3 + \mathcal{O}(1/\overline{a})$ in each subdomain Ω_i on which $\mathcal{A}$ is constant.
From Theorem 2.2.33, using uniform refinement, the rate of convergence for eigenvalues should be at least $\mathcal{O}(H_n^{\max})^{2s}$ or equivalently $\mathcal{O}(N^{-\beta})$, where N is the number of DOFs. In Table 5.8 there are the results of the computations using a sequence of uniform meshes; the value of β is computed as explained before and it could be considered an approximation of s.
Instead, using our method we obtain greater orders of convergence for some value of θ and $\tilde{\theta}$, as can be seen from Table 5.9. In fact the rate of convergence for $\theta = \tilde{\theta} = 0.8$ is close to the rate of convergence for smooth problems. In this case the exact eigenvalue λ is unknown, but we approximate it by computing the eigenvalue on a very fine mesh involving about a million of DOFs. To get easier the comparison between our method

n	$\lvert\lambda - \lambda_n\rvert$	N	β
1	1.1071	81	-
2	0.3521	289	0.9005
3	0.1168	1089	0.8316
4	0.0399	4225	0.7924
5	0.0136	16641	0.7874
6	0.0042	66049	0.8537

Table 5.3: Uniform refinement for the smallest eigenvalue of the generic elliptic eigenvalue problem with discontinuous coefficients.

	$\theta = \tilde{\theta} = 0.2$			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
n	$\lvert\lambda - \lambda_n\rvert$	N	β	$\lvert\lambda - \lambda_n\rvert$	N	β	$\lvert\lambda - \lambda_n\rvert$	N	β
1	1.1071	81	-	1.1071	81	-	1.1071	81	-
2	1.0200	103	0.3410	0.8738	199	0.2632	0.4834	356	0.5597
3	1.0105	129	0.0416	0.5848	314	0.8805	0.2244	799	0.9494
4	1.0039	147	0.0498	0.3983	491	0.8591	0.0990	2235	0.7957
5	0.8968	167	0.8843	0.2766	673	1.1564	0.0401	4764	1.1932
6	0.8076	194	0.6996	0.1933	975	0.9665	0.0180	12375	0.8372
7	0.8008	217	0.0747	0.1346	1476	0.8722	0.0065	29148	1.1888
8	0.7502	237	0.7401	0.0948	2080	1.0237	0.0020	65387	1.4482

Table 5.4: Comparison of the reduction of the error and DOFs of the adaptive method for the smallest eigenvalue of the generic elliptic eigenvalue problem with discontinuous coefficients.

Uniform			**Adaptive**					
			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
$\lvert\lambda - \lambda_n\rvert$	N	n	$\lvert\lambda - \lambda_n\rvert$	N	n	$\lvert\lambda - \lambda_n\rvert$	N	n
1.1071	81	1	1.1071	81	1	1.1071	81	1
0.3521	289	2	0.2766	673	5	0.2244	799	3
0.1168	1089	3	0.0948	2080	8	0.0990	2235	4
0.0399	4225	4	0.0315	6039	11	0.0180	12375	6
0.0136	16641	5	0.0148	12607	13	0.0065	29148	7
0.0042	66049	6	0.0038	37126	16	0.0020	65387	8

Table 5.5: Comparison between uniform refinement and the adaptive method for the smallest eigenvalue of the generic elliptic eigenvalue problem with discontinuous coefficients.

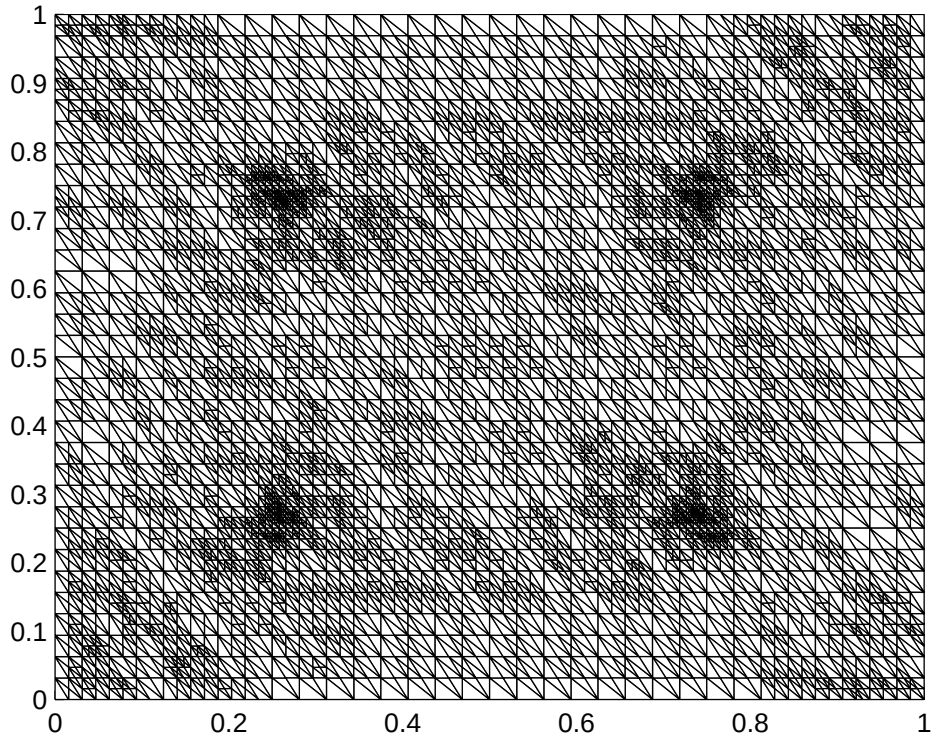

Figure 5-3: A refined mesh from the adaptive method corresponding to the smallest eigenvalue of the generic elliptic eigenvalue problem with discontinuous coefficients.

	$\theta = \tilde{\theta} = 0.2$			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
n	$\lvert\lambda - \lambda_n\rvert$	η_n	C_r	$\lvert\lambda - \lambda_n\rvert$	η_n	C_r	$\lvert\lambda - \lambda_n\rvert$	η_n	C_r
1	1.1071	6.5037	0.1618	1.1071	6.5037	0.1618	1.1071	6.5037	0.1618
2	1.0200	6.1186	0.1651	0.8738	5.3345	0.1752	0.4834	3.9436	0.1763
3	1.0105	5.9781	0.1681	0.5848	4.3535	0.1757	0.2244	2.6795	0.1768
4	1.0039	5.8811	0.1704	0.3983	3.5011	0.1803	0.0990	1.7435	0.1804
5	0.8968	5.6211	0.1685	0.2766	2.9665	0.1773	0.0401	1.16448	0.1720
6	0.8076	5.3577	0.1677	0.1933	2.5043	0.1756	0.0180	0.7496	0.1792
7	0.8008	5.1562	0.1736	0.1346	2.0853	0.1760	0.0065	0.4925	0.1639
8	0.7502	4.9499	0.1750	0.0948	1.7230	0.1787	0.0020	0.3223	0.1395

Table 5.6: Comparison between the reduction of the error and the computed residual for the adaptive method for the smallest eigenvalue of the generic elliptic eigenvalue problem with discontinuous coefficients.

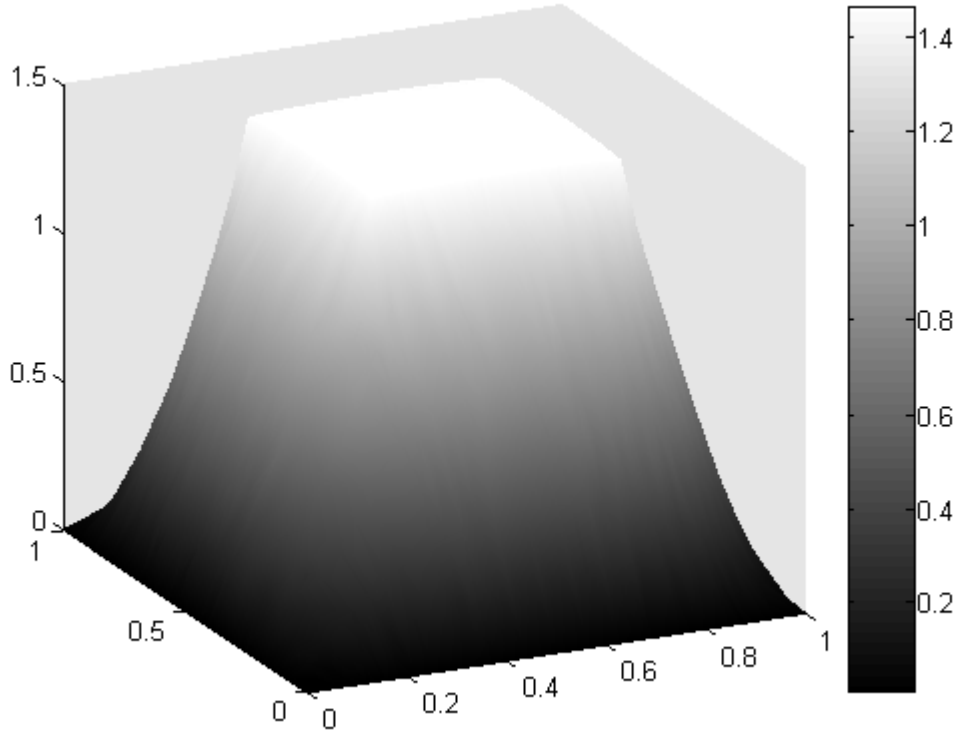

Figure 5-4: The eigenfunction corresponding to the smallest eigenvalue of the generic elliptic eigenvalue problem with discontinuous coefficients.

