MONOGRAPHS AND RESEARCH NOTES IN MATHEMATICS

Finite Element Methods for Eigenvalue Problems

Jiguang Sun
Aihui Zhou

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Finite Element Methods for Eigenvalue Problems

MONOGRAPHS AND RESEARCH NOTES IN MATHEMATICS

Published Titles

Application of Fuzzy Logic to Social Choice Theory, John N. Mordeson, Davender S. Malik and Terry D. Clark

Blow-up Patterns for Higher-Order: Nonlinear Parabolic, Hyperbolic Dispersion and Schrödinger Equations, Victor A. Galaktionov, Enzo L. Mitidieri, and Stanislav Pohozaev

Complex Analysis: Conformal Inequalities and the Bieberbach Conjecture, Prem K. Kythe

Computational Aspects of Polynomial Identities: Volume l, Kemer's Theorems, 2nd Edition Alexei Kanel-Belov, Yakov Karasik, and Louis Halle Rowen

A Concise Introduction to Geometric Numerical Integration, Fernando Casas and Sergio Blanes

Cremona Groups and Icosahedron, Ivan Cheltsov and Constantin Shramov

Delay Differential Evolutions Subjected to Nonlocal Initial Conditions Monica-Dana Burlică, Mihai Necula, Daniela Roșu, and Ioan I. Vrabie

Diagram Genus, Generators, and Applications, Alexander Stoimenow

Difference Equations: Theory, Applications and Advanced Topics, Third Edition Ronald E. Mickens

Dictionary of Inequalities, Second Edition, Peter Bullen

Finite Element Methods for Eigenvalue Problems, Jiguang Sun and Aihui Zhou

Introduction to Abelian Model Structures and Gorenstein Homological Dimensions Marco A. Pérez

Iterative Optimization in Inverse Problems, Charles L. Byrne

Line Integral Methods for Conservative Problems, Luigi Brugnano and Felice Iavernaro

Lineability: The Search for Linearity in Mathematics, Richard M. Aron, Luis Bernal González, Daniel M. Pellegrino, and Juan B. Seoane Sepúlveda

Modeling and Inverse Problems in the Presence of Uncertainty, H. T. Banks, Shuhua Hu, and W. Clayton Thompson

Monomial Algebras, Second Edition, Rafael H. Villarreal

Nonlinear Functional Analysis in Banach Spaces and Banach Algebras: Fixed Point Theory Under Weak Topology for Nonlinear Operators and Block Operator Matrices with Applications, Aref Jeribi and Bilel Krichen

Partial Differential Equations with Variable Exponents: Variational Methods and Qualitative Analysis, Vicenţiu D. Rădulescu and Dušan D. Repovš

A Practical Guide to Geometric Regulation for Distributed Parameter Systems Eugenio Aulisa and David Gilliam

Published Titles Continued

Reconstruction from Integral Data, Victor Palamodov

Signal Processing: A Mathematical Approach, *Second Edition*, Charles L. Byrne

Sinusoids: Theory and Technological Applications, Prem K. Kythe

Special Integrals of Gradshteyn and Ryzhik: the Proofs – Volume I, Victor H. Moll

Special Integrals of Gradshteyn and Ryzhik: the Proofs – Volume II, Victor H. Moll

Stochastic Cauchy Problems in Infinite Dimensions: Generalized and Regularized Solutions, Irina V. Melnikova

Submanifolds and Holonomy, Second Edition, Jürgen Berndt, Sergio Console, and Carlos Enrique Olmos

The Truth Value Algebra of Type-2 Fuzzy Sets: Order Convolutions of Functions on the Unit Interval, John Harding, Carol Walker, and Elbert Walker

Forthcoming Titles

Actions and Invariants of Algebraic Groups, Second Edition, Walter Ferrer Santos and Alvaro Rittatore

Analytical Methods for Kolmogorov Equations, Second Edition, Luca Lorenzi

Geometric Modeling and Mesh Generation from Scanned Images, Yongjie Zhang

Groups, Designs, and Linear Algebra, Donald L. Kreher

Handbook of the Tutte Polynomial, Joanna Anthony Ellis-Monaghan and Iain Moffat

Microlocal Analysis on Rˆn and on NonCompact Manifolds, Sandro Coriasco

Practical Guide to Geometric Regulation for Distributed Parameter Systems, Eugenio Aulisa and David S. Gilliam

Symmetry and Quantum Mechanics, Scott Corry

MONOGRAPHS AND RESEARCH NOTES IN MATHEMATICS

Finite Element Methods for Eigenvalue Problems

Jiguang Sun

Michigan Technological University

Houghton, USA

Aihui Zhou

Chinese Academy of Sciences

Beijing, China

CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the

Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20160511

International Standard Book Number-13: 978-1-4822-5464-8 (Hardback)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

The numerical solution of eigenvalue problems is of fundamental importance in many scientific and engineering applications, such as structural dynamics, quantum chemistry, electrical networks, magnetohydrodynamics, and control theory [77, 215, 110, 12, 28, 42]. Due to the flexibility in treating complex structures and rigorous theoretical justification, finite element methods, including conforming finite elements, nonconforming finite elements, mixed finite elements, discontinuous Galerkin methods, etc., have been popular for eigenvalue problems of partial differential equations.

There are many excellent references on finite element methods for eigenvalue problems [236, 46, 121, 78, 242, 211, 136, 137, 138, 175, 70, 114, 115, 142, 200, 167, 21, 22, 79, 28, 179, 148, 45, 41, 174, 239, 126, 11, 57, 35, 74, 75, 218, 73], in particular, the book chapter by Babuška and Osborn [23]. However, to the authors' opinion, there is a need for a self-contained, systematic, and up-to-date treatment. This is the motivation for this book.

We start with functional analysis including operator perturbation theory in Chapter 1. For fundamental materials such as Banach spaces, we present only the results and point out the references for their derivation and/or proofs. Advanced results, which are needed to treat a particular eigenvalue problem, are discussed in respective chapters. However, we give a detailed account of those that will be used quite often in later chapters. In particular, we include the proofs for the abstract convergence theory of Babuška and Osborn [23], which serves as a major tool for convergence analysis of many eigenvalue problems.

We introduce basics of finite element methods in Chapter 2. Again, other than a complete account, we keep the introduction concise and refer the readers to classical textbooks. For example, we only choose the typical triangular mesh in two dimensions and tetrahedral mesh in three dimensions. There are many other meshes such as rectangular meshes or hexahedral meshes. They are important topics in finite elements. However, since the focus of this book is the eigenvalue problem, we believe they are less relevant and left them out. Some implementation aspects are discussed at the end of this chapter.

In Chapter 3, the Laplace eigenvalue problem is treated using the Lagrange elements. The convergence analysis follows directly the theory of Babuška and Osborn [23]. The materials are classical and serve well as a model problem. In fact, the results for the Laplace eigenvalue problem are useful in the analysis of many other eigenvalue problems. Note that many other methods have been proposed for the Laplace eigenvalue problem, for example, mixed methods [36], and the discontinuous Galerkin method [11]. However, to make the first treatment of the eigenvalue problem easy to follow, we do not discuss those more technical methods.

Chapter 4 is on biharmonic eigenvalue problems. Different methods are discussed and compared in this chapter, including the conforming Argyris element, the non-conforming Morley element, the Ciarlet-Raviart mixed method, and an interior penalty discontinuous Galerkin method. Accordingly, different techniques are necessary for the convergence proofs. The biharmonic eigenvalue problem is of fourth order. It is a good model problem for the readers to see that different methods have cons and pros, respectively.

Chapter 5 contains the Maxwell's eigenvalue problem. We introduce a mixed method using the edge element, which is curl-conforming and spectrally correct. A rather detailed treatment of this problem can be found in [35]. At the end of the chapter, we discuss a mixed finite element method for the quad-curl eigenvalue problem. This is a fourth order problem. The study of finite element methods for it has barely started.

Chapter 6 is on the transmission eigenvalue problem, a new research topic arising from the inverse scattering theory. The problem is extremely challenging since it is nonlinear and nonself-adjoint. Only very recently, the problem drew some attention of numerical analysts. In fact, the theory of the problem is not complete yet. We present several methods including iterative methods and two mixed methods. Special treatment is needed for the convergence analysis due to the nonself-adjointness. We believe a lot of works can be done for this new problem. The problem can be written as a quadratic eigenvalue problem and techniques for nonlinear eigenvalue problems may be helpful. The problem is essentially a fourth order problem and most methods for the biharmonic eigenvalue problems might work. In addition, transmission eigenvalue problems of electromagnetics and elasticity are largely untouched.

Chapter 7 is on the Schrödinger eigenvalue problem. We first study the standard finite element method for a nonlinear eigenvalue problem, the Gross–Pitaevskii equation, which models a Bose–Einstein condensation. Both convergence and error estimate are addressed. To efficiently solve the resulting linear Schrödinger equation in electronic structure calculations, we present and analyze a two-scale finite element discretization.

Adaptive finite element methods have been an important topic, which is discussed in Chapter 8. The Laplace eigenvalue problem is used as a model problem to illustrate the basics. In particular, we focus on construction and analysis of the residual based a posteriori error estimators. The analysis starts from the approximation to the Poisson equation and moves on to the Laplace eigenvalue problem based on a so-called perturbation argument.

Finite element discretization inevitably leads to matrix eigenvalue problems. In general, one uses existing matrix eigenvalue solvers as a black box. However, we feel it is beneficial to introduce some effective methods, such as the QR method, the power iteration, the Arnoldi method, etc. These are the topics of Chapter 9.

In Chapter 10, we introduce integral based eigenvalue solvers, which are quite popular recently. In particular, we present a recursive eigenvalue solver based on the spectrum projection. An application of the new method to a nonlinear eigenvalue problem is presented. The methods of last two sections of Chapter 10 can be viewed as eigensolvers without actually computing the eigenvalues. We believe these non-classical methods are a promising research direction, specially for problems to which classical methods are handicapped. One example of such problems is the non-Hermitian eigenvalue problem for large sparse matrices. Some interior eigenvalues are needed and there is little a priori spectrum information.

There are many important and interesting works on finite element methods for eigenvalue problems. We made the choices based on two criteria. The first one is that the problem should be fundamental and can be used to illustrate the basic theory. The second is our own research interests. The Laplace eigenvalue problem, the biharmonic eigenvalue problem, and the Maxwell's eigenvalue problem meet the first criterion. The transmission eigenvalue problem, the Schrödinger eigenvalue problem, adaptive finite element approximations, and the integral based eigensolvers are chosen based on the second criterion.

Consequently, there are many eigenvalue problems not covered in this book, e.g., the Steklov eigenvalue problem [47, 9, 14, 71], eigenvalue problems of elasticity [199], waveguide band structures [48, 127, 207], Stokes eigenvalue problems [212, 197, 91, 143, 195, 118, 154, 247], etc.

Of course, many interesting topics are not discussed or fully discussed. We list some of them here: discontinuous Galerkin methods [11, 58], the bounds on eigenvalues approximated by finite element methods [76, 151, 25, 154, 152], multi-level/multi-grid methods [249, 160, 162, 257, 247, 193], superconvergence [153, 192, 195], computation of a large number of eigenvalues or eigenvalue cluster [146, 40, 122], spectra pollution [37, 36, 109, 186], nonlinear eigenvalue problems [234, 237, 86, 69, 84], etc.

This book can be used as a graduate textbook for a course on finite element methods for eigenvalue problems. The manuscript was used for a graduate course, Topics on Computational Mathematics, at Michigan Technological University. A one-semester course can be arranged as follows: Functional Analysis, Finite Elements, Laplace Eigenvalue Problem, Biharmonic Eigenvalue Problems, Maxwell's Eigenvalue Problem. If time permits, the instructors can choose to cover either Matrix Eigenvalue Problems or one of the remaining chapters.

The book can also serve as a self-contained reference for researchers who are interested in finite element methods for eigenvalue problems. In fact, we try to make every single chapter self-contained by minimizing the cross-references between chapters. Thus the readers can work on their interested eigenvalue problems without going back and forth too much in the book. Most materials on transmission eigenvalues are recent research results. The study of the quad-curl eigenvalue problem has just started. The last two sections of Chapter 10 were investigated within the last two years. We hope the presentation can draw some attention of researchers to these interesting research topics.

We would like to thank many people who helped us in the preparation of this book. The class of Topics on Computational Mathematics at Michigan Technological University (MTU) suggested many corrections and improvements. Graduate students at the Chinese Academy of Sciences (CAS) proofread the book. We would also like to thank the editors and staff from CRC Press, Taylor & Francis Group, for their great help. Finally, we would like to thank the National Science Foundation, the Funds for Creative Research Groups, the National Basic Research Program, and the National Science Foundation of China for their support. The Department of Mathematics at Michigan Technological University(MTU), MTU Research Excellence Fund, and Institute of Computational Mathematics and Scientific/Engineering Computing at the Chinese Academy of Sciences provided travel funds for the authors during the preparation of this book.

List of Figures

List of Tables

Symbol Description

$\|\cdot\|$	L^2-norm
Y^c	complement of set Y
$Y^\perp$	orthogonal complement of set Y
M^a	annihilator of M
$\mathcal{C}^k(\Omega)$	the set of k times continuously differentiable functions on Ω
$\mathcal{C}^k_0(\Omega)$	the set of k times continuously differentiable functions with compact support in Ω
$\mathcal{C}^k_0(\overline{\Omega})$	the set of k times continuously differentiable functions which have bounded and uniformly continuous derivatives up to order k with compact support in Ω
$\mathcal{C}^\infty_0(\Omega)$	the set of smooth function with compact support in Ω
$L^p(\Omega)$	the set of functions such that $\|\phi\|^p$ is integrable on Ω ($1 \leq p < \infty$)
$\boldsymbol{\alpha}$	multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$
$W^{s,p}(\Omega)$	Sobolev space of functions with L^p-integrable derivatives up to order s
$H^s(\Omega)$	$W^{s,2}(\Omega)$
$\mathcal{T}_h$	a mesh with size h
$\mathcal{P}_k$	the set of all polynomials of degree at most k
$\hookrightarrow$	compact embedding
$x_n \rightharpoonup x$	$\{x_n\}$ converges to x weakly
$\|\cdot\|_{H^s(\Omega)}$	the semi-norm in $H^s(\Omega)$
$\mathcal{E}$	the electric field
$\mathcal{H}$	the magnetic field
$\mathcal{D}$	the electric displacement
$\mathcal{B}$	the magnetic induction
$\sigma(T)$	the spectrum of T
$\rho(T)$	the resolvent set of T
$\sigma_p(T)$	the point spectrum of T
$\sigma_c(T)$	the continuous spectrum of T
$\sigma_r(T)$	the residual spectrum of T
j_m	the mth order spherical Bessel function
$\mathbb{S}$	the unit circle in $\mathbb{R}^2$

Chapter 1

Functional Analysis

1.1 Basics

The analysis of finite element methods relies on results from functional analysis. In this section, we collect some fundamental results which will be used in this book. Most proofs are not provided since the materials can be found in classical textbooks such as [178, 251, 79], which are the major sources of this chapter.

1.1.1 Metric Spaces, Banach Spaces and Hilbert Spaces

Definition 1.1.1. *A metric space is a set X together with a metric $d(\cdot,\cdot)$ defined on $X \times X$ such that for all $x, y, z \in X$*

(1) $d(\cdot,\cdot)$ is real-valued, finite and non-negative, $d(x,y) = 0$ if and only if $x = y$;

(2) $d(x,y) = d(y,x)$;

(3) $d(x,y) \le d(x,z) + d(z,y)$.

We also call d a distance function on X. Sometimes we write d_X to emphasize that it is a distance function related to space X. Property (3) is called the triangle inequality.

Given a point $x_0 \in X$ and a real number $r > 0$, we define open ball: $B(x_0; r) = \{x \in X \mid d(x, x_0) < r\}$;

closed ball: $\overline{B(x_0; r)} = \{x \in X \mid d(x, x_0) \leq r\}$;

sphere: $S(x_0; r) = \{x \in X \mid d(x, x_0) = r\}$.

Definition 1.1.2. *A subset Y of a metric space X is said to be open if it contains an open ball at each of its points. A subset Y of X is said to be closed if its complement in X is open, i.e., $Y^c = X \setminus Y$ is open.*

We call x_0 an interior point of a set Y if Y contains an open ball $B(x_0; \epsilon)$ for some $\epsilon > 0$. The interior of Y is the set of all interior points of Y. Now we can define the continuous mapping.

Definition 1.1.3. *Let (X, d_X) and (Y, d_Y) be metric spaces. A mapping $T : X \to Y$ is said to be continuous at a point $x_0 \in X$ if for every $\epsilon > 0$ there exists a $\delta > 0$ such that*

$$d_Y(Tx, Tx_0) < \epsilon \quad \textit{if} \quad d_X(x, x_0) < \delta.$$

T is said to be continuous if it is continuous at every point of X.

A point $x \in X$ is called an accumulation point of Y if, for any $\epsilon > 0$, there exists at least one point $y \in Y, y \neq x$ such that $d(x, y) < \epsilon$. Note that it is not necessary that $x \in Y$. The union of Y and all accumulation points of Y is called the closure of Y, written as $\overline{Y}$.

Definition 1.1.4. *A subset Y of a metric space X is said to be dense in X if the closure of Y is X, i.e., $\overline{Y} = X$. If X has a countable dense subset, X is said to be separable.*

Definition 1.1.5. *A sequence $\{x_n\}$ in a metric space X is said to be convergent if there is an $x \in X$, called the limit of $\{x_n\}$, such that*

$$\lim_{n\to\infty} d(x_n, x) = 0.$$

Definition 1.1.6. *A sequence $\{x_n\}$ in X is said to be a Cauchy sequence if for every $\epsilon > 0$ there is an integer N depending on ϵ such that*

$$d(x_m, x_n) < \epsilon \quad \textit{for every } m, n > N.$$

The space X is said to be complete if every Cauchy sequence in X converges to an element of X.

Definition 1.1.7. *A metric space X is said to be compact if every bounded sequence in X has a convergent subsequence. A subset M of X is said to be compact if every sequence in M has a convergent subsequence whose limit is an element of M.*

We move on to introduce vector spaces.

Definition 1.1.8. *A vector space over a field K is a nonempty set X of elements $x, y, \ldots$ together with two algebraic operations: vector addition and vector multiplication of vectors by scalars in K.*

(1) Vector addition associates with an ordered pair (x, y) for $x, y \in X$ a vector $x + y$, called the sum of x and y, such that

$$x + y = y + x \quad \textit{and} \quad x + (y + z) = (x + y) + z.$$

In addition, there exists a zero vector 0 such that for every vector x, there exists a vector, denoted by $-x$, satisfying

$$x + 0 = x \quad \textit{and} \quad x + (-x) = 0.$$

(2) Multiplication by scalars associates every vector x and scalar α a vector αx such that for all vectors $x, y \in X$ and $\alpha, \beta \in K$

$$\alpha(\beta x) = (\alpha\beta)x \quad \textit{and} \quad 1 \cdot x = x.$$

In addition, the following distributive laws hold

$$\alpha(x + y) = \alpha x + \alpha y \quad \textit{and} \quad (\alpha + \beta)x = \alpha x + \beta x.$$

A set of vectors $\{x_1, \ldots, x_n\}$ is said to be linearly independent if

$$\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n = 0$$

holds only for

$$\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0.$$

Otherwise, $\{x_1, \ldots, x_n\}$ is said to be linearly dependent.

Definition 1.1.9. *A normed space X is a vector space on which a real-valued function $\|\cdot\|$, called norm, is defined such that*

(1) $\|x\| \geq 0$, $\|x\| = 0$ *if and only if* $x = 0$,

(2) $\|\alpha x\| = |\alpha|\|x\|$,

(3) $\|x + y\| \leq \|x\| + \|y\|$,

where $x, y \in X$ and α is any scalar.

Sometimes we write $\|\cdot\|_X$ to emphasize it is a norm on X. A norm $\|\cdot\|$ on X induces a metric d on X:

$$d(x, y) = \|x - y\| \quad \text{for } x, y \in X.$$

Definition 1.1.10. *Let X be an infinite dimensional normed space. We say that X has a countably-infinite basis if there is a sequence $\{x_i\}_{i\geq 1} \subset X$ for which the following holds. For each $x \in X$, there exist $\{\alpha_{n,i}\}_{i=1}^n$, $n = 1, 2, \ldots$, such that*

$$\left\| x - \sum_{i=1}^{n} \alpha_{n,i} x_i \right\| \to 0 \quad \textit{as } n \to \infty.$$

The space X is also said to be separable. The sequence $\{x_i\}_{i\geq 1}$ is called a basis if any finite subset of the sequence is linearly independent. We say that X has a Schauder basis $\{x_i\}_{i\geq 1}$ if for each $x \in X$, it is possible to write $x = \sum_{i=1}^{\infty} \alpha_i x_i$ as a convergent series in X for a unique choice of scalars $\{\alpha_i\}_{i\geq 1}$.

Definition 1.1.11. *A complete normed space X is called a Banach space.*

Definition 1.1.12. *A norm $\|\cdot\|_0$ on a vector space X is said to be equivalent to a norm $\|\cdot\|_1$ on X if there exist $a, b > 0$ such that*

$$a\|x\|_0 \leq \|x\|_1 \leq b\|x\|_0 \quad \textit{for all } x \in X.$$

Over a finite dimensional space, any two norms are equivalent.

Definition 1.1.13. *Let Y be a subset of a normed space X. The set Y is said to be dense in X if for any $x \in X$ and any $\epsilon > 0$, there is a $y \in Y$ such that $\|x - y\| < \epsilon$.*

We will encounter semi-norms when we study the Sobolev spaces.

Definition 1.1.14. *Given a vector space X, a semi-norm $|\cdot|$ is a function from X to $\mathbb{R}$ with the following properties*

(1) $|x| \geq 0$;

(2) $|\alpha x| = |\alpha||x|$;

(3) $|x + y| \leq |x| + |y|$.

Note that $|x| = 0$ does not necessarily imply $x = 0$.

We move on to introduce Hilbert spaces. The eigenvalue problems discussed in this book are posed in Hilbert spaces.

Definition 1.1.15. *Let X be a vector space over the complex numbers $\mathbb{C}$. An inner product on X is a mapping $(\cdot,\cdot)_X : X \times X \to \mathbb{C}$ such that*

(1) $(x, x)_X \geq 0$, $(x, x)_X = 0$ *if and only if* $x = 0$;

(2) $\overline{(x, y)_X} = (y, x)_X$ *for all* $x, y \in X$;

(3) for all $x, y, z \in X$ *and* $\alpha, \beta \in \mathbb{C}$ *we have that*

$$(\alpha x + \beta y, z)_X = \alpha(x, z)_X + \beta(y, z)_X.$$

For simplicity, we write an inner product on X as $(\cdot,\cdot)$ when there is no confusion from context. Sometimes we refer to inner product as scalar product. The inner product induces a norm on X:

$$\|x\|_X = \sqrt{(x, x)_X} \quad \text{for all } x \in X.$$

For all $x, y \in X$, the Cauchy-Schwarz inequality holds

$$|(x, y)_X| \leq \|x\|_X \|y\|_X.$$

Definition 1.1.16. *A complete inner product space X is called a Hilbert space.*

Definition 1.1.17. *A vector space X is said to be the direct sum of two subspaces Y and Z, written as $X = Y \oplus Z$, if each $x \in X$ has a unique representation*

$$x = y + z, \quad y \in Y, \, z \in Z.$$

Two vectors x and y are said to be orthogonal if $(x, y) = 0$. An element $x \in X$ is said to be orthogonal to a subset $Y \subset X$ if $(x, y) = 0$ for all $y \in Y$. Let Y be a closed subspace of a Hilbert space X. The orthogonal complement of Y, denoted by $Y^\perp$, is the closed subspace given by

$$Y^\perp = \{x \in X \,|\, (x, y)_X = 0 \quad \text{for all } y \in Y\}.$$

The following theorem is useful when we study the Maxwell's eigenvalue problem.

Theorem 1.1.1. *Let Y be a closed subspace of a Hilbert space X. For every $x \in X$, there exist unique $y \in Y$ and $z \in Y^\perp$ such that*

$$x = y + z \tag{1.1}$$

and $X = Y \oplus Y^\perp$.

Definition 1.1.18. *Let X be an inner product space and $\{x_i\}_{i \geq 1}$ is a subset of X. We call $\{x_i\}_{i \geq 1}$ an orthonormal system if*

$$(x_i, x_j) = \delta_{i,j}, \quad i, j \geq 1.$$

If the orthonormal system is a basis of X, we call it an orthonormal basis for X.

An orthonormal system $\{x_i\}_{i \geq 1}$ for X satisfies the Bessel's inequality:

$$\sum_{i=1}^{\infty} |(x, x_i)|^2 \leq \|x\|^2 \quad \text{for all } x \in X.$$

For any $x \in X$, the series $\sum_{i=1}^{\infty} (x, x_i) x_i$ converges in X. If $x = \sum_{i=1}^{\infty} a_i x_i \in X$, then $a_i = (x, x_i)$.

1.1.2 Linear Operators

Let X and Y be normed spaces. An operator $T : X \to Y$ is said to be linear if

$$T(\alpha x_1 + \beta x_2) = \alpha T x_1 + \beta T x_2 \quad \text{for all } \alpha, \beta \in \mathbb{C}, x_1, x_2 \in X$$

and bounded if

$$\|Tx\|_Y \leq C \|x\|_X \quad \text{for all } x \in X$$

for some constant C. We say T is continuous if, for every convergent sequence $\{x_n\}$ in X with limit x, we have

$$T x_n \to T x \quad \text{in } Y \text{ as } n \to \infty.$$

A linear operator is continuous if and only if it is bounded.

Definition 1.1.19. *We denote the set of all the continuous linear operators from a normed space X to a normed space Y by $\mathcal{L}(X,Y)$. When $Y = X$, we simply write $\mathcal{L}(X)$. The set $\mathcal{L}(X,Y)$ is a linear space. The norm of a bounded linear operator $T : X \to Y$ is defined as*

$$\|T\|_{\mathcal{L}(X,Y)} = \sup_{x\neq 0, x\in X} \frac{\|Ax\|_Y}{\|x\|_X}.$$

For simplicity, we write $\|T\|$ when it leads to no confusion from context.

Theorem 1.1.2. *If Y is a Banach space, $\mathcal{L}(X,Y)$ is a Banach space.*

Definition 1.1.20. *Let X and Y be normed spaces. A sequence of linear operators $\{T_n\}$ from X to Y is said to converge uniformly to a linear operator $T \in \mathcal{L}(X,Y)$ if*

$$\lim_{n\to\infty} \|T - T_n\| = 0.$$

The range of an operator $T : X \to Y$ is denoted by $T(X)$:

$$T(X) = \{y \in Y \,|\, y = Tx \text{ for some } x \in X\}.$$

The null space of T, a subspace of X, is defined as

$$N(T) = \{x \in X \,|\, Tx = 0\}.$$

Definition 1.1.21. *Let X be a normed space. A linear functional $f : X \to K$ is a linear operator such that $K = \mathbb{R}$ if X is a real vector space or $K = \mathbb{C}$ if X is a complex vector space.*

The set of all bounded linear functionals on X, denoted by X', is a normed space. It is called the dual space of X.

Definition 1.1.22. *Let $f \in X'$. For $x \in X$, we write $f(x) = \langle f, x\rangle$ and call it duality pairing.*

The norm of f, $\|f\|_{X'}$, or simply $\|f\|$, is defined as

$$\|f\| = \sup_{x\in X, x\neq 0} \frac{|f(x)|}{\|x\|_X} = \sup_{x\in X, \|x\|_X = 1} |f(x)|.$$

From Theorem 1.1.2, it is easy to see that the dual space X' of a normed space X is a Banach space. In fact, X' is just $\mathcal{L}(X,\mathbb{R})$ or $\mathcal{L}(X,\mathbb{C})$. Note that both $\mathbb{R}$ and $\mathbb{C}$ are Banach spaces.

Let Y be a normed space and $T : X \to Y$ be a bounded linear operator. Let $g \in Y'$. For any $x \in X$, there exists a functional f on X by

$$f(x) = g(Tx). \tag{1.2}$$

It is easy to see that f is linear since g and T are linear, respectively. In addition,

$$|f(x)| = |g(T(x))| \leq \|g\|\|Tx\| \leq \|g\|\|T\|\|x\|,$$

implying that f is bounded. Hence $f \in X'$.

The dual space of X' is denoted by X''. For each $x \in X$, we define a mapping S from X to X'' such that $Sx = g_x$ given by

$$g_x(f) = f(x) \quad f \in X'.$$

The mapping S is called the canonical mapping of X into X''. If the range of S is X'', i.e., $R(S) = X''$, we say X is reflexive.

Definition 1.1.23. *Let $T : X \to Y$ be a bounded linear operator. The adjoint operator of T, denoted by T', is from Y' to X' such that*

$$f(x) = (T'g)(x) = g(Tx). \tag{1.3}$$

The following theorem from [178] states an important property of T'.

Theorem 1.1.3. *The adjoint operator T' is linear and bounded. Furthermore,*

$$\|T'\| = \|T\|.$$

Next we introduce the concept of the dual basis, which is important for the abstract convergence theory for eigenvalue problems. Let M be a finite-dimensional subspace of X such that $X = M \oplus N$. The annihilator M^a of M is a closed subspace of X' defined as

$$M^a := \{f \in X' \,|\, \langle f, x\rangle = 0 \quad \text{for all } x \in M\}.$$

Let M' be the dual space of M. Let $\{x_i\}, i = 1, \ldots m$, be a basis of M. For $j = 1, \ldots, m$, let

$$N_j := \text{span}\{x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_m\} \oplus N.$$

Then there exists an $x_j^* \in X'$ such that

$$\langle x_j^*, x_j\rangle = 1, \quad \langle x_j^*, x_i\rangle = 0, \quad i \neq j,$$

and

$$\|x_j^*\| = \frac{1}{\mathrm{d}(x_j, N_j)},$$

where $\mathrm{d}(x_j, N_j)$ denotes the distance from x_j to N_j defined as

$$d(x_j, N_j) = \inf_{y \in N_j} d(x_j, y).$$

Furthermore,

$$\langle x_j^*, y\rangle = 0, \quad y \in N,$$

i.e., $x_j^* \in N^a$ and

$$\langle x_j^*, x_i\rangle = \delta_{i,j}, \quad i, j = 1, \ldots, m.$$

The set $\{x_j^*\}, j = 1, \ldots, m$, is a basis of M' such that

$$\langle x_j^*, x_i\rangle = \delta_{i,j}, \quad i, j = 1, \ldots, m.$$

It is called the dual basis of the basis $\{x_i\}, i = 1, \ldots, m$, of M.

For Hilbert spaces, the Riesz Representation Theorem holds (see, for example, Theorem 2.30 of [198]).

Theorem 1.1.4. *(Riesz Representation Theorem) Let X be a Hilbert space. For each $g \in X'$ there exists a unique $u \in X$ such that*

$$(u, v) = g(v) \quad \text{for all } v \in X.$$

Furthermore, $\|g\| = \|u\|_X$.

Definition 1.1.24. *A sequence $\{x_n\}$ in a normed space X is said to weakly converge to an $x \in X$, written as $x_n \rightharpoonup x$, if*

$$\lim_{n\to\infty} f(x_n) = f(x) \quad \text{for every } f \in X'.$$

Definition 1.1.25. *Let X and Y be normed spaces. A sequence of bounded operators $\{T_n\}$ is said to be*

(1) strongly convergent if $\|T_n x - Tx\| \to 0$ for all $x \in X$,

(2) weakly convergent if $|f(T_n x) - f(Tx)| \to 0$ for all $x \in X$ and $f \in Y'$.

Definition 1.1.26. *Let X and Y be Hilbert spaces and $T : X \to Y$ be a bounded linear operator. The Hilbert adjoint operator T^* is defined as $T^* : Y \to X$ such that for all $x \in X$ and $y \in Y$*

$$(Tx, y)_Y = (x, T^* y)_X.$$

Definition 1.1.27. *A bounded linear operator $T : X \to X$ is said to be*

(1) self-adjoint or Hermitian if $T^* = T$,

(2) unitary if T is bijective and $T^* = T^{-1}$,

(3) normal if $TT^* = T^*T$.

Let X be a Hilbert space and Y be a closed subspace of X. Then (1.1) defines a mapping

$$P : X \to Y \quad \text{such that} \quad y = Px.$$

The mapping P is called a projection of X onto Y. The projection P has the following properties:

(1) $P^2 = P$.

(2) $N(P) = Y^\perp$.

(3) A bounded linear operator $P : X \to X$ on a Hilbert space X is a projection if and only if P is self-adjoint and $P^2 = P$.

1.1.3 Spectral Theory of Linear Operators

Let X be a complex normed space and $T : X \to X$ be a bounded linear operator. The following theorem gives the definition of the spectral radius of T (see, e.g., Theorem 2.7 of [79]).

Theorem 1.1.5. *Let $T \in \mathcal{L}(X)$. The limit*

$$r_\sigma(T) := \lim_{k\to\infty} \|T^k\|^{1/k}$$

exists and is called the spectral radius of T.

Let the operator be defined as

$$T_z = T - zI,$$

where $z \in \mathbb{C}$ and I is the identity operator. If T_z has an inverse, denoted by

$$R_z(T) = (T - zI)^{-1},$$

it is called the resolvent operator of T.

Definition 1.1.28. *Let X be a complex normed space and $T : X \to X$ a linear operator. A regular value z of T is a complex number such that*

(1) $R_z(T)$ exist,

(2) $R_z(T)$ is bounded, and

(3) $R_z(T)$ is defined on a set which is dense in X.

The resolvent set $\rho(T)$ of T is the set of all regular values z of T. Its complement $\sigma(T) = \mathbb{C} \setminus \rho(T)$ is called the spectrum of T. The spectrum $\sigma(T)$ can be partitioned into three disjoint sets:

(1) point spectrum $\sigma_p(T)$ is the set of z such that $R_z(T)$ does not exist. We write $z \in \sigma_p(T)$ and call it an eigenvalue of T,

(2) continuous spectrum $\sigma_c(T)$ is the set of z such that $R_z(T)$ exists and is defined on a dense set in X, but $R_z(T)$ is unbounded,

(3) residual spectrum $\sigma_r(T)$ is the set of z such that $R_z(T)$ exists and the domain of $R_z(T)$ is not dense in X.

For $z_1, z_2 \in \rho(T)$, the first resolvent equation is given by (see, for example, [79])

$$\begin{aligned} R_{z_1} - R_{z_2} &= (z_1 - z_2)R_{z_1}R_{z_2} \\ &= (z_1 - z_2)R_{z_2}R_{z_1}. \end{aligned} \tag{1.4}$$

For $z \in \rho(T_1) \cap \rho(T_2)$, the second resolvent equation is given by

$$\begin{aligned} R_z(T_1) - R_z(T_2) &= R_z(T_1)(T_2 - T_1)R_z(T_2) \\ &= R_z(T_2)(T_2 - T_1)R_z(T_1). \end{aligned} \tag{1.5}$$

Theorem 1.1.6. *(Theorems 2.21 of [79]) For $T \in \mathcal{L}(X)$, the following properties hold.*

(1) If $|z| > r_\sigma(T)$, $R_z(T)$ exists and has the series expansion

$$R_z(T) = -\sum_{k=0}^{\infty} z^{-k-1} T^k.$$

(2) $\rho(T)$ and $\sigma(T)$ are nonempty. $\sigma(T)$ is compact.

(3) $r_\sigma(T) = \max_{z \in \sigma(T)} |z|$.

Definition 1.1.29. *Let $z \in \sigma_p(T)$ be an eigenvalue of T. If*

$$T_z x := Tx - zx = 0$$

for some $x \neq 0$, x is called an eigenfunction of T associated to z.

A subspace M of X is called an invariant subspace under T if $T(M) \subset M$. We write $T_M := T|_M$ for the restriction of T on M. If $X = M \oplus N$, where M, N are closed subspaces of X and invariant under T, we say that T is completely reduced by (M, N). The study of the spectrum of T can be reduced to the study of the spectra of T_M and T_N, respectively.

Let λ be an isolated eigenvalue of T such that there exist simple closed curves $\Gamma, \Gamma' \subset \rho(T)$ enclosing λ. Furthermore, both Γ and Γ' enclose no other eigenvalues of T. We define

$$P := \frac{1}{2\pi i} \int_\Gamma R(z)\, \mathrm{d}z. \tag{1.6}$$

It is clear that $P \in \mathcal{L}(X)$. Furthermore, we have that

$$\begin{aligned} P^2 &= \frac{1}{(2\pi i)^2} \int_\Gamma \int_{\Gamma'} R(z) R(z') \mathrm{d}z' \mathrm{d}z \\ &= \frac{1}{(2\pi i)^2} \int_\Gamma \int_{\Gamma'} \frac{R(z') - R(z)}{z' - z} \mathrm{d}z' \mathrm{d}z. \end{aligned}$$

Since, for $z \in \Gamma$ and $z' \in \Gamma'$,

$$\int_{\Gamma'} \frac{1}{z' - z} \mathrm{d}z' = 2\pi i$$

and

$$\int_\Gamma \frac{1}{z' - z} \mathrm{d}z = 2\pi i,$$

we obtain

$$P^2 = \frac{1}{2\pi i} \int_\Gamma R(z)\, \mathrm{d}z = P.$$

Thus P is a projection. In fact, P is the projection from X to the generalized eigenspace associated with λ when T is a compact operator (see Definition 1.1.31).

The eigenvalue problems we discuss in this book are closely related to compact operators. We summarize some properties of compact linear operators in the following from [178].

Definition 1.1.30. *Let X and Y be normed spaces. An operator $T : X \to Y$ is called a compact linear operator if T is linear and for every bounded subset M of X, $T(M)$ is relatively compact, i.e., $\overline{T(M)}$ is compact.*

We have the following criterion for compact operators.

Theorem 1.1.7. *Let X and Y be normed spaces and $T : X \to Y$ be a linear operator. Then T is compact if and only if for every bounded sequence $\{x_n\} \subset X$, $\{Tx_n\}$ has a convergent subsequence.*

Let $T \in \mathcal{L}(X, Y)$ and $S \in \mathcal{L}(Y, Z)$. If either T or S is compact, TS is compact from X to Z.

Lemma 1.1.8. *Let X and Y be normed spaces. Then*

(1) Every compact linear operator $T : X \to Y$ is bounded, hence continuous.

(2) If dim$X = \infty$, the identity operator $I : X \to X$ is not compact.

Theorem 1.1.9. *Let X and Y be normed spaces and $T : X \to Y$ be a linear operator. Then*

(1) If T is bounded and dim$T(X) < \infty$, T is compact.

(2) If dim$X < \infty$, T is compact.

Theorem 1.1.10. *Let $\{T_n : X \to Y\}$ be a sequence of compact operators. If $\{T_n\}$ is uniformly convergent, i.e., $\|T_n - T\| \to 0$, then the limit operator T is compact.*

Theorem 1.1.11. *Let $T : X \to Y$ be a linear operator. If T is compact, its adjoint operator $T' : Y' \to X'$ is compact.*

For compact operators, one has the so-called Fredholm Alternative (see [16]).

Theorem 1.1.12. *(Fredholm Alternative) Let X be a Banach space and $T : X \to X$ be compact. Then the equation*

$$(z - T)u = f, \quad z \neq 0$$

has a unique solution $u \in X$ for any $f \in Y$ if and only if the homogeneous equation

$$(z - T)u = 0$$

has only the trivial solution $u = 0$. In such a case, the operator $z - T$ has a bounded inverse.

Let $T : X \to X$ be a compact linear operator. The set of eigenvalues of T is at most countable and 0 is the only possible accumulation point. Every spectral value $\lambda \neq 0$ is an eigenvalue. If X is infinite dimensional, then $0 \in \sigma(T)$.

For an eigenvalue $\lambda \neq 0$, the dimension of any eigenspace of T is finite and the null spaces of $T_\lambda, T_\lambda^2, T_\lambda^3, \ldots$ are finite dimensional. There is a number r depending on $\lambda \neq 0$ such that

$$X = N(T_\lambda^r) \oplus T_\lambda^r(X).$$

Furthermore, the null spaces satisfy

$$N(T_\lambda^r) = N(T_\lambda^{r+1}) = N(T_\lambda^{r+2}) = \ldots$$

and the ranges satisfy

$$T_\lambda^r(X) = T_\lambda^{r+1}(X) = T_\lambda^{r+2}(X) = \ldots.$$

If $r > 0$, the following inclusions are proper

$$N(T_\lambda^0) \subset N(T_\lambda) \subset \ldots \subset N(T_\lambda^r)$$

and

$$T_\lambda^0(X) \supset T_\lambda(X) \supset \ldots \supset T_\lambda^r(X).$$

Definition 1.1.31. *The space $N(T_\lambda^r)$ is called the generalized eigenspace of T associated to the eigenvalue λ. The algebraic multiplicity of λ is defined as dim$N(T_\lambda^r)$. The geometric multiplicity is defined as dim$N(T_\lambda)$.*

Let $T : X \to X$ be a bounded self-adjoint operator on a complex Hilbert space X. Then

(1) all the eigenvalues of T (if they exist) are real,

(2) eigenfunctions corresponding to different eigenvalues of T are orthogonal with respect to the inner product on X,

(3) $\|T\| = \sup_{\|x\|=1} |(Tx, x)_X|$.

If, in addition, T is compact, we have the Hilbert–Schmidt theory (see, for example, Theorem 2.36 in [202]).

Theorem 1.1.13. *Let $T : X \to X$ be a compact, self-adjoint, linear operator on a Hilbert space X. Then there exist at most a countable set of real eigenvalues $\lambda_1, \lambda_2, \ldots$ and corresponding eigenfunctions $x_1, x_2, \ldots$ such that*

(1) $Tx_j = \lambda_j x_j$ and $x_j \neq 0, j = 1, 2, \ldots$,

(2) x_m is orthogonal to x_n if $m \neq n$,

(3) $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq 0$,

(4) if the sequence of eigenvalues is infinite, $\lim_{j\to\infty} \lambda_j = 0$,

(5) $Tx = \sum_{j\geq 1} \lambda_j (x, x_j)_X x_j$ with convergence in X when the sum has infinitely many terms,

(6) letting $W = span\{x_1, x_2, \ldots\}$, then $X = \overline{W} \oplus N(T)$.

1.2 Sobolev Spaces

The variational theory and convergence analysis of finite element methods relies on the notions of Sobolev spaces. In this section, we introduce the basic concepts and results to analyze the eigenvalue problems in this book. We refer the readers to Adams [3] for a complete treatment.

1.2.1 Basic Concepts

Let $\Omega \subset \mathbb{R}^n, n = 1, 2, 3$, be a Lipschitz domain which is defined as follows (Definition 3.1 of [202]).

Definition 1.2.1. *Let Ω be a bounded domain in $\mathbb{R}^n$ and denote its boundary by $\partial\Omega$. Ω is called a Lipschitz domain if $\partial\Omega$ is Lipschitz continuous, i.e., for every $x \in \partial\Omega$, there exists an open set $\mathcal{O} \subset \mathbb{R}^n$ with $x \in \mathcal{O}$ and an orthogonal coordinate system with coordinate $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ having the following properties. There is a vector $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{a} = (a_1, a_2, \ldots, a_n)$, with*

$$\mathcal{O} = \{\boldsymbol{\xi} \mid -a_j < \xi_j < a_j, 1 \le j \le n\}$$

and a Lipschitz continuous function ϕ defined on

$$\mathcal{O}' = \{\boldsymbol{\xi}' \in \mathbb{R}^{n-1} \mid -a_j < \xi'_j < a_j, 1 \le j \le n-1\}$$

with $|\phi(\boldsymbol{\xi}')| \le a_n/2$ for all $\boldsymbol{\xi}' \in \mathcal{O}'$ such that

$$\Omega \cap \mathcal{O} = \{\boldsymbol{\xi} \mid \xi_n < \phi(\boldsymbol{\xi}'), \boldsymbol{\xi}' \in \mathcal{O}'\}$$

and

$$\partial\Omega \cap \mathcal{O} = \{\boldsymbol{\xi} \mid \xi_n = \phi(\boldsymbol{\xi}'), \boldsymbol{\xi}' \in \mathcal{O}'\}.$$

In this book, we consider eigenvalue problems of partial different equations defined on Lipschitz domains. In particular, we restrict Ω to be either a Lipschitz polygon in $\mathbb{R}^2$ or a Lipschitz polyhedron in $\mathbb{R}^3$. We refer the readers to [198, 202] for more details and discussions on Lipschitz domains.

We need notations for several standard function spaces:

(1) $\mathcal{C}^k(\Omega)$: the set of k times continuously differentiable functions on Ω;

(2) $\mathcal{C}^k_0(\Omega)$: the set of k times continuously differentiable functions with compact support in Ω;

(3) $\mathcal{C}^k_0(\overline{\Omega})$: the set of k times continuously differentiable functions which have bounded and uniformly continuous derivatives up to order k with compact support in Ω;

(4) $\mathcal{C}_0^\infty(\Omega)$: the set of smooth function, i.e., infinite times continuously differentiable functions with compact support in Ω;

(5) $L^p(\Omega), 1 \leq p < \infty$: the set of functions such that $|\phi|^p$ is integrable on Ω, i.e.,

$$\int_\Omega |\phi|^p \,\mathrm{d}x < \infty.$$

When $p = 2$, we have $L^2(\Omega)$ equipped with the inner product

$$(u, v)_{L^2(\Omega)} = \int_\Omega uv \,\mathrm{d}x$$

and the induced norm $\|\cdot\|_{L^2(\Omega)}$. For simplicity, we use $\|\cdot\|$ instead of $\|\cdot\|_{L^2(\Omega)}$ when it leads to no confusion from the context.

Let $C(\overline{\Omega})$ be the set of bounded and continuous function $f : \Omega \to \mathbb{R}$ with the norm defined as

$$\|f\|_{C(\overline{\Omega})} := \sup_{x\in\Omega} |f(x)|.$$

We call f a Lipschitz continuous function on Ω if

$$|f(x) - f(y)| \leq C|x - y| \quad \text{for all } x, y \in \Omega \tag{1.7}$$

for some constant $C > 0$.

Let $0 < \gamma < 1$. A function f is said to be Hölder continuous with exponent γ if

$$|f(x) - f(y)| \leq C|x - y|^\gamma \quad \text{for all } x, y \in \Omega$$

for some constant $C > 0$. The γth Hölder semi-norm of f is defined as

$$|f|_{C^{0,\gamma}(\overline{\Omega})} = \sum_{x,y\in\Omega, x\neq y} \frac{|f(x) - f(y)|}{|x - y|^\gamma},$$

and the γth Hölder norm as

$$\|f\|_{C^{0,\gamma}(\overline{\Omega})} := \|f\|_{C(\overline{\Omega})} + |f|_{C^{0,\gamma}(\overline{\Omega})}.$$

The multi-index $\boldsymbol{\alpha}$ is defined as

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$$

with non-negative integer components $\alpha_i, i = 1, \ldots, n$. The order of $\boldsymbol{\alpha}$ is defined as

$$|\boldsymbol{\alpha}| = \sum_{i=1}^n \alpha_i.$$

For $f \in \mathcal{C}^k(\Omega)$, we define

$$\frac{\partial^{\boldsymbol{\alpha}} f}{\partial \mathbf{x}^{\boldsymbol{\alpha}}} = \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}.$$

Definition 1.2.2. *The Hölder space $C^{k,\gamma}(\overline{\Omega})$ consists of all functions $f \in C^k(\overline{\Omega})$ such that*

$$\|f\|_{C^{k,\gamma}(\overline{\Omega})} := \sum_{|\boldsymbol{\alpha}|\le k} \|\partial^{\boldsymbol{\alpha}} f\|_{C(\overline{\Omega})} + \sum_{|\boldsymbol{\alpha}|=k} |\partial^{\boldsymbol{\alpha}} f|_{C^{0,\gamma}(\overline{\Omega})} < \infty.$$

The Hölder space $C^{k,\gamma}(\overline{\Omega})$ is a Banach space.

Let s be a non-negative integer and $1 \le p < \infty$. The Sobolev spaces are defined as

$$W^{s,p}(\Omega) = \left\{ f \in L^p(\Omega) \,\middle|\, \partial^{\boldsymbol{\alpha}} f \in L^p(\Omega) \text{ for all } |\boldsymbol{\alpha}| \le s \right\}$$

associated with the norm

$$\|f\|_{W^{s,p}(\Omega)} = \left(\sum_{|\boldsymbol{\alpha}|\le s} \int_\Omega |\partial^{\boldsymbol{\alpha}} f(x)|^p \,\mathrm{d}x \right)^{1/p}.$$

The corresponding semi-norm is defined as

$$|f|_{W^{s,p}(\Omega)} = \left(\sum_{|\boldsymbol{\alpha}|= s} \int_\Omega |\partial^{\boldsymbol{\alpha}} \phi(\mathbf{x})|^p \,\mathrm{d}x \right)^{1/p}.$$

We denote by $W_0^{s,p}(\Omega)$ the closure of $C_0^\infty(\Omega)$ in the $W^{s,p}$ norm. When $p = 2$, we usually write

$$H^s(\Omega) = W^{s,2}(\Omega)$$

and

$$H_0^s(\Omega) = W_0^{s,2}(\Omega).$$

We write $\|\cdot\|_{W^{s,2}(\Omega)}$ as $\|\cdot\|_{H^s(\Omega)}$ or simply $\|\cdot\|_{H^s}$ if the domain is clear from context.

Definition 1.2.3. *If $W^{s,p}(\Omega)$ is a subset of space X and the identity map I from $W^{s,p}(\Omega)$ to X is continuous, we say $W^{s,p}(\Omega)$ is embedded in X. An embedding is compact if I is compact, written as $W^{s,p}(\Omega) \hookrightarrow X$.*

Compact embeddings play an important role in the analysis of eigenvalue problems of partial differential equations. The following theorem on compact embedding can be found in [3] (see also [202]).

Theorem 1.2.1. *Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain and let Ω_0 be any subdomain of Ω. Let Ω_0^l denote the intersection of Ω_0 with an l-dimensional hyperplane in $\mathbb{R}^n$. Let j, m be integers with $m \ge 1$ and $j \ge 0$ and let $p \in \mathbb{R}$ with $1 \le p < \infty$. Then the following embeddings are compact:*

(1) $mp \le n$: the embedding of $W^{j+m,p}(\Omega)$ in $W^{j,q}(\Omega_0^l)$ is compact if

$$0 < n - mp < l \le n \quad \textit{and} \quad 1 \le q < lp/(n - mp),$$

(2) $mp \leq n$: the embedding of $W^{j+m,p}(\Omega)$ in $W^{j,q}(\Omega_0^l)$ is compact if

$$mp = n,\ 1 \leq l \leq n \quad \textit{and} \quad 1 \leq q < \infty,$$

(3) $mp > n$: the embedding of $W^{j+m,p}(\Omega)$ in $C^j(\overline{\Omega}_0)$ is compact.

The Sobolev spaces of fractional order can be defined as follows. Let $s \geq 0$ and $1 \leq p < \infty$. Define $\lfloor s \rfloor$ to be the non-negative integer such that $s = \lfloor s \rfloor + \sigma$ for $0 < \sigma < 1$. Then $W^{s,p}(\Omega)$ is the space of distributions $u \in \mathcal{C}_0^\infty(\Omega)'$ such that $u \in W^{\lfloor s \rfloor,p}(\Omega)$ and

$$\int_\Omega \int_\Omega \frac{|\partial^{\boldsymbol{\alpha}} u(x) - \partial^{\boldsymbol{\alpha}} u(y)|^p}{|x-y|^{n+\sigma p}}\, \mathrm{d}x\mathrm{d}y < \infty \quad \text{for all } |\boldsymbol{\alpha}| = \lfloor s \rfloor,$$

facilitated with the norm

$$\|u\|_{W^{s,p}(\Omega)} = \left\{ \|u\|^p_{W^{\lfloor s \rfloor,p}(\Omega)} + \sum_{|\boldsymbol{\alpha}|=\lfloor s \rfloor} \int_\Omega \int_\Omega \frac{|\partial^{\boldsymbol{\alpha}} u(x) - \partial^{\boldsymbol{\alpha}} u(y)|^p}{|x-y|^{n+\sigma p}}\, \mathrm{d}x\mathrm{d}y \right\}^{1/p}.$$

The space $W^{s,p}(\Omega)$ is a separable, reflexive Banach space. The space $W_0^{s,p}(\Omega)$ is defined as the closure of $\mathcal{C}_0^\infty(\Omega)$ in $W^{s,p}(\Omega)$ with respect to the norm $\|\cdot\|_{W^{s,p}(\Omega)}$ and $H^s(\Omega) = W^{s,2}(\Omega), s \geq 0$. Furthermore, the following embedding theorem holds.

Theorem 1.2.2. *(Theorem 3.7 of [202]) Let Ω be a bounded Lipschitz domain. Then, if $0 \leq t < s$ such that $s - 3/p = t - 3/q$, the embedding of $W^{s,p}(\Omega)$ in $W^{t,q}(\Omega)$ holds. Furthermore, if $0 \leq t < s < \infty$ and $p = q = 2$, the embedding is compact.*

1.2.2 Negative Norm

The negative Sobolev norm [180] is useful to study the regularity of the solutions of partial differential equations. We present some results following [251].

Any $f \in L^2(\Omega)$ defines a continuous linear functional on $H_0^s(\Omega), s \geq 0$ by

$$f(u) := (f, u), \quad u \in H_0^s(\Omega).$$

The negative norm of f is defined as

$$\|f\|_{-s} = \sup_{u \in H_0^s(\Omega), \|u\|_{H^s(\Omega)} \leq 1} |f(u)| = \sup_{u \in H_0^s(\Omega), \|u\|_{H^s(\Omega)} \leq 1} |(f, u)|.$$

By Schwarz's inequality, we have

$$|(f, u)| \leq \|f\| \cdot \|u\| \leq \|f\| \cdot \|u\|_{H^s(\Omega)}.$$

Thus we immediately have that

$$\|f\|_{-s} \leq \|f\|.$$

We denote by $H^{-s}(\Omega)$ the completion of $L^2(\Omega)$ with respect to the negative norm $\|\cdot\|_{-s}$. Sobolev spaces of negative indices have the following property.

Theorem 1.2.3. *(Section III.10 of [251]) The dual space $H_0^s(\Omega)'$ of $H_0^s(\Omega)$ may be identified with the completion of $L^2(\Omega)$ with respect to the negative norm, i.e.,*

$$H_0^s(\Omega)' = H^{-s}(\Omega).$$

Furthermore, any continuous linear functional on $H^{-s}(\Omega)$ can be represented by an element in $H_0^s(\Omega)$, i.e.,

$$H^{-s}(\Omega)' = H_0^s(\Omega).$$

1.2.3 Trace Spaces

Eigenvalue problems of partial differential equations involve boundary conditions. We now discuss Sobolev spaces related to boundary values. Recalling the definition of the Lipschitz domain, $\partial\Omega$ is locally an $n-1$ dimensional hyper-surface in $\mathbb{R}^n$.

Definition 1.2.4. *Let $\phi, \boldsymbol{\xi}'$ be defined as in Definition 1.2.1 and $\boldsymbol{\phi}(\boldsymbol{\xi}') = (\boldsymbol{\xi}', \phi(\boldsymbol{\xi}'))$. Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain with boundary $\partial\Omega$. A distribution u defined on $\partial\Omega$ belongs to $W^{s,p}(\partial\Omega)$ for $|s| \leq 1$ if the composition*

$$u \circ \boldsymbol{\phi} \in W^{s,p}(\mathcal{O}' \cap \boldsymbol{\phi}^{-1}(\partial\Omega \cap \mathcal{O})).$$

If $u \in \mathcal{C}^\infty(\overline{\Omega})$, the restriction of u on $\partial\Omega$, called the trace operator, is defined as

$$\gamma_0(u) = u|_{\partial\Omega}. \tag{1.8}$$

The following theorem from [202] shows that γ_0 can be extended to certain Sobolev spaces.

Theorem 1.2.4. *Let Ω be a bounded Lipschitz domain and $1/p < s \leq 1$. The mapping γ_0 defined on $C^\infty(\overline{\Omega})$ has a unique continuous extension as a linear operator from $W^{s,p}(\Omega)$ onto $W^{s-1/p,p}(\partial\Omega)$. In addition,*

$$W_0^{1,p}(\Omega) = \{u \in W^{1,p}(\Omega) \,|\, \gamma_0(u) = 0\}.$$

When $p = 2$, we have $H^s(\partial\Omega) = W^{s,2}(\partial\Omega)$ for $0 \leq s \leq 1$. When $s = 1/2$, the trace space is given by $H^{1/2}(\partial\Omega) = W^{1/2,2}(\partial\Omega)$ which is important in the analysis of the Laplacian eigenvalue problem. For the biharmonic eigenvalue problem, we need $s > 1$. We define the normed space

$$H^s(\partial\Omega) = \left\{u \in L^2(\partial\Omega) \,|\, u = U|_{\partial\Omega} \text{ for some } U \in H^{s+1/2}(\Omega)\right\}$$

whose norm is defined as

$$\|u\|_{H^s(\partial\Omega)} = \inf_{U \in H^{s+1/2}(\Omega), u = U|_{\partial\Omega}} \|U\|_{H^{s+1/2}(\Omega)}.$$

The Poincaré-Friedrichs inequality is of fundamental importance for the well-posedness of many elliptic problems.

Definition 1.2.5. *Let $\Omega \subset \mathbb{R}^n$ be an open set with piecewise smooth boundary. We denote the completion of $C_0^\infty(\Omega)$ with respect to the Sobolev norm $\|\cdot\|_{H^m(\Omega)}$ by $H_0^m(\Omega)$.*

One has the following Poincaré-Friedrichs inequality (see [44]).

Theorem 1.2.5. *If Ω is bounded, then $|\cdot|_{H^m(\Omega)}$ is a norm on $H_0^m(\Omega)$, which is equivalent to $\|\cdot\|_{H^m(\Omega)}$. If Ω is contained in a cube with side length l, then*

$$|v|_{H^m(\Omega)} \leq \|v\|_{H^m(\Omega)} \leq (1+l)^m |v|_{H^m(\Omega)} \quad \textit{for all } v \in H_0^m(\Omega).$$

Taking $m = 1$, we have the Poincaré-Friedrichs inequality for $H_0^1(\Omega)$, i.e.,

$$|v|_{H^1(\Omega)} \leq \|v\|_{H^1(\Omega)} \leq (1+l)|v|_{H^1(\Omega)} \quad \text{for all } v \in H_0^1(\Omega).$$

1.3 Variational Formulation

The eigenvalue problems considered in this book are posed in the variational formulations of partial differential equations. The materials in this section are based on Section 2.2.3 of [202]. In the rest of this section, we restrict the discussion on Hilbert spaces.

Definition 1.3.1. *Let X and Y be Hilbert spaces. A mapping $a : X \times Y \to \mathbb{C}$ is called a sesquilinear form if*

$$a(\alpha_1 u + \alpha_2 v, \phi) = \alpha_1 a(u, \phi) + \alpha_2 (v, \phi) \quad \textit{for all } u, v \in X, \phi \in Y, \alpha_1, \alpha_2 \in \mathbb{C},$$
$$a(u, \alpha_1 \phi + \alpha_2 \psi) = \bar{\alpha}_1 a(u, \phi) + \bar{\alpha}_2 (u, \psi) \quad \textit{for all } u \in X, \phi, \psi \in Y, \alpha_1, \alpha_2 \in \mathbb{C},$$

where $\bar{\alpha}$ denotes the complex conjugation of α.

A simple example of sesquilinear forms is the inner product

$$a(u, \phi) := (u, \phi) = \int_\Omega u \bar{\phi} \, \mathrm{d}x$$

defined on $L^2(\Omega) \times L^2(\Omega)$.

A sesquilinear form is said to be bounded if there exists a constant C such that

$$|a(u, \phi)| \leq C \|u\|_X \|\phi\|_Y \quad \text{for all } u \in X, \phi \in Y. \tag{1.9}$$

The following property of sesquilinear forms on $X \times X$ is essential to the well-posedness of many problems.

Definition 1.3.2. *A sesquilinear form $a(\cdot, \cdot)$ on $X \times X$ is said to be coercive if there exists a constant $\alpha > 0$ satisfying*

$$a(u, u) \geq \alpha \|u\|_X^2 \quad \textit{for all } u \in X. \tag{1.10}$$

Let $a(\cdot,\cdot)$ be a bounded coercive sesquilinear form defined on $X \times X$. Given $f \in X'$, we consider a variationally posed problem of finding $u \in X$ such that

$$a(u, \phi) = f(\phi) \quad \text{for all } \phi \in X. \tag{1.11}$$

The well-posedness of the above problem follows the Lax-Milgram Lemma.

Lemma 1.3.1. *(Lax-Milgram Lemma) Let $a : X \times X \to \mathbb{C}$ be a bounded coercive sesquilinear form. There exists a unique solution $u \in X$ to* (1.11) *for $f \in X'$ satisfying*

$$\|u\|_X \leq \frac{C}{\alpha}\|f\|_{X'},$$

where C and α are the constants of boundedness and coercivity in (1.9) *and* (1.10)*, respectively.*

The Lax-Milgram Lemma has the following generalized form (Theorem 2.22 of [202]).

Theorem 1.3.2. *Let $a : X \times Y \to \mathbb{C}$ be a bounded sesquilinear form and have the following properties.*

(1) There exists a constant α such that

$$\inf_{u \in X, \|u\|_X = 1} \sup_{v \in Y, \|v\|_Y \leq 1} |a(u, v)| \geq \alpha > 0;$$

(2) For every $v \in Y$, $v \neq 0$,

$$\sup_{u \in X} |a(u, v)| > 0.$$

Suppose $g \in Y'$, then there exists a unique $u \in X$ such that

$$a(u, \phi) = g(\phi) \quad \text{for all } \phi \in Y.$$

Furthermore,

$$\|u\|_X \leq \frac{C}{\alpha}\|g\|_{Y'}.$$

To treat problems posed in mixed form, e.g., the Maxwell's eigenvalue problem, we need the following Babuška-Brezzi condition or inf-sup condition.

Let X and S be two Hilbert spaces and $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ be two bounded sesquilinear forms

$$a : X \times X \to \mathbb{C}$$

and

$$b : X \times S \to \mathbb{C}.$$

In general, $a(\cdot,\cdot)$ is not coercive on X. However, it is sufficient for some problems if $a(\cdot,\cdot)$ is coercive on a suitable subspace of X. Let

$$Z = \{u \in X \,|\, b(u, \xi) = 0 \quad \text{for all } \xi \in S\}.$$

Definition 1.3.3. *A sesquilinear form $a(\cdot,\cdot)$ is said to be Z-coercive if there exists a constant $\alpha > 0$ such that*

$$|a(u,u)| \geq \alpha \|u\|_X \quad \text{for all } u \in Z, \tag{1.12}$$

where α is independent of u.

The following condition is called the Babuška-Brezzi condition.

Definition 1.3.4. *A sesquilinear form $b(\cdot,\cdot)$ is said to satisfy the Babuška-Brezzi condition if there exists a constant $\beta > 0$ such that, for all $p \in S$,*

$$\sup_{w \in X} \frac{|b(w,p)|}{\|w\|_X} \geq \beta \|p\|_S, \tag{1.13}$$

where β is independent of p.

If the conditions (1.12) and (1.13) are satisfied, then the following well-posedness result holds (see, for example, Theorem 2.5 of [202]).

Theorem 1.3.3. *Let X and S be Hilbert spaces. Let $a : X \times X \to \mathbb{C}$ and $b : X \times S \to \mathbb{C}$ be bounded sesquilinear forms satisfying the Z-coercivity and Babuška-Brezzi condition, respectively. Suppose $f \in X'$ and $g \in S'$ and consider the problem of finding $u \in X$ and $p \in S$ such that*

$$a(u,\phi) + b(\phi,p) = f(\phi) \qquad \text{for all } \phi \in X, \tag{1.14a}$$

$$b(u,\xi) = g(\xi) \qquad \text{for all } \xi \in S. \tag{1.14b}$$

Then there exists a unique solution (u,p) to (1.14) *and*

$$\|u\|_X + \|p\|_S \leq C(\|f\|_{X'} + \|g\|_{S'}).$$

1.4 Abstract Spectral Approximation Theories

Let X be a complex Banach space with norm $\|\cdot\|$ and T be a compact operator on X. Let $\{X_h\}$ be a sequence of finite dimensional subspaces of X and $\{T_h : X_h \to X_h\}$ be a sequence of linear operators. In many cases, T_h is the restriction of an operator $B_h : X \to X_h$ on X_h. Let $\sigma(T)$ and $\rho(T)$ be the spectrum and the resolvent set of T, respectively. For $z \in \rho(T)$, we recall the resolvent operator is from X to X given by

$$R_z(T) = (z - T)^{-1}.$$

Similarly, we have $\sigma(T_h)$, $\rho(T_h)$, and $R_z(T_h) = (z - T_h)^{-1}$ for $z \in \rho(T_h)$.

Let Y and Z be closed subspaces of X. For $x \in X$, we define the distance from x to Y

$$d(x, Y) = \inf_{y \in Y} \|x - y\|$$

and the distance from Y to Z

$$d(Y, Z) = \sup_{y \in Y, \|y\|=1} d(y, Z).$$

The gap between Y and Z is defined as

$$\delta(Y, Z) = \max\{d(Y, Z), d(Z, Y)\}.$$

The following inequality is useful to prove the convergence of finite element approximation of eigenvalues and eigenfunctions.

Lemma 1.4.1. *If* $dimY = dimZ < \infty$*, then*

$$d(Y, Z) \leq d(Z, Y) \left[1 - d(Y, Z)\right]^{-1}. \tag{1.15}$$

Let Γ be a simple closed curve in $\rho(T) \cap \rho(T_h)$. The spectral projection E from X into X is defined as (see [165])

$$E := \frac{1}{2\pi i} \int_\Gamma R_z(T) \, \mathrm{d}z \tag{1.16}$$

and E_h from X_h to X_h is defined as

$$E_h := \frac{1}{2\pi i} \int_\Gamma R_z(T_h) \, \mathrm{d}z. \tag{1.17}$$

Note that if T_h converges to T as $h \to 0$, E_h is well defined for h small enough. The ranges $E(X) := R(E)$ and $E_h(X) := R(E_h)$ are invariant subspaces for T and T_h, respectively.

1.4.1 Theory of Descloux, Nassif, and Rappaz

We first present the abstract converge theory due to Descloux, Nassif, and Rappaz [114, 115]. A spectrally correct approximation T_h of T should have the following properties:

(1) for any compact set $K \subset \rho(T)$, there exists h_0 such that

$$K \subset \rho(T_h) \quad \text{for all } h < h_0; \tag{1.18}$$

(2) for all $z \in \sigma(T)$,

$$\lim_{h \to 0} d(z, \sigma(T_h)) = 0; \tag{1.19}$$

(3) for all $x \in E(X)$
$$\lim_{h\to 0} d(x, E_h(X_h)) = 0, \tag{1.20}$$
in particular, if $\mathring{\Gamma} \cap \sigma(T) \neq \emptyset$, then for h small enough, $\mathring{\Gamma} \cap \sigma(T_h) \neq \emptyset$, where $\mathring{\Gamma}$ is the interior of Γ;

(4) $\lim_{h\to 0} d(E_h(X_h), E(X)) = 0$;

(5) for h small enough, the sums of the algebraic multiplicities of the eigenvalues of T and T_h in Γ are the same.

Conditions (1) and (2) imply non-pollution and completeness of the spectrum, i.e., there are no discrete spurious eigenvalues and all eigenvalues are approximated correctly. Conditions (3), (4), and (5) imply non-pollution and completeness of the eigenspaces, i.e., there are no spurious eigenfunctions and the eigenspace approximation has the right dimension.

It is desirable to see what conditions are necessary for the above properties to hold. Define the h-norm of an operator T as
$$\|T\|_h = \sup_{x\in X_h, \|x\|=1} \|Tx\|.$$

Descloux, Nassif, and Rappaz list two conditions in [114]:

P1. $\lim_{h\to 0} \|T - T_h\|_h = 0$;

P2. for all $x \in X$, $\lim_{h\to 0} d(x, X_h) = 0$.

In the Banach case, i.e., X is a Banach space, they prove the following results.

Theorem 1.4.2. *Assume that condition P1 is satisfied.*

(a) Let $F \subset \rho(A)$ be closed. Then there exists a constant C independent of h such that
$$\|R_z(T_h)\|_h \leq C \quad \textit{for all } z \in F$$
provided h is small enough.

(b) Let $\Omega \subset \mathbb{C}$ be an open set such that $\sigma(T) \subset \Omega$. Then there exists $h_0 > 0$ such that
$$\sigma(T_h) \subset \Omega, \quad \textit{for all } h < h_0.$$

(c) One has that
$$\lim_{h\to 0} \|E - E_h\|_h = 0 \quad \textit{and} \quad \lim_{h\to 0} d(E_h(X_h), E(X)) = 0.$$

(d) If, in addition, we assume that P2 is also satisfied, we have that for all $x \in E(X)$
$$\lim_{h\to 0} d(x, E_h(X_h)) = 0. \tag{1.21}$$

We present the proof of the above theorem from [114] since the techniques will be used later.

Proof. (a) Let $z \in F \subset \rho(T)$. Then for any $x \in X$, there exists a constant $C > 0$ such that

$$\|(z - T)x\| \geq 2C\|x\|.$$

For h small enough, **P1** implies

$$\|(T - T_h)x\| \leq C\|x\| \quad \text{for all } x \in X_h.$$

Hence for $x \in X_h$ and $z \in F$, we have that

$$\|(z - T_h)x\| \geq \|(z - T)x\| - \|(T - T_h)x\| \geq C\|x\|.$$

Since X_h is finite dimensional, $R_z(T_h)$ exists and

$$\|R_z(T_h)\|_h \leq C.$$

(b) It is a direct consequence of (a).

(c) For h small enough, one has that

$$\begin{aligned}
\|E - E_h\|_h &\leq \frac{1}{2\pi}\int_\Gamma \|R_z(T) - R_z(T_h)\|_h |\mathrm{d}z| \\
&= \frac{1}{2\pi}\int_\Gamma \|R_z(T)(T - T_h)R_z(T_h)\|_h |\mathrm{d}z| \\
&= \frac{1}{2\pi}\int_\Gamma \|R_z(T)\| \cdot \|T - T_h\|_h \|R_z(T_h)\|_h |\mathrm{d}z|.
\end{aligned}$$

Combination of **P1** and (a) implies $\lim_{h\to 0} \|E - E_h\|_h = 0$. Then

$$\lim_{h\to 0} d(E_h(X_h), E(X)) = 0$$

follows immediately.

(d) Let $x \in E(X)$. From **P2**, we conclude that there exists a sequence $\{x_h \in X_h\}$ such that

$$\lim_{h\to 0} \|x - x_h\| = 0.$$

Thus we have that

$$\begin{aligned}
\|x - E_h x_h\| &= \|Ex - E_h x_h\| \\
&\leq \|E(x - x_h)\| + \|(E - E_h)x_h\| \\
&\leq \|E\| \cdot \|x - x_h\| + \|E - E_h\|_h \|x_h\|.
\end{aligned}$$

Since E is continuous, (1.21) follows (c). □

When $E(X)$ is finite dimensional, the above theorem implies that

$$\lim_{h\to 0} \delta(E(X), E_h(X_h)) = 0.$$

In addition, $\dim E_h(X_h) = \dim E(X)$ when h is small enough.

1.4.2 Theory of Babuška and Osborn

In this section, we introduce the abstract convergence theory due to Babuška and Osborn [23]. It plays a key role in the convergence analysis for several finite element methods for the Laplace eigenvalue problem, the biharmonic eigenvalue, and the Maxwell's eigenvalue problem. The materials presented here are taken from Sections 7 and 8 of [23].

We assume that T is a compact operator from X to X and $T_h, 0 < h \leq 1$, is a family of compact operators also from X to X. In addition, $T_h \to T$ in norm as $h \to 0$.

Let $\lambda \in \sigma(T)$, i.e., λ is an eigenvalue of T. Then there exists a smallest integer r, called the ascent of $\lambda I - T$, such that

$$N((\lambda I - T)^r) = N((\lambda I - T)^{r+1}).$$

Recall that the space $N((\lambda I - T)^r)$ is finite dimensional and its dimension $m = \dim N((\lambda I - T)^r)$ is called the algebraic multiplicity of λ. The geometric multiplicity n of λ is the dimension of $N(\lambda I - T)$, i.e., $n = \dim N(\lambda I - T)$. Obviously, we have that $n \leq m$. A vector u in $N((\lambda I - T)^r)$ is called a generalized eigenvector of T and its order is the smallest integer j such that $u \in N((\lambda I - T)^j)$.

If X is a Hilbert Space and T is self-adjoint, the ascent of $\lambda - T$ is one and the algebraic multiplicity equals the geometric multiplicity (see, e.g., Hilbert–Schmidt theory (Theorem 1.1.13)).

Since T_h converges to T in norm, E_h converges to E in norm and

$$\dim(E_h(X_h)) = \dim(E(X)) = m.$$

In addition, there exist exactly m eigenvalues of T_h inside Γ if h is small enough. We denote these values by $\lambda_{1,h}, \ldots, \lambda_{m,h}$. Consequently,

$$\lim_{h\to 0} \lambda_{j,h} \to \lambda \quad \text{as } h \to 0 \text{ for } j = 1, \ldots, m. \tag{1.22}$$

Next consider the adjoint operator T' on the dual space X'. If λ is an eigenvalue with algebraic multiplicity m, then λ is an eigenvalue of T' with the same algebraic multiplicity m. The ascent of $\lambda - T'$ is also r. Let E' be the projection operator associated with T' and λ and E'_h be the discrete projection operator associated with T'_h and $\lambda_{1,h}, \ldots, \lambda_{m,h}$. Note that when X is a Hilbert space, it is natural to work with the Hilbert adjoint T^*.

Now we are ready to present main results based on Babuška and Osborn [23]. We choose to include the proofs of some theorems in order to show how adjoint problems play the role in the theory.

Let λ be a nonzero eigenvalue of T with algebraic multiplicity m and ascent r. Let $\lambda_{1,h}, \ldots, \lambda_{m,h}$ be the eigenvalues of T_h that converge to λ. Let $\phi_1, \ldots, \phi_m$ be a basis for $R(E)$ and $\phi'_1, \ldots, \phi'_m$ be the dual basis to $\phi_1, \ldots, \phi_m$, i.e., a basis of $R(E)'$. The following theorem from [23] claims that $R(E)$ can be approximated by $R(E_h)$ correctly.

Theorem 1.4.3. *(Theorem 7.1 in [23]) There is a constant C independent of h such that, for h small enough,*

$$\delta(R(E), R(E_h) \leq C\|(T - T_h)|_{R(E)}\|,$$

where $(T - T_h)|_{R(E)}$ denotes the restriction of $T - T_h$ to $R(E)$.

Proof. Let $f \in R(E)$ such that $\|f\| = 1$. Since $Ef = f$,

$$\|f - E_h f\| = \|Ef - E_h f\| \leq \|E - E_h\|,$$

and thus

$$\lim_{h \to 0} \delta(R(E), R(E_h)) = 0.$$

For h small enough, $\delta((R(E), R(E_h)) \leq 1/2$. Using (1.15), we obtain

$$\delta(R(E_h) - R(E)) \leq 2\delta(R(E), R(E_h)),$$

which implies that

$$d(R(E), R(E_h)) \leq 2\delta(R(E), R(E_h)).$$

By the definition of spectral projection,

$$\begin{aligned} \|f - E_h f\| &= \left\| \frac{1}{2\pi i} \int_\Gamma [R_z(T) - R_z(T_h)] f dz \right\| \\ &= \left\| \frac{1}{2\pi i} \int_\Gamma R_z(T_h)[T - T_h] R_z(T) f dz \right\|. \end{aligned}$$

Hence one has

$$\|f - E_h f\| \leq \frac{1}{2\pi} |\Gamma| \sup_{z \in \Gamma} \|R_z(T_h)\| \|(T - T_h)|_{R(E)}\| \sup_{z \in \Gamma} \|R_z(T)\| \|f\|,$$

where $|\Gamma|$ denotes the length of Γ. The proof is complete by noting that $\sup_{z \in \Gamma} \|R_z(T_h)\|$ and $\sup_{z \in \Gamma} \|R_z(T)\|$ are bounded and setting

$$C = \frac{1}{2\pi} |\Gamma| \sup_{z \in \Gamma} \|R_z(T_h)\| \sup_{z \in \Gamma} \|R_z(T)\|.$$

□

Due to the fact of (1.22) we define the average of the discrete eigenvalues

$$\hat{\lambda}_h = \frac{1}{m} \sum_{j=1}^{m} \lambda_{j,h}.$$

The following theorem gives the convergence of $\hat{\lambda}_h$ to λ.

Theorem 1.4.4. *(Theorem 7.2 in [23]) Let $\phi_1, \ldots, \phi_m$ be a basis for $R(E)$ and $\phi'_1, \ldots, \phi'_m$ be the dual basis. Then there exists a constant C, independent of h, such that*

$$|\lambda - \hat{\lambda}_h| \leq \frac{1}{m} \sum_{j=1}^{m} |\langle (T - T_h)\phi_j, \phi'_j \rangle| + C \|(T - T_h)|_{R(E)}\| \|(T' - T'_h)|_{R(E)}\|.$$

Proof. Note that the operator $E_h|_{R(E)} : R(E) \to R(E_h)$ is injective since

$$\|E - E_h\| \to 0.$$

In addition, $E_h|_{R(E)} : R(E) \to R(E_h)$ is surjective since

$$\mathrm{dim} R(E) = \mathrm{dim} R(E_h) = m.$$

Hence $(E_h|_{R(E)})^{-1}$ is well defined. For h sufficiently small and $f \in R(E)$ with $\|f\| = 1$, we have that

$$1 - \|E_h f\| = \|Ef\| - \|E_h f\| \leq \|Ef - E_h f\| \leq \|E - E_h\| \|f\| \leq \frac{1}{2},$$

which implies $\|E_h f\| \geq \|f\|/2$. Hence $(E_h|_{R(E)})^{-1}$ is bounded in h.

We define

$$\hat{T}_h = (E_h|_{R(E)})^{-1} T_h E_h|_{R(E)} : R(E) \to R(E)$$

and

$$\hat{T} = T|_{R(E)}.$$

Note that $\lambda_{j,h}, j = 1, \ldots, m$, are eigenvalues of $\hat{T}_h$. We have that

$$\mathrm{trace} \hat{T} = m\lambda, \quad \mathrm{trace} \hat{T}_h = m\hat{\lambda}_h,$$

and

$$\lambda - \hat{\lambda}_h = \frac{1}{m} \mathrm{trace}(\hat{T} - \hat{T}_h).$$

Let $\phi_1, \ldots, \phi_m$ be a basis for $R(E)$ and let $\phi'_1, \ldots, \phi'_m$ be the dual basis to $\phi_1, \ldots, \phi_m$. We obtain

$$\lambda - \hat{\lambda}_h = \frac{1}{m} \mathrm{trace}(\hat{T} - \hat{T}_h) = \frac{1}{m} \sum_{j=1}^{m} \langle (\hat{T} - \hat{T}_h)\phi_j, \phi'_j \rangle. \tag{1.23}$$

Here $\phi'_j \in R(E)'$, the dual space of $R(E)$.

Note that ϕ'_j can be extended to X as follows. Since $X = R(E) \oplus N(E)$, for $f \in X$, we write $f = g + h$ with $g \in R(E)$ and $h \in N(E)$. Define

$$\langle f, \phi'_n \rangle = \langle g, \phi'_j \rangle.$$

It is clear that ϕ'_j on X is bounded and thus $\phi'_h \in X'$. Note that

$$\langle f, (\lambda - T')^\alpha \phi'_j \rangle = \langle (\lambda - T)^\alpha f, \phi'_j \rangle$$

vanishes for all f. Thus $\phi'_1, \ldots, \phi'_m \in R(E')$. Since

$$T_h E_h = E_h T_h \quad \text{and} \quad (E_h|_{R(E)})^{-1} E_h = I|_{R(E)},$$

one has that

$$\begin{aligned} & \langle (\hat{T} - \hat{T}_h)\phi_j, \phi'_j \rangle \\ = \; & \langle T\phi_j - (E_h|_{R(E)})^{-1} T_h E_h \phi_j, \phi'_j \rangle \\ = \; & \langle (E_h|_{R(E)})^{-1} E_h (T - T_h)\phi_j, \phi'_j \rangle \\ = \; & \langle (T - T_h)\phi_j, \phi'_j \rangle + \langle ((E_h|_{R(E)})^{-1} E_h - I)(T - T_h)\phi_j, \phi'_j \rangle. \end{aligned}$$

Let $L_h = (E_h|_{R(E)})^{-1} E_h$. L_h is the projection on $R(E)$ along $N(E_h)$. Then L'_h is the projection on $N(E_h)^\perp = R(E'_h)$ along $R(E)^\perp = N(E')$. Consequently,

$$\langle ((E_h|_{R(E)})^{-1} E_h - I)(T - T_h)\phi_j, \phi'_j \rangle = \langle (L_n - I)(T - T_h)\phi_j, (E' - E'_h)\phi'_j \rangle.$$

Thus the following holds

$$\begin{aligned} & \left| \langle ((E_h|_{R(E)})^{-1} E_h - I)(T - T_h)\phi_j, \phi'_j \rangle \right| \\ \le \; & \left(\sup_h \|L_h - I\| \right) \|(T - T_h)|_{R(E)}\| \|(E' - E'_h)|_{R(E)}\| \|\phi_h\| \|\phi'_j\| \\ \le \; & C \|(T - T_h)|_{R(E)}\| \|(E' - E'_h)|_{R(E)}\|. \end{aligned}$$

The proof is complete by combining the above results for (1.23). □

For a particular $\lambda_{j,h}$ with the ascent r, the following estimate is from [23] (Theorem 7.3 therein).

Theorem 1.4.5. *Let r be the ascent of $\lambda - T$ and $\phi_1, \ldots, \phi_m$ be any basis for $R(E)$ and $\phi'_1, \ldots, \phi'_m$ be the dual basis. Then there is a constant C such that*

$$\begin{aligned} & |\lambda - \lambda_{j,h}| \\ \le \; & C \left\{ \sum_{j,k=1}^{m} |\langle (T - T_h)\phi_i, \phi'_k \rangle| + \|(T - T_h)|_{R(E)}\| \|(T' - T'_h)|_{R(E')\|} \right\}^{1/r}. \end{aligned}$$

Proof. Let $\lambda_{j,h}$ be an eigenvalue of $\hat{T}_h$ and $\hat{T}_h w_h = \lambda_{j,h} w_h$, $\|w_h\| = 1$. Choose $w'_h \in N((\lambda - T')^r)$ such that $\langle w_h, w'_h \rangle = 1$ and the norms $\|w'_h\|$ are bounded in h.