	$\theta = \tilde{\theta} = 0.2$		$\theta = \tilde{\theta} = 0.5$		$\theta = \tilde{\theta} = 0.8$	
n	$\lvert\lambda - \lambda_n\rvert$	p	$\lvert\lambda - \lambda_n\rvert$	p	$\lvert\lambda - \lambda_n\rvert$	p
1	1.1071	-	1.1071	-	1.1071	-
2	1.0200	0.9599	0.8738	0.8884	0.4834	0.6608
3	1.0105	0.9953	0.5848	0.8181	0.2244	0.6813
4	1.0039	0.9968	0.3983	0.8253	0.0990	0.6642
5	0.8968	0.9452	0.2766	0.8333	0.0401	0.6367
6	0.8076	0.9489	0.1933	0.8360	0.0180	0.6706
7	0.8008	0.9958	0.1346	0.8346	0.0065	0.6010
8	0.7502	0.9679	0.0948	0.8390	0.0020	0.5571

Table 5.7: Comparison between the values of p for different values of θ and $\tilde{\theta}$ for the smallest eigenvalue of the generic elliptic eigenvalue problem with discontinuous coefficients.

and uniform refinement, we dedicated Table 5.10 to this point.
In view of Theorem 5.1.4, we have constructed Table 5.11 where in the columns C_r we have estimated numerically the value of the hidden constant in the result of Theorem 5.1.4. The same consideration from the previous example can be applicable here.
In Figure 5-5 we depict the mesh coming from the fourth iteration of Algorithm 2 with $\theta = \tilde{\theta} = 0.8$. This mesh is the result of multiple refinements using both marking strategies 1 and 2 each time. As can be seen the corners of the subdomain are much more refined than the rest of the domain.
In Table 5.12, we compare computational estimations of the value of p considered in Theorem 4.2.16. To compute the values p, we used the formula $p = \sqrt{|\lambda - \lambda_n|/|\lambda - \lambda_{n-1}|}$. It is clear that the values of p, and then the rate of convergence, is sensitive to the values of θ and $\tilde{\theta}$. In particular, greater values of θ and $\tilde{\theta}$ lead to smaller p and consequentially to a faster convergence. Another interesting thing to notice is that the value of p remains almost constant during each run of the algorithm, this is a consequence of the monotone decay of the error that we experienced in our simulations.
Finally in Figure 5-6 we depict the eigenfunction corresponding to the smallest eigenvalue of the problem with discontinuous coefficients. This eigenfunction is the one used to refine the mesh in Figure 5-5.

5.1.5 A more efficient way to compute a bundle of eigenvalues for the TE case problem

In this subsection we are going to present a more efficient way to compute many eigenvalues for the TE case problem on periodic medium. The improved efficiency comes from the fact that we use just one sequence of adapted meshes for all eigenvalues. The idea presented below can be used with any kind of elliptic eigenvalue problem.
Suppose that you want to compute the smallest r eigenvalues for a fixed quasimomentum, then you can use our adaptive method on the r-th eigenvalue to construct a finite sequence of adapted meshes. Then, you can use the same sequence of meshes to compute with a quite good accuracy all the eigenvalues smaller than the r-th one. This very simple technique works very often because the eigenfunctions higher in the spectrum have also higher frequencies, so a mesh, that can resolve well such high frequencies, it can also resolve well the lower frequencies of the eigenfunctions lower in the spectrum. Moreover, when we are in presence of singularities in the gradient that are localized always in the same places for all eigenfunctions, as in this case, the mesh computed for the r-th eigenvalue resolves well also the singularities in the gradient of all other eigenfunctions.
In Table 5.13 we compared the errors on two sequences of meshes relative to the smallest eigenvalue for the TE case problem on the same periodic cell as in the previous section and with quasimomentum equal to $\vec{\kappa} = (\pi/4, \pi/4)$. On the left we have the results

n	$\lvert\lambda - \lambda_n\rvert$	N	β
1	6.1948	64	-
2	1.9462	256	0.8352
3	0.6458	1024	0.7957
4	0.2242	4096	0.7632
5	0.0797	16384	0.7458
6	0.0280	65536	0.7540

Table 5.8: Uniform refinement for the second smallest eigenvalue of the TE case problem on a periodic medium with quasimomentum to $\vec{\kappa} = (0, 0)$.

	$\theta = \tilde{\theta} = 0.2$			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
n	$\lvert\lambda - \lambda_n\rvert$	N	β	$\lvert\lambda - \lambda_n\rvert$	N	β	$\lvert\lambda - \lambda_n\rvert$	N	β
1	6.1948	64	-	6.1948	64	-	6.1948	64	-
2	5.7120	76	0.4722	4.0876	131	0.5804	2.2780	229	0.7848
3	4.8996	96	0.6567	2.4078	247	0.8345	0.8771	642	0.9258
4	3.9523	188	0.3197	1.3960	536	0.7036	0.3468	2117	0.7777
5	3.4904	199	2.1855	0.8976	712	1.5553	0.1373	5859	0.9098
6	2.9544	223	1.4642	0.5491	1248	0.8758	0.0603	13791	0.9622
7	2.5152	270	0.8415	0.3664	1884	0.9819	0.0252	31067	1.0743
8	2.2882	308	0.7182	0.2795	2972	0.5939	0.0105	70523	1.0667

Table 5.9: Comparison of the reduction of the error and DOFs of the adaptive method for second smallest eigenvalue of the TE case problem on a periodic medium with quasimomentum to $\vec{\kappa} = (0, 0)$.

Uniform			**Adaptive**					
			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
$\lvert\lambda - \lambda_n\rvert$	N	n	$\lvert\lambda - \lambda_n\rvert$	N	n	$\lvert\lambda - \lambda_n\rvert$	N	n
6.1890	64	1	6.1890	64	1	6.1890	64	1
1.9404	256	2	1.3960	535	4	0.8771	642	3
0.6400	1024	3	0.5491	1248	6	0.3468	2117	4
0.2184	4096	4	0.2795	2972	8	0.1373	5859	5
0.0739	16384	5	0.0771	11025	11	0.0603	13791	6
0.0222	65536	6	0.0195	47035	15	0.0252	31067	7

Table 5.10: Comparison between uniform refinement and the adaptive method for the second smallest eigenvalue of the TE problem on a periodic medium with quasimomentum to $\vec{\kappa} = (0, 0)$.

	$\theta = \widetilde{\theta} = 0.2$			$\theta = \widetilde{\theta} = 0.5$			$\theta = \widetilde{\theta} = 0.8$		
n	$\lvert\lambda - \lambda_n\rvert$	η_n	C_r	$\lvert\lambda - \lambda_n\rvert$	η_n	C_r	$\lvert\lambda - \lambda_n\rvert$	η_n	C_r
1	6.1948	12.5299	0.1986	6.1948	12.5299	0.1986	6.1948	12.5299	0.1986
2	5.7120	11.6360	0.2054	4.0876	9.4685	0.2135	2.2780	7.2670	0.2077
3	4.8996	10.9426	0.2023	2.4078	7.5190	0.2064	0.8771	4.5452	0.2061
4	3.9523	9.3597	0.2124	1.3960	5.3257	0.2219	0.3468	2.8269	0.2083
5	3.4904	9.0548	0.2063	0.8976	4.5155	0.2098	0.1373	1.8748	0.1977
6	2.9544	8.5901	0.2001	0.5491	3.7234	0.1990	0.0603	1.3077	0.1877
7	2.5152	7.8811	0.2012	0.3664	3.1270	0.1936	0.0252	0.9238	0.1718
8	2.2882	7.5483	0.2004	0.2795	2.6477	0.1997	0.0105	0.6462	0.1586

Table 5.11: Comparison between the reduction of the error and the computed residual for the adaptive method for the second smallest eigenvalue of the TE problem on a periodic medium with quasimomentum to $\vec{\kappa} = (0, 0)$.

	$\theta = \tilde{\theta} = 0.2$		$\theta = \tilde{\theta} = 0.5$		$\theta = \tilde{\theta} = 0.8$	
n	$\lvert\lambda - \lambda_n\rvert$	p	$\lvert\lambda - \lambda_n\rvert$	p	$\lvert\lambda - \lambda_n\rvert$	p
1	6.1948	-	6.1948	-	6.1948	-
2	5.7120	0.9602	4.0876	0.8123	2.2780	0.6064
3	4.8996	0.9262	2.4078	0.7675	0.8771	0.6205
4	3.9523	0.8981	1.3960	0.7614	0.3468	0.6288
5	3.4904	0.9398	0.8976	0.8019	0.1373	0.6293
6	2.9544	0.9200	0.5491	0.7821	0.0603	0.6624
7	2.5152	0.9227	0.3664	0.8169	0.0252	0.6465
8	2.2882	0.9538	0.2795	0.8734	0.0105	0.6458

Table 5.12: Comparison between the values of p for different values of θ and $\tilde{\theta}$ for the second smallest eigenvalue of the TE problem on a periodic medium with quasimomentum to $\vec{\kappa} = (0, 0)$.

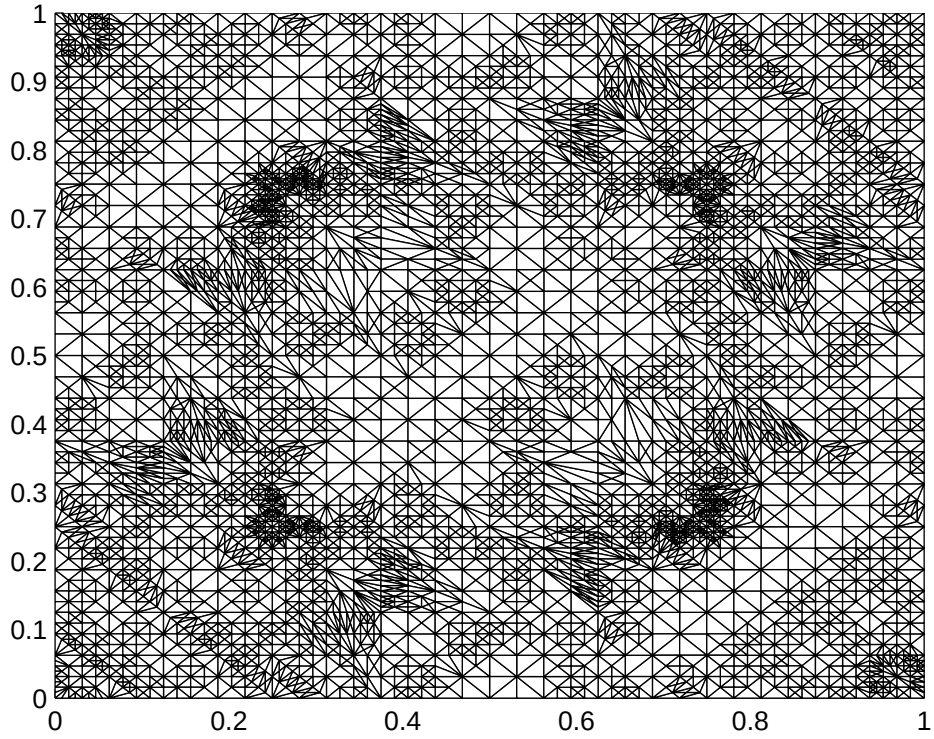

Figure 5-5: A refined mesh from the adaptive method corresponding to the second smallest eigenvalue of the TE problem on a periodic medium with quasimomentum to $\vec{\kappa} = (0, 0)$.