By the Hahn-Banach theorem, let $w'_h \in R(E)'$ such that $\langle w_h, w'_h \rangle = 1$ and

$\|w_h'\| = 1$. Extend w_h' to all of X. Hence $w_h' \in R(E')$ and $\|w_h'\| \le \|E\|$. Noting that $(T' - \lambda)^r w_h' = 0$, we obtain

$$\begin{aligned}
& |\lambda - \lambda_h(h)|^r \\
= \quad & |\langle (\lambda - \lambda_j(h))^r w_h, w_h' \rangle| \\
= \quad & |\langle ((\lambda - \lambda_{j,h})^r - (\lambda - T)^r) w_h, w_h' \rangle| \\
= \quad & \left| \left\langle \sum_{k=0}^{r-1} (\lambda - \lambda_{j,h})^k (\lambda - T)^{r-1-k} (\lambda_{j,h} - T) w_h, w_h' \right\rangle \right| \\
\le \quad & \sum_{k=0}^{r-1} |\lambda - \lambda_{j,h}|^k |\langle (\lambda_{j,h} - T) w_h, (\lambda - T')^{r-1-k} w_h' \rangle| \\
\le \quad & \sum_{k=0}^{r-1} |\lambda - \lambda_{j,h}|^k \max_{\phi' \in R(E'), \|\phi'\|=1} |\langle (\lambda_{j,h} - T) w_h, \phi' \rangle| \\
& \qquad \cdot \|\lambda - T'\|^{r-1-k} \|w_h'\|. \qquad (1.24)
\end{aligned}$$

For any $\phi' \in R(E')$ with $\|\phi'\| = 1$,

$$\begin{aligned}
& |\langle (\lambda_{j,h} - T) w_h, \phi' \rangle| \\
= \quad & |\langle (\hat{T} - T) w_h, \phi' \rangle| \\
= \quad & |\langle E_h^{-1} E_h (T_h - T) w_h, \phi' \rangle| \\
= \quad & |\langle (T - T_h) w_h, \phi' \rangle + \langle (E_h^{-1} E_h - I)(T - T_h) w_h, \phi' \rangle| \\
= \quad & |\langle (T - T_h) w_h, \phi' \rangle| + C \|(T - T_h)|_{R(E)}\| \|(T' - T_h')|_{R(E')}\|. \qquad (1.25)
\end{aligned}$$

There exists a constant C' such that

$$|\langle (T - T_h) w_h, \phi' \rangle| \le C' \sum_{i,k=1}^{m} |\langle (T_h - T) \phi_i, \phi_k' \rangle| \qquad (1.26)$$

for all $w_h \in R(E)$ and $\phi' \in R(E')$ with $\|w_h\| = \|\phi'\| = 1$. Then (1.24), (1.25), and (1.26) prove the theorem. □

Theorem 1.4.6. *(Theorem 7.4 in [23]) Let λ_h be an eigenvalue of T_h such that $\lim_{h \to 0} \lambda_h = \lambda$. Suppose for each h that w_h is a unit vector satisfying*

$$(\lambda_h - T_h)^k w_h = 0$$

for some positive integer $k \le r$. Then, for any integer l with $k \le l \le r$, there is a vector u_h such that $(\lambda - T)^l u_h = 0$ and

$$\|u_h - w_h\| \le C \|(T - T_h)|_{R(E)}\|^{(l-k+1)/r}. \qquad (1.27)$$

Proof. Since $N((\lambda - T)^l)$ is finite-dimensional, there exists a closed subspace M of X such that

$$X = N((\lambda - T)^l) \oplus M.$$

For $y \in R((\lambda - T)^l)$, the equation $(\lambda - T)^l x = y$ is uniquely solvable in M. Thus

$$(\lambda - T)^l|_M : M \to R((\lambda - T)^l)$$

is one-to-one and onto. Hence

$$(\lambda - T)^l|_M^{-1} : R((\lambda - T)^l) \to M$$

exists and, by the closed graph theorem, is bounded. Thus there is a constant C such that

$$\|f\| \le C\|(\lambda - T)^l f\| \quad \text{for all } f \in M.$$

Set $u_h = Pw_h$, where P is the projection on $N((\lambda - T)^l)$ along M. Then $(\lambda - T)^l u_h = 0$ and $w_h - u_h \in M$, and hence

$$\|w_h - u_h\| \le C\|(\lambda - T)^l (w_h - u_h)\|.$$

By Theorem 1.4.3 there are vectors $\tilde{u}_h \in R(E)$ such that

$$\|w_h - \tilde{u}_h\| \le C'\|(T - T_h)|_{R(E)}\|.$$

Hence there is a constant C_2 such that

$$\begin{aligned}
&\|[(\lambda - T)^l - (\lambda - T_h)^l]w_h\| \\
= &\left\| \sum_{j=0}^{l-1} (\lambda - T_h)^j (T - T_h)(\lambda - T)^{l-j-1}[(w_h - \tilde{u}_h) + \tilde{u}_h] \right\| \\
\le &\; C_2\|(T - T_h)|_{R(E)}\|.
\end{aligned}$$

Since $k \le l$,

$$\begin{aligned}
\|(\lambda - T_h)^l w_h\| &= \left\| \sum_{j=0}^{l-1} \binom{l}{j} (\lambda - \lambda_h)^j (\lambda_h - T_h)^{l-j} w_h \right\| \\
&= \left\| \sum_{j=l-k+1}^{l} \binom{l}{j} (\lambda - \lambda_h)^j (\lambda_h - T_h)^{l-j} w_h \right\| \\
&\le C_3|\lambda - \lambda_h|^{l-k+1}.
\end{aligned}$$

Combining the above equations, we obtain

$$\begin{aligned}
\|w_h - u_h\| &\le C\|(\lambda - T)^l (w_h - u_h)\| \\
&\le C\|(\lambda - T)^l w_h\| \\
&= C\|[(\lambda - T)^l - (\lambda - T_h)^l]w_h + (\lambda - T_h)^l w_h\| \\
&\le C[C_2\|(T - T_h)|_{R(E)}\| + C_3|\lambda - \lambda_h|^{l-k+1}].
\end{aligned}$$

Application of Theorem 1.4.5 completes the proof. $\square$

When X is a Hilbert space and T, T_h are self-adjoint, one actually has that, for $j = 1, \ldots, m$,

$$|\lambda - \lambda_{j,h}| \leq C \left\{ \sum_{i,j=1}^{m} |\langle (T - T_h)\phi_i, \phi_j^* \rangle| + \|(T - T_h)|_{R(E)}\|^2 \right\}. \tag{1.28}$$

1.4.3 Variationally Formulated Eigenvalue Problems

Now we consider the variationally formulated eigenvalue problems. The material in this section is based on Section 8 of [23]. Let H_1 and H_2 be complex Hilbert spaces and $a(\cdot, \cdot)$ be a sesquilinear form on $H_1 \times H_2$ such that

$$|a(u, v)| \leq C\|u\|_1 \|v\|_2 \quad \text{for all } u \in H_1, v \in H_2,$$

where $\|\cdot\|_1$ is the induced norm by the inner product $(\cdot, \cdot)_1$ on H_1 and $\|\cdot\|_2$ is the induced norm by the inner product $(\cdot, \cdot)_2$ on H_2. Furthermore, we assume that

$$\inf_{u \in H_1, \|u\|_1 = 1} \sup_{v \in H_2, \|v\|_2 = 1} |a(u, v)| = \delta > 0$$

and

$$\sup_{v \in H_2} |a(u, v)| > 0 \quad \text{for all } 0 \neq u \in H_1.$$

Let $\|\cdot\|_1'$ be a second norm on H_1 such that every bounded sequence in $\|\cdot\|_1$ norm has a convergent subsequence in $\|\cdot\|_1'$. We say $\|\cdot\|_1'$ is compact with respect to $\|\cdot\|_1$ norm. For example, if $H_1 = H^1(\Omega)$, the L^2 norm is compact with respect to the H^1-norm. Let $b(u, v)$ be a bilinear form on $H_1 \times H_2$ such that

$$|b(u, v)| \leq C_2 \|u\|_1' \|v\|_2 \quad \text{for all } u \in H_1, v \in H_2.$$

For many variationally posed eigenvalue problems, the form $b(u, v)$ is defined on $H_1 \times H_2$ such that H_1 and H_2 are compactly embedded in some spaces W_1 and W_2, respectively, and

$$|b(u, v)| \leq C_2 \|u\|_{W_1} \|v\|_{W_2} \quad \text{for all } u \in W_1, v \in W_2.$$

It can be shown that there exist bounded compact operators (solution operators) $T : H_1 \to H_2$ satisfying

$$a(Tu, v) = b(u, v) \quad \text{for all } u \in H_1, v \in H_2$$

and $T_* : H_2 \to H_2$ satisfying

$$a(u, T_* v) = b(u, v) \quad \text{for all } u \in H_1, v \in H_2.$$

A complex number λ is called an eigenvalue of $a(\cdot, \cdot)$ with respect to $b(\cdot, \cdot)$ if there exists a nonzero $u \in H_1$ such that

$$a(u, v) = \lambda b(u, v) \quad \text{for all } v \in H_2. \tag{1.29}$$

Obviously, (λ, u) is an eigenpair if and only if $\lambda T u = u$.

Remark 1.4.1. *Here we use λ again. It should be noted that it corresponds to $1/\lambda$ in previous sections.*

Next we consider the discrete approximation of (1.29). To this end, let $S_{1,h} \subset H_1$ and $S_{2,h} \subset H_2$ be two finite dimensional spaces which satisfy the inf-sup condition

$$\inf_{u\in S_{1,h},\|u\|=1} \sup_{v\in S_{2,h},\|v\|=1} |a(u,v)| \geq \beta = \beta(h) > 0$$

and

$$\sup_{u\in S_{1,h}} |a(u,v)| > 0 \quad \text{for all } v \in S_{2,h}, v \neq 0.$$

We also assume that for any $u \in H_1$,

$$\lim_{h\to 0} \beta(h)^{-1} \inf_{w\in S_{1,h}} \|u-w\| = 0.$$

Then the discrete form is to find λ_h and $u_h \in S_{1,h}, u_h \neq 0$, such that

$$a(u_h, v) = \lambda_h b(u_h, v) \quad \text{for all } v \in S_{2,h}. \tag{1.30}$$

We define the discrete solution operator $T_h : H_1 \to S_{1,h}$ such that

$$a(T_h u, v) = b(u,v) \quad \text{for all } u \in H_1, v \in S_{2,h}.$$

Thus (λ_h, u_h) is an eigenpair of (1.30) if and only if (λ_h^{-1}, u_h) is an eigenpair of T_h.

A generalized eigenvector u^j is said to be of order $j > 1$ corresponding to λ if and only if

$$a(u^j, v) = \lambda b(u^j, v) + \lambda a(u^{j-1}, v) \quad \text{for all } v \in H_2,$$

where u^{j-1} is a generalized eigenvector of order $j-1$.

Let λ be an eigenvalue of (1.29) with algebraic multiplicity m. Let r be the ascent of $\lambda^{-1} - T$. If $T_h \to T$ in norm, m eigenvalues $\lambda_{1,h}, \ldots, \lambda_{m,h}$ converge to λ. We define

$$\begin{aligned}
M(\lambda) &= \{u : u \text{ is a generalized eigenvector of (1.29)}, \|u\|_1 = 1\},\\
M^*(\lambda) &= \{v : v \text{ is a generalized adjoint eigenvector of (1.29)}, \|v\|_2 = 1\},\\
M_h(\lambda) &= \{u : u \in \text{span}\{u_{1,h}, \ldots, u_{m,h}\}, \|u\|_1 = 1\},
\end{aligned}$$

and

$$\begin{aligned}
\epsilon_{h,\lambda} &= \sup_{u\in M(\lambda)} \inf_{\phi\in S_{1,h}} \|u-\phi\|_1,\\
\epsilon^*_{h,\lambda} &= \sup_{v\in M*(\lambda)} \inf_{\psi\in S_{2,h}} \|v-\psi\|_2.
\end{aligned}$$

Theorem 1.4.7. *Let $\overline{M}(\lambda) = R(E)$ and $\overline{M}_h(\lambda) = R(E_h)$. There are constants C_1, C_2, C_3 such that*

$$\begin{aligned} \delta(\overline{M}(\lambda), \overline{M}_h(\lambda)) &\leq C_1\beta(h)^{-1}\epsilon_{h,\lambda}, \\ \left| \lambda - \left(\frac{1}{m}\sum_{j=1}^{m} \lambda_{j,h}^{-1} \right)^{-1} \right| &\leq C_2\beta(h)^{-1}\epsilon_{h,\lambda}\epsilon^*_{h,\lambda}, \\ |\lambda - \lambda_{j,h}| &\leq C_3\left(\beta(h)^{-1}\epsilon_{h,\lambda}\epsilon^*_{h,\lambda}\right)^{1/r}. \end{aligned}$$

For eigenvectors, we have the following result.

Theorem 1.4.8. *Let λ_h be an eigenvalue of* (1.30) *such that* $\lim_{h\to 0}\lambda_h = \lambda$. *Suppose for each h that w_h is a unit vector satisfying $(\lambda_h^{-1} - T)^k w_h = 0$ for some positive integer $k \leq r$. Then, for any integer j with $k \leq j \leq r$, there is a vector u_h such that $(\lambda^{-1} - T)^j u_h = 0$ and*

$$\|u - u_h\|_1 \leq C(\beta(h)^{-1}\epsilon_{h,\lambda})^{(l-k+1)/r}.$$

We devote the rest of this section to some discussion of the Ritz method for self-adjoint positive definite eigenvalue problems. Let H be a Hilbert space and $a(\cdot,\cdot)$ be a symmetric bilinear form on H such that

$$a(u,u) \geq \alpha\|u\|^2 \quad \text{for all } u \in H,$$

where α is a positive constant. The energy norm $\|\cdot\|_a$, which is equivalent to the usual norm on H, is defined as

$$\|u\|_a^2 = a(u,u) \quad \text{for all } u \in H.$$

Therefore, T is self-adjoint and positive definite.

Let $\{S_h \subset H, h > 0\}$ be a family of finite element spaces approximating H. Then (1.30) is called the Ritz method. The problem (1.29) has a countable sequence of eigenvalues

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

with $+\infty$ as the limit point. It can be chosen that the corresponding eigenvectors $u_1, u_2, \ldots$ satisfy

$$a(u_i, u_j) = \lambda_j b(u_i, u_j) = \delta_{i,j}.$$

We define the Rayleigh quotient as

$$R(u) = \frac{a(u,u)}{b(u,u)}.$$

The following results hold.

(1) Minimum principle:

$$\begin{aligned} \lambda_1 &= \min_{u\in H} R(u) = R(u_1), \\ \lambda_k &= \min_{u\in H, a(u,u_i)=0, i=1,\ldots,k} R(u) = R(u_k), k = 2, 3, \ldots. \end{aligned}$$

(2) Minimum-maximum principle:

$$\begin{aligned}\lambda_k &= \min_{V_K \subset H, \dim V_k = k} \max_{u \in V_k} R(u) \\ &= \max_{u \in \text{span}\{u_1, \dots, u_k\}} R(u), k = 1, 2, \dots.\end{aligned}$$

(3) Maximum-minimum principle:

$$\begin{aligned}\lambda_k &= \max_{z_1, \dots, z_{k-1}} \min_{u \in H, a(u, z_i) = 0, i = 1, \dots, k-1} R(u) \\ &= \min_{u \in H, a(u, u_i) = 0, i = 1, \dots, k-1} R(u), k = 1, 2, \dots.\end{aligned}$$

Similar results hold for the discrete problem (1.30). An important observation is that

$$\lambda_k \le \lambda_{k,h}, \quad k = 1, \dots, N, N = \dim S_h,$$

which explains conforming finite element methods for positive definite self-adjoint problems always approximate eigenvalues from above.

Assume that λ_k has geometric multiplicity n. Let $E = E(\lambda_k^{-1})$ be the orthogonal projection of H onto span$\{u_k, \dots, u_{k+n-1}\}$ and $E_h = E_h(\lambda_k^{-1})$ be the orthogonal projection of H onto span$\{u_{k,h}, \dots, u_{k+n-1,h}\}$. Then we have the following estimates:

$$\begin{aligned}&\|u - E_h u\|_1 = r_h^{(a)} \inf_{\phi \in S_h} \|u - \phi\|_a, \quad \text{for all } u \in \text{span}\{u_k, \dots, u_{k+n-1}\}, \\ &\|u_{j,h} - E u_{j,h}\|_1 = r_h^{(b)} \inf_{\phi \in S_h} \|E u_{j,h} - \phi\|_a, \quad j = k, \dots, k+n-1, \\ &(\lambda_{j,h} - \lambda_k)/\lambda_k = r_h^{(c)} \inf_{\phi \in S_h} \|E u_{j,h} - \phi\|_a^2, \quad j = k, \dots, k+n-1,\end{aligned}$$

where $r_h^{(a)}, r_h^{(b)}, r_h^{(c)} \to 0$ as $h \to 0$. Let

$$\eta(h) = \sup_{b(u,u)=1} \inf_{\phi \in S_h} \|Tu - \phi\|_a.$$

Then the following estimate holds:

$$|r_h^{(l)} - 1| \le C\eta^2(h), \quad l = a, b, c.$$

For more discussion of the results in this section, we refer the readers to [23] and references therein.

Chapter 2

Finite Elements

2.1 Introduction

In this chapter, we introduce fundamental concepts of finite elements. There are ample excellent books, for example, [88, 54, 44]. We try to give a concise self-contained introduction which suffices the consequent discussions for the eigenvalue problems.

We start with the definition of finite elements following [88].

Definition 2.1.1. *A finite element is a triple* $(K, \mathcal{P}, \mathcal{N})$ *such that*

(1) $K \subset \mathbb{R}^n$ *is a geometric domain (e.g., triangle, tetrahedron),*

(2) $\mathcal{P}$ *is a space of functions (e.g., polynomials) on* K*,*

(3) $\mathcal{N} = \{N_1, \dots, N_s\}$ *is a set of linear functionals on* $\mathcal{P}$*, called degrees of freedom.*

The finite element $(K, \mathcal{P}, \mathcal{N})$ is said to be unisolvent if the degrees of freedom of $\mathcal{N}$ uniquely determine a function in $\mathcal{P}$.

Definition 2.1.2. *Let* $(K, \mathcal{P}, \mathcal{N})$ *be a finite element. The basis* $\{\phi_1, \phi_2, \dots, \phi_s\}$ *of* $\mathcal{P}$ *dual to* $\mathcal{N}$ *(i.e.,* $N_i(\phi_j) = \delta_{ij}$*) is called the nodal basis of* $\mathcal{P}$*.*

Given a finite element $(K, \mathcal{P}, \mathcal{N})$, let v be a function such that $N_i(v), i = 1, \ldots, s$, are well defined. The local interpolant is defined as

$$I_K v := \sum_{i=1}^{s} N_i(v)\phi_i. \tag{2.1}$$

Let $\mathcal{T}$ be a subdivision for Ω, e.g., a triangular mesh in two dimensions. For $f \in C^m(\overline{\Omega})$, the global interpolant is denoted by $I_h f$ such that

$$I_h f|_K = I_K f \tag{2.2}$$

for each $K \in \mathcal{T}$.

2.1.1 Meshes

We assume that Ω is partitioned into a collection of simple geometric domains. To focus on the eigenvalue problems other than finite elements, we will mainly consider triangles in two dimensions and tetrahedra in three dimensions. There are many other alternative choices such as quadrilaterals in two dimensions and prisms in three dimensions. We refer the readers to [88, 54, 44, 202].

We start with some definitions of meshes following [44].

Definition 2.1.3. *(1) A partition $\mathcal{T} = \{K_1, \ldots, K_M\}$ of Ω into triangle (tetrahedron) elements is called admissible provided the following properties hold*

(i) $\overline{\Omega} = \cup_i^M K_i$,

(ii) If $K_i \cap K_j$ consists of exactly one point, then it is a common vertex of K_i and K_j,

(iii) If for $i \neq j$, $K_i \cap K_j$ consists of a line segment, then $K_i \cap K_j$ is a common edge of K_i and K_j,

(iv) If for $i \neq j$, $K_i \cap K_j$ consists of a triangle, then $K_i \cap K_j$ is a common face of K_i and K_j.

(2) We write $\mathcal{T}_h, h > 0$, implying every element has diameter at most $2h$.

(3) A family of partitions $\{\mathcal{T}_h\}$ is called shape regular provided there exists a number $\kappa > 0$ such that every K in $\mathcal{T}_h$ contains a circle of radius ρ_K with $\rho_K \geq h_K/\kappa$ where h_K is half the diameter of K (K contains a ball of radius ρ_K in three dimensions).

(4) A family of partitions $\{\mathcal{T}_h\}$ is called uniform provided that there exists a number $\kappa > 0$ such that every element K in $\mathcal{T}_h$ contains a circle with radius $\rho_K \geq h/\kappa$ (a ball of radius ρ_K in three dimensions).

(5) A family of partitions $\{\mathcal{T}_h\}$ is called quasi-uniform if there exists $\tau > 0$ such that

$$\min\{d_K : K \in \mathcal{T}_h\} \geq \tau d_\Omega,$$

where d_K is the diameter of the largest ball contained in K and d_Ω is the diameter of Ω.

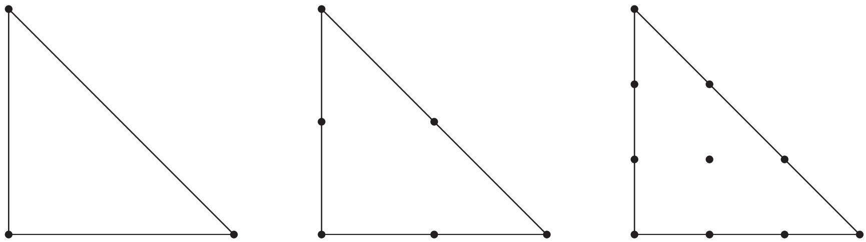


Figure 2.1: Left: Linear Lagrange element. Middle: Quadratic Lagrange element. Right: Cubic Lagrange element.

Let $\hat{K}$ be the reference element, i.e., the triangle whose vertices are $(0,0)$, $(1,0)$, and $(0,1)$ in $\mathbb{R}^2$, or the tetrahedron whose vertices are $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$. For any $K \in \mathcal{T}$, there is an affine mapping $F_K : \hat{K} \to K$ such that $F(\hat{K}) = K$ given by

$$F_K \hat{\mathbf{x}} = B_K \hat{\mathbf{x}} + \hat{\mathbf{b}}. \tag{2.3}$$

The reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ is affine equivalent to the finite element $(K, \mathcal{P}, \mathcal{N})$ if the following hold

1. $F_K(\hat{K}) = K$,
2. $F_K \circ \hat{\mathcal{P}} = \mathcal{P}$,
3. $\mathcal{N} \circ F_K = \hat{\mathcal{N}}$.

One reason to introduce the reference element is to simplify the implementation. Affine equivalence would allow us to build the local matrices on the reference element and transform them to the actual elements.

2.1.2 Lagrange Elements

Let K be a triangle in $\mathcal{T}$ and $\mathcal{P}_k := \mathcal{P}_k(K)$ denote the set of all polynomials of degree at most k. The dimension of $\mathcal{P}_k$ is

$$s = \dim(\mathcal{P}_k) = \frac{(k+1)(k+2)}{2}.$$

Let $z_1, z_2, \ldots, z_s$ be s points in K which lie on $k+1$ lines. The values on these points of a polynomial $p \in \mathcal{P}_k$, i.e., $p(z_1), \ldots, p(z_s)$, uniquely determine p. The set of functions in $\mathcal{P}_k$ which takes a nonzero value at exactly one point forms a basis of of $\mathcal{P}_k$, called the nodal basis.

When $k = 1$, we have $s = 3$. $\mathcal{P}_1$ contains linear polynomials. Let z_1, z_2, z_3 be the vertices of K and $\mathcal{N}_1 = \{N_1, N_2, N_3\}$ such that $N_i(v) = v(z_i)$ (see Fig. 2.1). It is easy to show that $\mathcal{N}_1$ determines $\mathcal{P}_1$. In particular, when $z_1 = (0,0)$, $z_2 = (1,0)$, and

$z_3 = (0,1)$ (the vertices of the reference triangle), we have the linear basis functions for $\mathcal{P}_1$:

$$L_1 = 1 - x - y, \quad L_2 = x, \quad L_3 = y,$$

such that $N_i(L_j) = \delta_{i,j}, i, j = 1, 2, 3$.

When $k = 2$, we have $\dim(\mathcal{P}_2) = 6$. In addition to z_1, z_2, z_3, let z_4, z_5, z_6 be the middle points of the edges $\overline{z_1 z_2}$, $\overline{z_1 z_3}$, $\overline{z_2 z_3}$, respectively. Let $\mathcal{N}_2 = \{N_1, \ldots, N_6\}$ such that $N_i(v) = v(z_i), v \in \mathcal{P}_2$. $\mathcal{N}_2$ determines $\mathcal{P}_2$. On the reference triangle, $N_i(L_j) = \delta_{ij}, i, j = 1, \ldots, 6$, give quadratic basis functions for $\mathcal{P}_2$.

For $k > 2$,

$$\mathcal{N}_k = \left\{N_1, \ldots, N_{\frac{(k+1)(k+2)}{2}}\right\}$$

and the evaluation points are

(1) 3 vertex nodes,

(2) $3(k-1)$ distinct edge nodes,

(3) $\frac{1}{2}(k-2)(k-1)$ interior points arranged as in Fig.2.1.

Definition 2.1.4. *Given a finite element $(K, \mathcal{P}, \mathcal{N})$, let the set $\{\phi_i\}$ be the nodal basis for $\mathcal{P}$ dual to $\mathcal{N}$. If v is a function for which all $N_i \in \mathcal{N}$ are defined, the local interpolant on K is given by*

$$\mathcal{I}_K v := \sum_{i=1}^{dim(\mathcal{P})} N_i(v)\phi_i.$$

The global interpolant on Ω is given by

$$\mathcal{I}_\mathcal{T}|_{K_i} = \mathcal{I}_{K_i} f.$$

The following theorem guarantees the unique interpolation polynomial using the Lagrange element (Remark 5.4 from [44]).

Theorem 2.1.1. *Let $k \geq 0$ and K be a triangle. Suppose $z_1, \ldots, z_s$ are the $s = (k+1)(k+2)/2$ interpolation points for Lagrange elements (see Fig. 2.1). Then for every continuous function $f \in C(K)$, there is a unique polynomial p of degree up to k satisfying the interpolation condition*

$$p(z_i) = f(z_i), \quad i = 1, 2, \ldots, s.$$

Proof. The theorem can be proved by induction. The result is trivial when $k = 0$. We assume that it holds for $k - 1$. Without loss of generality, we assume that one edge of K lies on the x-axis and it contains the points $z_1, \ldots, z_{k+1}$. Then there is a univariate polynomial $p_0(x)$ with

$$p_0(z_i) = f(z_i), \quad i = 1, 2, \ldots, k+1.$$

By induction, there exists a polynomial $q(x, y)$ of degree $k-1$ with

$$q(z_i) = \frac{1}{y_i}[f(z_i) - p_0(z_i)], \quad i = k+2, \dots, s.$$

The proof is complete. $\square$

To define the conforming elements, we need the follow theorem from [44].

Theorem 2.1.2. *Let $k \geq 1$ and Ω is bounded. Then a piecewise infinitely differentiable function $v : \overline{\Omega} \to \mathbb{R}$ belongs to $H^k(\Omega)$ if and only if $v \in C^{k-1}(\overline{\Omega})$.*

The finite elements above using nodal values are obviously continuous, i.e., the functions in

$$V_h := \{v \in L^2(\mathcal{T}), v\,|_K \in \mathcal{P}_k \text{ for every } K \in \mathcal{T}\}$$

are continuous. Thus $V_h \subset H^1(\Omega)$ and we call the finite element space H^1-conforming.

2.2 Quadrature Rules

To assemble finite element matrices, one needs quadrature rules to integrate functions over certain domains such as line segment, triangle, tetrahedron, etc. In general, a quadrature is stated as a weighted sum of function values at specified points (quadrature points). In this section, we present some commonly used quadrature rules for line segment, triangle, and tetrahedron.

2.2.1 Gaussian Quadratures

We present the n-point Gaussian quadrature rule, named after Carl Friedrich Gauss, which is exact for polynomials of degree $2n-1$ or less by a suitable choice of the quadrature points x_i and weights w_i for $i = 1, ..., n$. Taking the integration domain as $[-1, 1]$, the rule is stated as

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \sum_{i=1}^{n} w_i f(x_i).$$

The quadrature points x_i are just the roots of Legendre polynomials, $P_n(x)$, which can be expressed using Rodrigues' formula

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}[(x^2-1)^n].$$

With the nth order polynomial normalized to give $P_n(1) = 1$, the ith Gaussian node, x_i, is the ith root of $P_n(x)$. Its weight is given by [1]

$$w_i = \frac{2}{(1-x_i)^2(P_n'(x_i))^2}.$$

n	Points x_i	Weights w_i
1	0	2
2	$\pm\sqrt{1/3}$	1
3	$0, \pm\sqrt{3/5}$	$8/9, 5/9$
4	$\pm\sqrt{(3-2\sqrt{6/5})/7}, \pm\sqrt{(3-2\sqrt{6/5})/7}$	$\frac{18+\sqrt{30}}{36}, \frac{18-\sqrt{30}}{36}$
5	$0, \pm\frac{1}{3}\sqrt{5-2\sqrt{10/7}}, \pm\frac{1}{3}\sqrt{5-2\sqrt{10/7}}$	$\frac{128}{255}, \frac{322+13\sqrt{70}}{900}, \frac{322-13\sqrt{70}}{900}$

Table 2.1: Some low-order Gaussian quadratures on $[-1, 1]$ which are accurate for polynomials up to order $2n-1$. The weights are the same for the quadrature points with a "$\pm$" sign.

For integral over an arbitrary line segment $[a, b]$, a simple change of variable shows that

$$\int_a^b f(x)\mathrm{d}x = \frac{b-a}{2}\int_{-1}^1 f\left(\frac{b-a}{2}z + \frac{a+b}{2}\right)\mathrm{d}z.$$

A Gaussian quadrature provides the approximation:

$$\int_a^b f(x)\mathrm{d}x \approx \frac{b-a}{2}\sum_{i=1}^n w_i f\left(\frac{b-a}{2}z_i + \frac{a+b}{2}\right).$$

When the line segment is in $\mathbb{R}^n, n > 1$, a parametric representation of the line segment with parameter from $[-1, 1]$ suffices.

2.2.2 Quadratures for a Triangle

The integral over a triangle is approximated by the following quadrature rule

$$\int_{\hat{K}} f(x)\mathrm{d}x \approx \frac{1}{2}\sum_{i=1}^q w_i f(x_i), \tag{2.4}$$

where w_i's are weights and x_i's are quadrature points for the reference triangle $\hat{K}$. If (2.4) is exact for $p \in \mathcal{P}_k(\hat{K})$, then the interpolation error can be estimated as the following:

$$\left|\int_{\hat{K}} f(x)\mathrm{d}x - \sum_{j=1}^q w_i f(y_i)\right| \leq Ch^{k+1}\sum_{|\boldsymbol{\alpha}|=k+1}\int_{\hat{K}} |D^{\boldsymbol{\alpha}} f|\mathrm{d}x. \tag{2.5}$$

We would like to have quadrature rules on a triangle which is exact for polynomials of order up to k. When k is small, say $k \leq 3$, it is simple to find quadrature rules which are efficient.

Let $a^i, i = 1, 2, 3$, be the vertices of the triangle K and a^{ij} be the midpoint of the edge $\overline{a^i a^j}$ where $1 \leq i < j \leq 3$. Let a^{123} be the barycenter of K. Let $|K|$ denote the area of K. The following are quadrature rules which are exact for polynomials of order up to $1, 2, 3$, respectively.

(1) $k = 1$:

$$\int_K f(x)\mathrm{d}x \approx |K| f(a^{123}).$$

(2) $k = 2$:

$$\int_K f(x)\mathrm{d}x \approx \sum_{1 \le i < j \le 3} f(a^{ij}) \frac{|K|}{3}.$$

(3) $k = 3$:

$$\int_K f(x)\mathrm{d}x \approx \sum_{i=1}^{3} f(a^i) \frac{|K|}{20} + \sum_{1 \le i < j \le 3} f(a^{ij}) \frac{2|K|}{15} + f(a^{123}) \frac{9|K|}{20}.$$

Development of higher order efficient quadrature rules is not as straightforward. In Table 2.2, we give the symmetric Gaussian quadrature rules from [116], which are exact for polynomials of order up to 5. The readers can find quadrature rules for polynomial of higher order up to 20 in [116].

2.2.3 Quadrature Rules for Tetrahedra

For three-dimensional problems, we need quadrature rules for a tetrahedron. Again, we use the reference tetrahedron $\hat{K}$ whose vertices are $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$. In the following, we list the quadrature points and weights from [166]. See also:
people.sc.fsu.edu/~jburkardt/datasets/quadrature_rules_tet/quadrature_rules_tet.html

---

2.3 Abstract Convergence Theory

The abstract finite element convergence theory is critical to the error analysis for eigenvalue problems. We present some fundamentals here and put more technical results to pertinent chapters later. The materials in this section are classical and can be found in many finite element books, e.g., [88, 44, 54, 16, 202]. The presentation closely follows Section 2.3 of [202].

2.3.1 Céa's Lemma

Lemma 2.3.1. *Let $\{X_h\}$, $h > 0$, be a family of finite dimensional subspaces of a Hilbert space X. Suppose the sesquilinear form $a : X \times X \to \mathbb{C}$ is bounded and coercive. Let $f \in X'$. Then the problem of finding $u_h \in X_h$ such that*

$$a(u_h, v_h) = f(v_h) \quad \textit{for all } v_h \in X_h \tag{2.6}$$

k	Points x_i	Weights w_i
1	(0.33333333333333, 0.33333333333333)	1.00000000000000
2	(0.16666666666667, 0.16666666666667)	0.33333333333333
	(0.16666666666667, 0.66666666666667)	0.33333333333333
	(0.66666666666667, 0.16666666666667)	0.33333333333333
3	(0.33333333333333, 0.33333333333333)	-0.56250000000000
	(0.20000000000000, 0.20000000000000)	0.52083333333333
	(0.20000000000000, 0.60000000000000)	0.52083333333333
	(0.60000000000000, 0.20000000000000)	0.52083333333333
4	(0.44594849091597, 0.44594849091597)	0.22338158967801
	(0.44594849091597, 0.10810301816807)	0.22338158967801
	(0.10810301816807, 0.44594849091597)	0.22338158967801
	(0.09157621350977, 0.09157621350977)	0.10995174365532
	(0.09157621350977, 0.81684757298046)	0.10995174365532
	(0.81684757298046, 0.09157621350977)	0.10995174365532
5	(0.33333333333333, 0.33333333333333)	0.22500000000000
	(0.47014206410511, 0.47014206410511)	0.13239415278851
	(0.47014206410511, 0.05971587178977)	0.13239415278851
	(0.05971587178977, 0.47014206410511)	0.13239415278851
	(0.10128650732346, 0.10128650732346)	0.12593918054483
	(0.10128650732346, 0.79742698535309)	0.12593918054483
	(0.79742698535309, 0.10128650732346)	0.12593918054483

Table 2.2: Symmetric Gaussian quadratures on the reference triangle $\hat{K}$ which are accurate for polynomials up to degree k. $k = 1$: 1 point, $k = 2$: 3 points, $k = 3$: 4 points, $k = 4$: 6 points, $k = 5$: 7 points.

has a unique solution. In addition, if u is the exact solution of finding $u \in X$ such that

$$a(u, v) = f(v) \quad \textit{for all } v \in X, \tag{2.7}$$

then there is a constant C independent of u and u_h such that

$$\|u - u_h\|_X \leq C \inf_{v_h \in X_h} \|u - v_h\|_X. \tag{2.8}$$

Proof. Since $X_h \subset X$ and the sesquilinear form $a : X_h \times X_h \to \mathbb{C}$ is bounded and coercive, then the first part of the theorem follows directly from the Lax-Milgram Lemma 1.3.1.

From (2.7) and (2.6), the Galerkin orthogonality holds, i.e.,

$$a(u - u_h, v_h) = 0 \quad \text{for all } v_h \in X_h,$$

which implies that

$$a(u - u_h, u_h - v_h) = 0 \quad \text{for all } v_h \in X_h.$$

k	Points x_i	Weights w_i
0	(0.250000000000 0.250000000000 0.250000000000)	1.000000000000
1	(0.585410196624 0.138196601125 0.138196601125)	0.250000000000
	(0.138196601125 0.138196601125 0.138196601125)	0.250000000000
	(0.138196601125 0.138196601125 0.585410196625)	0.250000000000
	(0.138196601125 0.585410196625 0.138196601125)	0.250000000000
2	(0.250000000000 0.250000000000 0.250000000000)	-0.800000000000
	(0.500000000000 0.166666666667 0.166666666667)	0.450000000000
	(0.166666666667 0.166666666667 0.166666666667)	0.450000000000
	(0.166666666667 0.166666666667 0.500000000000)	0.450000000000
	(0.166666666667 0.500000000000 0.166666666667)	0.450000000000
3	(0.568430584197 0.143856471934 0.143856471934)	0.217765069880
	(0.143856471934 0.143856471934 0.143856471934)	0.217765069880
	(0.143856471934 0.143856471934 0.568430584197)	0.217765069880
	(0.143856471934 0.568430584197 0.143856471934)	0.217765069880
	(0.000000000000 0.500000000000 0.500000000000)	0.217765069880
	(0.500000000000 0.000000000000 0.500000000000)	0.021489953413
	(0.500000000000 0.500000000000 0.000000000000)	0.021489953413
	(0.500000000000 0.000000000000 0.000000000000)	0.021489953413
	(0.000000000000 0.500000000000 0.000000000000)	0.021489953413
	(0.000000000000 0.000000000000 0.500000000000)	0.021489953413
4	(0.250000000000 0.250000000000 0.250000000000)	-0.078933333333
	(0.785714285714 0.071428571429 0.071428571429)	0.045733333333
	(0.071428571429 0.071428571429 0.071428571429)	0.045733333333
	(0.071428571429 0.071428571429 0.785714285714)	0.045733333333
	(0.071428571429 0.785714285714 0.071428571429)	0.045733333333
	(0.100596423833 0.399403576169 0.399403576169)	0.149333333333
	(0.399403576167 0.100596423833 0.399403576169)	0.149333333333
	(0.399403576167 0.399403576169 0.100596423833)	0.149333333333
	(0.399403576167 0.100596423833 0.100596423833)	0.149333333333
	(0.100596423833 0.399403576167 0.100596423833)	0.149333333333
	(0.100596423833 0.100596423833 0.399403576167)	0.149333333333

Table 2.3: Quadrature rules for the reference tetrahedron $\hat{K}$ which are accurate for polynomials up to degree k. $k = 0$: 1 point, $k = 1$: 4 points, $k = 2$: 5 points, $k = 3$: 10 points, $k = 4$: 11 points.

Employing the boundedness and coercivity of $a(\cdot, \cdot)$, we have that

$$\begin{aligned}
\alpha \|u - u_h\|_X^2 &\leq |a(u - u_h, u - u_h)| \\
&= a(u - u_h, u - v_h) + a(u - u_h, u_h - v_h) \\
&= a(u - u_h, u - v_h) \\
&\leq C \|u - u_h\|_X \|u - v_h\|_X
\end{aligned}$$

and (2.8) follows immediately. □

The error estimate (2.8) is called quasi-optimal since the actual error is bounded by the multiplication of the best approximation and a constant C. An optimal error estimate has $C = 1$.

2.3.2 Discrete Mixed Problems

Let $X_h \subset X$ and $S_h \subset S$. We consider the conforming discrete mixed formulation to find $u_h \in X_h$ and $p_h \in S_h$ such that

$$a(u_h, \phi_h) + b(\phi_h, p_h) = f(\phi_h) \qquad \text{for all } \phi_h \in X_h, \tag{2.9a}$$

$$b(u_h, \xi_h) = g(\xi_h) \qquad \text{for all } \xi_h \in S_h, \tag{2.9b}$$

where $f \in X'$ and $g \in S'$. Similar to the continuous case, we define a space

$$Z_h = \{u_h \in X_h |\, b(u_h, \xi_h) = 0 \quad \text{for all } \xi_h \in S_h\}. \tag{2.10}$$

We assume that $a(\cdot, \cdot)$ is coercive on Z_h, i.e., there exists a constant $\alpha > 0$ independent of h such that

$$|a(u_h, u_h)| \geq \alpha \|u_h\|_X^2 \quad \text{for all } u_h \in Z_h. \tag{2.11}$$

Furthermore, we assume that the discrete Babuška-Brezzi condition holds, i.e., there exists a constant $\beta > 0$ independent of h and p_h such that

$$\sup_{\phi_h \in X_h} \frac{|b(\phi_h, p_h)|}{\|\phi_h\|_X} \geq \beta \|p_h\|_S. \tag{2.12}$$

The following theorem gives the existence and uniqueness of a solution for (2.9).

Theorem 2.3.2. *(Theorem 2.39 of [202]) Assume that $a : X \times X \to \mathbb{C}$ and $b : X \times S \to \mathbb{C}$ are bounded sesquilinear forms satisfying the discrete coercivity condition* (2.11) *and the discrete Babuška-Brezzi condition* (2.12)*, respectively. Then provided the space*

$$Z_h(g) = \{u_h \in X_h | b(u_h, \xi_h) = g(\xi_h) \quad \textit{for all } \xi_h \in S_h\} \tag{2.13}$$

is not empty, there exists a unique solution to (2.9).

Proof. Let $u_h^0 \in Z_h(g)$ and write $u_h = u_h^0 + u_h^1$ with $u_h^1 \in Z_h$. Substituting u_h in (2.9a), we have that

$$a(u_h^0 + u_h^1, \phi_h) + b(\phi_h, p_h) = f(\phi_h) \quad \text{for all } \phi_h \in X_h.$$

If $\phi_h \in Z_h$, i.e., $b(\phi_h, p_h) = 0$, we obtain

$$a(u_h^1, \phi_h) = f(\phi_h) - a(u_h^0, \phi_h) \quad \text{for all } \phi_h \in X_h. \tag{2.14}$$

By the Z_h-coercivity of $a(\cdot,\cdot)$ and the Lax-Milgram Lemma 1.3.1, there exists a unique solution $u_h^1 \in Z_h$ to (2.14).

Next we consider the problem of finding $p_h \in S_h$ such that

$$b(\phi_h, p_h) = -a(u_h, \phi_h) + f(\phi_h) \quad \text{for all } \phi_h \in X_h. \tag{2.15}$$

Let $X_h = Z_h \oplus Z_h^{\perp}$. If $\phi_h \in Z_h$, $b(\phi_h, p_h) = 0$ and

$$-a(u_h, \phi_h) + f(\phi_h) = -a(u_h, \phi_h) - b(\phi_h, p_h) + f(\phi_h) = 0,$$

i.e., the equation is trivial. Thus we only need to find $p_h \in S_h$ such that

$$b(\phi_h, p_h) = -a(u_h, \phi_h) + f(\phi_h) \quad \text{for all } \phi_h \in Z_h^{\perp}. \tag{2.16}$$

By the discrete Babuška-Brezzi condition

$$\sup_{\phi_h \in Z_h^{\perp}} \frac{|b(\phi_h, q_h)|}{\|\phi_h\|_X} \geq \alpha \|q_h\|_S$$

and

$$\sup_{q_h \in S_h} |b(\phi_h, q_h)| > 0 \quad \text{for } \phi_h \in Z_h^{\perp},$$

there exists a unique solution p_h to (2.16) due to the generalized Lax-Milgram Lemma (Theorem 1.3.2).

To show the uniqueness, we set $f = g = 0$. We see that $u_h \in Z_h$ since $g = 0$. Letting $\phi_h = u_h$ and $\xi_h = p_h$, one gets $a(u_h, u_h) = 0$. Since $a(\cdot,\cdot)$ is Z_h-coercive, $u_h = 0$. Furthermore, $b(\phi_h, p_h) = 0$ for all $\phi_h \in X_h$. The discrete Babuška-Brezzi condition (2.12) implies that $p_h = 0$. The uniqueness is verified. $\square$

We have the well-posedness of both the continuous and discrete problems (Theorems 1.3.3 and 2.3.2). We will move on to prove the error estimates.

Lemma 2.3.3. *Suppose the bounded sesquilinear form $b : X \times Y \to \mathbb{C}$ satisfies the discrete Babuška-Brezzi condition* (2.12). *Then for any function $v \in X$ there exists a unique function $v_h \in Z_h^{\perp}$ such that*

$$b(v - v_h, \phi_h) = 0 \quad \textit{for all } \phi_h \in S_h.$$

Furthermore,

$$\|v_h\|_X \leq \frac{C}{\alpha}\|v\|_X.$$

Proof. The problem can be written as follows. For $v \in X$, find $v_h \in Z_h^{\perp}$ such that

$$b(v_h, \phi_h) = b(v, \phi_h) \quad \text{for all } \phi_h \in S_h.$$

The lemma holds by the generalized Lax-Milgram Lemma (Theorem 1.3.2) since $b(\cdot,\cdot)$ satisfies the discrete Babuška-Brezzi condition (2.12) and for $v_h \in Z_h^{\perp}$ we have that

$$\sup_{q_h \in S_h} |b(\phi_h, q_h)| > 0 \quad \text{for } \phi_h \in Z_h^{\perp}.$$

$\square$

The following theorem provides the estimate for $u - u_h$.

Theorem 2.3.4. *Suppose that $b : X \times S \to \mathbb{C}$ is bounded and $a : X \times X \to \mathbb{C}$ is bounded and Z_h-coercive. Let (u, p) be the unique solution of the continuous problem* (1.14) *and (u_h, p_h) be the unique solution of the discrete problem* (2.9). *Then the following estimate holds*

$$\|u - u_h\|_X \leq C \left\{ \inf_{v_h \in Z_h(g)} \|u - v_h\|_X + \inf_{q_h \in S_h} \|p - q_h\|_S \right\} \tag{2.17}$$

for some constant C independent of h.