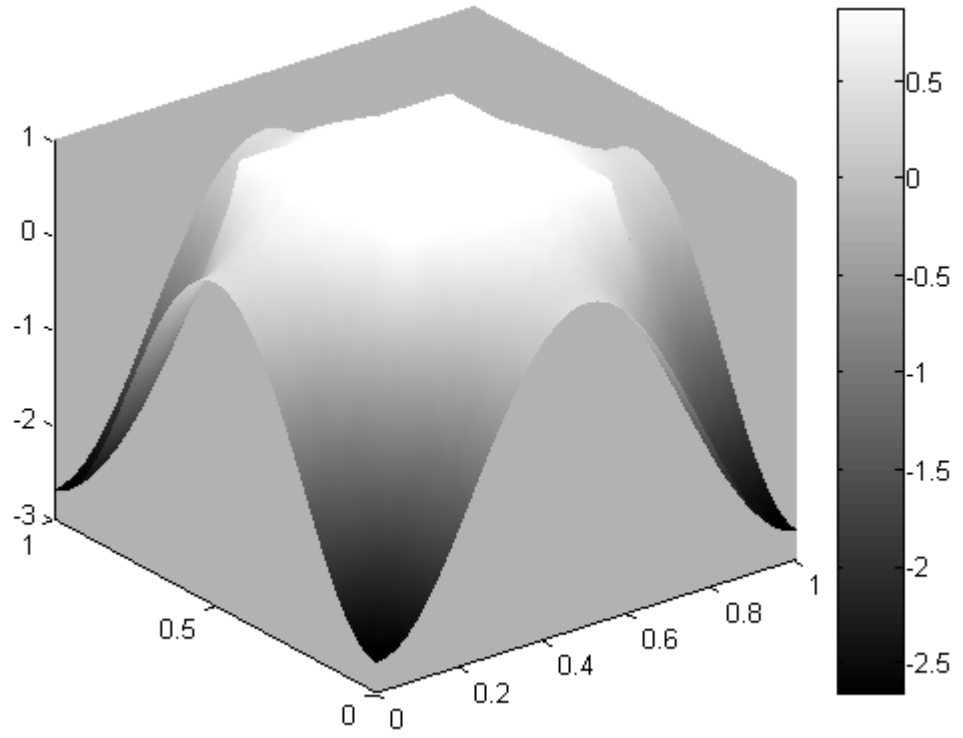

Figure 5-6: The eigenfunction corresponding to the second smallest eigenvalue of the TE problem on a periodic medium with quasimomentum to $\vec{\kappa} = (0, 0)$.

computed refining the meshes according to the first smallest eigenvalue, instead on the right we have the results computed refining the meshes accordingly the sixth smallest eigenvalue.
In Table 5.14 we have done the same comparison considering the second smallest eigenvalue of the same problem.
In conclusion we have that more than on eigenvalue can be computed on the same adapted mesh with good accuracy. But, on the other hand, it is straightforward that in general this method will only converge for the eigenvalue used to refine the meshes.

5.1.6 TE mode problem on supercell

Now it is time to consider a different and more interesting problem coming from applications. In this section we are going to hunt for frequencies of light trapped in the defect of a PCF. We continue to work with the TE case problem and the periodic structure, surrounding the defect, will be the same as the one analysed in the previous section. The defect will be a missing inclusion in the center of the section of the PCF. As explained in Chapter 1, we are going to use the supercell framework [49] to compute the modes coming from the defect. The supercell that we use has two layers of periodic structure surrounding the defect, as depicted in Figure 5-7.
Since the jumps of the coefficient $\mathcal{A}$ are the same as in the previous example, we have that also the regularity of the eigenfunction trapped in the defect is, in each subdomain, $u \in H^{1+s}(\Omega_i)$, with $s > 2/3 + \mathcal{O}(1/\overline{a})$. In Table 5.15 we can see the result using uniform refinement, the values of β are pretty similar to the ones in Table 5.8, as predicted.
Instead, using our method we obtain greater orders of convergence, as can be seen from Table 5.16. For trapped modes is usual to have peaks in the values of β that could exceed easily 1. For this problem the difference in the accuracy between our method and the uniform refinement method is much more profound than before. The reason is not only that we refine around the corners, where the singularities are, but also, because the most part of the "energy" of the solution is inside the defect, which is a very small region. Also for this case we computed the "exact" value of the eigenvalue λ using more than one million of DOFs. To get easier the comparison between our method and uniform refinement, we dedicated Table 5.17 to this point.
In view of Theorem 5.1.4, we have constructed Table 5.18 where in the columns C_r we have estimated numerically the value of the hidden constant in the result of Theorem 5.1.4. This time the values C_r seems not to be yet settled down.
In Figure 5-8 we depict the mesh coming from the fourth iteration of Algorithm 2 with $\theta = \tilde{\theta} = 0.8$. As can be seen there is a lot of refinement in the defect and just outside it, especially around the corners of the inclusions. Away from the defect there is just a bit of refinement which is again around the corners of the inclusions, the reason why the refinement is so concentrated in the defect and the reason why the corners of the

n	$\lvert\lambda - \lambda_n\rvert$	N	β	$\lvert\lambda - \lambda_n\rvert$	N	β
1	2.3399	64	-	2.3399	64	-
2	1.0810	231	0.6016	1.7819	277	0.1860
3	0.4505	637	0.8630	0.4583	941	1.1104
4	0.1621	2279	0.8019	0.4386	2239	0.0507
5	0.0411	7038	1.2169	0.3791	7177	0.1252
6	0.0108	22724	1.1377	0.1027	14560	1.8461
7	0.0028	80181	1.0730	0.0838	35861	0.2339

Table 5.13: Comparison of the reduction of the error and DOFs using different sequences of refined meshes of the adaptive method for first smallest eigenvalue of the TE case problem on a periodic medium with quasimomentum equal to $\vec{\kappa} = (\pi/4, \pi/4)$. The columns on the left are computed refining the meshes accordingly the first smallest eigenvalue, instead the columns on the right are computed refining the meshes accordingly the sixth smallest eigenvalue.

n	$\lvert\lambda - \lambda_n\rvert$	N	β	$\lvert\lambda - \lambda_n\rvert$	N	β
1	7.9082	64	-	7.9082	64	-
2	3.8633	210	0.6029	3.0248	277	0.6559
3	2.1012	644	0.5435	1.1982	941	0.7572
4	1.3480	2311	0.3474	0.7021	2239	0.6166
5	0.3841	8106	1.0004	0.4161	7177	0.4492
6	0.1760	26196	0.6654	0.1477	14560	1.4639
7	0.0477	90790	1.0505	0.0947	35861	0.4936

Table 5.14: Comparison of the reduction of the error and DOFs using different sequences of refined meshes of the adaptive method for second smallest eigenvalue of the TE case problem on a periodic medium with quasimomentum to $\vec{\kappa} = (\pi/4, \pi/4)$. The columns on the left are computed refining the meshes accordingly the second smallest eigenvalue, instead the columns on the right are computed refining the meshes accordingly the sixth smallest eigenvalue.

n	$\lvert\lambda - \lambda_n\rvert$	N	β
1	0.5858	10000	-
2	0.1966	40000	0.7876
3	0.0653	160000	0.7951
4	0.0188	640000	0.8982

Table 5.15: Uniform refinement for a trapped eigenvalue of the TE case problem on a supercell with quasimomentum $\vec{\kappa} = (0, 0)$.

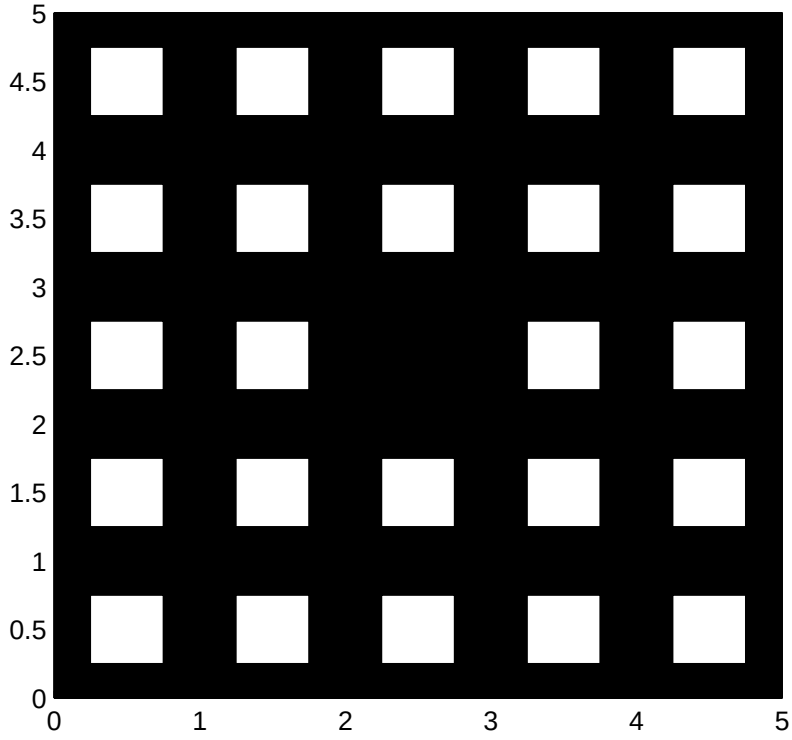

Figure 5-7: The structure of the supercell used for the computations.

inclusions away from the defect seem to not show important singularities, is because the trapped mode has a fast decay outside the defect that flatten down the singularities that it encounters, see Picture 5-9.
In Table 5.19, we compare computational estimations of the value of p considered in Theorem 4.2.16. As we have already noticed in the other examples before, the values of p is sensitive to the values of θ and $\tilde{\theta}$. Again as before, greater values of θ and $\tilde{\theta}$ lead to smaller p. The fact that the value of p remains almost constant during each run of the algorithm is a consequence of the monotone decay of the error that we experienced in our simulations.
Finally in Figure 5-9 we depict the eigenfunction corresponding to the mode trapped inside the defect. This eigenfunction is the one used to refine the mesh in Figure 5-8.