Proof. Let $v_h \in Z_h(g)$. Using the triangle inequality, Z_h-coercivity, and the boundedness of $a(\cdot, \cdot)$, we have that

$$\begin{aligned}
& \|u - u_h\|_X \\
\leq\; & \|u - v_h\|_X + \|v_h - u_h\|_X \\
\leq\; & \|u - v_h\|_X + \frac{1}{\alpha} \frac{a(v_h - u_h, v_h - u_h)}{\|v_h - u_h\|_X} \\
\leq\; & \|u - v_h\|_X + \frac{1}{\alpha} \sup_{w_h \in Z_h} \frac{a(v_h - u_h, w_h)}{\|w_h\|_X} \\
\leq\; & \|u - v_h\|_X + \frac{1}{\alpha} \left\{ \sup_{w_h \in Z_h} \frac{a(v_h - u, w_h)}{\|w_h\|_X} + \sup_{w_h \in Z_h} \frac{a(u - u_h, w_h)}{\|w_h\|_X} \right\} \\
\leq\; & \left(1 + \frac{C}{\alpha}\right) \|u - v_h\|_X + \frac{1}{\alpha} \sup_{w_h \in Z_h} \frac{a(u - u_h, w_h)}{\|w_h\|_X}.
\end{aligned}$$

Using the fact that $w_h \in Z_h$, we derive the following

$$\begin{aligned}
|a(u - u_h, w_h)| &= |a(u, w_h) - a(u_h, w_h)| \\
&= |a(u, w_h) - f(w_h)| \\
&= |-b(w_h, p)| \\
&= |-b(w_h, p - q_h)| \\
&\leq C\|w_h\|\|p - q_h\|_S
\end{aligned}$$

for all $q_h \in S_h$. Then (2.17) follows immediately since v_h and q_h can be any element in $Z_h(g)$ and S_h, respectively. □

Note that we do not need the discrete Babuška-Brezzi condition in the above proof. However, we do need it for the estimate of $\|p - p_h\|_S$.

Theorem 2.3.5. *Suppose that $b : X \times S \to \mathbb{C}$ is bounded and $a : X \times X \to \mathbb{C}$ is bounded and Z_h-coercive. In addition $b(\cdot, \cdot)$ satisfies the discrete Babuška-Brezzi condition* (2.12). *Let (u, p) be the unique solution of the continuous problem* (1.14) *and (u_h, p_h) be the unique solution of the discrete problem* (2.9). *Then the following estimate holds*

$$\|p - p_h\|_S \leq \frac{C}{\beta} \|u - u_h\|_X + \left(1 + \frac{C}{\beta}\right) \inf_{q_h \in S_h} \|p - q_h\|_S. \tag{2.18}$$

Proof. Setting $\phi = \phi_h$ in (1.14a), it holds that

$$a(u, \phi_h) + b(\phi_h, p) = f(\phi_h) \quad \text{for all } \phi_h \in X_h.$$

Subtracting (2.9a) from the above equation, we obtain

$$b(\phi_h, p - p_h) = -a(u - u_h, \phi_h) \quad \text{for all } \phi_h \in X_h.$$

By the discrete Babuška-Brezzi condition (2.12) and the boundedness of $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$,

$$\begin{aligned}
\beta \|q_h - p_h\|_S &\leq \sup_{\phi_h \in X_h} \frac{|b(\phi_h, q_h - p_h)|}{\|\phi_h\|_X} \\
&= \sup_{\phi_h \in X_h} \frac{|b(\phi_h, p - p_h) + b(\phi_h, q_h - p)|}{\|\phi_h\|_X} \\
&= \sup_{\phi_h \in X_h} \frac{|-a(u - u_h, \phi_h) + b(\phi_h, q_h - p)|}{\|\phi_h\|_X} \\
&\leq C\left(\|u - u_h\|_X + \|q_h - p\|_S\right).
\end{aligned}$$

The proof is complete by noting that

$$\|p - p_h\|_S \leq \|p - q_h\|_S + \|q_h - p_h\|_S.$$

□

The next theorem summarizes the above error estimates.

Theorem 2.3.6. *Assume that* $(u, p) \in X \times S$ *is the unique solution satisfying* (1.14) *and* $(u_h, p_h) \in X_h \times S_h$ *is the unique solution satisfying* (2.9) *such that the continuous and discrete coercivity conditions (*(1.12) *and* (2.11)*) and the continuous and discrete Babuška-Brezzi conditions (*(1.13) *and* (2.12)*) are satisfied. The following error estimate holds*

$$\|u - u_h\|_X + \|p - p_h\|_S \leq C \left\{ \inf_{v_h \in X_h} \|u - v_h\|_X + \inf_{q_h \in S_h} \|p - q_h\|_S \right\} \tag{2.19}$$

for some constant C *independent of* h.

Proof. For Theorems 2.3.4 and 2.3.5, we obtain

$$\|u - u_h\|_X + \|p - p_h\|_S \leq C \left\{ \inf_{v_h \in Z_h(g)} \|u - v_h\|_X + \inf_{q_h \in S_h} \|p - q_h\|_S \right\}. \tag{2.20}$$

For any $v_h \in X_h$, let $w_h \in Z_h(g)^T$ such that

$$b(w_h, q_h) = b(u - v_h, q_h) \quad \text{for all } q_h \in S_h.$$

The existence and uniqueness of w_h follows Lemma 2.3.3, which leads to

$$b(w_h + v_h, q_h) = b(u, q_h) = g(q_h) \quad \text{for all } w_h \in S_h.$$

This implies that $w_h + v_h \in Z_h(g)$. Furthermore,

$$\begin{aligned}\|u-(v_h+w_h)\|_X &\leq \|u-v_h\|_X + \|w_h\|_X \\ &\leq \left(1+\frac{C}{\alpha}\right)\|u-v_h\|_X.\end{aligned}$$

Hence

$$\inf_{v_h \in Z_h(g)} \|u-v_h\|_X \leq \left(1+\frac{C}{\alpha}\right)\|u-v_h\|_X. \tag{2.21}$$

The error estimate (2.19) follows readily by inserting (2.21) in (2.20). □

2.3.3 Inverse Estimates

Let K be a bounded domain. Let v be a function defined on K and define $\hat{v}$ as

$$\hat{v}(\hat{x}) = v((\text{diam}K)\hat{x}) \quad \text{for all } \hat{x} \in \hat{K},$$

where $\hat{K} = \{(1/\text{diam}K)x | x \in K\}$ and diam$K$ is the diameter of K. It is obvious that $v \in W_r^k(K)$ if and only if $\hat{v} \in W_r^k(\hat{K})$ and

$$|\hat{v}|_{W_r^k(\hat{K})} = (\text{diam}K)^{k-(n/r)}|v|_{W_r^k(K)}. \tag{2.22}$$

Let $\mathcal{P}$ be a vector space of functions defined on K and define $\hat{\mathcal{P}} := \{\hat{v} | v \in \mathcal{P}\}$. The following theorem is from [54].

Theorem 2.3.7. *Let $\rho h \leq diamK \leq h$, where $0 < h \leq 1$, and $\mathcal{P}$ be a finite dimensional subspace of $W_p^l(K) \cap W_q^m(K)$, where $1 \leq p \leq \infty$, $1 \leq q \leq \infty$, and $0 \leq m \leq l$. Then there exists $C = C(\hat{\mathcal{P}}, \hat{K}, l, p, q, \rho)$ such that for all $v \in \mathcal{P}$, we have that*

$$\|v\|_{W_p^l(K)} \leq Ch^{m-l+n/p-n/q}\|v\|_{W_q^m(K)}. \tag{2.23}$$

Proof. We first consider the case of $m = 0$. For any finite-dimensional space $\mathcal{P}$, we have that

$$\|\hat{v}\|_{W_p^l(\hat{K})} \leq C\|\hat{v}\|_{L^q(\hat{K})} \quad \text{for all } v \in \mathcal{P}.$$

Then (2.22) implies that

$$|v|_{W_p^j(K)}(\text{diam}K)^{j-n/p} \leq C\|v\|_{L^q(K)}(\text{diam}K)^{-n/q}, \quad 0 \leq j \leq l.$$

Thus one has that

$$|v|_{W_p^j(K)} \leq Ch^{-j+n/p-n/q}\|v\|_{L^q(K)}, \quad 0 \leq j \leq l.$$

Since $h \leq 1$, taking $j = l$, we obtain

$$\|v\|_{W_p^l(K)} \leq Ch^{-l+n/p-n/q}\|v\|_{W_q^m(K)}. \tag{2.24}$$

We assume $0 < m \le l$. For $l-m \le k \le l$ and $|\boldsymbol{\alpha}| = k$, let $D^{\boldsymbol{\alpha}} v = D^{\boldsymbol{\beta}} D^{\boldsymbol{\gamma}} v$ for $|\boldsymbol{\beta}| = l-m$ and $|\boldsymbol{\gamma}| = k+m-l$:

$$\begin{aligned} \|D^{\boldsymbol{\alpha}} v\|_{L^p(K)} &\le \|D^{\boldsymbol{\gamma}}\|_{W_p^{l-m}(K)} \\ &\le Ch^{-(l-m)+n/p-n/q} \|D^{\boldsymbol{\gamma}} v\|_{L^q(K)} \\ &\le Ch^{-(l-m)+n/p-n/q} |v|_{W_p^{k+m-l}(K)}. \end{aligned}$$

Note that $|\boldsymbol{\alpha}| = k$ is arbitrary. For any k such that $l-m \le k \le l$, we have that

$$|v|_{W_p^k(K)} \le Ch^{-(l-m)+n/p-n/q} |v|_{W_p^{k+m-l}(K)}.$$

In particular,

$$|v|_{W_p^k(K)} \le Ch^{-(l-m)+n/p-n/q} \|v\|_{W_q^m(K)} \tag{2.25}$$

for k such that $l-m \le k \le l$. This implies $k+m-l \le m$. Combination of (2.24) with $j = l-m$ and (2.25) proves (2.23). $\square$

In the case of $p = q = 2$, we have

$$\|v\|_{H^l(K)} \le Ch^{m-l} \|v\|_{H^m(K)}.$$

In particular, we have that

$$\|v\|_{H^1(K)} \le Ch^{-1} \|v\|_{L^2(K)}$$

and

$$\|v\|_{H^2(K)} \le Ch^{-2} \|v\|_{L^2(K)}.$$

Next we present inverse trace inequalities from [240] without proofs.

Theorem 2.3.8. *Let $K = [a, b]$ and $\mathcal{P}_k(K)$ be the space of kth order polynomials defined in K. For $u \in P_k(K)$ we have that*

$$|u(a)| \le \frac{p+1}{|b-a|} \|u\|_{L^2(K)}.$$

Theorem 2.3.9. *Let K be a triangle and $\mathcal{P}_k(K)$ be the space of polynomials of order at most k defined on K. In addition, let S be the perimeter length of K and A be the area of K. For $u \in \mathcal{P}_k(K)$ we have that*

$$\|u\|_{L^2(\partial K)} \le \sqrt{\frac{(p+1)(p+2)}{2} \frac{S}{A}} \|u\|_{L^2(K)}.$$

Theorem 2.3.10. *Let K be a tetrahedron and $\mathcal{P}_k(K)$ be the space of polynomials of order at most k defined on K. Denote the surface area of K by A and the volume of K by V. For $u \in \mathcal{P}_k(K)$ we have that*

$$\|u\|_{L^2(\partial K)} \le \sqrt{\frac{(p+1)(p+3)}{3} \frac{A}{V}} \|u\|_{L^2(K)}.$$

2.4 Approximation Properties

One important piece of the convergence analysis of finite element methods is the approximation property of the finite element space X_h. Essentially, it is the polynomial approximation theory in Sobolev spaces (see, e.g., Chapter 4 of [54]). We only sketch some basic results related to the Lagrange elements for triangular meshes in this section. Approximation properties of other finite element spaces will be discussed in the respective chapters later. The following materials are adapted from Section 2.6 of [44].

In view of Céa's Lemma, we would like to know how well the finite element space approximates the function space. To this end, we define the mesh dependent norm.

Definition 2.4.1. *Give a triangular mesh $\mathcal{T}_h = \{K_1, K_2, \dots, K_M\}$ of Ω, the mesh dependent norm is defined as*

$$\|v\|_{m,h} := \left(\sum_{K_j \in \mathcal{T}} \|v\|^2_{H^m(K_j)} \right)^{1/2}, \quad m \geq 1. \tag{2.26}$$

Definition 2.4.2. *A Lipschitz domain is said to satisfy a cone condition if the interior angles at each vertex are positive, so that a nontrivial cone can be positioned in Ω with its tip at the vertex.*

For each $v \in H^m(\Omega)$, there exists a uniquely defined interpolant $I_h v$ in the Lagrange element space. We would like to estimate $\|v - I_h v\|_{m,h}$ by $\|v\|_{H^t(\Omega)}$ for $m \leq t$. We first state a theorem on the interpolation operator (Lemma 6.2 of [44]).

Theorem 2.4.1. *Let $\Omega \subset \mathbb{R}^2$ be a Lipschitz domain which satisfies the cone condition. In addition, let $t \geq 2$ and suppose $z_1, z_2, \dots, z_s$ are $s := t(t+1)/2$ prescribed points in $\overline{\Omega}$ such that the interpolant operator $I : H^t \to P_{t-1}$ is well defined for polynomials of degree $\leq t-1$. Then there exists a constant C depending on Ω and $z_i, i = 1, \dots, s$, such that*

$$\|u - Iu\|_{H^t(\Omega)} \leq C|u|_{H^t(\Omega)} \quad \textit{for all } u \in H^t(\Omega). \tag{2.27}$$

Proof. We define a norm on $H^t(\Omega)$

$$|||v||| := |v|_{H^t(\Omega)} + \sum_{i=1}^{s} |v(z_i)|.$$

We first show that $|||\cdot|||$ is equivalent to $\|\cdot\|_{H^t(\Omega)}$. It is easy to verify that $|||\cdot|||$ is a norm. Note that $H^t(\Omega) \hookrightarrow C^0(\Omega)$, which implies

$$|v(z_i)| \leq C\|v\|_{H^t(\Omega)}, \quad i = 1, 2, \dots, s.$$

Thus $|||v||| \leq (1 + Cs)\|v\|_{H^t(\Omega)}$.

On the other hand, suppose that there does not exist a constant C such that

$$\|v\|_{H^t(\Omega)} \leq C\,|||v||| \quad \text{for all } v \in H^t(\Omega).$$

Then there exists a sequence $\{v_n\} \subset H^t(\Omega)$ such that

$$\|v_n\| = 1, \quad |||v_n||| \leq 1/n, \quad n = 1, 2, \ldots.$$

Due to the compact embedding of $H^t(\Omega)$ into $H^{t-1}(\Omega)$ (see Theorem 1.2.2), there exists a subsequence of $\{v_n\}$, still denoted by $\{v_n\}$, that converges in $H^{t-1}(\Omega)$. Since $|v_n|_{H^t(\Omega)} \to 0$ and

$$\|v_m - v_n\|^2_{H^t(\Omega)} \leq \|v_m - v_n\|^2_{H^{t-1}(\Omega)} + \left(|v_m|_{H^t(\Omega)} + |v_n|_{H^t(\Omega)}\right)^2,$$

the sequence $\{v_n\}$ is a Cauchy sequence in $H^t(\Omega)$. There exists $v^* \in H^t(\Omega)$ such that

$$\|v^*\|_{H^t(\Omega)} = 1 \quad \text{and} \quad |||v^*||| = 0.$$

This leads to contradiction. Hence $|||\cdot|||$ is equivalent to $\|\cdot\|_{H^t(\Omega)}$.

Since Iu takes the same values as u at the interpolation points z_i's, we have that

$$\begin{aligned}
\|u - Iu\|_{H^t(\Omega)} &\leq C\,|||u - Iu||| \\
&= C\left(|u - Iu|_{H^t(\Omega)} + \sum_{i=1}^{s} |(u - Iu)(z_i)|\right) \\
&= C|u - Iu|_{H^t(\Omega)} \\
&= C|u|_{H^t(\Omega)}.
\end{aligned}$$

Eqn. (2.27) follows directly. □

As a consequence, we have the following Bramble-Hilbert Lemma (see Section 2.6 of [44]).

Lemma 2.4.2. *Let $\Omega \subset \mathbb{R}^2$ be a Lipschitz domain. Suppose $t \geq 2$ and that L is a bounded linear mapping from $H^1(\Omega)$ into a normed linear space Y. If $\mathcal{P}_{t-1} \subset$ ker$L$, the kernel of L, then there exists a constant $C = C(\Omega) \geq 0$ such that*

$$\|Lv\| \leq C|v|_{H^t(\Omega)} \quad \textit{for all } v \in H^t(\Omega).$$

Proof. Let $I : H^t(\Omega) \to \mathcal{P}_{t-1}$ be an interpolation operator satisfying the properties in Theorem 2.4.1. Noting that $Iv \in$ ker$L$, the kernel of L, we have

$$\begin{aligned}
\|Lv\| &= \|L(v - Iv)\| \\
&\leq \|L\| \cdot \|v - Iv\|_{H^t(\Omega)} \\
&\leq C\|L\| \cdot |v|_{H^t(\Omega)}.
\end{aligned}$$

□

The following discussion is on the approximation property of the Lagrange elements. Let $\mathcal{T}$ be a triangulation for Ω. Define

$$V^{t-1} := V^{t-1}(\mathcal{T}) = \left\{ v \in L^2(\Omega) \,|\, v|_K \in \mathcal{P}_{t-1} \text{ for every } K \in \mathcal{T} \right\}.$$

By Theorem 2.1.2, there exists a unique interpolation operator

$$I_h : H^t(\Omega) \to V^{t-1}, \quad t \geq 2.$$

The following estimate holds for I_h (Theorem 6.4 of [44]).

Theorem 2.4.3. *Let $t \geq 2$, and suppose $\mathcal{T}_h$ is a shape-regular triangulation of Ω. Then there exists a constant C such that*

$$\|u - I_h u\|_{m,h} \leq C h^{t-m} |u|_{H^t(\Omega)} \quad \textit{for } u \in H^t(\Omega), \quad 0 \leq m \leq t.$$

Before we prove the theorem, let us check the transformation formula for affine mappings. Let K and $\hat{K}$ be affine equivalent, i.e., there exists a bijective affine mapping $F : \hat{K} \to K$ such that

$$F\hat{x} = B\hat{x} + x_0$$

with a nonsingular matrix B. If $v \in H^m(K)$, then $\hat{v} := v \circ F \in H^m(\hat{K})$, and there exists a content C depending only on the domain $\hat{T}$ and m such that

$$|\hat{v}|_{H^m(\hat{K})} \leq C\|B\|^m |\text{det}B|^{-1/2} |v|_{H^m(K)}, \tag{2.28}$$

where det$B$ denotes the determinant of B.

Proof. Let $\mathcal{T}_h$ be a shape-regular triangulation for Ω. For $K \in \mathcal{T}_h$, let $\rho(K)$ be the radius of the largest circle inscribed in K and $r(K)$ be the radius of the smallest circle containing K. For the reference triangle $\hat{K}$, we choose

$$r(\hat{K}) = 2^{-1/2}$$

and

$$\rho(\hat{K}) = (2 + \sqrt{2})^{-1} \geq 2/7.$$

Let $F : \hat{K} \to K$ for $K \in \mathcal{T}_h$ be the affine mapping. On the reference triangle $\hat{K}$, by Theorem 2.4.1, we have

$$\begin{aligned} |u - I_h u|_{m,K} &\leq C\|B\|^{-m} |\text{det}B|^{1/2} |\hat{u} - I_h \hat{u}|_{H^m(\hat{K})} \\ &\leq C\|B\|^{-m} |\text{det}B|^{1/2} \cdot C|\hat{u}|_{H^m(\hat{K})} \\ &\leq C\|B\|^{-m} |\text{det}B|^{1/2} \cdot C\|B\|^t \cdot |\text{det}B|^{-1/2} |u|_{H^t(K)} \\ &\leq C(\|B\| \cdot \|B^{-1}\|)^m \|B\|^{t-m} |u|_{H^t(K)}. \end{aligned}$$

Since $\mathcal{T}_h$ is shape-regular, $r/\rho \leq \kappa$ for some $\kappa > 0$. In addition,

$$\|B\| \cdot \|B^{-1}\| \leq (2 + \sqrt{2})\kappa$$

and

$$\|B\| \le r(K)/\rho(\hat{K}),$$

which implies

$$\|B\| \le h/\rho(\hat{K}) \le 4h.$$

Thus the following holds

$$|u - I_h u|_{H^l(K)} \le Ch^{t-l}|u|_{H^t(K)}.$$

Summing over l from 0 to m, we obtain that

$$\|u - I_h u\|_{H^m(K)} \le Ch^{t-m}|u|_{H^t(K)} \quad \text{for all } u \in H^t(K), K \in \mathcal{T}_h.$$

The theorem follows immediately. □

Finally we present a classical result of polynomial interpolation, which can be found in many classical finite element books (e.g., [88, 54]).

Theorem 2.4.4. *Let $v \in H^{t+1}(K)$. There exists a constant C such that*

$$\inf_{p \in \mathcal{P}_t} \|v + p\|_{H^{t+1}(K)} \le C|v|_{H^{t+1}(K)}.$$

2.5 Appendix: Implementation of Finite Elements in 1D

In this section, we discuss some implementing issues for finite element methods in one dimension. It aims to provide the readers a quick start on coding finite element and shed some light on the implementation of two- and three- dimensional problems in later chapters.

Let $\Omega = (0, 1)$. The model problem is to find $u \in H_0^1(\Omega)$ such that

$$a(u, v) := (u', v') = \lambda(u, v) \quad \text{for all } v \in H_0^1(\Omega), \tag{2.29}$$

where u' and v' denote the derivatives of u and v, respectively.

As we mentioned before, mesh generation has been an important research area. However, for one-dimensional problems, it is straightforward. The mesh contains two data structures. One is the $n + 1$ nodes $\{x_i\}, i = 1, \ldots, n + 1$, such that

$$0 = x_1 < x_2 < \ldots < x_{n+1} = 1.$$

The other one is n intervals $K_i, i = 1, \ldots, n$, such that $K_i = [x_i, x_{i+1}]$ and $h_i = x_{i+1} - x_i$.

Suppose we use linear Lagrange elements, i.e., there are two basis functions involving interval K_i: ϕ_1 is 1 at x_i and 0 at x_{i+1}, ϕ_2 is 0 at x_i and 1 at x_{i+1}. In fact, ϕ_1 and ϕ_2 are the restrictions of the global basis functions ϕ_i and ϕ_{i+1} on K_i. Thus the local index 1 corresponds to global index i and the local index 2 corresponds to $i+1$. It is sometimes termed as local to global index mapping. In Fig. 2.2, we plot a part of the mesh and basis functions. Basis function ϕ_i is 1 at x_i and 0 at all other nodes, i.e., $\phi_i(x_j) = \delta_{ij}, i, j = 1, \ldots, n+1$.

$$\phi_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & x \in K_{i-1}, \\ 1 - \frac{x - x_i}{x_{i+1} - x_i} & x \in K_i, \\ 0 & \text{otherwise}. \end{cases} \tag{2.30}$$

Linear Lagrange basis functions are also called the hat functions. In Fig. 2.3 we show the quadratic basis functions on K_i only for the readers' information.

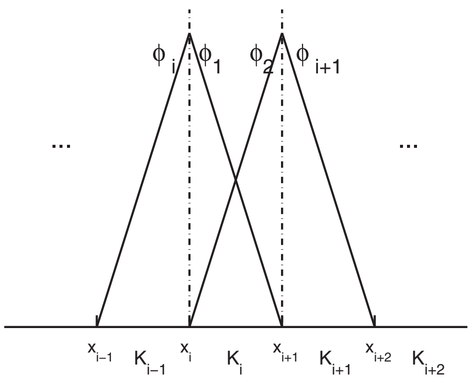


Figure 2.2: Linear Lagrange basis functions in one dimension.

We move on to assemble the so-called stiffness matrix S corresponding to (u', v') and the mass matrix M corresponding to (u, v):

$$S_{i,j} = (\phi_j', \phi_i') \quad \text{and} \quad M_{i,j} = (\phi_j, \phi_i), \quad i, j = 2, \ldots, n.$$

This is done by looping through all the intervals. On interval K_i, we need to evaluate 4 integrals for S locally

$$(\phi_1', \phi_1'), \quad (\phi_1', \phi_2'), \quad (\phi_2', \phi_1'), \quad (\phi_2', \phi_2'),$$

which contribute to the global entries $S_{i,i}$, $S_{i,i+1}$, $S_{i+1,i}$, and $S_{i+1,i+1}$, respectively. Note that each basis function ϕ_i is non-zero on K_{i-1} and K_i. We only need to compute the integrals when $|i - j| \leq 1$ due to the overlapping of basis functions.

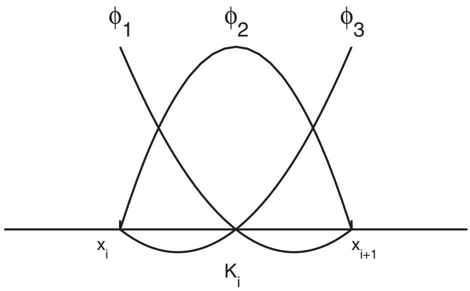


Figure 2.3: Quadratic Lagrange basis functions in one dimension.

Using (2.30), on K_i, the entries of the local stiffness matrix are given by

$$\begin{aligned}
(\phi_i', \phi_i') &= \int_{x_i}^{x_{i+1}} \frac{-1}{x_{i+1}-x_i} \cdot \frac{-1}{x_{i+1}-x_i}\, \mathrm{d}x, \\
(\phi_i', \phi_{i+1}') &= \int_{x_i}^{x_{i+1}} \frac{-1}{x_{i+1}-x_i} \cdot \frac{1}{x_{i+1}-x_i}\, \mathrm{d}x, \\
(\phi_{i+1}', \phi_i') &= \int_{x_i}^{x_{i+1}} \frac{1}{x_{i+1}-x_i} \cdot \frac{-1}{x_{i+1}-x_i}\, \mathrm{d}x, \\
(\phi_{i+1}', \phi_{i+1}') &= \int_{x_i}^{x_{i+1}} \frac{1}{x_{i+1}-x_i} \cdot \frac{1}{x_{i+1}-x_i}\, \mathrm{d}x.
\end{aligned}$$

The entries of the local mass matrix are given by

$$\begin{aligned}
(\phi_i, \phi_i) &= \int_{x_i}^{x_{i+1}} \left(1 - \frac{x-x_i}{x_{i+1}-x_i}\right) \cdot \left(1 - \frac{x-x_i}{x_{i+1}-x_i}\right) \mathrm{d}x, \\
(\phi_i, \phi_{i+1}) &= \int_{x_i}^{x_{i+1}} \left(1 - \frac{x-x_i}{x_{i+1}-x_i}\right) \cdot \left(\frac{x-x_i}{x_{i+1}-x_i}\right) \mathrm{d}x, \\
(\phi_{i+1}, \phi_i) &= \int_{x_i}^{x_{i+1}} \left(\frac{x-x_i}{x_{i+1}-x_i}\right) \cdot \left(1 - \frac{x-x_i}{x_{i+1}-x_i}\right) \mathrm{d}x, \\
(\phi_{i+1}, \phi_{i+1}) &= \int_{x_i}^{x_{i+1}} \left(\frac{x-x_i}{x_{i+1}-x_i}\right) \cdot \left(\frac{x-x_i}{x_{i+1}-x_i}\right) \mathrm{d}x.
\end{aligned}$$

Now expand u_h in terms of basis functions

$$u_h = \sum_{i=2}^{n} u_i \phi_i,$$

where we have taken the homogeneous Dirichlet boundary condition into account. Let $\mathbf{u} = (u_2, \ldots, u_n)^T$. The final matrix eigenvalue problem is

$$S(2:n, 2:n)\mathbf{u} = \lambda_h M(2:n, 2:n)\mathbf{u}, \tag{2.31}$$

where $S(2:n, 2:n)$ and $M(2:n, 2:n)$ are obtained by deleting the 1st row and the 1st column and the $(n+1)$th row and the $(n+1)$th column of S and M, respectively.

A simple MATLAB code is as follows. It has only about a dozen lines. However, it contains all the necessary elements to implement a finite element method.

```
1.  clear all
% number of subintervals for (0, 1)
2.  N = 20;
3.  h = 1/N;
% uniform mesh with h=1/N
4.  x = linspace(0, 1, N+1);
% initialization
5.  S = sparse(N+1, N+1); M = sparse(N+1, N+1);
6.  for it = 1:N
7.     index = [it it+1];
       % local stiffness matrix
8.     Sloc = [1/h -1/h; -1/h, 1/h];
9.     S(index, index) = S(index, index) + Sloc;
       % local mass matrix
10.    Mloc = [1/3*h 1/6*h; 1/6*h 1/3*h];
11.    M(index, index) = M(index, index) + Mloc;
12. end
13. eigs(S(2:N, 2:N), M(2:N, 2:N), 6, 'sm')
```

Some brief comments are given below.

1. Line 1 simply clears the workspace.
2. Line 2 gives the number of intervals (mesh).
3. Line 3 is the length of each interval assuming we use a uniform mesh.
4. Line 4 generates the actual mesh.
5. Line 6 to Line 12 loop through all the elements (intervals), generate the local stiffness matrix and the local mass matrix, and distribute the entries to the global matrices.
6. Line 7 is the local to global index mapping, i.e., the two local basis functions involving the interval K_i have the global indices i and $i+1$.
7. Line 8 and Line 10 compute the local stiffness matrix and the local mass matrix, respectively. Since we use a uniform mesh, they can be computed easily. Line 9 and Line 11 distribute the local contributions to global matrices according to the local to global index mapping.

8. Line 13 calls "*eigs*" to compute six smallest Dirichlet eigenvalues.

We conclude this section by commenting on some aspects which the one-dimensional problems might miss.

1. Mesh generation is an important part for the implementation of the finite element method. There are many publications and excellent softwares for it. Here in one dimension, it can be done easily. However, for higher dimensional problem with complex geometry, it needs to be treated carefully.

2. The local to global index mapping could be much more complicated in higher dimensions.

3. The local matrices are computed exactly. However, for higher dimensional problems, techniques such as affine mapping are needed.

4. There are no quadrate rules involved in the above code. However, in the case when exact evaluation of the integrals is not possible, quadrate rules are necessary.

5. Some additional data structures need to be constructed in higher dimensions. For example, for tetrahedral meshes, the generation software usually gives the data structures for nodes and tetrahedra. One needs to generate additional data structures for edges and faces.

Chapter 3

The Laplace Eigenvalue Problem

3.1 Introduction

The Laplace eigenvalue problem appears in many applications such as vibration modes in acoustics, nuclear magnetic resonance measurements of diffusive transport, electron wave functions in quantum waveguides, construction of heat kernels in the theory of diffusion, etc. [135].

The problem has been studied by many researchers; see, e.g., [102]. The theory and numerical methods are well developed. Due to the simplicity of both theory and implementation, it serves well as the first model problem to study finite element methods for eigenvalue problems. There are many finite element methods proposed for the Laplace eigenvalue problem in literature [36, 35, 11, 196]. In this chapter, we discuss the H^1-conforming Lagrange elements. The results will be frequently used in later chapters when we consider more difficult eigenvalue problems and complicated finite elements.

We assume that Ω is a Lipschitz polygon in $\mathbb{R}^2$. Note that similar results hold for three-dimensional cases. We begin with the source problem, i.e., Poisson's equation. Given a function f, find u such that

$$-\triangle u = f \quad \text{in } \Omega \tag{3.1}$$

with the homogeneous Dirichlet boundary condition

$$u = 0 \quad \text{on } \partial\Omega. \tag{3.2}$$

The weak formulation is obtained by multiplying (3.1) by a test function v and integrating by parts using the boundary condition (3.2): for $f \in H^{-1}(\Omega)$, find $u \in H_0^1(\Omega)$ such that

$$a(u,v) = (f,v) \quad \text{for all } v \in H_0^1(\Omega), \tag{3.3}$$

where

$$a(u,v) := (\nabla u, \nabla v) \quad u, v \in H_0^1(\Omega).$$

It is easy to show that the bilinear form $a(\cdot,\cdot)$ is bounded in $H^1(\Omega)$. Employing the Cauchy-Schwarz inequality, we have the boundedness of $a(\cdot,\cdot)$:

$$\begin{aligned} |a(u,v)| &= |(\nabla u, \nabla v)| \\ &\leq \|\nabla u\| \|\nabla v\| \\ &\leq \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \end{aligned}$$

for all $u, v \in H_0^1(\Omega)$. Recall that $\|\cdot\|$ denotes the $L^2(\Omega)$ norm.

Next we show that the bilinear form $a(\cdot,\cdot)$ is coercive. As a special case of Theorem 1.2.5, the following Poincaré-Friedrichs inequality holds for functions in $H_0^1(\Omega)$ (see also Chapter 2, Section 1 of [44]).

Theorem 3.1.1. *Suppose Ω is contained in an n-dimensional cube with side length s. Then*

$$\|v\| \leq s|v|_{H^1(\Omega)} \quad \textit{for all } v \in H_0^1(\Omega).$$

Consequently, the coercivity of $a(\cdot,\cdot)$ holds:

$$a(u,u) = \|\nabla u\|^2 \geq \alpha \|u\|^2_{H^1(\Omega)} \quad \text{for all } u \in H_0^1(\Omega), \tag{3.4}$$

where α is a positive constant. Thus by the Lax-Milgram Lemma 1.3.1, we obtain the following theorem.

Theorem 3.1.2. *There exists a unique solution $u \in H_0^1(\Omega)$ to* (3.3) *such that*

$$\|u\|_{H^1(\Omega)} \leq C \|f\|_{H^{-1}(\Omega)}.$$

The regularity of u plays an important role in error estimates for the finite element methods. It depends not only on the data f but also on Ω. In general, the weak solution $u \notin H^2(\Omega)$ if Ω is a non-convex polygon. The following regularity result is from [68] (see also Chapter 8 of [139]).

Theorem 3.1.3. *Let Ω be a bounded Lipschitz polygon. There exists an $\alpha_0 > 1/2$ depending on the interior angles of Ω. For α such that $\frac{1}{2} \leq \alpha \leq \alpha_0$, the solution u of* (3.3) *satisfies*

$$\|u\|_{H^{1+\alpha}(\Omega)} \leq C \|f\|_{H^{-1+\alpha}(\Omega)}.$$

In particular, α_0 is at least 1 when Ω is convex.

We define the solution operator $T : L^2(\Omega) \to L^2(\Omega)$ which maps f to the solution u, i.e., $Tf = u$ and consequently,

$$a(Tf, v) = (f, v) \quad \text{for all } v \in H_0^1(\Omega).$$

Due to the Sobolev Embedding Theorem 1.2.1 for $H^1(\Omega)$ into $L^2(\Omega)$, T is a compact operator. It is easy to see that T is self-adjoint:

$$\begin{aligned}(Tu, v)_{L^2(\Omega)} &= (v, Tu)_{L^2(\Omega)} \\ &= a(Tv, Tu) \\ &= a(Tu, Tv) \\ &= (u, Tv)_{L^2(\Omega)}.\end{aligned}$$

We are now ready to discuss the Laplace eigenvalue problem. When the boundary condition is given by the Dirichlet boundary condition (3.2), we call it the Dirichlet eigenvalue problem. Although not included in this book, there are other boundary conditions as well, for example, the Neumann boundary condition, which leads to the Neumann eigenvalue problem.

The Dirichlet eigenvalue problem is to find $\lambda \in \mathbb{R}$ and u such that

$$-\triangle u = \lambda u \quad \text{in } \Omega \tag{3.5}$$

with u satisfying the boundary condition (3.2). The variational formulation is to find $\lambda \in \mathbb{R}$ and a non-trivial $u \in H_0^1(\Omega)$ such that

$$a(u, v) = \lambda(u, v) \quad \text{for all } v \in H_0^1(\Omega). \tag{3.6}$$

Using the operator T, the problem is equivalent to the operator eigenvalue problem:

$$\lambda T u = u.$$

Thus, λ is a Dirichlet eigenvalue if and only if $\mu := 1/\lambda$ is an eigenvalue of the compact self-adjoint operator T.

3.2 Lagrange Elements for the Source Problem

In this section, we consider the finite element method for the source problem, i.e., Poisson's equation.

Assume that Ω is covered by a regular triangular mesh $\mathcal{T}$ (see Fig. 3.1). Let V_h be the finite element space of the Lagrange element of order k with zero values for the nodes on $\partial\Omega$. From Chapter 2, we know that $V_h \subset H^1(\Omega)$, i.e., V_h is H^1-conforming. Furthermore, the following approximation results hold provided that $u \in H^r(\Omega)$ (see Section 2.4)

$$\inf_{v_h \in V_h} \|u - v_h\| \le Ch^{\min\{k+1,r\}} \|u\|_{H^r(\Omega)}, \tag{3.7}$$

$$\inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \le Ch^{\min\{k,r-1\}} \|u\|_{H^r(\Omega)}. \tag{3.8}$$

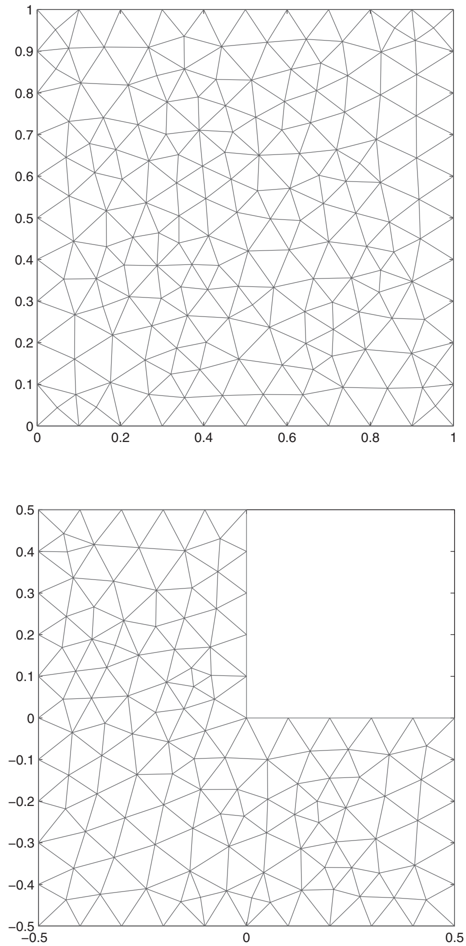


Figure 3.1: Two polygonal domains with triangular meshes. Top: unit square (convex). Bottom: L-shaped domain (non-convex).

Recall that when Ω is a non-convex polygon, the solution u belongs to a Sobolev space of fractional order. We assume u and α satisfy the condition of Theorem 3.1.3. Letting $P_h u$ be the $H_0^1(\Omega)$ projection onto V_h, one has the following standard approximation estimates:

$$\|u - P_h u\|_{H^1(\Omega)} \le Ch^\alpha \|u\|_{H^{1+\alpha}(\Omega)} \le Ch^\alpha \|f\|_{H^{-1+\alpha}(\Omega)}. \tag{3.9}$$

The discrete problem for Poisson's equation is to find $u_h \in V_h$ such that

$$a(u_h, v_h) = f(v_h) \quad \text{for all } v_h \in V_h. \tag{3.10}$$

The well-posedness of the discrete problem can be obtained the same way as the continuous case since we are using conforming finite elements, i.e., there exists a unique solution $u_h \in V_h$ for (3.10).

Consequently, we can define a discrete solution operator

$$T_h : L^2(\Omega) \to V_h \subset L^2(\Omega)$$

such that

$$a(T_h f, v_h) = f(v_h) \quad \text{for all } v_h \in V_h.$$

It is clear that T_h is self-adjoint since $a(\cdot, \cdot)$ is symmetric. From (3.3) and (3.10), we have the following Galerkin orthogonality.

Theorem 3.2.1. *Let u and u_h be the solutions of* (3.3) *and* (3.10)*, respectively. Then the following Galerkin orthogonality holds*

$$a(u - u_h, v_h) = 0 \quad \textit{for all } v_h \in V_h. \tag{3.11}$$

We proceed to study the error estimate in H^1-norm. The following theorem is classic. For example, a simpler version is Theorem 7.3 of [44].

Theorem 3.2.2. *Suppose $\mathcal{T}_h$ is a family of shape-regular triangulations of Ω. Let u be the solution of Poisson's equation such that $u \in H^s(\Omega), s > 1$. Let $\tau = \min\{k, s-1\}$, where k is the order of the Lagrange elements. Then the finite element approximation u_h of u satisfies*

$$\|u - u_h\|_{H^1(\Omega)} \le Ch^\tau \|f\|. \tag{3.12}$$

Proof. From Céa's Lemma 2.3.1,

$$\|u - u_h\|_{H^1(\Omega)} \le C \inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}.$$

Then (3.8) implies that

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} &\le Ch^\tau \|u\|_{H^{\tau+1}(\Omega)} \\ &\le Ch^\tau \|f\|_{H^{-1+\tau}(\Omega)}, \end{aligned}$$

where we have used (3.9). By the result on negative norm (Section 1.2.2), we have that

$$\|f\|_{H^{-1+\tau}(\Omega)} \le \|f\|,$$

and thus

$$\|u - u_h\|_{H^1(\Omega)} \le Ch^\tau \|f\|.$$

□

Since $\|f\| \leq \|f\|_{H^1(\Omega)}$, a consequence of the above theorem is the uniform convergence of T_h to T.

Corollary 3.2.3. *Let $f \in H^1(\Omega)$. We have that*

$$\|Tf - T_h f\|_{H^1(\Omega)} \leq Ch^\tau \|f\|_{H^1(\Omega)}. \tag{3.13}$$

Next we would like to show the error estimate in the L^2-norm. It is done by a duality argument called the Nitsche's trick. We present the Aubin-Nitsche Lemma in the abstract formulation in the spirit of [18, 209]. The following theorem is taken from [44] (Theorem 7.6 therein).

Theorem 3.2.4. Aubin-Nitsche Lemma *Let H be a Hilbert space with the norm $\|\cdot\|_H$ and the scalar product $(\cdot,\cdot)$. Let V be a subspace which is also a Hilbert space with norm $\|\cdot\|_V$. Let $a(\cdot,\cdot)$ be a bounded coercive sesquilinear form on $V \times V$. In addition, the embedding of V to H is continuous. Given $f \in V'$, let u and u_h be the solutions of*

$$a(u, v) = f(v) \quad \text{for all } v \in V$$

and

$$a(u_h, v_h) = f(v_h) \quad \text{for all } v_h \in V_h,$$

respectively. Then the finite element solution $u_h \in V_h \subset V$ satisfies

$$\|u - u_h\|_H \leq C\|u - u_h\|_V \sup_{g \in H, g \neq 0} \left\{ \frac{1}{\|g\|_H} \inf_{v \in V_h} \|\phi_g - v\|_V \right\},$$

where, for every $g \in H$, $\phi_g \in V$ denotes the corresponding unique solution of the equation

$$a(w, \phi_g) = (g, w) \quad \text{for all } w \in V. \tag{3.14}$$

Proof. By Riesz Representation Theorem 1.1.4, the norm of an element in a Hilbert space can be defined as

$$\|w\|_H = \sup_{g \in H, g \neq 0} \frac{(g, w)}{\|g\|_H}. \tag{3.15}$$

Letting $w = u - u_h$ in (3.14), we obtain

$$\begin{aligned} (g, u - u_h) &= a(u - u_h, \phi_g) \\ &= a(u - u_h, \phi_g - v_h) \\ &\leq C\|u - u_h\|_V \|\phi_g - v_h\|_V, \end{aligned}$$

where we have used the Galerkin orthogonality. It follows that

$$(g, u - u_h) \leq C\|u - u_h\|_V \inf_{v_h \in V_h} \|\phi_g - v_h\|_V.$$

The duality argument (3.15) implies that

$$\begin{aligned} \|u - u_h\|_H &= \sup_{g \in H, g \neq 0} \frac{(g, u - u_h)}{\|g\|_H} \\ &\leq C\|u - u_h\|_V \sup_{g \in H, g \neq 0} \left\{ \inf_{v_h \in V_h} \frac{\|\phi_g - v_h\|_V}{\|g\|_H} \right\}. \end{aligned}$$

□

By applying of the above theorem to Poisson's equation, we obtain the following corollary.

Corollary 3.2.5. *Let $\mathcal{T}_h$ be a family of shape-regular triangulation of Ω and V_h be the Lagrange finite element space of order k associated with $\mathcal{T}_h$. Let u and u_h be the solutions of* (3.3) *and* (3.10)*, respectively. Assume that $u \in H^s(\Omega), 1 \le s \le 2$ and $\tau = \min\{k, s-1\}$. Then*

$$\|u - u_h\| \le Ch^\tau \|u - u_h\|_{H^1(\Omega)}.$$

Furthermore, if $f \in H^{-1+\tau}(\Omega)$ so that $u \in H^{1+\tau}(\Omega)$,

$$\|u - u_h\| \le Ch^{2\tau} \|f\|_{H^{-1+\tau}(\Omega)} \le Ch^{2\tau} \|f\|.$$

Proof. Let $H = L^2(\Omega)$ with $\|\cdot\|_H = \|\cdot\|$ and $V = H_0^1(\Omega)$ with $\|\cdot\|_V = \|\cdot\|_{H^1(\Omega)}$. It is obvious that $V \subset H$ and the embedding is continuous. Since ϕ_g solves (3.14), the estimate in (3.12) implies

$$\sup_{g \in H, g \neq 0} \left\{ \inf_{v_h \in V_h} \frac{\|\phi_g - v_h\|_{H^1(\Omega)}}{\|g\|} \right\} \le Ch^\tau.$$

Applying the Aubin-Nitsche Lemma (Theorem 3.2.4) and (3.7), we obtain that

$$\|u - u_h\| \le Ch^\tau \|u - u_h\|_{H^1(\Omega)}.$$

The corollary is proved by using (3.12) once more. □

3.3 Convergence Analysis

The discrete Dirichlet eigenvalue problem is to find $(\lambda_h, u_h) \in \mathbb{R} \times V_h$ such that

$$a(u_h, v_h) = \lambda_h (u_h, v_h) \quad \text{for all } v_h \in V_h. \tag{3.16}$$

The problem is equivalent to the operator eigenvalue problem:

$$\lambda_h T_h u_h = u_h.$$

Similar to the continuous case, λ_h is an eigenvalue if and only if $\mu_h := 1/\lambda_h$ is an eigenvalue of T_h.

We view T_h as an operator from $L^2(\Omega)$ to $L^2(\Omega)$. From Corollary 3.2.5,

$$\|Tf - T_h f\| \le Ch^{2\tau} \|f\|,$$

which implies

$$\|T - T_h\| \le Ch^{2\tau}.$$

Thus we immediately have the following theorem for the optimal convergence order for the eigenfunctions.

Theorem 3.3.1. *Let u be an eigenfunction associated with the eigenvalue λ of multiplicity m. Let $w_h^1, \ldots, w_h^m$ be the eigenfunctions associated with the m discrete eigenvalues $\lambda_h^1, \ldots, \lambda_h^m$ approximating λ. Then there exists $u_h \in$ span$\{w_h^1, \ldots, w_h^m\}$ such that*

$$\|u - u_h\| \leq Ch^{2\tau}\|u\|.$$

Let Γ be a simple closed curve which encloses λ of algebraic multiplicity m and no other eigenvalues. Provided h is small enough, there are m discrete eigenvalues of T_h inside Γ approximating λ. Let E be the spectral projection defined in (1.16). The following theorem gives the convergence rate of the eigenvalue approximation.

Theorem 3.3.2. *Let $\hat{\lambda}_h = \frac{1}{m}\sum_{j=1}^m \lambda_h^j$ where $\lambda_h^1, \ldots, \lambda_h^m$ are the discrete eigenvalues approximating λ. Then the following convergence rate holds*

$$|\lambda - \hat{\lambda}_h| \leq Ch^{2\tau}.$$

Proof. Due to the fact that both T and T_h are self-adjoint and in view of Theorem 1.4.4, we only need to approximate

$$\sum_{j,k=1}^{m} |((T - T_h)\phi_j, \phi_k)|,$$

where $\{\phi_1, \ldots, \phi_m\}$ is a basis for the generalized eigenspace $R(E)$ corresponding to λ. Recall that $R(E)$ is the range of the eigenvalue projection E (see (1.16)).

Using the definition of T and T_h, symmetry of $a(\cdot,\cdot)$, Galerkin orthogonality, and the estimate of $T - T_h$, we have that

$$\begin{aligned}
|((T - T_h)u, v)| &= |(v, (T - T_h)u)| \\
&= |a(Tv, (T - T_h)u)| \\
&= |a((T - T_h)u, Tv)| \\
&= |a((T - T_h)u, (T - T_h)v)| \\
&\leq \|(T - T_h)u\|_{H^1(\Omega)}\|(T - T_h)v\|_{H^1(\Omega)} \\
&\leq Ch^{2\tau},
\end{aligned}$$

which holds for any $u, v \in R(E)$ with $\|u\| = \|v\| = 1$. The theorem follows immediately. □

It is also possible to obtain the error estimates using $H_0^1(\Omega)$ (see, e.g., Section 10 of [35]). Let $T_{H_0^1} : H_0^1(\Omega) \to H_0^1(\Omega)$ be the restriction of T on $H_0^1(\Omega)$ such that

$$a(T_{H_0^1} f, v) = (f, v) \quad \text{for all } v \in H_0^1(\Omega).$$

Theorem 3.3.3. *The operator $T_{H_0^1}$ from $H_0^1(\Omega)$ to $H_0^1(\Omega)$ is compact.*

Proof. Let $\{u_n\}$ be a bounded sequence in $H_0^1(\Omega)$. Due to the compact embedding of $H_0^1(\Omega)$ to $L^2(\Omega)$, there exists a convergent subsequence of $\{u_n\}$, still denoted by $\{u_n\}$, in $L^2(\Omega)$. Let $u = \lim_{n\to\infty} u_n$ such that $u \in L^2(\Omega)$. Then $Tu \in H_0^1(\Omega)$ such that

$$a(Tu, v) = (u, v) \quad \text{for all } v \in H_0^1(\Omega).$$

On the other hand, we have that

$$a(T_{H_0^1} u_n, v) = (u_n, v).$$

Therefore

$$a(Tu - T_{H_0^1} u_n, v) = (u - u_n, v) \to 0 \quad \text{as } n \to \infty,$$

for all $v \in H_0^1(\Omega)$. Note that $a(\cdot,\cdot)$ defines an inner product on $H_0^1(\Omega)$. Thus we have that

$$T_{H_0^1} u_n \to Tu \quad \text{as } n \to \infty.$$

Hence $T_{H_0^1}$ is compact. □

Similarly, we define the discrete operator $T^h_{H_0^1} : H_0^1(\Omega) \to H_0^1(\Omega)$ as

$$a(T^h_{H_0^1} f, v) = (f, v) \quad \text{for all } v_h \in V_h \subset H_0^1(\Omega).$$

The self-adjointness of $T_{H_0^1}$ and $T^h_{H_0^1}$ can be derived in the same way as above. From Corollary 3.2.3, we see that $T^h_{H_0^1}$ converges to $T_{H_0^1}$ uniformly. In addition, one has that

$$\|T_{H_0^1} - T^h_{H_0^1}\| \le Ch^\tau.$$

The following argument is an alternative proof for Theorem 3.3.2. Since T and T_h are self-adjoint, again from Theorem 1.4.4, we only need to approximate

$$\sum_{j,k=1}^{m} \left| ((T_{H_0^1} - T^h_{H_0^1})\phi_j, \phi_k)_{H_0^1(\Omega)} \right|,$$

where $\{\phi_1, \dots, \phi_m\}$ is a basis for the eigenspace $R(E)$. Let $u, v \in R(E)$ corresponding to the eigenvalue λ. Since $v = \lambda T_{H_0^1} v$, one has that

$$\|v\|_{H^{1+\tau}(\Omega)} \le C\|v\|_{H^1(\Omega)}.$$

Thus we have that

$$\begin{aligned}
\left| ((T_{H_0^1} - T^h_{H_0^1})u, v))_{H_0^1(\Omega)} \right| &= C \left| a((T_{H_0^1} - T^h_{H_0^1})u, v) \right| \\
&= C \inf_{v_h \in V_h} |a((T_{H_0^1} - T^h_{H_0^1})u, v - v_h)| \\
&\le C\|(T_{H_0^1} - T^h_{H_0^1})u\|_{H^1(\Omega)} \inf_{v_h \in V_h} \|v - v_h\|_{H^1(\Omega)} \\
&\le Ch^\tau \|u\|_{H^1(\Omega)} h^\tau \|v\|_{H^{1+\tau}(\Omega)} \\
&\le Ch^{2\tau} \|u\|_{H^1(\Omega)} \|v\|_{H^{1+\tau}(\Omega)}.
\end{aligned}$$

The error estimate follows immediately.

3.4 Numerical Examples

We consider the Dirichlet eigenvalue problem of two simple polygonal domains in $\mathbb{R}^2$ to verify the theory developed above. The first one is the unit square given by $(0,1) \times (0,1)$. The second one is the L-shaped domain given by

$$(0,1) \times (0,1) \setminus (1/2,1) \times (0,1/2).$$

The Dirichlet eigenvalues of the unit square are known analytically

$$(m^2+n^2)\pi^2, \quad m,n \in \mathbb{Z}^+$$

with the corresponding eigenfunctions

$$\sin(m\pi x)\sin(n\pi y), \quad m,n \in \mathbb{Z}^+.$$

Here $\mathbb{Z}^+$ denotes the set of positive integers.

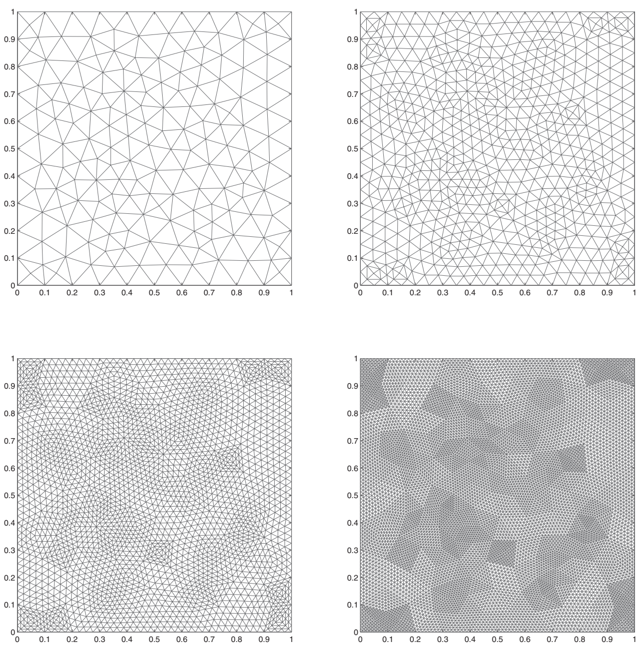


Figure 3.2: Sample uniformly refined unstructured meshes for the unit square.

For simplicity, we only show the numerical results of the first eigenvalue, i.e., $2\pi^2$. We generate a series of uniformly refined unstructured meshes (see Fig.3.2) and use linear and quadratic Lagrange elements. In Table 3.1, for the unit square, we show the mesh sizes h (column 1), the computed eigenvalue (column 2), the error (column 3), and the convergence order (column 4). Since the domain is convex and we use linear Lagrange elements, we have $\tau = 1$ (see Corollary 3.2.5) and the second order convergence is observed (see Theorem 3.3.2).

h	λ_h	$\|\lambda_h - \lambda\|$	convergence order
1/10	19.928106244003025	0.188897441824309	-
1/20	19.787168473383172	0.047959671204456	1.9777
1/40	19.751276465091120	0.012067662912404	1.9907
1/80	19.742232591845479	0.003023789666763	1.9967
1/160	19.739965301539787	0.000756499361071	1.9989

Table 3.1: Convergence order for the first eigenvalue of the unit square (linear Lagrange element).

Next we use the quadratic Lagrange element and the result is shown in Table. 3.2. For this case, we have that $\tau = 2$ and the convergence rate is $O(h^4)$. In Fig. 3.3, we show the log-log plot of the error. The first two eigenfunctions are shown in Fig. 3.4.

h	λ_h	$\lambda_h - \lambda$	convergence order
1/10	19.739634731484767	0.000425929306051	-
1/20	19.739235736678957	0.000026934500241	3.9831
1/40	19.739210497897183	0.000001695718467	3.9895
1/80	19.739208908566553	0.000000106387837	3.9945
1/160	19.739208808844928	0.000000006666212	3.9963

Table 3.2: Convergence order for the first eigenvalue of the unit square (quadratic Lagrange element).

For the L-shaped domain, the first eigenvalue can not be obtained exactly. To study the convergence rate, we use the relative error defined as

$$\text{Rel. Err.} = \frac{|\lambda_{h,j-1} - \lambda_{h,j}|}{\lambda_{h,j}},$$

where $\lambda_{h,j}$ denotes the computed eigenvalue on mesh level j. For the linear Lagrange element, the convergence rate is less than 2 (see Table 3.3). The non-convexity of the domain affects the regularity of the eigenfunction since $\tau < 1$ (see Corollary 3.2.5). Using the quadratic element does not improve the convergence rate, which confirms the fact that the regularity of the eigenfunction dominates (see Table 3.4).

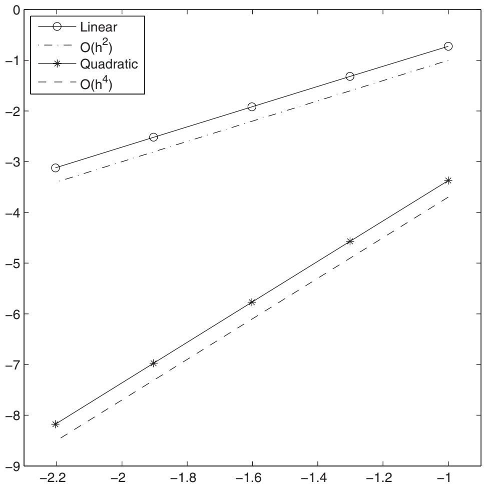


Figure 3.3: The log-log plot of the error of linear and quadratic Lagrange elements for the first eigenvalue of the unit square.

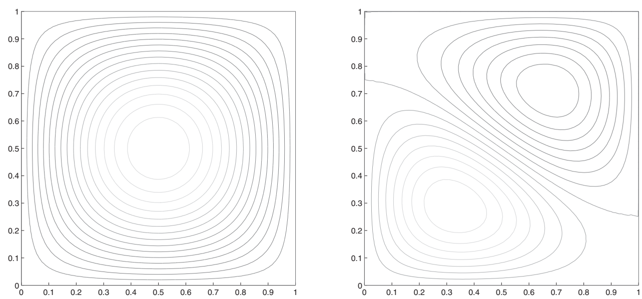


Figure 3.4: Eigenfunctions of the unit square. Left: the first eigenfunction. Right: the second eigenfunction.

It is easy to see that the eigenfunction $\sin(2\pi x)\sin(2\pi y)$ of the unit square is also an eigenfunction of the L-shaped domain. The corresponding eigenvalue, $8\pi^2$, turns out to be the third eigenvalue of the L-shaped domain. Tables 3.5 and 3.6 show the convergence rates are $O(h^2)$ and $O(h^4)$ for the linear and quadratic elements, respectively.

Remark 3.4.1. *Note that even when the domain is non-convex, the eigenfunctions*

h	λ_h	Rel. Err.	convergence order
1/10	39.946262635981505	-	-
1/20	39.012617299372167	0.023931881561414	-
1/40	38.714683702853314	0.007695622642964	1.6368
1/80	38.614656620170017	0.002590391613920	1.5709
1/160	38.579513835805820	0.000910918279420	1.5078

Table 3.3: Convergence order for the first eigenvalue of the L-shape domain (linear Lagrange element).

h	λ_h	Rel. Err.	convergence order
1/10	38.686756478047457	-	-
1/20	38.610227975933363	0.001982078483499	-
1/40	38.579357721059083	8.001754486811769e-04	1.3086
1/80	38.567026926676974	3.197237475824115e-04	1.3235
1/160	38.562123613420887	1.271536107617266e-04	1.3303

Table 3.4: Convergence order for the first eigenvalue of the L-shape domain (quadratic Lagrange element).

can have higher regularity than the solution of the source problem. The convergence order is determined by the regularity of the associated eigenspaces. The results verify the theory of Babuška and Osborn introduced in Chapter 1.

In Fig. 3.5, we show the contour plots of the first and the third eigenfunctions for the L-shaped domain. The log-log plot of the error is shown in Fig. 3.6.

h	λ_h	Rel. Err.	convergence order
1/10	81.931917460661182	2.975082251946318	-
1/20	79.705255772476349	0.748420563761485	1.9910
1/40	79.144599781181142	0.187764572466278	1.9949
1/80	79.003841330178417	0.047006121463554	1.9980
1/160	78.968592243605428	0.011757034890564	1.9993

Table 3.5: Convergence order for the third eigenvalue of the L-shape domain (linear Lagrange element).

3.5 Appendix: Implementation of the Linear Lagrange Element

h	λ_h	Rel. Err.	convergence order
1/10	78.979282322676966	0.022447113962102	-
1/20	78.958278044714859	0.001442835999995	3.9596
1/40	78.956926448545772	9.123983090830734e-05	3.9831
1/80	78.956840940848195	5.732133331548539e-06	3.9925
1/160	78.956835567836734	3.591218700194077e-07	3.9965

Table 3.6: Convergence order for the third eigenvalue of the L-shape domain (quadratic Lagrange element).

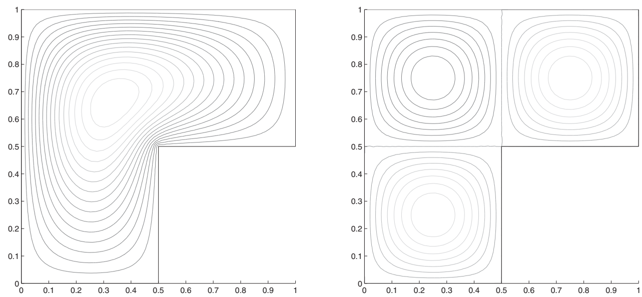


Figure 3.5: Dirichlet eigenfunctions of the L-shaped domain. Left: The first eigenfunction. Right: The third eigenfunction.

3.5.1 Triangular Meshes

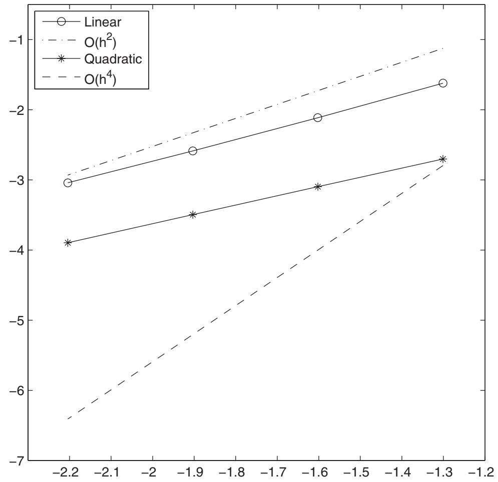
We illustrate how to use the MATLAB PDEtool to generate 2D triangular meshes using a simple example.

2dtriangle.m:

```
1. [pde_fig,ax]=pdeinit;
2. pdetool('appl_cb',1);
3. pderect([0  1 0 1],'R1');
4. set(findobj(get(pde_fig,'Children'),...
         'Tag','PDEEval'), 'String','R1');
5. setappdata(pde_fig,'Hgrad',1.3);
6. setappdata(pde_fig,'refinemethod','regular');
7. pdetool('initmesh')
8. pdetool('refine')
```

The above code generates a triangular mesh for the unit square.

a. Line 1 and 2 initiate the "pdetool" in MATLAB. Note that the MATLAB PDEtool can also be initiated by typing "pdetool" in the command window directly.



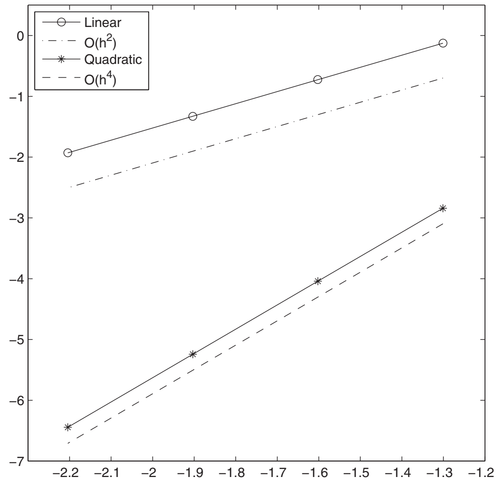


Figure 3.6: The log-log plot for the error for the L-shaped domain. Top: the first eigenvalue. Bottom: the third eigenvalue.

b. Line 3 defines a rectangular domain and labels it as "R1". The command

```
pderect([xmin xmax ymin ymax],LABEL)
```

defines a rectangle with dimensions given by the four values in the brackets. The label is optional. If omitted, a label will be automatically assigned.

Other commands are available to define different domains.

```
pdecirc(XC,YC,RADIUS,LABEL)
```

The command draws a circle with center at (XC,YC), RADIUS radius, and label. Label is optional. If omitted, a default label will be used.

```
pdeellip(XC,YC,RADIUSX,RADIUSY,ANGLE,LABEL)
```

The command draws an ellipse with center at (XC,YC), x- and y-axis radius (RADIUSX, RADIUSY), rotated counterclockwise by ANGLE radians. The ellipse is labeled using label (name) LABEL. LABEL and ANGLE are optional.

```
pdepoly(X,Y,LABEL)
```

The command draws a polygon with vertices determined by vectors X and Y and a label. Label is optional. A label will be assigned automatically if omitted.

c. Line 4 sets the object for partition. Lines 5 and 6 contain the command

```
setappdata(H, NAME, VALUE)
```

which sets application-defined data for the object with handle "H".

d. Line 7 partitions the object.

e. Finally, Line 8 refines the initial triangulation uniformly once.

In the "PDE Toolbox" window, one can choose "Mesh", then "Export Mesh ...", and accept the names of variables. The default names are "p, e, t". The following illustrates the data structure of the triangular mesh from the MATLAB PDE tool:

(1) the point matrix "p" is a $2 \times n$ matrix where n is the number of nodes (vertices) of the mesh. The first and second rows contain x- and y-coordinates of the nodes, respectively.

(2) the triangle matrix "t" is a $4 \times m$ matrix where m is the number of the triangles of the mesh. The first three rows contain indices of the vertices of the triangles, given in counterclockwise order. The fourth row contains the subdomain number.

(3) the edge matrix "e" is a $7 \times p$ matrix where p is the number of edges. The first and second rows of "e" contain indices of the starting and ending points of the edge, respectively. The third and fourth rows contain the starting and ending parameter values, respectively. The fifth row contains the edge segment number. And the sixth and seventh rows contain the left- and right-hand side subdomain numbers, respectively.

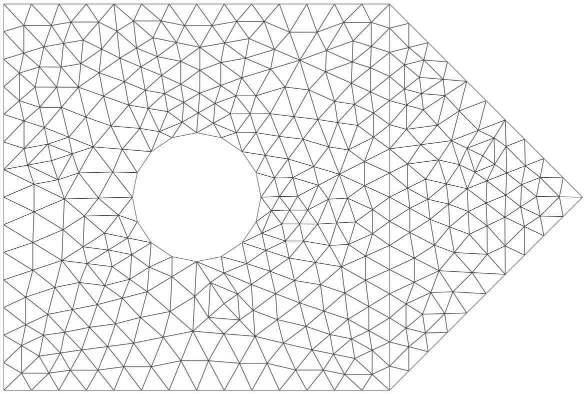


Figure 3.7: A domain and its triangular mesh obtained by the combination of simple geometries using MATLAB PDEtool.

Note that "e" only contains edges which coincide to the boundary of the domain (and subdomains). In general, we only need "p" and "t". The data structure "e" can be derived from "p" and "t".

One can use the combination of the above simple geometries to generate more complicated domains. For example, one can substitute Lines 3 and 4 with the following

```
pderect([0  1 0 1],'R1');
pdecirc(1/2,1/2,1/6, 'C1');
pdepoly([1, 3/2, 1],[0,1/2,1],'T1');
set(findobj(get(pde_fig,'Children'),'Tag','PDEEval'),...
         'String','R1-C1+T1');
```

The domain and the mesh are shown in Fig. 3.7.

3.5.2 Matrices Assembly

We consider the implementation of (3.16) using the linear Lagrange element. Let $\Omega = (0,1) \times (0,1)$. Assume that a triangular mesh $\mathcal{T}$ is given, i.e., we have nodes "p" and triangles "t". For linear Lagrange element, the degrees of freedom are the values on the nodes. The basis function at a node p_0 is a linear function which is 1 at p_0 and 0 at all other nodes. The support of the basis function is the union of all triangles sharing the vertex p_0. Such a function is called a hat function.

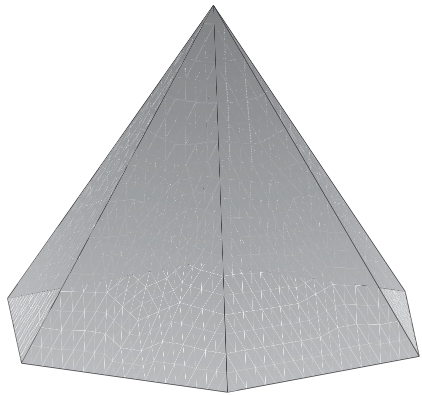


Figure 3.8: Linear Lagrange basis function.

Let $\{\phi_1, \phi_2, \dots, \phi_N\}$ be the basis functions of the linear Lagrange element space $V_h \subset H_0^1(\Omega)$ associated with the mesh $\mathcal{T}$. Let

$$u_h = \sum_{i=1}^{N} u_i \phi_i.$$

Substituting u_h in (3.16) and choosing $v_h = \phi_j$, we obtain

$$a\left(\sum_{i=1}^{N} u_i \phi_i, \phi_j\right) = \lambda_h \left(\sum_{i=1}^{N} u_i \phi_i, \phi_j\right), \quad j = 1, \dots, N.$$

Using the definition of $a(\cdot, \cdot)$, we have that

$$\sum_{i=1}^{N} (\nabla \phi_i, \nabla \phi_j) u_i = \lambda_h \sum_{i=1}^{N} (\phi_i, \phi_j) u_i, \quad j = 1, \dots, N.$$

The matrix form of the above linear system is given by

$$A\mathbf{u} = \lambda_h M \mathbf{u}, \tag{3.17}$$

where A and M are the $N \times N$ stiffness matrix and mass matrix given by

$$A_{i,j} = (\nabla \phi_j, \nabla \phi_i)$$

and

$$M_{i,j} = (\phi_j, \phi_i),$$

respectively. Here $\mathbf{u} = (u_1, u_2, \ldots, u_N)^T$.

Now we are facing the task of the construction of A and M. Note that the support of a nodal basis function usually spans several triangles sharing the vertex. Hence other than looping through all the basis functions (vertices), it is simpler to loop through the triangles, compute the local contribution (local stiffness and mass matrices), and distribute them to the global stiffness and mass matrices.

Let K be a triangle of the mesh whose vertices are I, J, L in "p", which are the global indices. In other words, the global basis functions ϕ_I, ϕ_J, ϕ_L have K as part of their support. Locally, we give indices $\{1, 2, 3\}$ to these vertices such that we have the so-called local-to-global mapping

$$1 \leftrightarrow I, \quad 2 \leftrightarrow J, \quad 3 \leftrightarrow L. \tag{3.18}$$

Denote the restriction of the basis function ϕ_I, ϕ_J, ϕ_L on K by ϕ_1, ϕ_2, ϕ_3, respectively. We construct the local matrices and distribute them to the global matrices. For example, the local stiffness matrix is given by

$$A_{loc} = \begin{pmatrix} (\nabla\phi_1, \nabla\phi_1)_K & (\nabla\phi_2, \nabla\phi_1)_K & (\nabla\phi_3, \nabla\phi_1)_K \\ (\nabla\phi_1, \nabla\phi_2)_K & (\nabla\phi_2, \nabla\phi_2)_K & (\nabla\phi_3, \nabla\phi_2)_K \\ (\nabla\phi_1, \nabla\phi_3)_K & (\nabla\phi_2, \nabla\phi_3)_K & (\nabla\phi_3, \nabla\phi_3)_K \end{pmatrix}.$$

Let the coordinates of the vertices of K be given by $(x_1, y_1), (x_2, y_2), (x_3, y_3)$. Then ϕ_1 is nothing but the linear function which is 1 at (x_1, y_1) and 0 at the other two points. The computation of A_{loc} is usually done by using the reference triangle $\hat{K}$ and the affine mapping. Recall that the vertices of $\hat{K}$ are $(0,0), (1,0), (0,1)$. The linear basis functions on $\hat{K}$ are simply

$$\hat{\phi}_1 = 1 - x - y, \quad \hat{\phi}_2 = x, \quad \hat{\phi}_3 = y.$$

Their gradients are given by

$$\nabla\hat{\phi}_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \nabla\hat{\phi}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \nabla\hat{\phi}_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Simple calculation shows that the local stiffness matrix and mass matrix for $\hat{K}$ are

$$\frac{1}{2}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \frac{1}{2}\begin{pmatrix} \frac{1}{6} & \frac{1}{12} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{6} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{12} & \frac{1}{6} \end{pmatrix},$$

respectively. Here $\frac{1}{2}$ is the area of the reference triangle $\hat{K}$.

The affine mapping from $\hat{K}$ to K is defined as $F : \hat{K} \to K$ such that

$$F\hat{x} := B\hat{x} + \mathbf{b},$$

where

$$B = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}.$$

If $\hat{p}$ is a scalar function, we obtain a function p on K by

$$p(F(\hat{x})) = \hat{p}(\hat{x}). \tag{3.19}$$

In particular, the basis functions $\hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3$ are transformed to ϕ_1, ϕ_2, ϕ_3, respectively. The gradient transforms as

$$(\nabla p) \circ F = (B^{-1})^T \hat{\nabla} \hat{p}, \tag{3.20}$$

where $\hat{\nabla}$ is with respect to $\hat{x}$.

To compute the local matrices, we need to evaluate the integrals related to the basis function on K. For the linear Lagrange element, it is enough to use the three points quadrature rule, which is exact for polynomials up to degree 2 (see Section 2.2.2). The quadrature points a^{12}, a^{23}, a^{31} are the middle points of three edges, respectively, with weight $1/3$.

For local stiffness matrix, the values of the gradients of the basis functions at a^{12}, a^{23}, a^{31} can be obtained from the corresponding values for the reference triangle $\hat{K}$ using (3.20). For example, we have that

$$\begin{aligned} (\nabla\phi_1, \nabla\phi_2) &= \int_K \nabla\phi_1 \cdot \nabla\phi_2 \, \mathrm{d}x \\ &= \frac{|K|}{3} \sum_{1 \le i < j \le 3} \nabla\phi_1(a^{ij}) \cdot \nabla\phi_2(a^{ij}) \\ &= \frac{|K|}{3} \sum_{1 \le i < j \le 3} \left[B^{-T} \nabla\hat{\phi}_1(\hat{a}^{ij}) \right] \cdot \left[B^{-T} \nabla\hat{\phi}_2(\hat{a}^{ij}) \right], \end{aligned}$$

where $|K|$ denotes the area of K and $\hat{a}^{ij}$'s are the middle points of the edges of $\hat{K}$. Note that $|K| = |\det(B)|/2$. For the linear Lagrange element, we have

$$(\nabla\phi_1, \nabla\phi_2) = |K| \left[B^{-T} \begin{pmatrix} -1 \\ -1 \end{pmatrix} \right] \cdot \left[B^{-T} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right].$$

The case for the local mass matrix is simpler. Since the values of the basis functions do not change (see (3.19)), we have that

$$(\phi_1, \phi_2) = \frac{|K|}{3} \sum_{1 \le i < j \le 3} \hat{\phi}_1(\hat{a}^{ij}) \cdot \hat{\phi}_2(\hat{a}^{ij}).$$

For the linear Lagrange element, it is simply

$$(\phi_1, \phi_2) = \frac{|K|}{3} \left(\frac{1}{2} \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} + \frac{1}{2} \cdot 0 \right) = \frac{|K|}{12}.$$

3.5.3 Boundary Conditions

For the linear Lagrange element, the zero boundary condition can be enforced by discarding all the nodes on the boundary. Or one can set the degrees of freedom associated with the boundary nodes to be zero. These boundary nodes can be found by various ways. For example, one can work only with "p" and "t" to generate a data structure for edges and search for boundary edges. The end points of these edges are the degrees of freedom on the boundary of the domain. If one has the exact information of the boundary, then a simple test can tell whether a node is on the boundary or not.

If the mesh is generated by MATLAB PDEtool and there are no interior boundaries, we can just take the data structure "e" and the end points are the boundary nodes.

3.5.4 Sample Codes

Assuming a triangular mesh for Ω is available in the MATLAB format, the following codes compute the Dirichlet eigenvalues. The main function is "DirichletEig.m". The inputs include the mesh "p", "t", "e", and "num", number of (smallest) eigenvalues to compute. The output is the computed eigenvalues stored in the vector 'lambda'.

DirichletEig.m

```
1. function lambda = DirichletEig(p, t, e, num)
2. [S, M]=assemble(p,t);
3. N=length(p);
%%%-------Find boundary nodes---------------------------
4. bdnodeE = unique([e(1,:),e(2,:)]);
5. Inode = setdiff(linspace(1,N,N), bdnodeE);
6. A = S(Inode, Inode); B = M(Inode, Inode);
7. [V,D]=eigs(A, B, num, 'sm');
8. lambda = diag(D);
```

a. Line 2 calls "assemble" to construct the stiffness and mass matrices. Note that 'assemble' returns matrices including the basis functions on the boundary of Ω.

b. Line 3 gives the number of nodes of the mesh.

c. Line 4 finds all the nodes on the boundary using "e".

d. Line 5 sets all the interior nodes "Inode" by subtracting the boundary nodes from the entire node sets.

e. Line 6 excludes the boundary nodes in the matrices.

f. Line 7 calls "eigs" to compute "num" eigenvalues.

g. Line 8 puts the computed eigenvalues in "lambda".

assemble.m

```
9.   function [S, M] = assemble(p, t)
     % 3 point quadrature rule
10. [weight,point]=quad_3;
11. nq=length(weight);
12. yloc=zeros(3,nq); gyloc=zeros(2,3,nq);
13. for r=1:nq
14.         [yloc(:,r),gyloc(:,:,r)]=phiRef(point(:,r));
15. end
16. nt = length(t); nv=length(p);
17. S=sparse(nv,nv); M=sparse(nv,nv);
18. for it=1:nt
19.     indices=t(1:3,it)';
20.     % The coordinates of the vertices of 'it'
21.     v=p(:,t(1:3,it));
22.     B=[v(:,2)-v(:,1),v(:,3)-v(:,1)];
23.     detB=abs(det(B))/2;
24.     for r=1:nq
25.         gphi(:,:,r)=(inv(B))'*gyloc(:,:,r);
26.     end
        % Stiffness matrix
27.     Sloc=zeros(3,3);
28.     for r=1:nq
29.       Sloc=Sloc+(gphi(:,:,r)'*gphi(:,:,r))*weight(r);
30.     end
31.     Sloc=Sloc*detB;
32.     S(indices, indices) = S(indices, indices)+Sloc;
        % Mass matrix
33.     Mloc=zeros(3,3);
34.     for r=1:nq
35.         Mloc=Mloc+((yloc(:,r)*yloc(:,r)'))*weight(r);
36.     end
37.     Mloc=Mloc*detB;
38.     M(indices, indices) = M(indices, indices)+Mloc;
39. end
```

We move on to explain the subroutine "assemble.m". It constructs the stiffness and mass matrices including basis functions on the boundary. It loops through all the triangles and uses the reference triangles to compute the local matrices. Then it distributes the local entries to the global matrices.

a. Line 10 calls "quad_3" to obtain the 3-point quadrature on a triangle.

b. Lines 12-15 call "phiRef" to compute values and gradients of basis functions at the quadrature points on the reference triangle.

c. Line 19 finds the global indices of the verities of triangle "it".

d. Line 21 gets the coordinates of the vertices of triangle "it".

e. Line 22 computes the affine transformation.

f. Line 23 computes the area of triangle "it".

g. Lines 24–26 compute the values of gradients of the basis functions of triangle "it".

h. Lines 27–31 compute the local stiffness matrix.

i. Line 32 distributes the local stiffness matrix to the global stiffness matrix.

j. Lines 33–37 compute the local mass matrix.

k. Line 38 distributes the local mass matrix to the global mass matrix.

The function "quad_3.m" simply gives the quadrature points and weights for the reference triangle.

quad_3.m

```
40. function [weight,point]=quad_3()
% 3 point quadrature
41. weight=[1/3, 1/3, 1/3];
42. point(:,1)=[0; 1/2];
43. point(:,2)=[1/2; 0];
44. point(:,3)=[1/2; 1/2];
```

The function "phiRef.m" computes the values and gradients of basis functions at "xhat" of the reference triangle.

phiRef.m

```
45. function [y,grady]=phiRef(xhat)
% Linear Basis Functions on the reference triangle
46. y=zeros(3,1); grady=zeros(2,3);
47. y(1) = 1 - xhat(1) - xhat(2);
48. y(2) = xhat(1);
49. y(3) = xhat(2);
50. grady(:,1) = [-1; -1];
51. grady(:,2) = [1; 0];
52. grady(:,3) = [0; 1];
```

Chapter 4

The Biharmonic Eigenvalue Problem

4.1 Introduction

The biharmonic eigenvalue problem is a fourth order eigenvalue problem appearing in many applications, e.g., mechanics (vibration and buckling of plates [157, 70, 158, 215]) and the inverse scattering theory (the transmission eigenvalue problem [62, 228]).

The source problem is the biharmonic equation. There are three classical approaches to discretize the biharmonic equation in literature. The first approach uses conforming finite elements, for example, the Argyris finite element method [13] or the partition of unity finite element method [230, 111]. These methods require globally continuously differentiable finite element spaces, which are difficult to implement (in particular for three dimensional problems). The second approach uses nonconforming finite elements such as the Adini element [4] or the Morley element [204, 216, 225]. A disadvantage is that such elements do not come in a natural hierarchy and existing nonconforming elements only involve low order polynomials that are not efficient for capturing smooth solutions. The third approach uses mixed finite element methods [89, 20, 20, 87] that only require continuous Lagrange finite element spaces. However, for the boundary conditions of simply supported plates, some mixed finite element methods can result in spurious solutions on non-convex domains (Sapondjan paradox [206]). This is also the case for the boundary conditions of the Cahn-Hilliard type that appear in mathematical models for phase separation phenomena. In addition, the solution of the saddle point problems resulting from the use of a mixed finite element method is more involved than that for a direct discretization of the fourth order operator.

An alternative to the three classical approaches is the C^0 interior penalty Galerkin (C^0 IPG) method developed in the last decade [117, 55, 49]. It is a discontinuous Galerkin method based on standard continuous Lagrange finite element spaces usually used for second order elliptic problems. The lowest order methods in this approach are almost as simple as classical nonconforming finite element methods and are much simpler than finite element methods using continuously differentiable basis functions. Unlike classical nonconforming finite element methods, higher order finite elements can be used in this approach to capture smooth solutions efficiently. Furthermore, the C^0 IPG method converges for the biharmonic source problem with boundary conditions of the clamped plate, the simply supported plate, and the Cahn-Hilliard type. It also preserves the symmetric positive-definiteness of the continuous problems. This last property is very attractive for eigenvalue problems since it means that the convergence for the eigenvalue problem can be derived from the convergence for the source problem by using the classical spectral approximation theory. In contrast, the convergence of mixed finite element methods for the source problem does not necessarily lead to convergence for the eigenvalue problem unless the mixed method is chosen carefully [36].

In this chapter, we discuss several finite element methods for the biharmonic eigenvalue problem. We first study the conforming Argyris element. Then we present a mixed finite element method followed by the nonconforming Morley element. Finally, we study the C^0 interior penalty discontinuous Galerkin method. Along the way, we will introduce additional abstract convergence theory needed for respective methods.

We begin with the source problem. Let Ω denote a bounded Lipschitz polygonal domain in $\mathbb{R}^2$ with boundary $\partial\Omega$. Let ν denote the unit outward normal to $\partial\Omega$. Given a function f, the biharmonic problem with clamped plate boundary condition is to find a function u such that

$$\Delta^2 u = f \qquad \text{in } \Omega, \tag{4.1a}$$

$$u = \frac{\partial u}{\partial \nu} = 0 \qquad \text{on } \partial\Omega. \tag{4.1b}$$

The biharmonic eigenvalue problem is as follows. Find λ and $u \neq 0$ such that

$$\Delta^2 u = \lambda u \qquad \text{in } \Omega, \tag{4.2a}$$

$$u = \frac{\partial u}{\partial \nu} = 0 \qquad \text{on } \partial\Omega. \tag{4.2b}$$

We proceed to derive the weak formulation for the biharmonic problem. Recalling the Sobolev space $H_0^2(\Omega)$ given by

$$H_0^2(\Omega) := \left\{ v \in H^2(\Omega) : v = \frac{\partial v}{\partial \nu} = 0 \text{ on } \partial\Omega \right\},$$

we define a sesquilinear form $a : H_0^2(\Omega) \times H_0^2(\Omega)$ such that

$$a(u, v) := (\triangle u, \Delta v). \tag{4.3}$$

Using the standard approach, a weak formulation for (4.1) is as follows. For $f \in L^2(\Omega)$, find $u \in H_0^2(\Omega)$ such that

$$a(u, v) = (f, v) \quad \text{for all } v \in H_0^2(\Omega). \tag{4.4}$$

The weak formulation for (4.2) is to find $\lambda \in \mathbb{R}$ and $u \in H_0^2(\Omega), u \neq 0$ such that

$$a(u, v) = \lambda(u, v) \quad \text{for all } v \in H_0^2(\Omega). \tag{4.5}$$

The well-posedness of the biharmonic equation can be obtained using the Poincaré-Friedrichs inequality Theorem 1.2.5 (see Eqn. 1.2.8 of [88]). The following theorem is from [88].