5.2 Spectral bands and trapped modes

In this section we describe how we applied our method to compute a band of the spectrum, instead of a single eigenpair for a fixed value of the quasimomentum. We analysed the band associated to a trapped mode in a supercell. We choose this problem because it is very relevant for applications.
In Chapter 1 we explained how a compact defect in a periodic structure could produce eigenvalues in the gaps between bands of essential spectrum. Also in Chapter 1, we anticipated that we were going to use the supercell framework to look for trapped mode in gaps and as consequence of this choice we have that the defects could produce narrow bands of essential spectra in the gaps, instead of eigenvalues. These narrow bands should eventually shrink to eigenvalues, if we increase the size of the supercell.
We used the supercell displayed in Figure 5-7. Since the shape of the cell is square of length 5, it follows that the first Brillouin zone associated to this supercell is $\mathcal{K} = [-\pi/5, \pi/5]^2$ as shown in Figure 5-10.
In order to approximate the band corresponding to a trapped mode, we used the values of the quasimomentum coming from a uniform grid of 13 points per side on the first Brillouin zone. There are standard arguments based on the symmetries of the operator for our problem, which are used also in [8, 16, 4], saying that it is not necessary to use all the values of the quasimomentum in the first Brillouin zone to analyse the bands. But it is enough to use the values for the quasimomentum inside the reduced first Brillouin zone (which is the grey region in Figure 5-10). Moreover, we are going to use only the points of the uniform grid inside the reduced first Brillouin zone. For each considered value of the quasimomentum, we have computed the corresponding eigenvalue, in the band of the trapped mode, using firstly a sequence of uniform meshes and then sequences of adapted meshes using different values for θ and $\tilde{\theta}$.
The most important piece of information, that is possible to get from this kind of

	$\theta = \tilde{\theta} = 0.2$			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
n	$\lvert\lambda - \lambda_n\rvert$	N	β	$\lvert\lambda - \lambda_n\rvert$	N	β	$\lvert\lambda - \lambda_n\rvert$	N	β
1	0.5886	10000	-	0.5886	10000	-	0.5886	10000	-
2	0.5108	10093	15.3015	0.3876	10866	5.0306	0.2340	15076	2.2467
3	0.4279	10340	7.3227	0.2590	14064	1.5622	0.1075	25716	1.4569
4	0.3945	10811	1.8266	0.1523	18612	1.8948	0.0473	64680	0.8902
5	0.3746	11357	1.0511	0.0952	23726	1.9349	0.0199	131440	1.2224

Table 5.16: Comparison of the reduction of the error and DOFs of the adaptive method for a trapped eigenvalue of the TE case problem on a supercell with quasimomentum $\vec{\kappa} = (0, 0)$.

Uniform			**Adaptive**					
			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
$\lvert\lambda - \lambda_n\rvert$	N	n	$\lvert\lambda - \lambda_n\rvert$	N	n	$\lvert\lambda - \lambda_n\rvert$	N	n
0.5858	10000	1	0.5858	10000	1	0.5858	10000	1
0.1966	40000	2	0.1523	18612	4	0.1075	25716	3
0.0653	160000	3	0.0570	51542	7	0.0473	64680	4
0.0188	640000	4	0.0115	218937	11	0.0199	131440	5

Table 5.17: Comparison between uniform refinement and the adaptive method for a trapped eigenvalue of the TE case problem on a supercell with quasimomentum $\vec{\kappa} = (0, 0)$.

	$\theta = \tilde{\theta} = 0.2$			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
n	$\lvert\lambda - \lambda_n\rvert$	η_n	C_r	$\lvert\lambda - \lambda_n\rvert$	η_n	C_r	$\lvert\lambda - \lambda_n\rvert$	η_n	C_r
1	0.5886	3.5771	0.2145	0.5886	3.5771	0.2145	0.5886	3.5771	0.2145
2	0.5108	3.4409	0.2077	0.3876	3.1316	0.1988	0.2340	2.3296	0.2077
3	0.4279	3.3280	0.1966	0.2590	2.6531	0.1918	0.1075	1.7441	0.1880
4	0.3945	3.2105	0.1956	0.1523	2.0561	0.1898	0.0473	1.2288	0.1770
5	0.3746	3.1288	0.1956	0.0952	1.7375	0.1776	0.0199	0.8892	0.1586

Table 5.18: Comparison of the reduction of the error and the residuals of the adaptive method for a trapped eigenvalue of the TE case problem on a supercell with quasimomentum $\vec{\kappa} = (0, 0)$.

	$\theta = \tilde{\theta} = 0.2$		$\theta = \tilde{\theta} = 0.5$		$\theta = \tilde{\theta} = 0.8$	
n	$\lvert\lambda - \lambda_n\rvert$	p	$\lvert\lambda - \lambda_n\rvert$	p	$\lvert\lambda - \lambda_n\rvert$	p
1	0.5886	-	0.5886	-	0.5886	-
2	0.5108	0.9316	0.3876	0.8115	0.2340	0.6306
3	0.4279	0.9153	0.2590	0.8175	0.1075	0.6777
4	0.3945	0.9601	0.1523	0.7669	0.0473	0.6633
5	0.3746	0.9744	0.0952	0.7907	0.0199	0.6483

Table 5.19: Comparison between the values of p for different values of θ and $\tilde{\theta}$ for a trapped eigenvalue of the TE case problem on a supercell with quasimomentum $\vec{\kappa} = (0, 0)$.

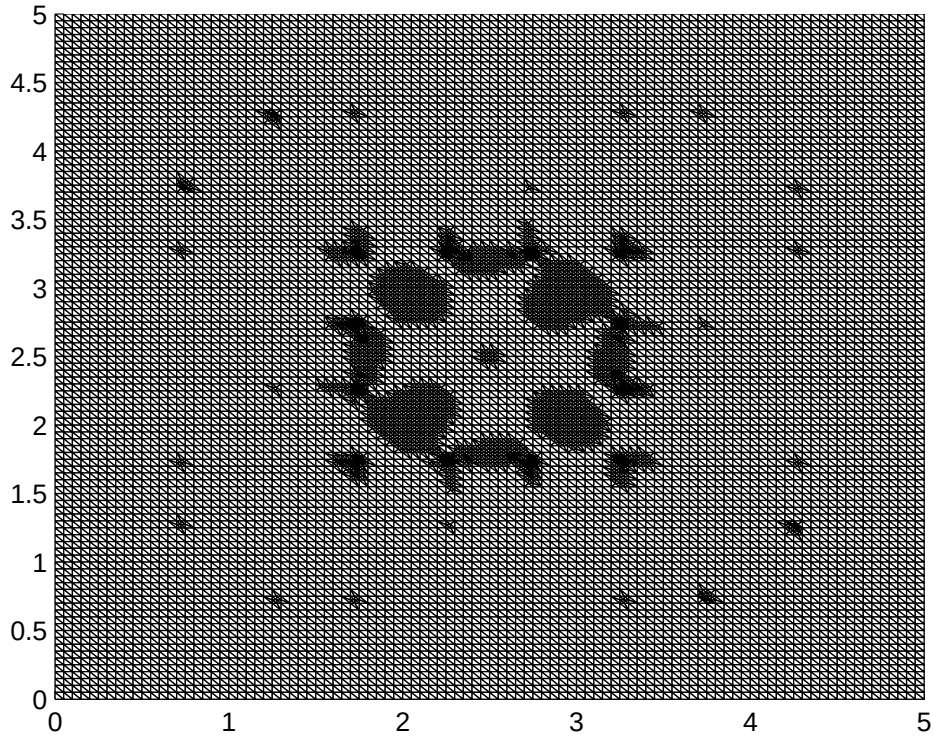

Figure 5-8: An adapted mesh for a trapped eigenvalue of the TE case problem on a supercell with quasimomentum $\vec{\kappa} = (0, 0)$.

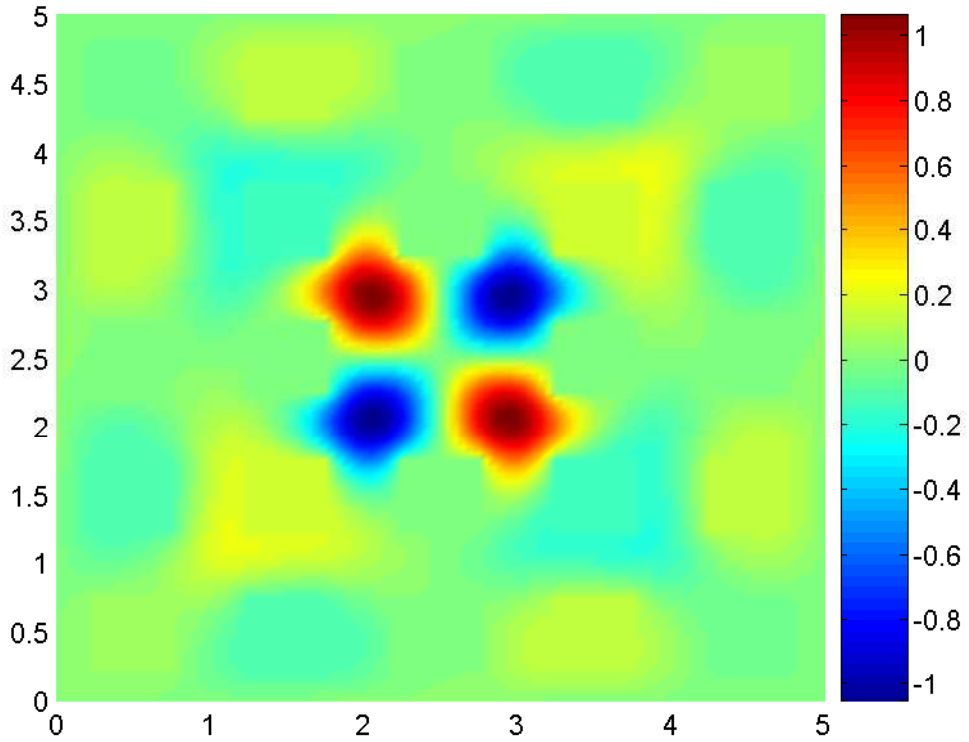

Figure 5-9: A picture of the eigenfunction trapped in the defect of the TE case problem on a supercell with quasimomentum $\vec{\kappa} = (0, 0)$.