Theorem 4.1.1. *There exists a unique solution $u \in H_0^2(\Omega)$ to the biharmonic equation* (4.4) *such that*

$$\|u\|_{H^2(\Omega)} \leq C\|f\|.$$

Proof. It is easy to see that $a(\cdot, \cdot)$ is bounded. For the well-posedness of (4.4), we only need to show the coercivity of $a(\cdot, \cdot)$ on $H_0^2(\Omega) \times H_0^2(\Omega)$.

Let $\nu = (\nu_1, \nu_2)$ be the unit outward normal to $\partial\Omega$. The normal derivative is given by $\partial_\nu := \sum_{i=1}^2 \nu_i \partial_i$. For $u, v \in H^1(\Omega)$, we have the Green's formula

$$\int_\Omega u \partial_i v \,\mathrm{d}x = -\int_\Omega \partial_i u v \,\mathrm{d}x + \int_{\partial\Omega} u v \nu_i \mathrm{d}s, \quad i = 1, 2. \tag{4.6}$$

Note that, for $v \in H_0^2(\Omega)$, we have that

$$|v|_{H^2(\Omega)}^2 = \int_\Omega \left\{ \sum_{i=1} (\partial_{ii} v)^2 + \sum_{i \neq j} (\partial_{ij} v)^2 \right\} \mathrm{d}x,$$

$$\|\Delta v\|^2 = \int_\Omega \left\{ \sum_i (\partial_{ii} v)^2 + \sum_{i \neq j} \partial_{ii} v \partial_{jj} v \right\} \mathrm{d}x.$$

Let $w \in \mathcal{C}_0^\infty(\Omega)$, the space of smooth functions with compact support in Ω. We have that

$$\begin{aligned}\int_\Omega (\partial_{ij} w)^2 \,\mathrm{d}x &= -\int_\Omega \partial_i w \partial_{ijj} w \,\mathrm{d}x \\ &= \int_\Omega \partial_{ii} w \partial_{jj} w \,\mathrm{d}x.\end{aligned}$$

Using a density argument, one obtains that

$$\|\Delta v\| = |v|_{H^2(\Omega)} \quad \text{for all } v \in H_0^2(\Omega).$$

The Poincaré-Friedrichs inequality implies that the semi-norm $\|\Delta u\|$ is equivalent to the norm $\|u\|_{H^2(\Omega)}$ on $H_0^2(\Omega)$. Hence $a(\cdot,\cdot)$ is coercive. The theorem is proved by applying the Lax-Milgram Lemma 1.3.1. □

Consequently, there exists a solution operator $T : L^2(\Omega) \to H_0^2(\Omega) \subset L^2(\Omega)$ such that, given $f \in L^2(\Omega)$,

$$a(Tf, v) = (f, v) \quad \text{for all } v \in H_0^2(\Omega). \tag{4.7}$$

It is obvious that T is self-adjoint due to the symmetry of $a(\cdot,\cdot)$ and compact due to the compact embedding of $H_0^2(\Omega)$ into $L^2(\Omega)$.

The operator eigenvalue problem is to find $\lambda \in \mathbb{R}$ and $u \in H_0^2(\Omega), u \neq 0$ such that

$$\lambda T u = u. \tag{4.8}$$

Remark 4.1.1. *The regularity of the biharmonic solution u is critical to the convergence analysis. For polygonal domains, the solution of the biharmonic equation* (4.4) *belongs to $H^{2+\alpha}(\Omega)$ if $f \in H^{-2+\alpha}$ only for some $\alpha \in (1/2, 1]$. Furthermore, one has that*

$$\|u\|_{H^{2+\alpha}(\Omega)} \leq C\|f\|_{H^{-2+\alpha}(\Omega)} \tag{4.9}$$

for some constant C depending only on Ω. When Ω is convex, $\alpha = 1$. In general, α is referred to as the index of elliptic regularity for the biharmonic equation. Detailed studies can be found in [139, 108].

4.2 The Argyris Element

To treat a high order eigenvalue problem with conforming elements, e.g., the biharmonic eigenvalue problem, one needs finite elements with high regularity. In general, it is difficult to construct such elements. In this section, we present a C^1-element, the Argyris element [13], which is H^2-conforming for triangular meshes.

Let $\mathcal{T}_h$ be a triangular mesh for Ω and $K \in \mathcal{T}_h$ be a triangle. The Argyris element

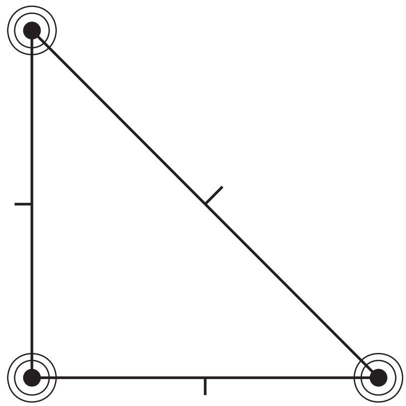


Figure 4.1: The Argyris element. There are 21 degrees of freedoms: 3 degrees of freedom are the values at three vertices, 6 degrees of freedom are the values of the first order partial derivatives at three vertices, 9 degrees of freedoms are the values of the second order derivatives at three vertices, and 3 degrees of freedom are the values of the normal derivatives at the midpoints of three edges.

uses the polynomials of degree 5. Note that $\dim(\mathcal{P}_5) = 21$. For $\mathcal{N} = \{N_1, \ldots, N_{21}\}$, 3 degrees of freedom are the values at the vertices of K, 6 degrees of freedom are the values of the first order partial derivatives at the vertices of K, 9 degrees of freedoms are the values of the second order derivatives at the vertices of K, and 3 degrees of freedom are the values of the normal derivatives at the midpoints of three edges.

Definition 4.2.1. *The Argyris element is the following triple* $(K, \mathcal{P}, \mathcal{N})$*:*

1. K *is a triangle with vertices* z_1, z_2, z_3*,*
2. $\mathcal{P} = \mathcal{P}_5(K)$*, the space of polynomials of order 5 on* K*,*
3. $\mathcal{N} = \{N_1, \ldots, N_{21}\}$ *is the set of degrees of freedom given by*

$$\begin{aligned}
& v(z_i), \quad i = 1, 2, 3 \\
& \frac{\partial v(z_i)}{\partial x}, \frac{\partial v(z_i)}{\partial y}, \quad i = 1, 2, 3, \\
& \frac{\partial^2 v(z_i)}{\partial x^2}, \frac{\partial^2 v(z_i)}{\partial x \partial y}, \frac{\partial^2 v(z_i)}{\partial y^2}, \quad i = 1, 2, 3, \\
& \frac{\partial v(z_4)}{\partial \nu}, \frac{\partial v(z_5)}{\partial \nu}, \frac{\partial v(z_6)}{\partial \nu}, z_4 = \frac{z_1 + z_2}{2}, z_5 = \frac{z_2 + z_3}{2}, z_6 = \frac{z_3 + z_1}{2},
\end{aligned}$$

where ν *is the unit outward normal to the edge (see Fig. 4.1).*

The following theorem (Theorem 2.2.13 of [88]) shows that the Argyris element is H^2-conforming.

Theorem 4.2.1. *Let V_h be the finite element space associated with the Argyris element and $\mathcal{T}_h$. Then the inclusion $V_h \subset C^1(\overline{\Omega}) \cap H^2(\Omega)$ holds.*

Note that the Argyris element does not belong to the affine families. This is due to the fact that normal derivatives are used as degrees of freedom. Fortunately, their interpolation properties are quite similar to those of affine families. Hence the Argyris element is referred to be almost-affine.

Definition 4.2.2. *(Section 6.1 of [88]) A family of finite element $(K, \mathcal{P}, \mathcal{N})$ is said to be almost-affine if, for any integer $k, m \geq 0$ and any number $p, q \in [1, \infty]$ compatible with the following inclusions:*

$$
\begin{aligned}
&W^{k+1,p}(K) \hookrightarrow \mathcal{C}^s(K),\\
&W^{k+1,p}(K) \hookrightarrow W^{m,q}(K),\\
&P_k(K) \subset P_K \subset W^{m,q}(K),
\end{aligned}
$$

there exists a constant C independent of K such that, for all $v \in W^{k+1,p}(K)$,

$$\|v - I_K v\|_{W^{m,q}(K)} \leq C(|K|))^{1/q-1/p} h_K^{k+1-m} |v|_{W^{k+1,p}(K)},$$

where h_K is the diameter of K, $I_K v$ is the interpolation of v, and $|K|$ is the measure of K.

Taking $p = q = 2$, a consequence of the above estimate is the following inequality

$$\|v - I_K v\|_{H^2(\Omega)} \leq C h^{k-1} |v|_{H^{k+1}(\Omega)}.$$

We present a theorem on the interpolation error for the Argyris element from [88] without proof.

Theorem 4.2.2. *A regular family of Argyris triangles is almost-affine. For all $p \in [1, \infty]$ and all pairs (m, q) with $m \geq 0$ and $q \in [1, \infty]$ compatible with the inclusion*

$$W^{6,p}(K) \to W^{m,q}(K),$$

there exists a constant C independent of K such that

$$\|v - I_K v\|_{W^{m,q}(K)} \leq C(|K|)^{1/q-1/p} h_K^{6-m} |v|_{W^{6,p}(K)},$$

where I_K denotes the associated $\mathcal{P}_5(K)$-interpolation operator.

We refer the readers to [44] for some discussion on other H^2-conforming elements including the triangular element of Bell, the Hsieh-Clough-Tocher element, and the reduced Hsieh-Clough-Tocher element. Construction of H^2-conforming finite element in $\mathbb{R}^3$ is even more difficult. According to [254], it requires 220 degrees of freedom per element. Any reasonably fine mesh would lead to a formidable number of degrees of freedom.

4.2.1 The Discrete Problem

The discrete problem for the source problem can be stated as follows. For $f \in H^{-2}(\Omega)$, the dual space of $H_0^2(\Omega)$, find $u_h \in V_h \subset H_0^2(\Omega)$ such that

$$a(u_h, v_h) = (f, v_h) \quad \text{for all } v_h \in V_h. \tag{4.10}$$

The discrete formulation for the eigenvalue problem (4.2) is to find $\lambda_h \in \mathbb{R}$ and $u_h \in V_h \subset H_0^2(\Omega)$, $u_h \neq 0$ such that

$$a(u_h, v_h) = \lambda_h(u_h, v_h) \quad \text{for all } v_h \in V_h. \tag{4.11}$$

The existence and uniqueness of a solution to the discrete problem (4.10) follow the continuous case. The following theorem, which is from [88], provides an error estimate for $\|u - u_h\|_{H^2(\Omega)}$.

Theorem 4.2.3. *Let u be the solution of* (4.4) *and u_h be the solution of* (4.10) *with V_h being the space of the Argyris element. If the solution $u \in H^{k+1}(\Omega) \cap H_0^2(\Omega)$ for $k \geq 1$, there exists a constant C independent of h such that*

$$\|u - u_h\|_{H^2(\Omega)} \leq Ch^{k-1}|u|_{H^{k+1}(\Omega)}. \tag{4.12}$$

The proof of the above theorem is standard. For completeness, we give a sketch as follows.

Proof. We have the Galerkin orthogonality

$$a(u - u_h, v_h) = 0 \quad \text{for all } v_h \in V_h.$$

Céa's Lemma 2.3.1 implies

$$\|u - u_h\|_{H^2(\Omega)} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{H^2(\Omega)}.$$

From Definition (4.2.2), we obtain that

$$\|v - I_K v\|_{H^2(\Omega)} \leq Ch^{k-1}|v|_{H^{k+1}(\Omega)},$$

which implies (4.12). □

Using the above theorem and Remark 4.1.1, one immediately gets that

$$\|u - u_h\|_{H^2(\Omega)} \leq Ch^{\alpha}\|f\|_{H^{-2}(\Omega)}, \tag{4.13}$$

where α is the index of elliptic regularity of the biharmonic equation.

To apply the abstract convergence theory, we restrict T and T_h on $H_0^2(\Omega) \subset L^2(\Omega)$ to $H_0^2(\Omega)$. We first show that T is a compact operator from $H_0^2(\Omega)$ to $H_0^2(\Omega)$.

Theorem 4.2.4. *The operator T from $H_0^2(\Omega)$ to $H_0^2(\Omega)$ is compact.*

Proof. Let $\{u_n\}$ be a bounded sequence in $H_0^2(\Omega)$. Due to the compact embedding of $H_0^2(\Omega)$ to $L^2(\Omega)$, there exists a convergent subsequence of $\{u_n\}$, still denoted by $\{u_n\}$, in $L^2(\Omega)$. Let $u = \lim_{n\to\infty} u_n$ such that $u \in L^2(\Omega)$. Then $Tu \in H_0^2(\Omega)$ such that

$$a(Tu, v) = (u, v) \quad \text{for all } v \in H_0^2(\Omega).$$

In addition, we have that

$$a(Tu_n, v) = (u_n, v).$$

Therefore,

$$a(Tu - Tu_n, v) = (u - u_n, v) \to 0 \quad \text{as } n \to \infty,$$

for all $v \in H_0^2(\Omega)$. Note that $a(\cdot,\cdot)$ defines an inner product on $H_0^2(\Omega)$. Thus we have that

$$Tu_n \to Tu \quad \text{as } n \to \infty.$$

Hence T is compact from $H_0^2(\Omega)$ to $H_0^2(\Omega)$. □

Theorem 4.2.5. *Let λ be a biharmonic eigenvalue with multiplicity m. Let $\hat{\lambda}_h = \frac{1}{m}\sum_{i=1}^m \lambda_h^i$, where $\lambda_h^1, \ldots, \lambda_h^m$ are the discrete eigenvalues approximating λ. Then the following convergence holds*

$$|\lambda - \hat{\lambda}_h| \le Ch^{2\alpha}, \tag{4.14}$$

where α is the index of elliptic regularity of the biharmonic equation in Remark 4.1.1.

Proof. Note that both T and T_h are self-adjoint. By (4.13), we have that

$$\|(T - T_h)|_{R(E)}\| \cdot \|(T' - T_h')|_{R(E')}\| \le Ch^{2\alpha},$$

where $R(E)$ and $R(E')$ denotes the eigenspace and the dual eigenspace associated with λ, respectively (see Section 1.4).

Let $\{\phi_1, \ldots, \phi_m\}$ be a basis for the eigenspace $R(E)$ associated with λ. In order to use Theorem 1.4.5, we need to estimate

$$\sum_{j,k=1}^m |((T - T_h)\phi_j, \phi_k)|.$$

By the definition of T and T_h, symmetry of $a(\cdot,\cdot)$, Galerkin orthogonality, and the estimate of $T - T_h$, one has

$$\begin{aligned} |((T - T_h)u, v)| &= |(v, (T - T_h)u)| \\ &= |a(Tv, (T - T_h)u)| \\ &= |a((T - T_h)u, Tv)| \\ &= |a((T - T_h)u, (T - T_h)v)| \\ &\le C\|(T - T_h)u\|_{H^2(\Omega)}\|(T - T_h)v\|_{H^2(\Omega)} \\ &\le Ch^{2\alpha}, \end{aligned}$$

which holds for any $u, v \in R(E)$ such that $\|u\|_{H_0^2(\Omega)} = \|v\|_{H_0^2(\Omega)} = 1$. Then (4.14) follows Theorem 1.4.5. □

4.2.2 Numerical Examples

We choose two polygonal domains: the unit square and the L-shaped domain given by
$$(0,1) \times (0,1) \setminus [1/2,1] \times [0,1/2].$$

For the biharmonic eigenvalue problem on the unit square, an accurate lower bound for the first eigenvalue is $1,294.933940$ given by Wieners [245]. An accurate upper bound is $1,294.9339796$ given by Bjørstad and Tjøstheim [32].

Recall that the relative error is defined as
$$\text{Rel. Err.} = \frac{|\lambda_{h,j-1} - \lambda_{h,j}|}{\lambda_{h,j}},$$

where $\lambda_{h,j}$ denotes the computed eigenvalue on mesh level j. In Table 4.1, we show the first and fourth biharmonic eigenvalues for the unit square on a few levels of uniformly refined meshes. The relative error is $O(h^2)$ for the first eigenvalue, which is consistent with the fact that the unit square is convex and the solution u of the biharmonic equation is at least in $H^3(\Omega)$. The relative error is $O(h^3)$ for the fourth eigenvalue indicating that the corresponding eigenfunction is smoother than the eigenfunction corresponding to the first eigenvalue.

h	1st	Rel. Err.	order
1/10	1.294934118026000e+03	-	-
1/20	1.294933983690248e+03	1.037394599480429e-07	-
1/40	1.294933949077454e+03	2.672938959503176e-08	1.9565
h	4th	Rel. Err.	order
1/10	1.171081740321877e+04	-	-
1/20	1.171081141097147e+04	5.116850652571765e-07	-
1/40	1.171081067647905e+04	6.271917803259636e-08	3.0283

Table 4.1: The convergence rates of the first and fourth biharmonic eigenvalues of the unit square using the Argyris element.

In Table 4.2, we show the first and second eigenvalues of the L-shaped domain. For the first eigenvalue, the convergence rate is less than $O(h^2)$ due to the fact that the regularity of the associated eigenfunction is lower than $H^3(\Omega)$ since the domain is nonconvex. Again, we obtain higher order of convergence for the second eigenvalue indicating that the corresponding eigenfunction is smoother.

Remark 4.2.1. *The convergence order depends on the regularity of the associated eigenspace. The regularity of the source problem is the lower bound for the regularity of the eigenfunctions.*

h	1st	Rel. Err.	order
1/10	7.118075264191162e+03	-	-
1/20	6.892658815967231e+03	0.032703845387174	-
1/40	6.790509619725640e+03	0.015042935208406	1.1204
1/80	6.744002412652421e+03	0.006896083991009	1.1252
h	2nd	Rel. Err.	order
1/10	1.113158721636632e+04	-	-
1/20	1.107644696967104e+04	0.004978152908262	-
1/40	1.106070418633956e+04	0.001423307509745	1.8064
1/80	1.105626514205896e+04	4.014958237312070e-04	1.8258

Table 4.2: The first and second biharmonic eigenvalues of the L-shaped domain.

4.3 A Mixed Finite Element Method

One way to avoid the complicated high regularity elements is to use the mixed method. In this section, we present a mixed finite element method for the biharmonic eigenvalue problem. A similar formulation will be used to treat the quad-curl eigenvalue problem later.

4.3.1 Abstract Framework

The convergence theory for the mixed method needs different techniques than conforming elements. We start with the abstract framework from [23] developed for certain mixed formulations.

Let V, W, H, and G be real Hilbert spaces with their respective inner products and induced norms. In addition, we assume that $V \subset H$ and $W \subset G$. Let $A(\cdot,\cdot)$ and $B(\cdot,\cdot)$ be bounded bilinear forms on $H \times H$ and $V \times W$, respectively. Assume that $A(\cdot,\cdot)$ is symmetric and satisfies

$$A(\sigma,\sigma) > 0 \quad \text{for all } 0 \neq \sigma \in H. \tag{4.15}$$

The bilinear form $B(\cdot,\cdot)$ statisfies

$$\sup_{\psi \in V} |B(\psi,u)| > 0 \quad \text{for all } 0 \neq u \in W. \tag{4.16}$$

The weakly posed eigenvalue problem is to find $\lambda \in \mathbb{R}$ and $(\sigma,u) \in V \times W$, $(\sigma,u) \neq (0,0)$ such that

$$A(\sigma,\psi) + B(\psi,u) = 0 \qquad \text{for all } \psi \in V, \tag{4.17a}$$

$$B(\sigma,v) = -\lambda(u,v)_G \qquad \text{for all } v \in W, \tag{4.17b}$$

where $(\cdot,\cdot)_G$ is the inner product on G.

Let $V_h \subset V$ and $W_h \subset W$. We consider the discrete eigenvalue problem of finding $\lambda_h \in \mathbb{R}$ and $(\sigma_h, u_h) \in V_h \times W_h$, $(\sigma_h, u_h) \neq (0,0)$ such that

$$A(\sigma_h, \psi_h) + B(\psi_h, u_h) = 0 \qquad \text{for all } \psi_h \in V_h, \tag{4.18a}$$

$$B(\sigma_h, v_h) = -\lambda_h (u_h, v_h)_G \qquad \text{for all } v_h \in W_h. \tag{4.18b}$$

As usual, the analysis starts with the source problem. Given $g \in G$, find $(\sigma, u) \in V \times W$ such that

$$A(\sigma, \psi) + B(\psi, u) = 0 \qquad \text{for all } \psi \in V, \tag{4.19a}$$

$$B(\sigma, v) = -(g, v)_G \qquad \text{for all } v \in W. \tag{4.19b}$$

The associated discrete problem is as follows. Given $g \in G$, find $(\sigma_h, u_h) \in V_h \times W_h$ such that

$$A(\sigma_h, \psi_h) + B(\psi_h, u_h) = 0 \qquad \text{for all } \psi_h \in V_h, \tag{4.20a}$$

$$B(\sigma_h, v_h) = -(g_h, v_h)_G \qquad \text{for all } v_h \in W_h. \tag{4.20b}$$

Assuming that both (4.19) and (4.20) are well-posed, there exist four solution operators

$$S : G \to V, \qquad Sg = \sigma, \tag{4.21a}$$

$$S_h : G \to V, \qquad S_h g = \sigma_h, \tag{4.21b}$$

$$T : G \to G, \qquad Tg = u, \tag{4.21c}$$

$$T_h : G \to G, \qquad T_h g = u_h, \tag{4.21d}$$

where (σ, u) and (σ_h, u_h) are solutions of (4.19) and (4.20), respectively.

Note that, from (4.17), both components of an eigenfunction (σ, u) cannot be zero. Taking $v = u$ in (4.17b) and $\psi = \sigma$ in (4.17a), we obtain that

$$\lambda = \frac{A(\sigma, \sigma)}{(u, u)_G},$$

which implies that $\lambda > 0$. If $(\lambda, (\sigma, u))$ is an eigenpair of (4.17), it is clear that

$$\lambda T u = u, \quad u \neq 0.$$

On the other hand, if $\lambda T u = u, u \neq 0$, there exists a $\sigma \in V$, $\sigma = S(\lambda u)$, such that $(\lambda, (\sigma, u))$ is an eigenpair of (4.17). Thus we have established the following result.

Lemma 4.3.1. *λ is an eigenvalue of* (4.17) *if and only if λ^{-1} is an eigenvalue of T.*

Similarly, λ_h is an eigenvalue of (4.18) if and only if λ_h^{-1} is an eigenvalue of T_h. We assume the uniform convergence of T_h to T, i.e.,

$$\|T - T_h\|_{\mathcal{L}(G,G)} \to 0 \quad \text{as } h \to 0.$$

It is obvious that T is compact since it is the limit of a sequence of operators on finite dimensional spaces (Theorem 1.1.10). Next, letting $v = Tf$ in (4.19b), we obtain

$$B(Sg, Tf) = -(g, Tf)_G.$$

Replacing g with f and setting $\psi = Sg$ in (4.19a),

$$A(Sf, Sg) + B(Sg, Tf) = 0.$$

Therefore,

$$(g, Tf)_G = A(Sf, Sg) \quad \text{for all } f, g \in G.$$

By the symmetry of $A(\cdot, \cdot)$, we obtain that

$$(Tg, f)_G = (f, Tg)_G = A(Sg, Sf) = A(Sf, Sg) = (g, Tf)_G.$$

Hence T is self-adjoint. Similar properties hold for the discrete operator T_h.

Let λ^{-1} be an eigenvalue of T with multiplicity m. Due to the uniform convergence of T_h to T in G, there exist m eigenvalues of T_h, $\lambda_{1,h}^{-1}, \ldots, \lambda_{m,h}^{-1}$, converging to λ^{-1}. Since T and T_h are self-adjoint, the relevant ascents are one and all eigenvalues have equal algebraic and geometric multiplicity. Let $\overline{M}$ be the eigenspace of T associated with λ^{-1}, i.e.,

$$\overline{M} := \overline{M}(\lambda^{-1}) = R(E).$$

The following theorem gives the convergence of the eigenvalues.

Theorem 4.3.2. *Under the above assumption, there exists a constant C such that*

$$\begin{aligned} |\lambda - \lambda_{l,h}| \quad \leq \quad & C\Big\{\|(S - S_h)|_{\overline{M}}\|^2_{\mathcal{L}(G,H)} + \|(T - T_h)|_{\overline{M}}\|^2_{\mathcal{L}(G,G)} \\ & + \|(S - S_h)|_{\overline{M}}\|_{\mathcal{L}(G,V)}\|(T - T_h)|_{\overline{M}}\|_{\mathcal{L}(G,W)}\Big\} \end{aligned} \tag{4.22}$$

for $l = 1, \ldots, m$.

Proof. Let $\{u\}_1^m$ be an orthonormal basis for $\overline{M}(\lambda^{-1})$. From Theorem 1.4.6 with $\alpha = 1$, we have that

$$|\lambda^{-1} - \lambda_{l,h}^{-1}| \leq C\left\{\sum_{i,j=1}^{m} |((T - T_h)u_j, u_j)_G| + \|(T - T_h)_{\overline{M}}\|^2_{\mathcal{L}(G,G)}\right\},$$

for $l = 1, \ldots, m$.

Let $g, f \in G$. From (4.19) and (4.21), it holds that

$$(g, v)_G = -A(Sg, \psi) - B(\psi, Tg) - B(Sg, v) \quad \text{for all } (\psi, v) \in V \times W.$$

Setting $v = (T - T_h)f$ and $\psi = (S - S_h)f$, we get

$$(g, (T - T_h)f)_G = -A(Sg, (S - S_h)f) - B((S - S_h)f, Tg) - B(Sg, (T - T_h)f).$$

Replacing g with f in (4.19) and subtracting (4.20) from (4.19), we obtain that

$$A((S-S_h)f,\psi)+B(\psi,(T-T_h)f)+B((S-S_h)f,v)=0$$

for all $(\psi,v)\in V_h\times W_h$. The above two equations imply that

$$\begin{aligned}
& |(g,(T-T_h)f)_G| \\
= \;& |A((S-S_h)f,Sg-\psi)+B((S-S_h)f,Tg-v) \\
& \qquad +B(Sg-\psi,(T-T_h)f)| \\
\le \;& C_1\|(S-S_h)f\|_H\|Sg-\psi\|_H+C_2\|(S-S_h)f\|_V\|Tg-v\|_W \\
& \qquad +C_2\|Sg-\psi\|_V\|(T-T_h)f\|_W,
\end{aligned}$$

where we have used the boundedness of $A(\cdot,\cdot)$ and $B(\cdot,\cdot)$. Letting $\psi=S_hg$ and $v=T_hg$, we obtain that

$$\begin{aligned}
& |((T-T_h)g,f)_G| \\
\le \;& C_1\|(S-S_h)f\|_H\|(S-S_h)g\|_H \\
& \quad +C_2\|(S-S_h)f\|_V\|(T-T_h)g\|_W+C_2\|(S-S_h)g\|_V\|(T-T_h)f\|_W.
\end{aligned}$$

Plugging $g=u_i$ and $f=u_j$ in the above equation, we have that

$$\begin{aligned}
& |((T-T_h)u_i,u_j)_G| \\
\le \;& C_1\|(S-S_h)|_M\|^2_{\mathcal{L}(G,H)}+2C_2\|(S-S_h)|_{\overline{M}}\|_{\mathcal{L}(G,V)}\|(T-T_h)|_{\overline{M}}\|_{\mathcal{L}(G,W)}.
\end{aligned}$$

Then (4.22) follows immediately. □

As a direct consequence of Theorems 1.4.3 and 1.4.6, the following result holds.

Theorem 4.3.3. *There exists a constant C such that*

$$\|u-u_h\|_G\le C\|(T-T_h)|_M\|_{\mathcal{L}(G,G)}.$$

4.3.2 The Ciarlet–Raviart Method

The mixed finite element method we will introduce is due to Ciarlet and Raviart [89]. The presentation here follows Section 7.1 of [88]. In the rest of this section, we assume that the solution of the biharmonic equation (4.62) u belongs to $H^3(\Omega)$. Note that this condition is satisfied if Ω is convex.

Introducing an auxiliary variable $\sigma=-\triangle u$, we obtain a second order system.

$$\sigma+\Delta u=0 \qquad \text{in } \Omega, \tag{4.23a}$$

$$-\Delta\sigma=f \qquad \text{in } \Omega, \tag{4.23b}$$

$$u=\frac{\partial u}{\partial\nu}=0 \qquad \text{on } \partial\Omega. \tag{4.23c}$$

The weak formulation is as follows. Given $f \in L^2(\Omega)$, find $(u, \sigma) \in H_0^1(\Omega) \times H^1(\Omega)$ such that

$$(\sigma, \psi) - (\nabla u, \nabla \psi) = 0 \qquad \text{for all } \psi \in H^1(\Omega), \tag{4.24a}$$

$$-(\nabla \sigma, \nabla v) = -(f, v) \qquad \text{for all } v \in H_0^1(\Omega). \tag{4.24b}$$

The biharmonic eigenvalue problem becomes the following.

$$\sigma + \Delta u = 0 \qquad \text{in } \Omega, \tag{4.25a}$$

$$-\Delta \sigma = \lambda u \qquad \text{in } \Omega, \tag{4.25b}$$

$$u = \frac{\partial u}{\partial \nu} = 0 \qquad \text{on } \partial\Omega. \tag{4.25c}$$

The variational formulation can be stated as follows. Find $\lambda \in \mathbb{R}$ and $(\sigma, u) \in H^1(\Omega) \times H_0^1(\Omega)$, $(\sigma, u) \neq (0, 0)$ such that

$$(\sigma, \psi) - (\nabla u, \nabla \psi) = 0 \qquad \text{for all } \psi \in H^1(\Omega), \tag{4.26a}$$

$$-(\nabla \sigma, \nabla v) = -\lambda(u, v) \qquad \text{for all } v \in H_0^1(\Omega). \tag{4.26b}$$

Let $V = H^1(\Omega)$, $W = H_0^1(\Omega)$, and $H = G = L^2(\Omega)$. We define two bilinear forms

$$A(\sigma, \psi) = (\sigma, \psi) \quad \text{for } \sigma, \psi \in H^1(\Omega)$$

and

$$B(\psi, u) = (\nabla \psi, \nabla u) \quad \text{for } \psi \in H^1(\Omega), u \in H_0^1(\Omega).$$

Problem (4.26) has exactly the same form as (4.18). Note that $A(\cdot, \cdot)$ is symmetric and conditions (4.15) and (4.16) are satisfied. We can see that, if (λ, u) is an eigenpair of (4.2) and $\sigma = -\triangle u$, then $(\lambda, (\sigma, u))$ is an eigenpair of (4.26). On the other hand, if $(\lambda, (\sigma, u))$ is an eigenpair of (4.26), then (λ, u) is an eigenpair of (4.2) and $\sigma = -\triangle u$.

We first study the mixed method of the associated source problem. The solution $u \in H_0^2(\Omega)$ of the biharmonic problem (4.1) satisfies the minimization problem

$$J(u) = \inf_{u \in H_0^2(\Omega)} J(v), \tag{4.27}$$

where

$$J(v) = \frac{1}{2}(\Delta v, \Delta v) - (f, v).$$

An equivalent problem is to minimize the following function

$$\mathcal{J}(v, \psi) = \frac{1}{2}(\psi, \psi) - (f, v)$$

over the pairs $(v, \psi) \in H_0^1(\Omega) \times L^2(\Omega)$ satisfying $-\Delta v = \psi$. The following theorem characterizes the space for (v, ψ).

Theorem 4.3.4. *Let Ω be convex. Define the space*

$$\mathcal{V} = \left\{(v,\psi) \in H_0^1(\Omega) \times L^2(\Omega) | \beta((v,\psi),\varphi) = 0 \, \textit{for all} \; \varphi \in H^1(\Omega)\right\}, \qquad (4.28)$$

where

$$\beta((v,\psi),\varphi) = (\nabla v, \nabla \varphi) - (\psi, \varphi). \qquad (4.29)$$

Then the mapping

$$(v,\psi) \in \mathcal{V} \to \|\psi\|$$

is a norm over the space $\mathcal{V}$, which is equivalent to the norm

$$(v,\psi) \in \mathcal{V} \to \left(|v|^2_{H^1(\Omega)} + \|\psi\|^2\right)^{1/2},$$

and which makes $\mathcal{V}$ a Hilbert space. In addition,

$$\mathcal{V} = \left\{(v,\psi) \in H_0^1(\Omega) \times L^2(\Omega) | -\Delta v = \psi\right\}.$$

Proof. It is obvious that $\mathcal{V}$ is a Hilbert space. Let $(v,\psi) \in \mathcal{V}$ and choose $\varphi = v$ in (4.28) to obtain

$$|v|^2_{H^1(\Omega)} = (\psi, v) \leq C\|\psi\||v|_{H^1(\Omega)}$$

for some constant C, where we have used the Poincaré inequality. Thus we have

$$\left(|v|^2_{H^1(\Omega)} + \|\psi\|^2\right)^{1/2} \leq (C+1)^{1/2}\|\psi\|,$$

which proves the first assertion of the theorem.

Since Ω is a Lipschitz domain, the Green's formula holds:

$$\int_\Omega \nabla v \cdot \nabla \varphi \, \mathrm{d}x = -\int_\Omega \Delta v \varphi \, \mathrm{d}x + \int_{\partial\Omega} \partial_\nu v \varphi \, \mathrm{d}s$$

for all $v \in H^2(\Omega)$ and $\varphi \in H^1(\Omega)$. Let the functions $v \in H_0^2(\Omega)$ and $\psi \in L^2(\Omega)$ be related through $-\Delta v = \psi$. For any function $\varphi \in H^1(\Omega)$, the above Green's formula implies $\beta((v,\psi),\varphi) = 0$ since $\partial_\nu v = 0$ on $\partial\Omega$.

Conversely, let the functions $v \in H_0^1(\Omega)$ and $\psi \in L^2(\Omega)$ satisfy $\beta((v,\psi),\varphi) = 0$, i.e.,

$$(\nabla v, \nabla \varphi) = (\psi, \varphi) \quad \text{for all } \varphi \in H_0^1(\Omega).$$

Thus v is the solution of a homogeneous Dirichlet problem for $-\Delta$ in Ω. Since Ω is convex, $v \in H^2(\Omega)$. Using the above Green's formula with functions $\varphi \in H_0^1(\Omega)$, we deduce that $-\Delta v = \psi$. Choosing $\varphi \in H^1(\Omega)$, we deduce that $\partial_\nu v = 0$ on $\partial\Omega$. The proof is complete. □

Theorem 4.3.5. *Let $u \in H_0^2(\Omega)$ denote the solution of the minimization problem* (4.27). *Then $(u, -\Delta u)$ is the unique solution of*

$$\mathcal{J}(u, -\Delta u) = \inf_{(v,\psi)\in\mathcal{V}} \mathcal{J}(v,\psi).$$

Proof. We consider the symmetric continuous bilinear form on $\mathcal{V} \times \mathcal{V}$ defined as

$$\mathcal{A}((u, \phi), (v, \psi)) := (\phi, \psi).$$

By Theorem 4.3.4, it is $\mathcal{V}$-elliptic, i.e., coercive on $\mathcal{V}$. The linear form

$$f(v, \psi) = (f, v)$$

is continuous. Hence the minimization problem: Find an element $(u^*, \phi) \in \mathcal{V}$ such that

$$\mathcal{J}(u^*, \phi) = \inf_{(v,\psi)\in\mathcal{V}} \mathcal{J}(v, \psi)$$

has a unique solution satisfying

$$(\phi, \psi) = (f, v). \tag{4.30}$$

Next we show that (u^*, ϕ) is the solution of (4.27). Since $(u^*, \phi) \in \mathcal{V}$, then $u^* \in H_0^2(\Omega)$ and $-\Delta u^* = \phi$. Using Theorem 4.3.4 and (4.30), we obtain that

$$(\Delta u^*, \Delta v) = (f, v) \quad \text{for all } v \in H_0^2(\Omega).$$

Thus the function u^* is the solution u of (4.27). $\square$

Now we consider a finite element discretization for it. Let $X_h \subset H^1(\Omega)$ be a finite element space, e.g., Lagrange finite element space. Let $X_{0,h}$ be defined as

$$X_{0,h} = \{v_h \in X_h | v_h = 0 \text{ on } \partial\Omega\}$$

and $\mathcal{V}_h$ be defined as

$$\mathcal{V}_h = \{(v_h, \psi_h) \in X_{0,h} \times X_h | \, \beta((v_h, \psi_h), \varphi_h) = 0 \text{ for all } \varphi_h \in X_h\}.$$

Then the discrete problem is as follows. Find $(u_h, \phi_h) \in \mathcal{V}_h$ such that

$$\mathcal{J}(u_h, \phi_h) = \inf_{(v_h,\psi_h)\in\mathcal{V}_h} \mathcal{J}(v_h, \psi_h). \tag{4.31}$$

Similar results hold as the continuous case. For example, we have the follow theorem.

Theorem 4.3.6. *The discrete problem* (4.31) *has a unique solution.*

With the well-posedness of both continuous and discrete mixed problems, we conclude that there exist solution operators T, S, T_h, S_h such that, given $f \in L^2(\Omega)$,

$$\begin{aligned} u &= Tf, \\ \phi &= Sf, \\ u_h &= T_h f, \\ \phi_h &= S_h f. \end{aligned}$$

For the error estimates, the following theorem holds.

Theorem 4.3.7. *There exists a constant C independent of the space X_h such that*

$$\begin{aligned} & |u-u_h|_{H^1(\Omega)} + \|\Delta u + \phi_h\| \\ \leq\ & C\left(\inf_{(v_h,\psi_h)\in\mathcal{V}_h}\big(|u-v_h|_{H^1(\Omega)} + \|\Delta u+\psi_h\|\big) + \inf_{\varphi_h\in X_h}\|\Delta u+\varphi_h\|_{H^1(\Omega)}\right). \end{aligned}$$

Proof. Since Ω is convex, $u \in H^3(\Omega)$,

$$-(\nabla v\cdot\nabla(\Delta u)) = (\Delta v, \Delta u) = (f, v) \quad \text{for all } v \in H_0^1(\Omega).$$

Given any function $v \in H_0^1(\Omega)$ and any function $\psi \in L^2(\Omega)$, one has that

$$\beta((v,\psi), -\Delta u) = (f, v) + (\psi, \Delta u).$$

Therefore,

$$(\Delta u + \phi_h, \phi_h - \psi_h) = -\beta((u_h - v_h, \phi_h - \psi_h), \Delta u + \varphi_h),$$

which implies

$$|(\Delta u + \phi_h, \phi_h - \psi_h)| \leq C\|\phi_h - \psi_h\|\|\Delta u + \varphi_h\|_{H^1(\Omega)}$$

for some constant C. As a consequence,

$$\begin{aligned} \|\phi_h - \psi_h\|^2 &= (\phi_h - \psi_h, \Delta u + \phi_h) - (\phi_h - \psi_h, \Delta u + \psi_h) \\ &\leq C\|\phi_h - \psi_h\|\,\|\Delta u + \varphi_h\|_{H^1(\Omega)} + \|\phi_h - \psi_h\|\|\Delta u + \psi\|. \end{aligned}$$

Therefore,

$$\|\phi_h - \psi_h\| \leq C\|\Delta u + \varphi_h\|_{H^1(\Omega)} + \|\Delta u + \psi_h\|.$$

In addition, we have that

$$\begin{aligned} & |u-u_h|_{H^1(\Omega)} + \|\Delta u + \phi_h\| \\ \leq\ & |u-v_h|_{H^1(\Omega)} + |v_h - u_h|_{H^1(\Omega)} + \|\Delta u + \psi_h\| + \|\psi_h - \phi_h\| \\ \leq\ & |u-v_h|_{H^1(\Omega)} + \|\Delta u + \psi_h\| + (1+C)\|\psi_h - \phi_h\|. \end{aligned}$$

Combination of the above two inequalities leads to

$$\begin{aligned} & |u-u_h|_{H^1(\Omega)} + \|\Delta u + \phi_h\| \\ \leq\ & |u-v_h|_{H^1(\Omega)} + (2+C)\|\Delta u + \psi_h\| + C(1+C)\|\Delta u + \varphi_h\|_{H^1(\Omega)}. \end{aligned}$$

Taking the infimum completes the proof. □

Given a family of regular triangulations $\mathcal{T}_h$ for Ω, we can derive the actual error estimate for Lagrange elements. The inverse estimate (2.3.7) implies

$$|\varphi_h|_{H^1(\Omega)} \leq \frac{C}{h}\|\varphi_h\|. \tag{4.32}$$

From (3.7) and (3.8), if $u \in H^{k+2}(\Omega)$, we have that

$$\inf_{v_h\in X_{0,h}} |u-v_h|_{H^1(\Omega)} \leq Ch^k|u|_{H^{k+1}(\Omega)}, \tag{4.33}$$

$$\inf_{\varphi_h\in X_h} \|\Delta u - \varphi_h\|_{H^1(\Omega)} \leq Ch^{k-1}|\Delta u|_{H^k(\Omega)}. \tag{4.34}$$

Theorem 4.3.8. *Let u be the solution of the biharmonic source problem such that $u \in H_0^2(\Omega)$. Let u_h and ϕ_h be the solution of the mixed finite element method using Lagrange elements. If $u \in H^{k+1}(\Omega)$, the following holds*

$$|u - u_h|_{H^1(\Omega)} + \|\Delta u + \phi_h\| \le Ch^{k-1}\left(|u|_{H^{k+1}(\Omega)} + |\Delta u|_{H^k(\Omega)}\right). \tag{4.35}$$

Proof. Let $(v_h, \psi_h) \in \mathcal{V}_h$ and $\varphi_h \in X_h$. We have that

$$\beta((v_h, \psi_h), \psi_h + \varphi_h) = 0$$

since $\psi_h + \varphi_h \in X_h$. Noting that

$$(\Delta u, \psi_h + \varphi_h) = -(\nabla u, \nabla(\psi_h + \varphi_h)),$$

we obtain

$$(\Delta u + \psi_h, \psi_h + \varphi_h) = (\nabla(v_h - u), \nabla(\psi_h + \varphi_h)).$$

Therefore,

$$\begin{aligned} |(\Delta u + \psi_h, \psi_h + \varphi_h)| &\le |u - v_h|_{H^1(\Omega)}|\psi_h + \varphi_h|_{H^1(\Omega)} \\ &\le \frac{C}{h}|u - v_h|_{H^1(\Omega)}\|\psi_h + \varphi_h\|, \end{aligned}$$

where we have used (4.32). It implies that

$$\begin{aligned} \|\psi_h + \varphi_h\|^2 &= (\varphi_h - \Delta u, \psi_h + \varphi_h) + (\Delta u + \psi_h, \psi_h + \varphi_h) \\ &\le \|\varphi_h - \Delta u\| \cdot \|\psi_h + \varphi_h\| + \frac{C}{h}|u - v_h|_{H^1(\Omega)}\|\psi_h + \varphi_h\|. \end{aligned}$$

Consequently, the following holds

$$\begin{aligned} \|\Delta u + \psi_h\| &\le \|\Delta u - \varphi_h\| + \|\psi_h + \varphi_h\| \\ &\le 2\|\Delta u - \varphi_h\| + \frac{C}{h}|u - v_h|_{H^1(\Omega)}, \end{aligned}$$

and, therefore,

$$\begin{aligned} &\inf_{(v_h,\psi_h)\in\mathcal{V}_h}\left(|u - v_h|_{H^1(\Omega)} + \|\Delta u + \psi_h\|\right) \\ \le\ &\left(1 + \frac{C}{h}\right)\inf_{v_h\in X_{0,h}}|u - v_h|_{H^1(\Omega)} + 2\inf_{\varphi_h\in X_h}\|\Delta u - \varphi_h\|. \end{aligned}$$

Combining the above inequality and Theorem 4.3.7, we obtain that

$$\begin{aligned} &|u - u_h|_{H^1(\Omega)} + \|\Delta u + \phi_h\| \\ \le\ &C\left(1 + \frac{C}{h}\right)\inf_{v_h\in X_{0,h}}|u - v_h|_{H^1(\Omega)} + 2\inf_{\varphi_h\in X_h}\|\Delta u - \varphi_h\|. \end{aligned}$$

Application of (4.33) and (4.34) proves (4.35). □

From above error estimates and Section 3 of [119], we have that

$$\|Tf - T_h f\| \leq Ch^2 \|Tf\|_{H^3(\Omega)} \tag{4.36}$$

and

$$\|Tf - T_h f\| \leq Ch^2 \|f\|. \tag{4.37}$$

The following results are also proved in [119]:

$$\begin{aligned}
\|(S - S_h)f\| &\leq Ch^{k-1}\|Tf\|_{H^{k+1}(\Omega)}, \\
\|(S - S_h)f\|_{H^1(\Omega)} &\leq Ch^{k-2}\|Tf\|_{H^{k+1}(\Omega)}, \\
\|(T - T_h)f\| &\leq Ch^{k}\|Tf\|_{H^{k+1}}, \\
\|(T - T_h)f\|_{H^1(\Omega)} &\leq Ch^{k}\|Tf\|_{H^{k+1}(\Omega)},
\end{aligned}$$

which imply

$$\|(S - S_h)|_{\overline{M}}\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega))} \leq Ch^{k-1}, \tag{4.38a}$$

$$\|(S - S_h)|_{\overline{M}}\|_{\mathcal{L}(L^2(\Omega),H^1(\Omega))} \leq Ch^{k-2}, \tag{4.38b}$$

$$\|(T - T_h)|_{\overline{M}}\|_{\mathcal{L}(L^2(\Omega),L^2(\Omega))} \leq Ch^{k}, \tag{4.38c}$$

$$\|(T - T_h)|_{\overline{M}}\|_{\mathcal{L}(L^2(\Omega),H^1(\Omega))} \leq Ch^{k}. \tag{4.38d}$$

Using the above inequalities and Theorems 4.3.2, 4.3.3, we actually proved the following result.