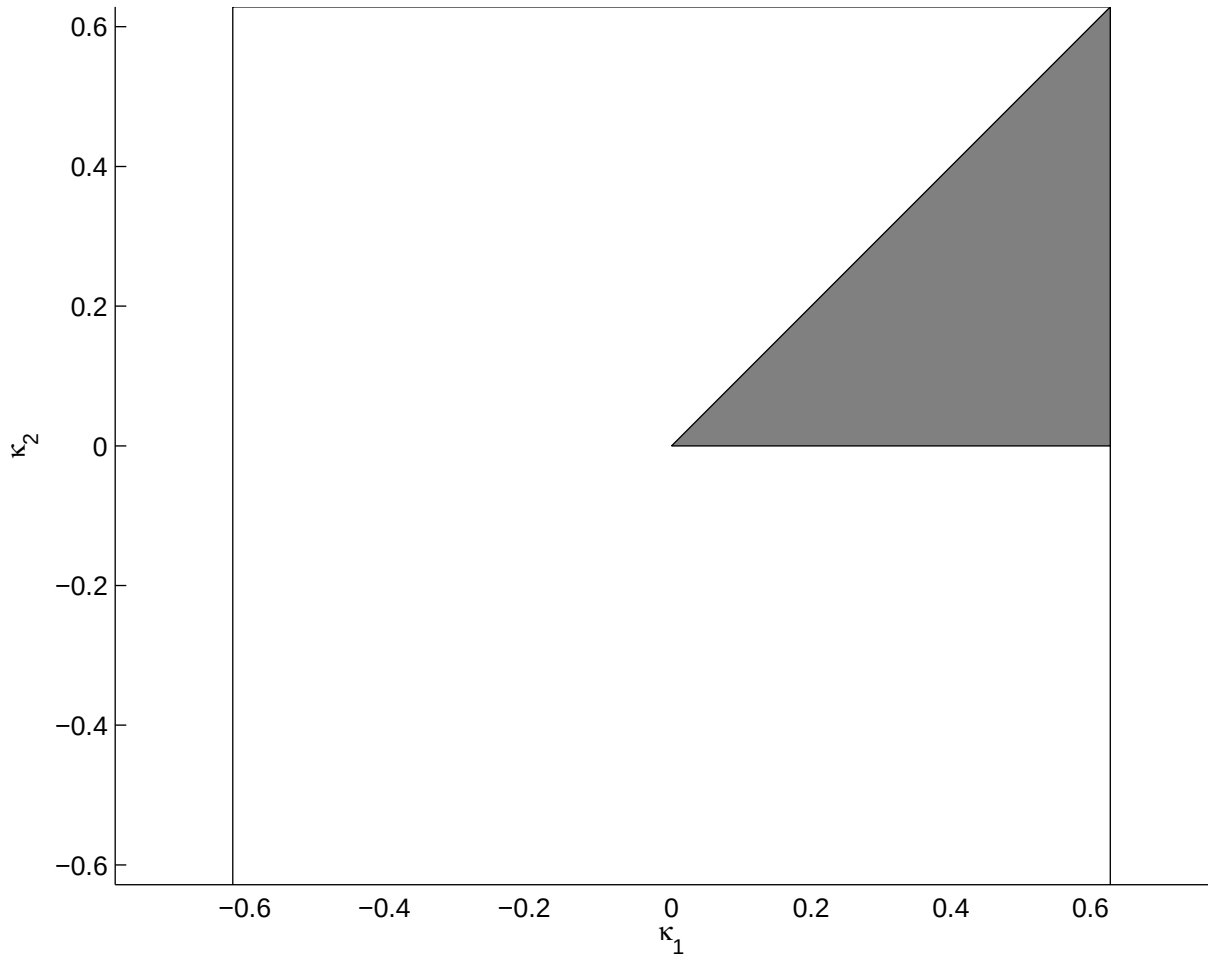

Figure 5-10: A picture of the first Brillouin zone associated to the used supercell and, in grey, the reduced Brillouin zone.

computation, is the position of the band of the trapped mode inside the gap. The position of the band is important because, if the computation is accurate, the physical frequency of the trapped mode would be near the center of the band. So, we decided to measure the error in the computations monitoring the absolute value of the error of the position of the center of the computed bands, with respect to the position of the center of the band computed using very fine meshes with more than one million of DOFs. In Table 5.20 there are the results using both the sequence of uniform meshes and adaptive method; as before n is the iteration number, moreover, we introduce the notation $\mathrm{err}_{\mathrm{pos}}$ to denote the error in the position of the band and $N^{\max}$ to denote the maximum number of DOFs used in a mesh for a fixed iteration.

Uniform			**Adaptive**					
			$\theta = \tilde{\theta} = 0.5$			$\theta = \tilde{\theta} = 0.8$		
$\mathrm{err}_{\mathrm{pos}}$	$N^{\max}$	n	$\mathrm{err}_{\mathrm{pos}}$	$N^{\max}$	n	$\mathrm{err}_{\mathrm{pos}}$	$N^{\max}$	n
0.6302	10000	1	0.6302	10000	1	0.6302	10000	1
0.2128	40000	2	0.2978	16121	3	0.2781	16581	2
0.0693	160000	3	0.0654	96147	7	0.0593	113276	4
0.0177	640000	4	0.0309	243674	9	0.0219	337072	5

Table 5.20: Comparison between uniform refinement and adaptive method applied to the band of the trapped mode for the TE case problem on a supercell.

5.3 An efficient and convergent method to compute the bands

In the last section we have approximated the band corresponding to a trapped mode in a supercell. In order to do that we choose many values of quasimomentum $\vec{\kappa}$ and for each value of $\vec{\kappa}$ we run Algorithm 2 starting from the same structured mesh. This method is very inefficient because, from the theory [15, 35] it is clear that the bands in the spectrum are continuous, in the sense that each eigenpair as a function of $\vec{\kappa}$ is continuous. So, it is reasonable to suppose that, for close values of $\vec{\kappa}$, the corresponding eigenpairs in the same band are very close, too. Moreover, the adaptive method should produce very similar meshes for close enough values of $\vec{\kappa}$. This should suggest a more efficient way to approximate bands, in which information from different runs of Algorithm 2 are shared. We would like to find a way to reuse the same adapted meshes for close values of $\vec{\kappa}$.

In this section we are going to describe such an efficient method to compute bands in the spectrum. By efficient we mean that the method needs fewer mesh refinements to reach the same approximation of a band as the adaptive method illustrated in the previous section. Moreover, we are going to show that the sequence of approximated bands $\mathcal{C}_m$ computed with this method converges to the true band $\mathcal{C}$.

Let $\mathcal{G}_0$ be a conforming and shape regular mesh of triangles constructed on the reduced first Brillouin zone $\mathcal{K}_{\text{red}}$ - see Figure 5-10. We are going to construct a sequence of meshes on $\mathcal{K}_{\text{red}}$ starting with the mesh $\mathcal{G}_0$ and where $\mathcal{G}_{m+1}$ is the resulting mesh after all the elements in $\mathcal{G}_m$ have been refined as described in Figure 5-11. It is important to understand that the meshes $\mathcal{G}_m$ are different from the meshes $\mathcal{T}_n$, since the formers are subdivision of the reduced first Brillouin zone $\mathcal{K}_{\text{red}}$, while $\mathcal{T}_n$ are subdivision of the primitive cell Ω. Moreover, we denote by $\mathcal{N}_m$ the set of all the nodes in the mesh $\mathcal{G}_m$.
In the method that we are going to present, we shall use Algorithm 2 as a subroutine, so let us define in Algorithm 3 the subroutine called AFEM implementing Algorithm 2. The subroutine AFEM is just a rewriting of Algorithm 2 in the form of a subroutine. AFEM takes as inputs the value of the quasimomentum $\vec{\kappa}$ for which compute the approximated eigenpair (λ_n, u_n), the initial mesh $\mathcal{T}_0$, the values of θ and $\tilde{\theta}$ for the marking strategies and the parameters for the stopping criteria tol and $\max_n$. The subroutine returns the eigenpair (λ_n, u_n) computed on the finest constructed mesh $\mathcal{T}_n$ and the mesh $\mathcal{T}_n$ itself.
Let's introduce the notation $(\lambda_m^\kappa, u_m^\kappa)$ and $\mathcal{T}_m^\kappa$ to denote the computed eigenpair and the mesh used to compute it for the value of the quasimomentum $\vec{\kappa} \in \mathcal{N}_m$. Thanks to the particular refining procedure that we have adopted to refine meshes $\mathcal{G}_m$, each point in $\vec{\kappa} \in \mathcal{N}_{m+1}$ has a unique "father" $\vec{\kappa'} \in \mathcal{N}_m$, where the father of the node $\vec{\kappa} \in \mathcal{N}_{m+1}$ is the node $\vec{\kappa'} \in \mathcal{N}_m$ closest to $\vec{\kappa}$. In the case that $\vec{\kappa} \in \mathcal{N}_{m+1} \cap \mathcal{N}_m$ then the father is $\vec{\kappa}' = \vec{\kappa}$. The relation is explained graphically in Figure 5-12.
Let's also define a function FatherMesh which takes as argument a point $\vec{\kappa} \in \mathcal{N}_{m+1}$ and it returns the mesh $\mathcal{T}_m^{\kappa'}$, where $\vec{\kappa}'$ is the father of $\vec{\kappa}$.
Now it is time to present our efficient method to approximate bands, which is illustrated in Algorithm 4. This algorithm works on two levels A and B. In the level A, which is implemented in the external repeat-until loop with counter m, the algorithm constructs the sequence of meshes $\mathcal{G}_m$ on the reduced first Brillouin zone $\mathcal{K}_{\text{red}}$. At each iteration a finer mesh $\mathcal{G}_{m+1}$ is constructed refining the previous mesh $\mathcal{G}_m$ by the refinement procedure illustrated in Figure 5-11. Moreover, each iteration of level A constructs an approximation $\mathcal{C}_m$ of the band of interest using level B, which is described next. In the level B, which is implemented in the inner for-all-do loop, many sequences of adapted meshes on the primitive cell Ω are constructed, each sequence corresponds to a different node $\vec{\kappa} \in \mathcal{N}_m$. The purpose of this level is to apply our AFEM to approximate the eigenpair of interest for each value of the quasimomentum $\vec{\kappa} \in \mathcal{N}_m$. Any run of Algorithm 4 may consist in many iteration of levels A and B. *The Algorithm 4 is more efficient in approximating bands than the adaptive algorithm presented in the previous section, since, for each child node $\vec{\kappa}$, the adaptive procedure in level B, which is used to approximate the eigenpair, starts from the already adapted mesh of the father node from the previous iteration of level A.* This exchange of infor-

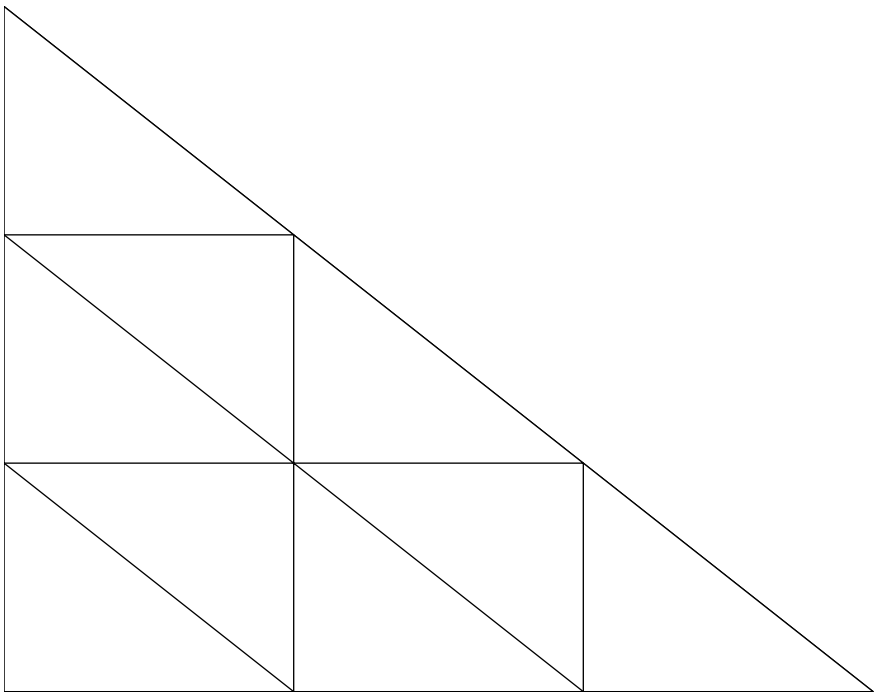

Figure 5-11: An element of a mesh $\mathcal{G}_m$ split in 9 elements.

Algorithm 3 The subroutine AFEM

$(\lambda_n, u_n, \mathcal{T}_n) :=$ AFEM$(\vec{\kappa}, \mathcal{T}_0, \theta, \tilde{\theta}, \text{tol}, \max_n)$
$n = 0$
repeat
  Compute (λ_n, u_n) on $\mathcal{T}_n$ with quasimomentum equal to $\vec{\kappa}$
  Mark the elements using the first marking strategy (Definition 4.1.1)
  Mark any additional unmarked elements using the second marking strategy (Definition 4.1.4)
  Refine the mesh $\mathcal{T}_n$ using bisection 5 and construct $\mathcal{T}_{n+1}$
  $n = n + 1$
until $\eta_n \geq$ tol AND $n \leq \max_n$

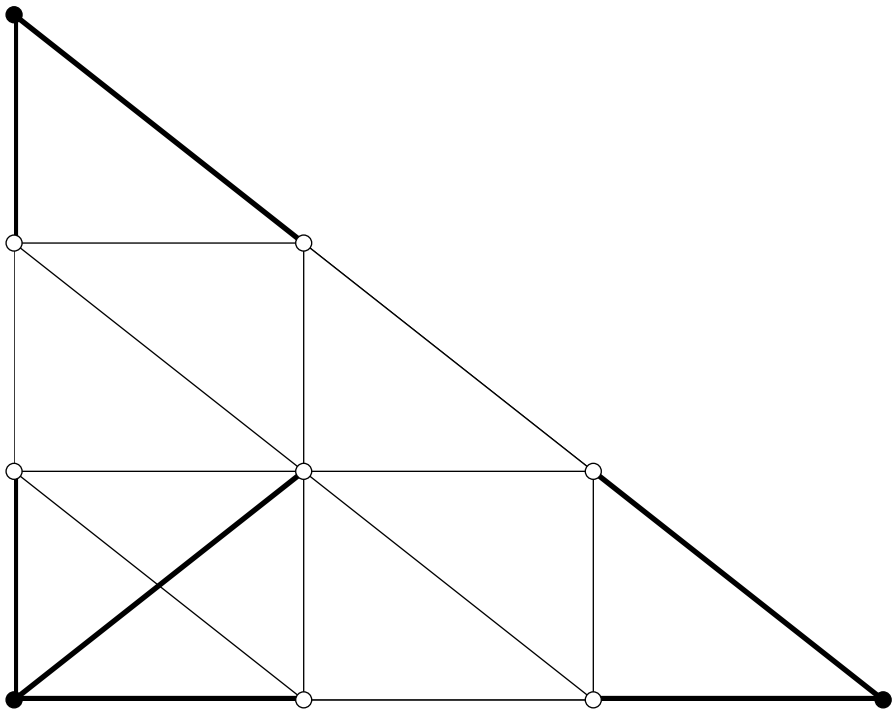

Figure 5-12: A refined element of a mesh $\mathcal{G}_m$. The black dots are the "father" nodes and the white dots are the "children". The thick lines links the children to their father.

mation from different values of the quasimomentum and from different iterations has been done by the function FatherMesh, which implements the relation father-children for the nodes of consecutive meshes $\mathcal{G}_m$ and $\mathcal{G}_{m+1}$ on the reduced first Brillouin zone $\mathcal{K}_{\text{red}}$. In this way we take advantage of the fact that eigenpairs in the same band for close values of the quasimomentum are very close, too. This is in contrast to what we have done in the previous section, where we restart the adaptive procedure always from the same structured mesh $\mathcal{T}_0$ for each value of the quasimomentum.
Finally, we have to define some parameters in order to use Algorithm 4. These parameters are: θ and $\tilde{\theta}$, which are already been introduced for Algorithm 3; an integer value $\max_{\text{it}}$ greater than 0, which sets the maximum number of refinements, it plays the role of $\max_n$ as in Algorithm 3; an initial mesh $\mathcal{T}_0$ on the primitive cell Ω; another integer value $\max_m$ greater than 0; and finally a finite subsequence of length $\max_m$ of real values tol_m, where $0 < \text{tol}_{m+1} < \text{tol}_m < \cdots < \text{tol}_0$, which prescribe the wanted tolerance for the approximated band $\mathcal{C}_m$, for each iteration of level A.
Algorithm 4 is convergent in the sense that, if its main repeat-until loop is run infinitely many times, $\mathcal{C}_m$ will converge to the true band. To prove this statement we are going to suppose to be able to run Algorithm 4 with $\max_m = \infty$ and with tol_m values forming a strictly monotone decreasing sequence converging to 0, in this way the main loop of Algorithm 4 becomes an infinite loop.
From a standard result in [15], it is well know that the bands of PCF problems are continuous, in view of this we wrote the following straightforward lemma:

Lemma 5.3.1. *Let $\mathcal{W}_m$ be the finite dimensional space of elementwise linear functions on the mesh $\mathcal{G}_m$, then $\mathcal{W}_\infty$, which is the limit of $\mathcal{W}_m$ when m goes to infinity, is dense in $C^0(\mathcal{K}_{\text{red}})$.*

Also the next lemma is straightforward. It is an application of Theorem 4.2.16.

Lemma 5.3.2. *For any value of m and for any $\vec{\kappa} \in \mathcal{N}_m$, we have that $\mathcal{C}_m(\vec{\kappa})$ converges to the true value $\mathcal{C}(\vec{\kappa})$.*

Proof. In Algorithm 4, with $\max_m = \infty$ we have that, for any value of m and for any $\vec{\kappa} \in \mathcal{N}_m$ the subroutine AFEM is applied infinitely many times to the point $\vec{\kappa}$. This is equivalent to apply Algorithm 1 to the point $\vec{\kappa}$, then the convergence of $\mathcal{C}_m(\vec{\kappa}) \equiv \lambda_m^\kappa$ to $\mathcal{C}(\vec{\kappa}) \equiv \lambda^\kappa$ comes as a consequence of Theorem 4.2.16. □

Theorem 5.3.3 (Convergence to the true band)**.** *Let suppose that $\mathcal{T}_0$ is fine enough in the sense of Theorem 4.2.16 for all λ^κ in the considered band, for all $\vec{\kappa} \in \mathcal{K}_{\text{red}}$. Then applying Algorithm 4 with $\max_m = \infty$ we have that $\mathcal{C}_m$ converges to the true band $\mathcal{C}$.*

Proof. Let define $\mathcal{N}_\infty := \bigcup_{m \geq 0} \mathcal{N}_m$. Then for any $\vec{\kappa} \in \mathcal{N}_\infty$ let us denote by m' the minimum value such that $\vec{\kappa} \in \mathcal{N}_{m'}$. Now, using Lemma 5.3.2, we have that the sequence

formed by $\mathcal{C}_m(\vec{\kappa})$, for any $m \geq m'$, converges to $\mathcal{C}(\vec{\kappa})$ when m goes to infinity. So this implies that, for any $\vec{\kappa} \in \mathcal{N}_\infty$, $\mathcal{C}_m(\vec{\kappa})$ converges to $\mathcal{C}(\vec{\kappa})$. Because the set of points $\mathcal{N}_\infty$ is dense in $\mathcal{K}_{\text{red}}$, we conclude that $\mathcal{C}_m$ converges pointwise almost everywhere to $\mathcal{C}$. Furthermore, $\mathcal{C}$ is a continuous function, as well as all the functions in the sequence $\mathcal{C}_m$, so the pointwise convergence on a dense set of points is enough to imply the uniform convergence.