Theorem 4.3.9. *Let $k \geq 2$ and suppose the biharmonic eigenfunctions belong to $H^{k+1}(\Omega)$. Then*

$$|\lambda - \lambda_{j,h}| \leq Ch^{2k-2}$$

and

$$\|u - u_{j,h}\| \leq Ch^{k}.$$

4.3.3 Numerical Examples

We present some numerical results for the mixed finite element method using Lagrange elements. The first domain is the unit square. Since the domain is convex, the solution of the biharmonic problem is in $H^3(\Omega)$. According to the theory, the convergence order is 2. This is verified in Table 4.3, where we show the first and fourth biharmonic eigenvalues, relative errors, and convergence orders.

The second example is the L-shaped domain (see Table 4.4). In this case, the domain is non-convex. Therefore, the solution of the biharmonic source problem does not satisfy the regularity assumption for the mixed finite element. However, the mixed method seems to compute the correct eigenvalues. Interestingly, the first eigenvalue has a higher order of convergence.

h	1st	Rel. Err.	order
1/10	1.328932304009413e+03	-	-
1/20	1.303528784816815e+03	0.019488268681514	-
1/40	1.297093057464753e+03	0.004961654304619	1.9737
1/80	1.295474902459790e+03	0.001249082480788	1.9899
h	4th	Rel. Err.	order
1/10	1.277606032749956e+04	-	-
1/20	1.198145970594394e+04	0.066319183226183	-
1/40	1.177913324026064e+04	0.017176685377134	1.9489
1/80	1.172799489054373e+04	0.004360365961461	1.9779

Table 4.3: The first and fourth biharmonic eigenvalues of the unit square using the mixed finite element.

h	1st	Rel. Err.	order
1/10	6.938665720164172e+03	-	-
1/20	6.732382926149918e+03	0.030640383394268	-
1/40	6.695979460358917e+03	0.005436615510324	2.4946
1/80	6.694553228693430e+03	2.130435918223471e-04	4.6735
h	2nd	Rel. Err.	order
1/10	1.158131183074591e+04	-	-
1/20	1.120233107309107e+04	0.033830526448659	-
1/40	1.109137353731609e+04	0.010003949051186	1.7577
1/80	1.106347130171947e+04	0.002522014550016	1.9879

Table 4.4: The first and second biharmonic eigenvalues of the L-shaped domain using the mixed finite element.

4.4 The Morley Element

There exist many nonconforming elements for the biharmonic problem, e.g., the Morley element [204], Adini element [4], the Zienkiewicz triangle [26], etc. In this section, we present the Morley element for triangular meshes.

4.4.1 Abstract Theory

The following basic finite element convergence theory for nonconforming finite elements is adapted from [44]. Consider the variational problem

$$a(u, v) = f(v) \quad \text{for all } v \in V, \tag{4.39}$$

where $H_0^m(\Omega) \subset V \subset H^m(\Omega)$. The discrete problem is to find $u_h \in V_h$ such that

$$a_h(u_h, v_h) = f_h(v_h) \quad \text{for all } v_h \in V_h. \tag{4.40}$$

Note that, for nonconforming finite elements, $V_h \not\subset V$.

Let $\|\cdot\|_h$ be a mesh-dependent norm on V_h. For nonconforming finite element methods, the H^m-norm might not be defined for all $v_h \in V_h$ and thus one needs to use the mesh-dependent norm. We assume that $a_h(\cdot,\cdot)$ is defined for $v \in V$ and $v_h \in V_h$. Furthermore, we assume the coercivity and boundedness hold, i.e., there exist some positive constants α and C independent of h such that

$$a_h(v_h, v_h) \geq \alpha \|v_h\|_h^2 \qquad \text{for all } v_h \in V_h, \tag{4.41a}$$

$$|a_h(u, v_h)| \leq C\|u\|_h \|v_h\|_h \qquad \text{for all } u \in V + V_h, v_h \in V_h. \tag{4.41b}$$

The following two Strang Lemmas are needed for the convergence analysis for the Morley element.

Lemma 4.4.1. *(First Lemma of Strang, Section 3.1, [44]) Let u and u_h be the solutions of* (4.39) *and* (4.40)*, respectively. There exists a constant C independent of h such that*

$$\begin{aligned} \|u - u_h\| \quad \leq \quad & C\left(\inf_{v_h \in V_h}\left\{\|u - v_h\| + \sup_{w \in V_h} \frac{a(v_h, w_h) - a_h(v_h, w_h)}{\|w_h\|}\right\}\right. \\ & \left. + \sup_{w \in V_h} \frac{f(w_h) - f_h(w_h)}{\|w_h\|}\right). \end{aligned} \tag{4.42}$$

Proof. Let $v_h \in V_h$. By (4.40) and (4.41a), we have

$$\begin{aligned} \alpha\|u_h - v_h\|^2 \quad &\leq \quad a_h(u_h - v_h, u_h - v_h) \\ &= \quad a(u - v_h, u_h - v_h) + [a(v_h, u_h - v_h) - a_h(v_h, u_h - v_h)] \\ &\qquad + [a_h(u_h, u_h - v_h) - a(u, u_h - v_h)] \\ &= \quad a(u - v_h, u_h - v_h) + [a(v_h, u_h - v_h) - a_h(v_h, u_h - v_h)] \\ &\qquad - [f(u_h - v_h) - f_h(u_h - v_h)]. \end{aligned}$$

Dividing the above equation by $\|u_h - v_h\|$, we obtain that

$$\begin{aligned} \|u_h - v_h\| \quad \leq \quad & C\left(\|u - v_h\| + \frac{|a(v_h, u_h - v_h) - a_h(v_h, u_h - v_h)|}{\|u_h - v_h\|}\right. \\ & \left. + \frac{|f(u_h - v_h) - f_h(u_h - v_h)|}{\|u_h - v_h\|}\right). \end{aligned}$$

Since $v_h \in V_h$ is arbitrary and

$$\|u - u_h\| \leq \|u - v_h\| + \|u_h - v_h\|,$$

the inequality (4.42) follows immediately. $\square$

The second lemma of Strang (Section 3.1 of [44]) is as follows.

Lemma 4.4.2. *Let u and u_h be the solutions of* (4.39) *and* (4.40)*, respectively. In addition, we assume that $a_h(\cdot,\cdot)$ satisfies the coercivity and boundedness conditions* (4.41a) *and* (4.41b)*, respectively. There exists a constant C independent of h such that*

$$\|u-u_h\|_h \le C\left(\inf_{v_h\in V_h}\|u-v_h\|_h + \sup_{w_h\in V_h}\frac{|a_h(u,w_h)-f_h(w_h)|}{\|w_h\|_h}\right). \tag{4.43}$$

Proof. Let $v_h \in V_h$. Using (4.41a) and (4.41b), the following holds

$$\begin{aligned}\alpha\|u_h-v_h\|_h^2 &\le a_h(u_h-v_h,u_h-v_h)\\ &= a_h(u-v_h,u-v_h)+[f_h(u_h-v_h)-a_h(u,u_h-v_h)].\end{aligned}$$

Dividing by $\|u_h-v_h\|$ and setting $w_h := u_h - v_h$, we obtain

$$\|u_h-v_h\|_h \le \frac{1}{\alpha}\left(C\|u-v_h\|_h + \frac{|a_h(u,w_h)-f_h(w_h)|}{\|w_h\|_h}\right).$$

Then (4.43) follows the triangle inequality $\|u-u_h\| \le \|u-v_h\| + \|u_h-v_h\|$. □

Remark 4.4.1. *The first term and second term on the right-hand side of* (4.43) *are called the approximation error and the consistency error, respectively.*

4.4.2 The Morley Element

The Morley element is defined for triangular meshes. It uses polynomial spaces of order 2. Let K be a triangle with vertices z_1, z_2, and z_3. For $\mathcal{N} = \{N_1,\ldots,N_6\}$, 3 degrees of freedom are the values at the three vertices and 3 degrees of freedom are the values of the normal derivatives at the midpoints of three edges, i.e.,

$$\begin{aligned}&v(z_i), i=1,2,3,\\ &\frac{\partial v(z_{12})}{\partial\nu}, \frac{\partial v(z_{23})}{\partial\nu}, \frac{\partial v(z_{13})}{\partial\nu}, z_{12}=\frac{z_1+z_2}{2}, z_{23}=\frac{z_2+z_3}{2}, z_{13}=\frac{z_3+z_1}{2}.\end{aligned}$$

Definition 4.4.1. *The Morley element is the following triple $(K,\mathcal{P},\mathcal{N})$:*

1. *K is a triangle,*
2. *$\mathcal{P}=\mathcal{P}_2(K)$, the space of quadratic polynomials,*
3. *$\mathcal{N}$ is the set of degrees of freedom given by*

$$v(z_i), i=1,2,3, \quad \textit{and} \quad \frac{\partial v}{\partial\nu}(z_{ij}), 1\le i<j\le 3, \tag{4.44}$$

where z_{ij} is the middle point of the edge whose end points are z_i and z_j.

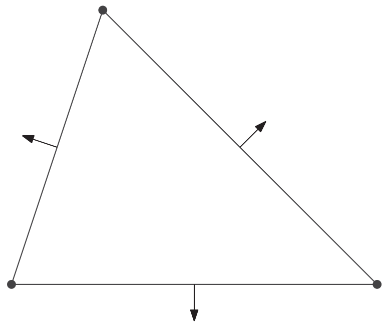


Figure 4.2: The Morley element: 6 degrees of freedoms are three values at the three vertices and three values of the normal derivatives at the midpoints of edges.

For $u, v \in L^2(\Omega)$ such that $u|_K, v|_K \in H^m(K), K \in \mathcal{T}_h$, the discrete inner product is defined as

$$(u, v)_{m,h} = \sum_{K \in \mathcal{T}_h} (u, v)_{H^m(K)},$$

which induces the discrete norm

$$\|u\|_{m,h} = (u, u)_{m,h}^{1/2}.$$

The associated mesh-dependent semi-norm is defined as

$$|u|_{m,h} = \left(\sum_{K \in \mathcal{T}_h} |u|_{H^m(K)}^2 \right)^{1/2}.$$

Let V_h be the Morley finite element space. For the clamped plate boundary conditions, we set all the degrees of freedom on $\partial\Omega$ to be zero to obtain the finite element space $V_{0,h}$. We have the following interpolation result and refer the readers to [238] for its proof.

Lemma 4.4.3. *(Lemma 6 of [238]) Let $v_h \in V_{0,h}$. There exist functions $w_{h,k} \in H_0^1(\Omega), 0 \le k \le 2$, such that $w_{h,k}|_K \in C^\infty(K)$ for all $K \in \mathcal{T}_h$ and*

$$|v_h - w_{h,0}|_{m,h} \le Ch^{2-m}|v_h|_{2,h}, \quad 0 \le m \le 2, \tag{4.45}$$

$$\left| \frac{\partial v_h}{\partial x_k} - w_{h,k} \right|_{m,h} \le Ch^{1-m}|v_h|_{2,h}, \quad 0 \le m \le 1, 1 \le k \le 2, \tag{4.46}$$

where C is a constant independent of h.

Let $v, w \in H^2(\Omega) + V_h$. We define

$$a_h(v,w) = \sum_{K\in\mathcal{T}_h} \int_K \sum_{i,j=1}^{2} \frac{\partial^2 v}{\partial x_i \partial x_j} \frac{\partial^2 w}{\partial x_i \partial x_j}. \tag{4.47}$$

The weak formulation for the discrete biharmonic problem is to find $u_h \in V_{0,h}$ such that

$$a_h(u_h, v_h) = (f, v_h) \qquad \text{for all } \ v_h \in V_{0,h}. \tag{4.48}$$

It is easy to verify that

$$a_h(v_h, v_h) \geq \alpha \|v_h\|_{2,h}^2 \qquad \text{for all } v_h \in V_{0,h}, \tag{4.49a}$$

$$|a_h(u_h, v_h)| \leq C\|u_h\|_{2,h}\|v_h\|_{2,h} \qquad \text{for all } u \in H^2(\Omega) + V_h, v_h \in V_h. \tag{4.49b}$$

By the Lax-Milgram Lemma 1.3.1, (4.48) has a unique solution.

The analysis of the Morley element was discussed in many papers, for example, [215, 225, 238]. The following error analysis is adapted from [238].

Let $K \in \mathcal{T}_h$ and I_K be the interpolation operator. The following lemmas give the interpolation property of the Morley elements (Lemma 3 of [238]).

Lemma 4.4.4. *There exists a constant C independent of h such that*

$$|u - I_K u|_{H^m(K)} \leq Ch^{3-m}|u|_{H^3(K)}, \quad 0 \leq m \leq 3, u \in H^3(K), K \in \mathcal{T}_h. \tag{4.50}$$

Lemma 4.4.5. *There exists a constant C independent of h such that, for*

$$v \in H^3(\Omega) \cap H_0^2(\Omega),$$

it holds that

$$|a_h(v, v_h) - (\triangle^2 v, v_h)| \leq Ch\left(|v|_{H^3(\Omega)} + h\|\triangle^2 v\|\right)|v_h|_{2,h} \tag{4.51}$$

for all $v_h \in V_{0,h}$.

Proof. Let $v \in H^3(\Omega) \cap H_0^2(\Omega)$ and $v_h \in V_{0,h}$. We have that

$$a_h(v, v_h) - (\triangle^2 v, v_h) = a_h(v, v_h) - (\triangle^2 v, w_h) + (\triangle^2 v, w_h - v_h).$$

For the second term, the following holds

$$|(\triangle^2 v, w_h - v_h)| \leq Ch^2\|\triangle^2 v\|\,|v_h|_{2,h}. \tag{4.52}$$

For the first term, integration by parts leads to

$$\begin{aligned} & a_h(v, v_h) - (\triangle^2 v, w_h) \\ = & \sum_{i,j=1}^{2} \sum_{K\in\mathcal{T}_h} \int_K \left(\frac{\partial^2 v}{\partial x_i \partial x_j} \cdot \frac{\partial v_h^2}{\partial x_i \partial x_j} + \frac{\partial^3 v}{\partial x_i \partial x_j^2} \cdot \frac{\partial v_h}{\partial x_i} \right) \mathrm{d}x \\ & + \sum_{i,j=1}^{2} \sum_{K\in\mathcal{T}_h} \int_K \frac{\partial^3 v}{\partial x_i \partial x_j^2} \cdot \frac{\partial (w_h - v_h)}{\partial x_i} \,\mathrm{d}x. \end{aligned}$$

Using the inequality

$$\left| \sum_{K \in \mathcal{T}_h} \int_K \frac{\partial^3 v}{\partial x_i \partial x_j^2} \cdot \frac{\partial (w_h - v_h)}{\partial x_i} \, \mathrm{d}x \right| \leq Ch |v|_{H^3(\Omega)} |v_h|_{2,h},$$

we obtain that

$$|a_h(v, v_h) - (\triangle^2 v, w_h)| \leq Ch |v|_{H^3(\Omega)} |v_h|_{2,h}. \tag{4.53}$$

Combination of (4.51) and (4.53) proves the lemma. □

Lemma 4.4.6. *For any $v_h \in V_{0,h}$, it holds that*

$$|v_h|_{2,h} \leq \|v_h\|_{2,h} \leq C |v_h|_{2,h} \tag{4.54}$$

for some constant C independent of h.

Proof. For $v_h \in V_{0,h}$, there exists functions $w_{h,k} \in H_0^1(\Omega), 0 \leq k \leq n$, such that Lemma 4.4.3 holds. It follows that

$$\begin{aligned} \|v_h\|_{0,h} &\leq \|v_h - w_{h,0}\|_{0,h} + \|w_{h,0}\| \\ &\leq C \left(|v_h|_{2,h} + |w_{h,0}|_{H^1(\Omega)} \right) \\ &\leq C \left(|v_h|_{2,h} + |v_h|_{1,h} \right), \end{aligned}$$

and

$$\begin{aligned} |v_h|_{1,h} &\leq \sum_{k=1}^{2} \left(\left| \frac{\partial v_h}{\partial x_k} - w_{h,k} \right|_{0,h} + \|w_{h,k}\| \right) \\ &\leq C \left(|v_h|_{2,h} + \sum_{k=1}^{2} \|w_{h,k}\|_{H^1(\Omega)} \right) \\ &\leq C |v_h|_{2,h}. \end{aligned}$$

Thus $\|v_h\|_{2,h} \leq C |v_h|_{2,h}$ and the proof is complete. □

The following theorem provides the error estimate of the source problem using the Morley element.

Theorem 4.4.7. *(Theorem 2 of [238]) Let u be the solution of the biharmonic equation* (4.4) *and u_h be the discrete solution of* (4.48) *using the Morley element. Then there exists a constant C independent of h such that*

$$\|u - u_h\|_{2,h} \leq Ch \left(|u|_{H^3(\Omega)} + h\|f\| \right) \tag{4.55}$$

provided the solution $u \in H^3(\Omega)$.

Proof. Applying the second Strang Lemma 4.4.2, we have that

$$|u - u_h|_{2,h} \le C \left(\inf_{w_h \in V_{0,h}} |u - w_h|_{2,h} + \sup_{w_h \in V_{0,h}, w_h \neq 0} \frac{a_h(u, w_h) - (f, w_h)}{|w_h|_{2,h}} \right).$$

Employing Lemma 4.4.6 to replace $|u - u_h|_{2,h}$ with $\|u - u_h\|_{2,h}$, and using (4.50) and (4.51), we immediately obtain (4.55). □

The Morley element method for the biharmonic eigenvalue problem can be stated as follows. Find $\lambda_h \in \mathbb{R}$ and $u_h \in V_h$ such that

$$a_h(u_h, v_h) = \lambda_h(u_h, v_h) \quad \text{for all } v_h \in V_h. \tag{4.56}$$

The following theorem on the error estimate of eigenvalues computed by the Morley element holds.

Theorem 4.4.8. *Let Ω be a convex polygonal domain. Then we have that*

$$|\lambda - \hat{\lambda}_{j,h}| \le Ch^2.$$

Proof. Let $T : L^2(\Omega) \to L^2(\Omega)$ be the solution operator for the biharmonic equation and $T_h : L^2(\Omega) \to L^2(\Omega)$ be the discrete solution operator using the Morley element. It is clear that T is compact and self-adjoint. Since Ω is a convex polygonal domain, the regularity result of biharmonic equation (4.9) implies that

$$\|u\|_{H^3(\Omega)} \le C\|f\|_{H^{-1}(\Omega)} \le C\|f\|.$$

From Theorem 4.4.7, we have that

$$\|Tf - T_h f\| \le \|u - u_h\|_{2,h} \le Ch \left(|u|_{H^3(\Omega)} + h\|f\| \right) \le Ch\|f\|.$$

Then the result follows Theorem 1.4.4 immediately. □

4.4.3 Numerical Examples

We present some numerical results for the Morley element. Again the unit square and the L-shaped domain are chosen as test domains. In Table 4.5, we show the computed biharmonic eigenvalues of the unit square. In this case, the convergence rate is roughly $O(h^2)$. Note that the eigenvalues are approximated from below (the computed eigenvalues are smaller than the exact eigenvalues).

In Table 4.6, we show the computed eigenvalues of the L-shaped domain. Note that the domain is non-convex. Thus the theory does not cover this case.

h	1st	Rel. Err.	order
1/10	1.191925453955360e+03	-	-
1/20	1.266623619822192e+03	0.058974240411937	-
1/40	1.287647517746532e+03	0.016327370366957	1.8527
1/80	1.293096908047353e+03	0.004214216480534	1.9540
1/160	1.294473622814774e+03	0.001063532499355	1.9864
h	4th	Rel. Err.	order
1/10	9.744622766202596e+03	-	-
1/20	1.111407850629441e+04	0.123218109294102	-
1/40	1.155245431175843e+04	0.037946551757216	1.6992
1/80	1.167055340203249e+04	0.010222857159786	1.8922
1/160	1.170070118066749e+04	0.002576578802372	1.9883

Table 4.5: The first and fourth biharmonic eigenvalues of the unit square using the Morley element.

h	1st	Rel. Err.	order
1/10	5.329047684697986e+03	-	-
1/20	6.184595525129695e+03	0.138335294031014	-
1/40	6.513273566166486e+03	0.050462803027977	1.4549
1/80	6.629917681896610e+03	0.017593599396962	1.5202
1/160	6.673100865519929e+03	0.006471231964505	1.4429
h	2nd	Rel. Err.	order
1/10	8.892614946897947e+03	-	-
1/20	1.037350341389021e+04	0.142756830350038	-
1/40	1.086642778617978e+04	0.045362135744047	1.6540
1/80	1.100464149022314e+04	0.012559582623946	1.8527
1/160	1.104140078314963e+04	0.003329223678085	1.9155

Table 4.6: The first and second biharmonic eigenvalues of the L-shaped domain.

4.5 A Discontinuous Galerkin Method

For the biharmonic equation, we have seen that the H^2-conforming Argyris element method is complicated and involves many degrees of freedom, the mixed method puts a restriction on the domain, and the nonconforming Morley element cannot capture the smooth solution effectively.

In this section, we present a C^0 interior penalty discontinuous Galerkin method (C^0 IPG) by Brenner et al. [52]. We will study several biharmonic eigenvalue problems of plate vibration and buckling with three types of boundary conditions. We show that C^0 IPG converges for all three types of boundary conditions. At the end of the section, we compare the performance of the C^0 IPG method, the Argyris C^1 finite element method, the Ciarlet-Raviart mixed finite element method, and the Morley nonconforming finite element method.

Note that some numerical results for a related C^0 discontinuous Galerkin method were presented in [243] for the plate vibration and buckling problems on a square with the boundary conditions of simply supported plates. However the convergence of the method for the eigenvalue problems was not addressed.

4.5.1 Biharmonic Eigenvalue Problems

We consider the biharmonic eigenvalue problems of plate vibration and buckling with three types of boundary conditions.

Clamped Plate (CP)

$$u = \frac{\partial u}{\partial \nu} = 0 \qquad \text{on } \partial\Omega. \tag{4.57}$$

Simply Supported Plate (SSP)

$$u = \triangle u = 0 \qquad \text{on } \partial\Omega. \tag{4.58}$$

Cahn-Hilliard Type (CH)

$$\frac{\partial u}{\partial \nu} = \frac{\partial \triangle u}{\partial \nu} = 0 \qquad \text{on } \partial\Omega. \tag{4.59}$$

The biharmonic eigenvalue problems for plate vibration are to find $u \neq 0$ and $\lambda \in \mathbb{R}$ such that

$$\Delta^2 u = \lambda u \qquad \text{in } \Omega$$

together with the boundary conditions (4.57), (4.58), or (4.59). We will refer to them as the V-CP problem, the V-SSP problem, and the V-CH problem, respectively. Note that the V-CP problem is the biharmonic eigenvalue problem discussed in the previous sections.

The biharmonic eigenvalue problem for plate buckling is to find $u \neq 0$ and $\lambda \in \mathbb{R}$ such that

$$\Delta^2 u = -\lambda \Delta u \qquad \text{in } \Omega$$

together with the boundary conditions (4.57), (4.58), or (4.59). We will refer to them as the B-CP problem, the B-SSP problem, and the B-CH problem, respectively.

Let the bilinear form $a(\cdot,\cdot)$ be defined by

$$a(u,v) = \int_\Omega D^2 u : D^2 v \, \mathrm{d}x, \tag{4.60}$$

where

$$D^2 u : D^2 v = \sum_{i,j=1}^{2} u_{x_i x_j} v_{x_i x_j}$$

is the Frobenius inner product of the Hessian matrices of u and v. We also define a bilinear form $b(\cdot,\cdot)$:

$$b(u,v) = \begin{cases} (u,v) = \displaystyle\int_\Omega uv \, \mathrm{d}x & \text{for plate vibration,} \\ (\nabla u, \nabla v) = \displaystyle\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x & \text{for plate buckling.} \end{cases} \tag{4.61}$$

Remark 4.5.1. *The bilinear form $a(\cdot,\cdot)$ is different from* (4.3)*. It is easy to show that they are equivalent on $H_0^2(\Omega) \times H_0^2(\Omega)$. The definition of* (4.60) *makes the error analysis of C^0 IPG simpler.*

The weak formulation of the biharmonic eigenvalue problem is to seek $(\lambda, u) \in \mathbb{R} \times V$ such that $u \neq 0$ and

$$a(u,v) = \lambda b(u,v) \qquad \text{for all } v \in V, \tag{4.62}$$

where

$$V = H_0^2(\Omega) \tag{4.63}$$

for V-CP and B-CP,

$$V = H^2(\Omega) \cap H_0^1(\Omega) \tag{4.64}$$

for V-SSP and B-SSP, and

$$V = \left\{ v \in H^2(\Omega) : \frac{\partial v}{\partial \nu} = 0 \text{ on } \partial\Omega \text{ and } (v,1) = 0 \right\} \tag{4.65}$$

for V-CH and B-CH.

Remark 4.5.2. *Since the bilinear form $a(\cdot,\cdot)$ is symmetric positive-definite on V for all three types of boundary conditions, the biharmonic eigenvalues are positive. Note that we have excluded the trivial eigenvalue 0 from the CH problem by imposing the zero mean constraint.*

4.5.2 C^0 Interior Penalty Galerkin Method

Let $\mathcal{T}_h$ be a regular triangulation of Ω with mesh size h and $\tilde{V}_h \subset H^1(\Omega)$ be the Lagrange finite element space of order $k \geq 2$ associated with $\mathcal{T}_h$. Let $\mathcal{E}_h$ be the set of the edges in $\mathcal{T}_h$. For an edge $e \in \mathcal{E}_h$ that is the common edge of two adjacent triangles $K_\pm \in \mathcal{T}_h$ and for $v \in \tilde{V}_h$, we define the jump of the flux to be

$$[\![\partial v/\partial \nu_e]\!] = \frac{\partial v_{K_+}}{\partial \nu_e}\bigg|_e - \frac{\partial v_{K_-}}{\partial \nu_e}\bigg|_e,$$

where ν_e is the unit normal pointing from K_- to K_+. We let

$$\frac{\partial^2 v}{\partial \nu_e^2} = \nu_e \cdot (D^2 v)\nu_e$$

and define the average normal-normal derivative to be

$$\left\{\!\!\left\{\frac{\partial^2 v}{\partial \nu_e^2}\right\}\!\!\right\} = \frac{1}{2}\left(\frac{\partial^2 v_{K_+}}{\partial \nu_e^2} + \frac{\partial^2 v_{K_-}}{\partial \nu_e^2}\right).$$

For $e \in \partial\Omega$, we take ν_e to be the unit outward normal and define

$$[\![\partial v/\partial \nu_e]\!] = -\frac{\partial v}{\partial \nu_e} \quad \text{and} \quad \left\{\!\!\left\{\frac{\partial^2 v}{\partial \nu_e^2}\right\}\!\!\right\} = \frac{\partial^2 v}{\partial \nu_e^2}.$$

Let $\mathbb{R}_+$ be the set of positive real numbers. The C^0 IPG method for the biharmonic eigenvalue problem is to find $(\lambda_h, u_h) \in \mathbb{R}_+ \times V_h$ such that $u_h \neq 0$ and

$$a_h(u_h, v) = \lambda_h b(u_h, v) \qquad \text{for all } v \in V_h, \tag{4.66}$$

where the choices of V_h and $a_h(\cdot,\cdot)$ depend on the boundary conditions.

1. For the CP boundary condition the choices for V_h and $a_h(\cdot,\cdot)$ are given by

$$V_h = \tilde{V}_h \cap H_0^1(\Omega), \tag{4.67}$$

$$\begin{aligned} a_h(w,v) = \sum_{K\in\mathcal{T}_h} \int_K & D^2 w : D^2 v \,\mathrm{d}x \\ & + \sum_{e\in\mathcal{E}_h} \int_e \left\{\!\!\left\{\frac{\partial^2 w}{\partial \nu_e^2}\right\}\!\!\right\} \left[\!\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\!\right] + \left\{\!\!\left\{\frac{\partial^2 v}{\partial \nu_e^2}\right\}\!\!\right\} \left[\!\!\left[\frac{\partial w}{\partial \nu_e}\right]\!\!\right] \mathrm{d}s \\ & + \sigma \sum_{e\in\mathcal{E}_h} \frac{1}{|e|} \int_e \left[\!\!\left[\frac{\partial w}{\partial \nu_e}\right]\!\!\right] \left[\!\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\!\right] \mathrm{d}s, \end{aligned} \tag{4.68}$$

   where $\sigma > 0$ is a (sufficiently large) penalty parameter.

2. For the SSP boundary condition we use the same V_h in (4.67) and the bilinear form

$$\begin{aligned} a_h(w,v) = \sum_{K\in\mathcal{T}_h} \int_K & D^2 w : D^2 v \,\mathrm{d}x \\ & + \sum_{e\in\mathcal{E}_h^i} \int_e \left\{\!\!\left\{\frac{\partial^2 w}{\partial \nu_e^2}\right\}\!\!\right\} \left[\!\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\!\right] + \left\{\!\!\left\{\frac{\partial^2 v}{\partial \nu_e^2}\right\}\!\!\right\} \left[\!\!\left[\frac{\partial w}{\partial \nu_e}\right]\!\!\right] \mathrm{d}s \\ & + \sigma \sum_{e\in\mathcal{E}_h^i} \frac{1}{|e|} \int_e \left[\!\!\left[\frac{\partial w}{\partial \nu_e}\right]\!\!\right] \left[\!\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\!\right] \mathrm{d}s, \end{aligned} \tag{4.69}$$

   where $\mathcal{E}_h^i$ is the set of the edges interior to Ω.

3. For the CH boundary condition we use the same bilinear form $a_h(\cdot,\cdot)$ defined in (4.68) and take

$$V_h = \left\{ v \in \tilde{V}_h : (v, 1) = 0 \right\}. \tag{4.70}$$

The convergence of the C^0 IPG method for these eigenvalue problems is based on the convergence of the C^0 IPG method for the corresponding source problems.

Let W be the space $L^2(\Omega)$ for the plate vibration problems, the space $H_0^1(\Omega)$ for the B-CP and B-SSP problems, and the space $\{v \in H^1(\Omega) : (v,1) = 0\}$ for the B-CH problem. We will denote by $\|f\|_b$ the norm induced by the bilinear form $b(\cdot,\cdot)$ defined in (4.61), i.e.,

$$\|f\|_b^2 = b(f,f).$$

Given $f \in W$, the weak formulation for the source problem is to find $u \in V$ such that

$$a(u,v) = b(f,v) \qquad \text{for all } v \in V, \tag{4.71}$$

where the bilinear form $a(\cdot,\cdot)$ is defined in (4.60). For the V-CH source problem, we also assume that f satisfies the constraint $(f,1) = 0$.

The corresponding C^0 IPG method for (4.71) is to find $u_h \in V_h$ such that

$$a_h(u_h,v) = b(f,v) \qquad \text{for all } v \in V_h, \tag{4.72}$$

where V_h and $a_h(\cdot,\cdot)$ are defined by

1. Equations (4.67) and (4.68), respectively, for the CP boundary conditions,
2. Equations (4.67) and (4.69), respectively, for the SSP boundary conditions, and
3. Equations (4.70) and (4.68), respectively, for the CH boundary conditions.

The following lemma summarizes the results for the source problems obtained in [55, 53, 50].

Lemma 4.5.1. *The biharmonic source problem* (4.71) *and the discrete source problem* (4.72) *are uniquely solvable for the boundary conditions of CP, SSP, and CH. In addition, there exists $\beta > 0$ such that*

$$\|u - u_h\|_h \le Ch^{\beta}\|f\|_b, \tag{4.73}$$

$$\|u - u_h\|_b \le Ch^{2\beta}\|f\|_b, \tag{4.74}$$

where $u \in V$ and $u_h \in V_h$ are the solutions of (4.71) *and* (4.72)*, respectively. The mesh-dependent energy norm $\|\cdot\|_h$ is defined by*

$$\|v\|_h^2 = \sum_{K\in\mathcal{T}_h} |v|_{H^2(K)}^2 + \sum_{e\in\mathcal{E}_h} |e|^{-1}\|[\![\partial v/\partial \nu_e]\!]\|_{L^2(e)}^2 \tag{4.75}$$

for the boundary conditions of CP and CH, and

$$\|v\|_h^2 = \sum_{K\in\mathcal{T}_h} |v|_{H^2(K)}^2 + \sum_{e\in\mathcal{E}_h^i} |e|^{-1}\|[\![\partial v/\partial \nu_e]\!]\|_{L^2(e)}^2 \tag{4.76}$$

for the boundary conditions of SSP.

Remark 4.5.3. *Let V be the Sobolev space for the biharmonic problem defined in* (4.63), (4.64), *or* (4.65) *and V_h be the corresponding finite element space. Then the discrete norm $\|\cdot\|_h$ defined by* (4.75) *is a norm on the space $V+V_h$ for the boundary conditions of CP and CH, and $\|\cdot\|_h$ defined by* (4.76) *is a norm on the space $V+V_h$ for the boundary conditions of SSP. Moreover in all three cases we have a Poincaré-Friedrichs inequality* [*56*]

$$\|v\|_b \leq C\|v\|_h \qquad \textit{for all } v \in V+V_h. \tag{4.77}$$

Remark 4.5.4. *The exponent β in* (4.73) *and* (4.74) *is given by $\beta = \min(\alpha, k-1)$, where α is the index of regularity that appears in the elliptic regularity estimate* [*33*, *139*, *140*]

$$\|u\|_{H^{2+\alpha}(\Omega)} \leq C\|f\|_b$$

for the solution u of the source problem (4.71). *It is determined by the angles at the corners of Ω and the boundary conditions. For the CP boundary conditions* (4.57), *α belongs to $(\frac{1}{2}, 1]$ and $\alpha = 1$ if Ω is convex. For the SSP boundary conditions* (4.58) *and the CH boundary conditions* (4.59), *α belongs to $(0, 1]$ in general, $\alpha = 2$ for a rectangular domain, and α is any number strictly less than $1/3$ for an L-shaped domain.*

For the convergence analysis of the C^0 IPG method for the biharmonic eigenvalue problems, we need two solution operators

$$T : W \to V\,(\subset W)$$

and

$$T_h : W \to V_h\,(\subset W)$$

on the Hilbert space $\big(W, b(\cdot,\cdot)\big)$, which are defined by

$$\begin{aligned} a(Tf, v) &= b(f, v) \qquad \text{for all } v \in V, \\ a_h(T_h f, v) &= b(f, v) \qquad \text{for all } v \in V_h. \end{aligned}$$

Note that (4.62) is equivalent to $Tu = (1/\lambda)u$, (4.66) is equivalent to $T_h u_h = (1/\lambda_h)u_h$, and the estimates (4.73)–(4.74) can be rewritten as

$$\|(T-T_h)f\|_h \leq Ch^{\beta}\|f\|_b \qquad \text{for all } f \in W, \tag{4.78}$$

$$\|(T-T_h)f\|_b \leq Ch^{2\beta}\|f\|_b \qquad \text{for all } f \in W. \tag{4.79}$$

The operator T is symmetric, positive-definite, and compact due to the compact embedding of V into W. Therefore the spectrum of T consists of a sequence of positive eigenvalues $\mu_1 \geq \mu_2 \geq \ldots$ decreasing to 0, and the numbers $\lambda_j = 1/\mu_j$ are the biharmonic eigenvalues, which have a limit ∞.

The convergence of the C^0 IPG method for the biharmonic eigenvalue problems follows from (4.78), (4.79), and the classical spectral approximation theory in Section 1.4 (see also [165, Section 5.4.3] and [23, Section 2.7]).

Theorem 4.5.2. *Let* $0 < \lambda_1 \leq \lambda_2 \leq \ldots$ *be the biharmonic eigenvalues,*

$$\lambda = \lambda_j = \ldots = \lambda_{j+m-1}$$

be a biharmonic eigenvalue with multiplicity m*, and* V_λ *be the corresponding* m*-dimensional eigenspace. Let* $0 < \lambda_{h,1} \leq \lambda_{h,2} \leq \ldots$ *be the discrete eigenvalues obtained by the* C^0 *IPG method. Then we have, as* $h \to 0$*,*

$$|\lambda_{h,l} - \lambda| \leq Ch^{2\beta}, \quad l = j, j+1, \ldots, j+m-1.$$

In addition, if $V_\lambda \subset V$ *is the space spanned by the eigenfunctions corresponding to the biharmonic eigenvalues* $\lambda_j, \ldots, \lambda_{j+m-1}$ *and* $V_{h,\lambda} \subset V_h$ *is the space spanned by the eigenfunctions corresponding to the discrete eigenvalues* $\lambda_{h,j}, \ldots, \lambda_{h,j+m-1}$*, then we have, as* $h \to 0$*,*

$$\delta(V_\lambda, V_{h,\lambda}) = O(h^\beta)$$

in $(W, \|\cdot\|_h)$ *and*

$$\delta(V_\lambda, V_{h,\lambda}) = O(h^{2\beta})$$

in $(W, \|\cdot\|_b)$*.*

Remark 4.5.5. *We can apply the classical theory because we use the Hilbert space* $\big(W, b(\cdot,\cdot)\big)$ *and* V_h *is a subspace of* W*. This would not be possible if we use the space* V *in* (4.63)–(4.65).

Remark 4.5.6. *The convergence of the method in* [243] *for eigenvalue problems can also be established analogously by the classical spectral approximation theory.*

4.5.3 Numerical Examples

In this section we present numerical results for two domains to illustrate the performance of the C^0 IPG method for the biharmonic eigenvalue problems. The penalty parameter σ is taken to be 50 in all the computations. Discussion on how to choose σ for the C^0 IPG can be found in [163].

We list some facts which are useful for the validation of different numerical methods. Recall that for the V-CP problem on the unit square, an accurate lower bound for the first eigenvalue is

$${}^{-}\lambda_1^{\text{V-CP}} = 1294.933940$$

given by Wieners [245]. An accurate upper bound is

$${}^{+}\lambda_1^{\text{V-CP}} = 1294.9339796$$

given by Bjørstad and Tjøstheim [32].

For the V-SSP problem on convex domains, the biharmonic eigenvalues are just the squares of the eigenvalues for the Laplace operator with the homogeneous Dirichlet boundary condition. The V-SSP eigenvalues for the unit square are therefore given by

$$4\pi^4, 25\pi^4, 25\pi^4, 64\pi^4, 100\pi^4, 100\pi^4, \ldots \tag{4.80}$$

with the corresponding eigenfunctions

$$\begin{aligned}
&\sin(\pi x_1)\sin(\pi x_2),\\
&\sin(2\pi x_1)\sin(\pi x_2),\\
&\sin(\pi x_1)\sin(2\pi x_2),\\
&\sin(2\pi x_1)\sin(2\pi x_2),\\
&\sin(3\pi x_1)\sin(\pi x_2),\\
&\sin(\pi x_1)\sin(3\pi x_2),\\
&\ldots
\end{aligned}$$

Similarly, for the V-CH problem on convex domains, the positive biharmonic eigenvalues are given by the square of the positive eigenvalues for the Laplace operator with the homogeneous Neumann boundary condition. Therefore the V-CH eigenvalues on the unit square are given by

$$\pi^4, \pi^4, 4\pi^4, 16\pi^4, 16\pi^4, 25\pi^4, 25\pi^4, \ldots \tag{4.81}$$

with the corresponding eigenfunctions

$$\begin{aligned}
&\cos(\pi x_1),\\
&\cos(\pi x_2),\\
&\cos(\pi x_1)\cos(\pi x_2),\\
&\cos(2\pi x_1),\\
&\cos(2\pi x_2),\\
&\cos(2\pi x_1)\cos(\pi x_2),\\
&\cos(\pi x_1)\cos(2\pi x_2),\\
&\ldots
\end{aligned}$$

We also consider the L-shaped domain defined as

$$(0,1)\times(0,1)\setminus[1/2,1)\times(0,1/2].$$

For the V-SSP problem on the L-shaped domain, some of the eigenvalues are from (4.80) because the restrictions of the corresponding eigenfunctions on the L-shaped domain also satisfy the boundary conditions in (4.58). For example, the eigenfunction for the unit square

$$\sin(2\pi x_1)\sin(2\pi x_2)$$

is also an eigenfunction for the L-shaped domain with the same eigenvalue. Similarly, for the V-CH problem, the eigenfunctions,

$$\cos(2\pi x_1) \quad \text{and} \quad \cos(2\pi x_2),$$

for the unit square are also eigenfunctions for the L-shaped domain.