□

Finally, we present some numerical results using Algorithm 4. We use the same supercell used in Section 5.2 and also we shall approximate the band of the trapped mode already analysed in that section. We are going to compare the results from Algorithm 4 against the results from the adaptive method presented in the previous section, which consists in applying Algorithm 3 to each considered value of the quasimomentum with always the same structured starting mesh. In particular we are interested in comparing the computational costs of these two approaches.

The starting mesh $\mathcal{G}_0$ contains just one element as big as $\mathcal{K}_{\text{red}}$ for the considered supercell. In this numerical experiment we are going to construct just two refinements of $\mathcal{G}_0$, namely: $\mathcal{G}_1$ and $\mathcal{G}_2$; so we set $\max_m = 2$. Moreover, we set $\max_{\text{it}} = 2$, which means that for any iteration of level A we are going to refine twice the meshes for each $\vec{\kappa} \in \mathcal{N}_m$ in level B. We also set $\theta = \tilde{\theta} = 0.5$. For the sake of clearness we are not going to consider all the nodes in the sequence of meshes $\mathcal{G}_m$, but just a subset of them showed in Figure 5-13. So, for $m = 0$ we are going to consider only the point $\vec{\kappa} = (0, 0)$; for $m = 1$ we are going to consider only the points $\vec{\kappa} = (\pi/15, 0)$ and $\vec{\kappa} = (\pi/15, \pi/15)$; finally for $m = 2$ we are going to consider only the points $\vec{\kappa} = (\pi/45, 0)$, $\vec{\kappa} = (2\pi/45, 0)$, $\vec{\kappa} = (\pi/45, \pi/45)$, $\vec{\kappa} = (2\pi/45, 2\pi/45)$, $\vec{\kappa} = (\pi/15, \pi/45)$, $\vec{\kappa} = (\pi/15, 2\pi/45)$ and $\vec{\kappa} = (2\pi/45, \pi/45)$.

In our simulation, due to the choice of $\max_{\text{it}}$, the meshes for all the points in $\mathcal{N}_0$ will be refined at maximum 6 times. For all the points in $\mathcal{N}_1/\mathcal{N}_0$, the meshes will be refined at maximum 4 times and for all the points in $\mathcal{N}_2/(\mathcal{N}_0 \bigcup \mathcal{N}_1)$, the meshes will be refined at maximum 2 times. In Table 5.21, we compare, for all the considered values of the quasimomentum, the results from Algorithm 4 against the approximations from Algorithm 2. In column m we put the minimum value of m such that each considered point $\vec{\kappa} \in \mathcal{N}_m$. In the columns #ref we put for each method the number of refinements of the mesh on Ω necessary to reach the same accuracy. In the run of Algorithm 4 a total number of 28 refinements and a total of 38 computations of discrete eigenpairs have been done. Instead, summing the values of columns #ref, it is clear that Algorithm 2 needed 60 refinements, which correspond to 70 computations of discrete eigenpairs, to reach the same accuracy. In conclusion, the saving of computational power is quite remarkable. However, the efficiency of Algorithm 4 may depend on how fine is the mesh $\mathcal{G}_0$ and also on all the other parameters that we set.

Algorithm 4 Efficient method to compute bands

Require: $\mathcal{G}_0$
Require: $\max_m > 0$
Require: $\mathrm{tol}_m > 0, \quad \forall 0 \leq m \leq \max_m$
Require: $0 < \theta < 1$
Require: $0 < \tilde{\theta} < 1$
Require: $\max_{\mathrm{it}} > 0$
Require: $\mathcal{T}_0$
  for all $\vec{\kappa} \in \mathcal{N}_0$ **do**
    $\mathcal{T}_0^{\kappa} := \mathcal{T}_0$
    $\mathcal{C}_0(\vec{\kappa}) := 0$
  end for
  $m = 0$
  repeat
    for all $\vec{\kappa} \in \mathcal{N}_m$ **do**
      $(\lambda_{m+1}^{\kappa}, u_{m+1}^{\kappa}, \mathcal{T}_{m+1}^{\kappa}) = \mathrm{AFEM}(\vec{\kappa}, \mathrm{FatherMesh}(\vec{\kappa}), \theta, \tilde{\theta}, \mathrm{tol}_m, \max_{\mathrm{it}})$
      $\mathcal{C}_{m+1}(\vec{\kappa}) := \lambda_{m+1}^{\kappa}$
    end for
    Refine the mesh $\mathcal{G}_m$ and construct $\mathcal{G}_{m+1}$
    $m = m + 1$
  until $m \leq \max_m$

Algorithm 4				**Standard adaptivity**	
m	$\vec{\kappa}$	$\lvert\lambda_m^{\kappa} - \lambda^{\kappa}\rvert$	#ref	$\lvert\lambda_m^{\kappa} - \lambda^{\kappa}\rvert$	#ref
0	$(0,0)$	0.0428	6	0.0428	6
1	$(\pi/15, 0)$	0.0373	4	0.0336	6
1	$(\pi/15, \pi/15)$	0.0598	4	0.0403	6
2	$(\pi/45, 0)$	0.0269	2	0.0252	7
2	$(2\pi/45, 0)$	0.0277	2	0.0261	6
2	$(\pi/45, \pi/45)$	0.0269	2	0.0331	6
2	$(2\pi/45, 2\pi/45)$	0.0488	2	0.0337	6
2	$(\pi/15, \pi/45)$	0.0407	2	0.0312	6
2	$(\pi/15, 2\pi/45)$	0.0517	2	0.0622	5
2	$(2\pi/45, \pi/45)$	0.0324	2	0.0259	6

Table 5.21: Comparison between Algorithm 4 and the standard adaptive method, both applied to the band of the trapped mode for the TE case problem on a supercell.

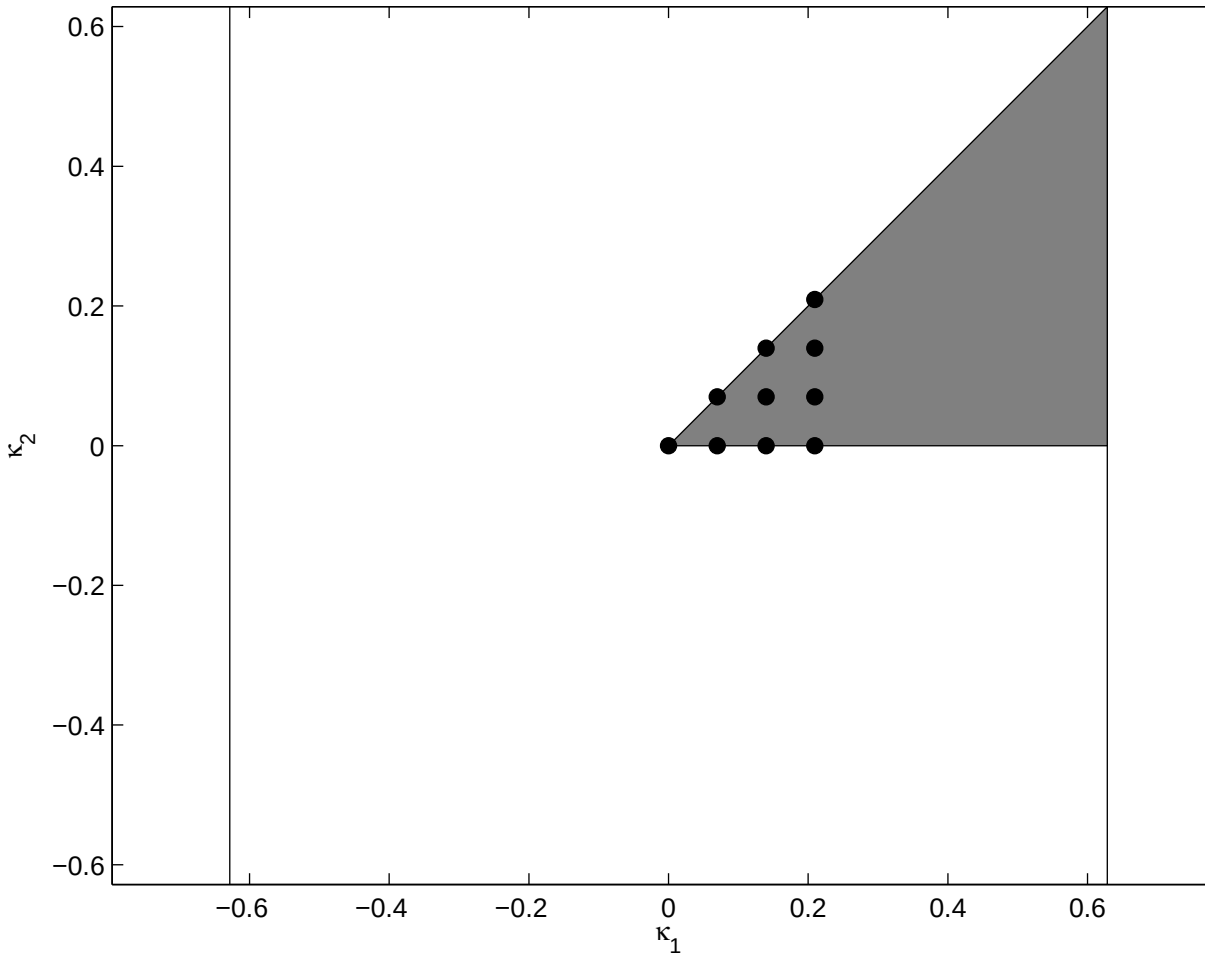

Figure 5-13: A picture of the reduced first Brillouin zone with the points considered in the simulations.

Conclusions

The main objective of this work is a new convergent and efficient method for PCFs based on an adaptive FEM. In order to reach our goal, we had also to prove some new results and to extend some existing theories. In Chapter 2, we extended the theory in [51] to the multiple eigenvalue case. This first extension leads us toward new results in Chapter 3 about a posteriori error estimators in the multiple eigenvalue case. We have also presented a new estimator for PCF eigenvalue problems and also we have fully embedded in the a posteriori theory for elliptic eigenvalue problems the fact that eigenvalues can have multiplicity greater than one. This is particularly clear in all the reliability and efficiency results which are stated in a way to consider any degree of multiplicity.

The central part of this work is, of course, the proof of convergence for adaptive finite element methods for elliptic eigenvalue problems and for PCF problems. At the moment, these results are stated only for simple eigenvalues, but we would like to extend them in the future to the multiple eigenvalue case.

Another aspect that we have planed to study further in the future is the dependence of the convergent results (Theorem 4.1.17 and Theorem 4.2.13) on the initial meshes and on the value of the considered eigenvalues. In particular, we would like to prove results showing that for a given problem, and considering an eigenvalue λ, a value $H_0^{\max}$ for the initial mesh is enough in order to trigger the convergence. Such a result will be very useful in practice, since it will ensure that the method is going to converge to the correct eigenpair.

In addition, we would like to extend the proof of convergence to higher order finite elements. When we tried to do this we found that the main difficulty was the extension of Lemma 4.1.11 to higher orders.

In addition, we would like to note the rich set of numerical and theoretical results collected in Chapter 5. In Section 5.3 we presented the *first* convergent adaptive method to compute bands of spectra for PCFs. Finally, we are proud of the numerical results in Section 5.1.6 about trapped modes, which are of great interest in applications.

Bibliography

[1] Adams, R. A. and J. J. F. Fournier (2003). *Sobolev Spaces*. Elsevier.

[2] Ainsworth, M. and T. J. Oden (2000). *A Posteriori Error Estimation in Finite Element Analysis*. Wiley.

[3] Ammari, H. and F. Santosa (2004). Guided waves in a photonic bandgap strucutre with a line defect. *SIAM J. Appl. Math. 64*(6), 2018–2033.

[4] Axmann, W. and P. Kuchment (1999). An efficient finite element method for computing spectra of photonic and acoustic band-gap materials. *J. Comput. Physics 150*, 468–481.

[5] Babuška, I. (1970). The finite element method for elliptic equations with discontinuous coefficients. *Computing 5*, 207–213.

[6] Babuška, I. and J. Osborn (1991). Eigenvalue problems. *in Handbook of Numerical Analysis Vol II, eds P.G. Cairlet and J.L. Lions, North Holland*, 641–787.

[7] Bernardi, C. and R. Verfürth (2000). Adaptive finite element methods for elliptic equations with non-smooth coefficients. *Numer. Math. 85*, 579–608.

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

[9] Brenner, S. C. and L. R. Scott (2002). *The Mathematical Theory of Finite Element Methods*. Springer.

[10] Cao, Y., Z. Hou, and Y. Liu (2004, June). Convergence problem of plane-wave expansion method for phononic crystals. *Physics Letters A 327*, 247–253.

[11] Carstensen, C. (2005). A unifying theory of a posteriori finite element error control. *Numer. Math. 100*, 617–637.

[12] Carstensen, C. and J. Gedicke (2008). An oscillation-free adaptive fem for symmetric eigenvalue problems. preprint.

[13] Carstensen, C. and R. H. W. Hoppe (2006a). Convergence analysis of an adaptive non-conforming finite element method. *Numer. Math. 103*, 251–266.

[14] Carstensen, C. and R. H. W. Hoppe (2006b). Error reduction and convergence for an adaptive mixed finite element method. *Math. Comp. 75*(255), 1033–1042.

[15] Cox, S. J. and D. C. Dobson (1999). Maximizing band gaps in two-dimesional photonic crystals. *SIAM J. Appl. Math. 59*(6), 2108–2120.

[16] Dobson, D. C. (1999). An efficient method for band structure calculations in 2D photonic crystals. *J. Comp. Phys. 149*, 363–376.

[17] Dobson, D. C., J. Gopalakrishnan, and J. E. Pasciak (2000). An efficient method for band structure calculations in 3D photonic crystals. *J. Comput. Phys. 161*(2), 668–679.

[18] Dobson, D. C. and F. Santosa (2004). Optimal loacalization of eigenfunctions in an inhomogeneous medium. *SIAM J. Appl. Math. 64*(3), 762–774.

[19] Dörfler, W. (1995). A robust adaptive strategy for the nonlinear poisson equation. *Computing 55*, 289–304.

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

[21] Durán, R. G., C. Padra, and R. Rodríguez (2003). A posteriori error estimates for the finite element approximation of eigenvalue problems. *Math. Mod. & Met.in Appl. Sc. 13*(8), 1219–1229.

[22] Figotin, A. and V. Goren (2001). Resolvent method for computations of localized defect modes of H-polarization in two-dimensional photonic crystals. *Phys. Rev. E 64*, 1–16.

[23] Figotin, A. and V. Gorentsveig (1998). Localized electromagnetic waves in a layered periodic dielectric medium with a defect. *Phys. Rev. B 58*(1), 180–188.

[24] Figotin, A. and A. Klein (1997). Localized classical waves created by defects. *J. Stat. Phys. 86*, 165–177.

[25] Figotin, A. and A. Klein (1998). Midgap defect modes in dielectric and acoustic media. *SIAM J. Appl. Math. 58*(6), 1748–1773.

[26] Giani, S. and I. G. Graham (2007). A convergent adaptive method for elliptic eigenvalue problems. *submitted for publication*.

[27] Hackbusch, W. (1992). *Elliptic Differential Equations*. Springer.

[28] Heuveline, V. and R. Rannacher (2001). A posteriori error control for finite element approximations of elliptic eigenvalue problems. *J. Adv. Comp. Math. 15*, 107–138.

[29] Hiett, B. (2000). Photonic crystal modelling using finite element analysis. Master's thesis, University of Southampton.

[30] Hislop, P. D. and I. M. Sigal (1996). *Introduction to Spectral Analysis*. Springer.

[31] Johnson, S. G. and J. D. Joannopoulos (2001). Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis. *Optics Express 8*, 173–190.

[32] Johnson, S. G. and J. D. Joannopoulos (2002). *Photonic Cristals. The Road from Theory to Practise*. Kluwer Acad. Publ.

[33] Klöckner, A. (2004). On the computation of maximally localized Wannier functions. Master's thesis, Karlsruhe University.

[34] Kuchment, P. (1993). *Floquet Theory for Partial Differential Equations*. Birkhauser Verlag.

[35] Kuchment, P. (2001). The mathematics of photonic crystals. *SIAM, Frontiers Appl. Math. 22*, 207–272.

[36] Kuchment, P. and S. Levendorskiĭ (1991). On the structure of spectra of periodic elliptic operators.

[37] Larson, M. G. (2000). A posteriori and a priori error analysis for finite element approximations of self-adjoint elliptic eigenvalue problems. *SIAM J. Numer. Anal 38*, 608–625.

[38] Lehoucq, R. B., D. C. Sorensen, and C. Yang (1998). *ARPACK Users' Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM.

[39] M. Bourland, M. Dauge, M. S. L. and S. Nicaise (1992). Coefficients of the singularities for elliptic boundary value problems on domains with conical points. iii: Finite element methods on polygonal domains. *SIAM J. Numer. Anal. 29*, 136–155.

[40] Mekchay, K. and R. H. Nochetto (2005). Convergence of adaptive finite element methods for general second order linear elliptic PDEs. *SIAM J. Numer. Anal. 43*, 1803–1827.

[41] Mommer, M. S. and R. Stevenson (2006). A goal-oriented adaptive finite element method with convergence rates. Technical Report 1357, Department of Mathematics, Utrecht University.

[42] Morin, P., R. H. Nochetto, and K. G. Siebert (2000). Data oscillation and convergence of adaptive FEM. *SIAM J. Numer. Anal. 38*, 466–488.

[43] Morin, P., R. H. Nochetto, and K. G. Siebert (2002). Local problems on stars: a posteriori error estimators, convergence, and performance. *SIAM J. Numer. Anal. 72*, 1067–1097.

[44] Pearce, G. J., T. D. Hedley, and D. M. Bird (2005, May). Adaptive curvilinear coordinates in a plane-wave solution of Maxwell's equations in photonic crystals. *Physical Review B 71*(19), 195108–+.

[45] Petzoldt, M. (2001). *Regularity and error estimators for elliptic problems with discontinuous coefficients*. Ph. D. thesis, Weierstraß Institut.

[46] Sakoda, K. (2001). *Optical Properties of Photonic Crystals*. Springer-Verlag.

[47] Scott, J. A. (2000). Sparse direct methods: An introduction. *Lecture Notes in Physics 401*(535).

[48] Scott, R. L. and S. Zhang (1990). Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math Comp 54*, 483–493.

[49] Soussi, S. (2005). Convergence of the supercell method for defect modes calculations in photonic crystals. *SIAM J. Numer. Anal. 43*(3), 1175–1201.

[50] Spence, A. and C. Poulton (2005). Photonic band structure calculations using nonlinear eigenvalue techniques. *J. Comp. Phys. 204*, 65–81.

[51] Strang, G. and G. J. Fix (1973). *An Analysis of Finite Element Method*. Prentice-Hall.

[52] Verfürth, R. (1996). *A Review of Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*. Wiley-Teubner.

[53] Walsh, T. F., G. M. Reese, and U. L. Hetmaniuk (2007). Explicit a posteriori error estimates for eigenvalue analysis of heterogeneous elastic structures. *CMAME 196*(37), 3614–3623.