For the B-CP problem on the unit square, an accurate approximation

$$\lambda_1^{\text{B-CP}} \approx 52.34469116$$

for the first eigenvalue is given in [32]. For the B-SSP problem on the unit square the first eigenvalue is the simple eigenvalue

$$\lambda_1^{\text{B-SSP}} = 2\pi^2 \approx 19.73920880$$

with eigenfunction $\sin(\pi x_1)\sin(\pi x_2)$. For the B-CH problem on the unit square, the first eigenvalue is the double eigenvalue $\pi^2 \approx 9.869604401$ whose eigenspace is spanned by the functions $\cos(\pi x_1)$ and $\cos(\pi x_2)$.

In Table 4.7 we display the first biharmonic eigenvalues for the V-CP problem, the V-SSP problem, and the V-CH problem, computed by the C^0 IPG method on a series of unstructured meshes generated by uniform refinement. We recall that the first V-CP eigenvalue obtained in [245] is 1294.93398. The first V-SSP eigenvalue is

$$\lambda_1^{\text{V-SSP}} = 4\pi^4 \approx 389.6363$$

and the first V-CH eigenvalue is

$$\lambda_1^{\text{V-CH}} = \pi^4 \approx 97.4091.$$

Therefore the C^0 IPG method provides good approximations in all three cases.

h	CP(1)	SSP(1)	CH(1)
1/10	1,377.1366	395.1181	98.2067
1/20	1,318.5091	391.1631	97.6410
1/40	1,301.3047	390.0452	97.4711
1/80	1,296.5904	389.7422	97.4251

Table 4.7: The first V-CP, V-SSP, and V-CH eigenvalues of the unit square.

The second domain is the L-shaped domain. In Table 4.8 we present the first biharmonic plate vibration eigenvalues computed by the C^0 IPG method on a series of uniformly refined unstructured meshes. We also include the results for the third eigenvalues of V-SSP and V-CH, whose exact values are

$$\lambda_3^{\text{V-SSP}} = 64\pi^4 \approx 6234.1818$$

and

$$\lambda_3^{\text{V-CH}} = 16\pi^4 \approx 1558.5455,$$

respectively. They are approximated correctly with less than 1% relative error at the finest meshes. Comparing Table 4.7 and Table 4.8, we see that the convergence for the L-shaped domain is slower.

h	CP(1)	SSP(1)	SSP(3)	CH(1)	CH(3)
1/10	7,834.5030	2,870.9514	6,327.5449	177.4750	1,603.9472
1/20	7,104.1915	2,748.1841	6,573.0063	174.1519	1,571.3380
1/40	6,854.7447	2,693.7255	6,259.2682	172.3741	1,562.0031
1/80	6,763.0157	2,663.3927	6,240.6958	171.1519	1,559.4471

Table 4.8: The first V-CP, V-SSP, and V-CH eigenvalues of the L-shaped domain.

We recall the relative error of the eigenvalue defined as

$$\text{Rel. Err.} = \frac{|\lambda_{h_i} - \lambda_{h_{i+1}}|}{\lambda_{h_{i+1}}},$$

where λ_{h_i} is a fixed eigenvalue computed by the C^0 IPG method on the mesh with mesh size h_i. In Fig. 4.3 we plot the convergence history of the C^0 IPG method. For the unit square, the convergence rates are $O(h^2)$ as predicted by the theory in the previous section. For the L-shaped domain, there is a decrease in the convergence rate due to the reentrant corner, which is also consistent with the theoretical result.

Next we present some numerical results for the B-CP problem, the B-SSP problem, and the B-CH problem. Tables 4.9 and 4.10 display the first eigenvalues on a series of uniformly refined meshes for the unit square and the L-shaped domain. The approximate eigenvalue for the B-CP on the unit square agrees with the approximation obtained in [32]. The approximate eigenvalues for B-SSP and B-CH problems on the unit square also agree with

$$\lambda_1^{\text{B-SSP}} = 2\pi^2 \approx 19.73920880$$

and

$$\lambda_1^{\text{B-CH}} = \pi^2 \approx 9.869604401,$$

respectively. The convergence histories for the plate buckling problem for the buckling problems are similar to the vibration problems.

h	BCP(1)	BSSP(1)	BCH(1)
1/10	55.4016	20.0244	9.9541
1/20	53.2067	19.8193	9.8930
1/40	52.5757	19.7607	9.8758
1/80	52.4045	19.7448	9.8712

Table 4.9: The first B-CP, B-SSP, and B-CH eigenvalues for the unit square.

In Fig. 4.4 we present the 2D contour plots of the eigenfunctions corresponding to the first biharmonic eigenvalues of the unit square and the L-shaped domain for the V-CP problem and the V-SSP problem. The eigenfunctions for V-CP exhibit the correct

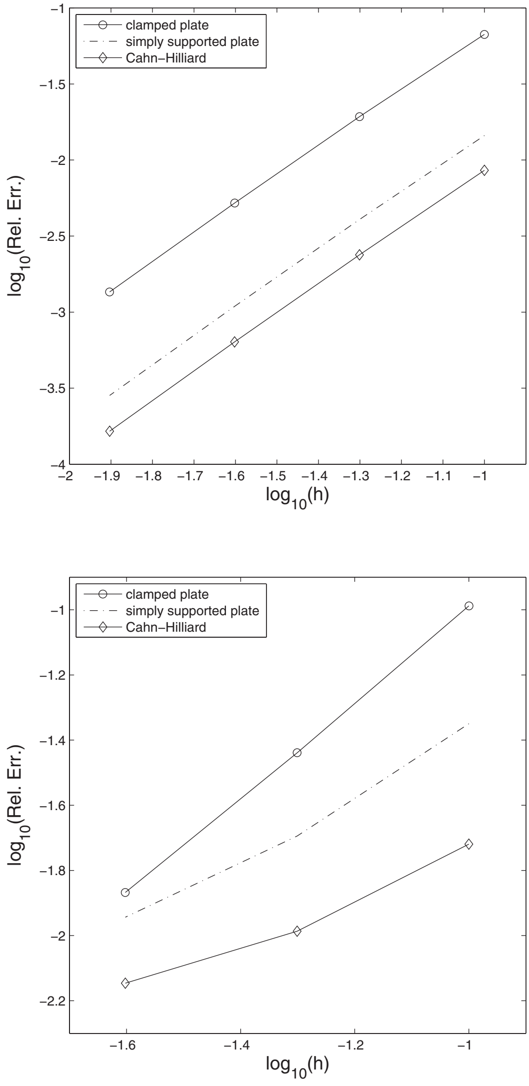


Figure 4.3: Relative errors of the first biharmonic plate vibration eigenvalues. Top: the unit square. Bottom: the L-shaped domain.

rotational symmetry, which is consistent with the fact that the first V-CP eigenvalue is a simple eigenvalue for both domains (see Table 4.12 and Table 4.13). This is also true for the V-SSP problem (see Table 4.14 and Table 4.15). Moreover, the computed

h	BCP(1)	BSSP(1)	BCH(1)
1/10	148.0750	65.8585	15.3809
1/20	135.0775	63.3735	14.8899
1/40	130.8045	62.2093	14.6087
1/80	129.3580	61.6123	14.4305

Table 4.10: The first B-CP, B-SSP, and B-CH eigenvalues of the L-shaped domain.

V-SSP eigenfunction for the first biharmonic eigenvalue on the unit square should approximate a multiple of $\sin(\pi x_1)\sin(\pi x_2)$ and this is observed.

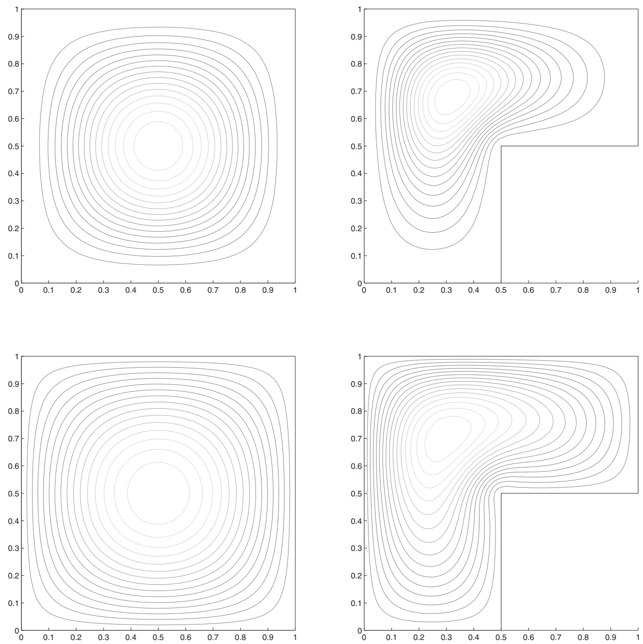


Figure 4.4: Eigenfunctions corresponding to the first biharmonic plate vibration eigenvalues. First row: V-CP eigenfunctions. Second row: V-SSP eigenfunctions.

In Fig. 4.5 we present the 2D surface plots of eigenfunctions for the V-CH problem. As was mentioned before, the first eigenvalue of the V-CH problem on the square is π^2 (with multiplicity 2) and the eigenspace is spanned by the two functions $\cos(\pi x_1)$ and $\cos(\pi x_2)$. In Fig. 4.5, we show the computed eigenfunctions.

Remark 4.5.7. *The computed eigenfunctions do not necessarily look the same as* $\cos(\pi x_1)$ *and* $\cos(\pi x_2)$. *In fact, they can be any linear combinations of* $\cos(\pi x_1)$ *and* $\cos(\pi x_2)$ *as long as they span the same eigenspace. In order to get better visu-*

alization of the computed V-CH eigenfunctions for this double eigenvalue, one can express $\cos(\pi x_1)$ *and* $\cos(\pi x_2)$ *approximately as linear combinations of the computed eigenfunctions. In other words, one can first find*

$$a_{11} = (\cos(\pi x_1), \phi_1),\ a_{12} = (\cos(\pi x_1), \phi_2),$$
$$a_{21} = (\cos(\pi x_2), \phi_1),\ a_{22} = (\cos(\pi x_2), \phi_2),$$

where ϕ_1 *and* ϕ_2 *are the computed eigenfunctions such that*

$$\|\phi_1\| = \|\phi_2\| = 1.$$

Then the eigenfunctions $\cos(\pi x_1)$ *and* $\cos(\pi x_2)$ *can be obtained by* $a_{11}\phi_1 + a_{12}\phi_2$ *and* $a_{21}\phi_1 + a_{22}\phi_2$.

The 2D contour plots of the computed eigenfunctions for the first and second V-CH eigenvalues for the L-shaped domain is displayed on the second row of Fig. 4.5. It is observed that the computed eigenfunction is anti-symmetric with respect to the line connecting the reentrant corner and the upper left corner, which is consistent with the zero mean constraint and with the fact that the first V-CH eigenvalue is a simple eigenvalue (cf. Table 4.17).

The 2D contour plots for some other V-SSP and **V**-CH eigenfunctions on the L-shaped domain are presented in Fig. 4.6. As was mentioned earlier, $\sin(2\pi x_1)\sin(2\pi x_2)$ is also a V-SSP eigenfunction for the L-shaped domain with the simple eigenvalue

$$\lambda_3^{\text{V-SSP}} = 64\pi^4 \approx 6234.18182618,$$

which turns out to be the 3rd eigenvalue. The 2D contour plot of the computed eigenfunction for this eigenvalue is displayed in the first row of Fig. 4.6, where the same symmetry as the function $\sin(2\pi x_1)\sin(2\pi x_2)$ is observed.

The functions $\cos(2\pi x_1)$ and $\cos(2\pi x_2)$ span the eigenspace of the double (3rd and 4th) V-CH eigenvalue

$$\lambda_{3,4}^{\text{V-CH}} = 16\pi^4 \approx 1558.54555654.$$

As in the case of the double eigenvalue π^2 of the V-CH problem on the unit square, we plot the computed eigenfunctions in the second row of Fig. 4.6.

Next we present some numerical results for the B-CP problem, the B-SSP problem, and the B-CH problem. Tables 4.9 and 4.10 display the first eigenvalues on a series of uniformly refined meshes for the unit square and the L-shaped domain. The approximate eigenvalue for the B-CP on the unit square agrees with the approximation obtained in [32], and the approximate eigenvalues for B-SSP and B-CH problems on the unit square also agree with

$$\lambda_1^{\text{B-SSP}} = 2\pi^2 \approx 19.73920880$$

and

$$\lambda_1^{\text{B-CH}} = \pi^2 \approx 9.86960440,$$

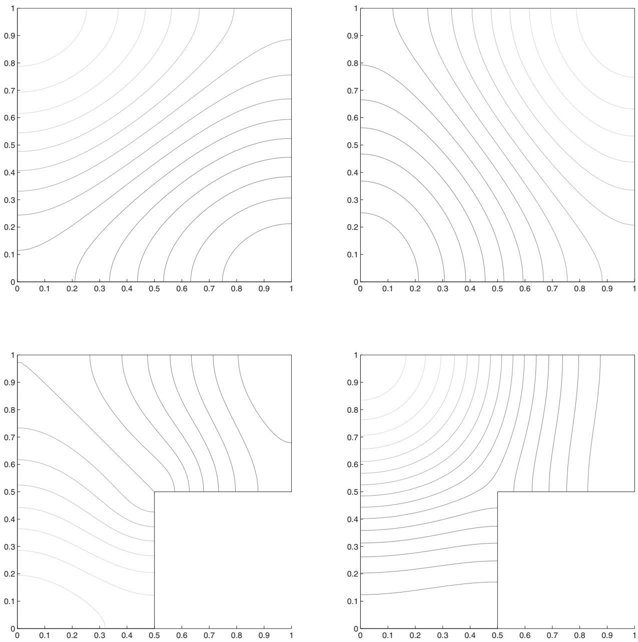


Figure 4.5: Eigenfunctions corresponding to the first two V-CH eigenvalues for the unit square (first row) and the first two V-CH eigenvalues for the L-shaped domain (second row).

respectively.

The convergence histories of the first eigenvalue for the plate buckling problem on the unit square and the L-shaped domain are presented in Fig. 4.7, which exhibit similar behavior as the plate vibration problem.

The behavior of the eigenfunctions for the buckling problems is similar to the eigenfunctions for vibration problems.

4.5.4 Comparisons of Different Methods

In this section we compare the quadratic C^0 IPG method with the Argyris C^1 finite element method [13], the Ciarlet-Raviart mixed finite element method [89], and the Morley nonconforming finite element method [204] discussed in the previous sections. Since we are considering all six biharmonic eigenvalue problems, we need the weak formulations here for the Argyris method, Ciarlet-Raviant mixed method, and the Morley method with different functional spaces.

Let V_h be the Argyris finite element space such that $V_h \subset V$, where V is defined in (4.63)–(4.65) for the three types of boundary conditions. The discrete biharmonic

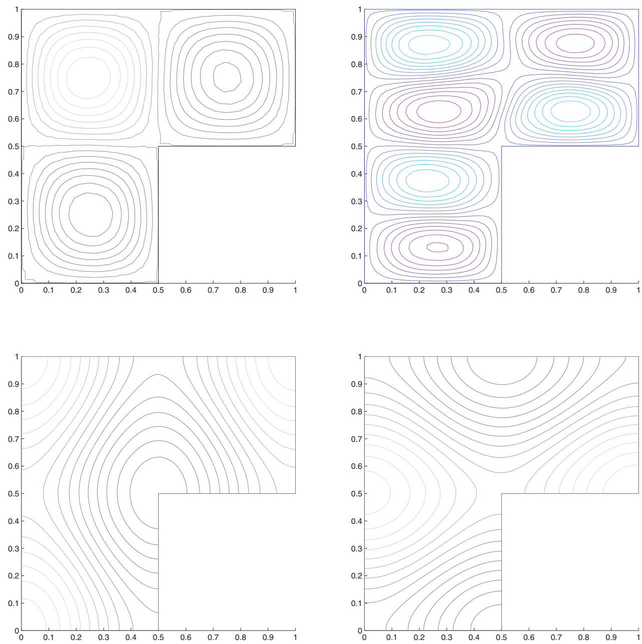


Figure 4.6: Eigenfunctions for the L-shaped domain. Top: the third and seventh V-SSP eigenfunctions. Bottom: the third and fourth V-CH eigenfunctions.

eigenvalue problem for the Argyris finite element method is to find $(\lambda_h, u_h) \in \mathbb{R}_+ \times V_h$ such that $u_h \neq 0$ and

$$(\triangle u_h, \triangle v_h) = \lambda_h b(u_h, v_h) \qquad \text{for all } \; v \in V_h.$$

The Ciarlet-Raviart mixed finite element method for the biharmonic eigenvalue problems is based on the following weak formulation: Find $\lambda \in \mathbb{R}_+$ and nontrivial $(p, u) \in Q \times V$ such that

$$\int_\Omega pq dx - \int_\Omega \nabla q \cdot \nabla u \, dx = 0 \qquad \text{for all } \; q \in Q, \tag{4.82a}$$

$$- \int_\Omega \nabla p \cdot \nabla v \, dx = -\lambda b(u, v) \qquad \text{for all } \; v \in V, \tag{4.82b}$$

where

$$Q = H^1(\Omega) \quad \text{and} \quad V = H_0^1(\Omega)$$

for the CP boundary conditions,

$$Q = H_0^1(\Omega) \quad \text{and} \quad V = H_0^1(\Omega)$$

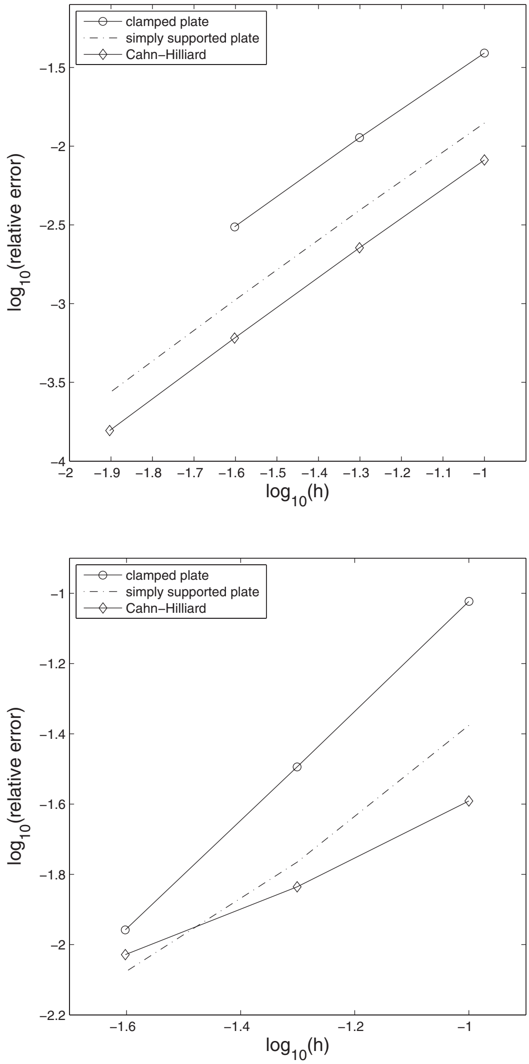


Figure 4.7: Convergence of the first B-CP, B-SSP, and B-CH eigenvalues. Top: the unit square. Bottom: the L-shaped domain.

for the SSP boundary conditions, and

$$Q = H^1(\Omega) \quad \text{and} \quad V = H^1(\Omega)$$

for the CH boundary conditions.

The discrete eigenvalue problem is to find $\lambda_h \in \mathbb{R}_+$ and nontrivial $(p_h, u_h) \in Q_h \times V_h$ such that

$$\int_\Omega p_h q dx - \int_\Omega \nabla q \cdot \nabla u_h dx = 0 \qquad \text{for all } q \in Q_h, \tag{4.83a}$$

$$- \int_\Omega \nabla p_h \cdot \nabla v dx = -\lambda_h b(u_h, v) \qquad \text{for all } v \in V_h, \tag{4.83b}$$

where $Q_h \subset Q$ and $V_h \subset V$ are standard linear Lagrange finite element spaces.

Let u be a biharmonic eigenfunction. If $p = -\Delta u$ belongs to $H^1(\Omega)$, then (p, u) satisfy the weak formulation (4.82). Therefore, the discrete eigenvalue problem (4.83) defines a Galerkin method for such an eigenfunction. This is the case for the CP, SSP, and CH biharmonic eigenvalue problems if Ω is convex [33, 139, 140]. However, as far as we know, only the convergence of the Ciarlet-Raviart finite element method for the CP eigenvalue problem on convex domains has been established in [200].

The Morley finite element space V_h is determined by the boundary conditions. For the CP boundary conditions, we set all the degrees of freedom on $\partial\Omega$ to be zero. For the SSP boundary conditions, we set only the degrees of freedom at the vertices in $\partial\Omega$ to be zero. For the CH boundary conditions we set only the degrees of freedom at the midpoints of the edges along $\partial\Omega$ to be zero, and we also impose the zero mean condition on V_h.

The Morley finite element method is to find $(\lambda_h, u_h) \in \mathbb{R}_+ \times V_h$ such that $u_h \neq 0$ and

$$\sum_{T \in \mathcal{T}_h} \int_T D^2 u_h : D^2 v \, dx = \lambda_h b_h(u_h, v) \qquad \text{for all } v \in V_h,$$

where

$$b_h(w, v) = \begin{cases} \displaystyle\int_\Omega wv \, \mathrm{d}x & \text{for plate vibration problems,} \\ \displaystyle\sum_{T \in \mathcal{T}_h} \int_T \nabla w \cdot \nabla v \, \mathrm{d}x & \text{for plate buckling problems.} \end{cases}$$

The numerical results of the four methods for the plate vibration problem on the unit square, the L-shaped domain, and with the three types of boundary conditions are presented in Tables 4.12–4.17. The mesh size used in the computations is $h \approx 0.0125$. In Table 4.11, we show the degrees of freedom (DoF) of different methods. The mixed method has the least degrees of freedom. The C^0 IPG and the Morley method have the same number of degrees of freedom. The Argyris element has the most degrees of freedom.

Remark 4.5.8. *The number of degrees of freedom is only a partial indicator of the complexity of the finite element method. In a comprehensive comparison one would also take into account the differences in the sparsity of the system matrices and the differences in the solution procedures for symmetric positive-definite systems and saddle point systems.*

	DoF (unit square)	ratio	DoF (L-shaped)	ratio
C^0IPG	16129	4.06	32705	4.04
Argyris	36226	9.13	73502	9.08
Mixed	3969	1.00	8097	1.00
Morley	16129	4.06	32705	4.04

Table 4.11: The degrees of freedom of different methods. The size of the triangular mesh is $h \approx 0.0125$.

	1st	2nd	3rd	4th	5th
C^0IPG	0.1299e4	0.5411e4	0.5411e4	1.1811e4	1.7412e4
Argyris	0.1304e4	0.5427e4	0.5427e4	1.1798e4	1.7443e4
Mixed	0.1309e4	0.5451e4	0.5451e4	1.1877e4	1.7548e4
Morley	0.1290e4	0.5349e4	0.5349e4	1.1607e4	1.7113e4

Table 4.12: The first 5 V-CP eigenvalues for the unit square.

	1st	2nd	3rd	4th	5th
C^0IPG	0.6694e4	1.0815e4	1.4655e4	2.5862e4	3.3418e4
Argyris	0.6775e4	1.1122e4	1.4985e4	2.6274e4	3.3686e4
Mixed	0.6695e4	1.1063e4	1.4925e4	2.6201e4	3.3499e4
Morley	0.6630e4	1.1004e4	1.4842e4	2.6018e4	3.3164e4

Table 4.13: The first 5 V-CP eigenvalues for the L-shaped domain.

	1st	2nd	3rd	4th	5th
C^0IPG	0.3896e3	2.4166e3	2.4166e3	6.1961e4	9.6768e3
Argyris	0.3896e3	2.4352e3	2.4352e3	6.2343e3	9.7409e3
Mixed	0.3900e3	2.4409e3	2.4409e3	6.2609e3	9.7806e3
Morley	0.3893e3	2.4295e3	2.4295e3	6.2143e4	9.6896e3

Table 4.14: The first 5 V-SSP eigenvalues for the unit square.

We observe that the numerical results for the quadratic C^0 IPG method and the Argyris method are comparable in all six cases. In view of the smooth nature of the eigenfunctions on the unit square and the high order of the finite element, the Argyris method provides very accurate approximation of the biharmonic eigenvalues on the unit square. Therefore the quadratic C^0 IPG method is quite efficient for the unit square. This can also be seen by comparing the eigenvalues in Table 4.12 with the ones in [245].

From Table 4.12, Table 4.14, and Table 4.16 we see that the Ciarlet-Raviart mixed finite element method converges on the unit square for all three types of boundary

	1st	2nd	3rd	4th	5th
C^0IPG	0.2718e4	0.3743e4	0.6061e4	1.3666e4	1.9156e4
Argyris	0.2692e4	0.3765e4	0.6234e4	1.3972e4	1.9375e4
Mixed	0.1491e4	0.3699e4	0.6242e4	1.3969e4	1.6354e4
Morley	0.2414e4	0.3663e4	0.6225e4	1.3904e4	1.8642e4

Table 4.15: The first 5 V-SSP eigenvalues for the L-shaped domain.

	1st	2nd	3rd	4th	5th
C^0IPG	0.0970e3	0.0970e3	0.3881e3	1.5524e3	1.5524e3
Argyris	0.0974e3	0.0974e3	0.3896e3	1.5585e3	1.5585e3
Mixed	0.0974e3	0.0974e3	0.3901e3	1.5606e3	1.5606e3
Morley	0.0974e3	0.0974e3	0.3893e3	1.5548e3	1.5548e3

Table 4.16: The first 5 V-CH eigenvalues for the unit square.

	1st	2nd	3rd	4th	5th
C^0IPG	0.1783e3	0.2089e3	1.5097e3	1.5138e3	2.0354e3
Argyris	0.1755e3	0.2068e3	1.5585e3	1.5585e3	2.0856e3
Mixed	0.0349e3	0.1998e3	1.5595e3	1.5595e3	2.0769e3
Morley	0.1498e3	0.1971e3	1.5575e3	1.5576e3	2.0701e3

Table 4.17: The first 5 V-CH eigenvalues for the L-shaped domain.

conditions. It is interesting to note that the eigenvalues computed by the Ciarlet-Raviart method are consistently larger than the corresponding eigenvalues computed by the C^0 IPG method, and the eigenvalues computed by the Argyris method are always between the other two with only one exception (the fourth eigenvalue in Table 4.12).

For the L-shaped domain, we observe from Table 4.13 that the Ciarlet-Raviart mixed finite element method also converges for the V-CP problem, and again the eigenvalues computed by the Ciarlet-Raviart mixed finite element method are consistently larger than the corresponding eigenvalues computed by the C^0 IPG method. For the boundary conditions of SSP and CH, the results in Table 4.15 and Table 4.17 show spurious eigenvalues generated by the Ciarlet-Raviart mixed finite element method.

Compared with the C^0 IPG method, the performance of the Morley finite element method is slightly better when the eigenfunction is very smooth and slightly worse when the eigenfunction is less smooth. The eigenvalues computed by the Morley finite element method are consistently less than the approximations generated by the Argyris finite element method (see [151]).

Finally numerical results for the first eigenvalues of the plate buckling problems are presented in Table 4.18 (unit square) and Table 4.19 (L-shaped domain). The mesh size h in the computations is roughly 0.0125. For the unit square, the results from all four methods with respect to all three boundary conditions are consistent. For the L-shaped domain, the results from the C^0IPG method, the Argyris finite element method, and the Morley finite element method are consistent for all three boundary conditions, whereas the Ciarlet-Raviart mixed finite element method is consistent with the other methods only for the CP boundary conditions and generates spurious eigenvalues for the other two boundary conditions.

	B-CP	B-SSP	B-CH
C^0IPG	52.4045	19.7448	9.8712
Argyris	52.3469	19.7392	9.8695
Mixed	52.3671	19.7422	9.8704
Morley	52.3301	19.7383	9.8694

Table 4.18: The first eigenvalues of the plate buckling eigenvalues for the unit square.

	B-CP	B-SSP	B-CH
C^0IPG	129.3580	61.6123	14.4305
Argyris	129.0132	61.9109	14.6288
Mixed	128.4905	38.6147	5.9099
Morley	127.7805	59.1396	13.9426

Table 4.19: The first eigenvalues of the plate buckling eigenvalues for the L-shaped domain.

The following remark in [52] gives some insight to why the Ciarlet-Raviart method works for the V-CP and B-CP eigenvalue problems on the L-shaped domain.

Remark 4.5.9. *The behavior of the Ciarlet-Raviart mixed finite element method on nonconvex domains with respect to the boundary conditions of CP, SSP, and CH can be given a heuristic explanation as follows. Since an eigenfunction u for the CP eigenvalue problem always belongs to $H^{2+\alpha}(\Omega)$ for some $\alpha \in (\frac{1}{2}, 1]$, we can replace* (4.82) *by another weak formulation: Find $\lambda \in \mathbb{R}_+$ and nontrivial $(p, u) \in H^\alpha(\Omega) \times H_0^1(\Omega)$ such that*

$$\int_\Omega pq dx - (\nabla q, \nabla u) = 0 \qquad \textit{for all } q \in H^1(\Omega), \tag{4.84a}$$

$$-\langle \nabla p, \nabla v \rangle = -\lambda b(u, v) \qquad \textit{for all } v \in H_0^{2-\alpha}(\Omega), \tag{4.84b}$$

where $\langle \cdot, \cdot \rangle$ is the duality pairing between $H^{\alpha-1}(\Omega)$ and $H^{1-\alpha}(\Omega)$. Since the P_1 finite element spaces satisfy

$$Q_h \subset H^1(\Omega) \subset H^\alpha(\Omega)$$

and

$$V_h \subset H_0^{2-\alpha}(\Omega) \subset H_0^1(\Omega),$$

we can treat (4.83) *as a Petrov-Galerkin method for the V-CP and B-CP eigenvalue problems based on* (4.84). *This explains why the Ciarlet-Raviart method converges for the V-CP and B-CP eigenvalue problems on the L-shaped domain. On the other hand, since* $p = -\Delta u$ *may only belong to* $H^\alpha(\Omega)$ *for some* $\alpha \in (0, \frac{1}{2})$ *if* u *is an eigenfunction for the biharmonic eigenvalue problems with the SSP or the CH boundary conditions, a similar Petrov-Galerkin interpretation for* (4.83) *is not valid because* V_h *is not a subspace of* $H^{2-\alpha}(\Omega)$ *when* $\alpha < \frac{1}{2}$.

4.6 C^0 IPG for a Fourth Order Problem

In this section, we consider a fourth order eigenvalue problem arising in the study of transmission eigenvalues, which have important applications in the inverse scattering theory [65, 228]. The transmission eigenvalue problem is the main topic in Chapter 6.

Due to the presence of lower order terms, the norm convergence of discrete operators is not readily available and thus the theory of Babuška-Osborn cannot be applied directly. Alternatively, we choose to employ the abstract theory by Descloux, Nassif, and Rappaz in Section 1.4 (see also [114, 115]). The material in this section is based on [159].

Let Ω be a bounded Lipschitz polygonal domain in $\mathbb{R}^2$ with unit outward normal ν. Let $m(x)$ be a bounded smooth function such that $m(x) > \gamma > 0$. Let τ be a positive constant and C, C_1, C_2 denote generic positive constants.

We consider a fourth order eigenvalue problem of finding μ and u such that

$$(\triangle + \tau)m(x)(\triangle + \tau)u + \tau^2 u = \mu \triangle u \quad \text{in } \Omega, \tag{4.85}$$

with the boundary conditions

$$u = 0, \quad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial\Omega. \tag{4.86}$$

The corresponding source problem can be stated as follows. Given f, find u such that

$$(\triangle + \tau)m(x)(\triangle + \tau)u + \tau^2 u = \triangle f, \tag{4.87}$$

with the boundary conditions (4.86).

Let $\mathcal{A} : H_0^2(\Omega) \times H_0^2(\Omega) \to \mathbb{C}$ and $\mathcal{B} : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{C}$ be defined as

$$\mathcal{A}(u, v) = (m(\triangle u + \tau u), (\triangle v + \tau v)) + \tau^2 (u, v), \tag{4.88a}$$

$$\mathcal{B}(u, v) = (\nabla u, \nabla v). \tag{4.88b}$$

The variational formulation for the eigenvalue problem (4.85) is to find $\mu \in \mathbb{R}$ and $u \in H_0^2(\Omega)$, such that

$$\mathcal{A}(u, v) - \mu \mathcal{B}(u, v) = 0 \quad \text{for all } v \in H_0^2(\Omega). \tag{4.89}$$

The variational formulation for the associated source problem is to find $u \in H_0^2(\Omega)$ for $f \in H_0^1(\Omega)$ such that

$$\mathcal{A}(u, v) = \mathcal{B}(f, v) \quad \text{for all } v \in H_0^2(\Omega). \tag{4.90}$$

For the source problem, the following result holds.

Theorem 4.6.1. *Let $f \in H_0^1(\Omega)$. There exists a unique solution $u \in H_0^2(\Omega)$ to* (4.90) *such that*

$$\|u\|_{H^2(\Omega)} \le C|f|_{H^1(\Omega)},$$

for some constant C independent of u and f.

Proof. We first show that $\mathcal{A}$ is a coercive sesquilinear form on $H_0^2(\Omega) \times H_0^2(\Omega)$. Simple calculation shows that

$$\begin{aligned} \mathcal{A}(u, u) &\ge \gamma\|\triangle u + \tau u\|^2 + \tau^2\|u\|^2 \\ &\ge \gamma\|\triangle u\|^2 - 2\gamma\tau\|\triangle u\|\|u\| + (\gamma + 1)\tau^2\|u\|^2 \\ &= \epsilon(\tau\|u\| - \gamma/\epsilon\|\triangle u\|)^2 + \gamma(1 - \gamma/\epsilon)\|\triangle u\|^2 + (1 + \gamma - \epsilon)\tau^2\|u\|^2 \\ &\ge \gamma(1 - \gamma/\epsilon)\|\triangle u\|^2 + (1 + \gamma - \epsilon)\tau^2\|u\|^2 \end{aligned} \tag{4.91}$$

for any ϵ such that $\gamma < \epsilon < \gamma + 1$. Moreover, since $u \in H_0^2(\Omega)$, it holds that

$$\|\nabla u\|^2 \le C\|\triangle u\|^2.$$

Together with the Poincaré inequality (Theorem 1.2.5), we obtain

$$\mathcal{A}(u, u) \ge C\|u\|^2_{H^2(\Omega)}$$

for some positive constant C.

For boundedness, employing the Cauchy-Schwatz inequality, we obtain that

$$\begin{aligned} |\mathcal{A}(u, v)| &\le C\|\triangle u + \tau u\|\|\triangle v + \tau v\| + \tau^2\|u\|\|v\| \\ &\le C(\|\triangle u\| + \tau\|u\|)(\|\triangle v\| + \tau\|v\|) + \tau^2\|u\|\|v\| \\ &\le C\|u\|_{H^2(\Omega)}\|v\|_{H^2(\Omega)} \end{aligned}$$

for some constant C. Then the theorem follows the Lax-Milgram Lemma 1.3.1. □

Remark 4.6.1. *Due to the fact that* (4.85) *is a fourth order problem with lower order perturbations, there exists an $\alpha > 0$ [139], such that*

$$u \in H^{2+\alpha}(\Omega). \tag{4.92}$$

The elliptic index α depends on the corner of Ω. Furthermore, $\alpha \in (\frac{1}{2}, 1]$ for a polygonal domain and $\alpha = 1$ if Ω is convex.

4.6.1 The Source Problem

In this section, we employ the C^0 IPG method for the source problem (4.87). Note that our formulation is different from that in [55] since we need to incorporate lower order terms. Let $\mathcal{T}_h$ be a regular triangulation for Ω and $V_h \subset H_0^1(\Omega)$ be the associated $\mathcal{P}_k$ $(k \geq 2)$ Lagrange finite element space with zero boundary condition on $\partial\Omega$.

Assuming the solution u is smooth enough, we start with the following integration by parts formula on a triangle $K \subset \mathcal{T}_h$

$$\begin{aligned} & \int_K \triangle(m\triangle u)v \,\mathrm{d}x \\ = \quad & \int_{\partial K} \left(\frac{\partial(m\triangle u)}{\partial\nu} v - m\triangle u \frac{\partial v}{\partial \nu} \right) \mathrm{d}s + \int_K m\triangle u \triangle v \,\mathrm{d}x. \end{aligned} \tag{4.93}$$

Summing up (4.93) over all the triangles in $\mathcal{T}_h$, with cancelations we get

$$\begin{aligned} & \sum_{K\in\mathcal{T}_h} \int_K \triangle(m\triangle u)v \,\mathrm{d}x \\ = \quad & -\sum_{K\in\mathcal{T}_h} \int_{\partial K} m\triangle u \frac{\partial v}{\partial \nu} \mathrm{d}s + \sum_{K\in\mathcal{T}_h} \int_K m\triangle u \triangle v \,\mathrm{d}x. \end{aligned} \tag{4.94}$$

For an interior edge e shared by two triangles $K_\pm$, we define the jumps and averages as

$$\left[\!\!\left[\frac{\partial u}{\partial \nu_e}\right]\!\!\right] = \nu_e \cdot (\nabla u_+ - \nabla u_-)$$

and

$$\{\!\{m\triangle u\}\!\} = \frac{1}{2}(m_-\triangle u_- + m_+\triangle u_+),$$

respectively, where $u_\pm = u|_{K_\pm}$ and ν_e points from K_- to K_+.

For a boundary edge e, we take ν_e to be the unit normal pointing towards the outside of Ω and define

$$\left[\!\!\left[\frac{\partial u}{\partial \nu_e}\right]\!\!\right] = -\nu_e \cdot \nabla u, \quad \{\!\{m\triangle u\}\!\} = m\triangle u.$$

We rewrite the first term on the right-hand side of (4.94) as a sum over the edges

$$-\sum_{K\in\mathcal{T}_h} \int_{\partial K} m\triangle u \frac{\partial v}{\partial \nu} \mathrm{d}s = \sum_{e\in\mathcal{E}_h} \int_e m\triangle u \left[\!\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\!\right] \mathrm{d}s,$$

where $\mathcal{E}_h$ is the set of all the edges of $\mathcal{T}_h$. Replacing $m\triangle u$ in the above equation by $\{\!\{m\triangle u\}\!\}$, introducing the symmetric term

$$\int_e \{\!\{m\triangle v\}\!\} \left[\!\!\left[\frac{\partial u}{\partial \nu_e}\right]\!\!\right] \mathrm{d}s,$$

and adding the penalty term

$$\frac{1}{|e|}\int_e \left[\!\left[\frac{\partial u}{\partial \nu_e}\right]\!\right]\left[\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\right] \mathrm{d}s,$$

we obtain the following discrete problem. For $f \in H_0^1(\Omega)$, find $u_h \in V_h$ such that

$$\mathcal{A}_h(u_h, v) = \mathcal{B}_h(f, v) \quad \text{for all } v \in V_h, \tag{4.95}$$

where

$$\mathcal{A}_h(w, v) = a_h(w, v) + b_h(w, v) + \sigma c_h(w, v), \tag{4.96}$$

$$\mathcal{B}_h(f, v) = \sum_{K \in \mathcal{T}_h} \int_K \nabla f \cdot \nabla v \, \mathrm{d}x, \tag{4.97}$$

and

$$a_h(w, v) = \sum_{K \in \mathcal{T}_h} \int_K m(\triangle + \tau)w(\triangle + \tau)v + \tau^2 wv \, \mathrm{d}x,$$

$$b_h(w, v) = \sum_{e \in \mathcal{E}_h} \int_e \{\!\{m\triangle w\}\!\}\left[\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\right] + \{\!\{m\triangle v\}\!\}\left[\!\left[\frac{\partial w}{\partial \nu_e}\right]\!\right] \mathrm{d}s,$$

$$c_h(w, v) = \sum_{e \in \mathcal{E}_h} \frac{1}{|e|}\int_e \left[\!\left[\frac{\partial w}{\partial \nu_e}\right]\!\right]\left[\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\right] \mathrm{d}s.$$

Here $\sigma > 0$ is the penalty parameter.

Let $V(h) = H_0^2(\Omega) + V_h$. We define the mesh dependent norm $\|\cdot\|_h$ on $V(h)$ as

$$\|v\|_h^2 = \sum_{K \in \mathcal{T}_h} \|\triangle v\|_{L^2(K)}^2 + \sigma \sum_{e \in \mathcal{E}_h} \frac{1}{|e|}\left\|\left[\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\right]\right\|_{L^2(e)}^2. \tag{4.98}$$

It is easy to see that the following Poincaré inequality holds.

Lemma 4.6.2. *For every $v \in V(h)$, $\|v\| \le C\|v\|_h$.*

The form $\mathcal{A}_h(\cdot,\cdot)$ is bounded, i.e.,

$$|\mathcal{A}_h(w, v)| \le C\|w\|_h\|v\|_h \quad \text{for all } w, v \in V_h.$$

From Lemma 4.6.2, standard inverse estimates, and the Cauchy-Schwarz inequality, one has that

$$\begin{aligned}
&\sum_{e\in\mathcal{E}_h}\left|\int_e \{\!\{m\triangle w\}\!\}\left[\!\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\!\right]\mathrm{d}s\right| \\
\leq\ & \left(\sum_{e\in\mathcal{E}_h}|e|\,\|\{\!\{m\triangle w\}\!\}\|^2_{L^2(e)}\right)^{\frac{1}{2}}\left(\sum_{e\in\mathcal{E}_h}|e|^{-1}\left\|\left[\!\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\!\right]\right\|^2_{L^2(e)}\right)^{\frac{1}{2}} \\
\leq\ & C\left(\sum_{e\in\mathcal{E}_h}\sum_{K\in\mathcal{T}_e}\|\triangle w\|^2_{L^2(K)}\right)^{\frac{1}{2}}\left(\sum_{e\in\mathcal{E}_h}|e|^{-1}\left\|\left[\!\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\!\right]\right\|^2_{L^2(e)}\right)^{\frac{1}{2}} \\
\leq\ & C\left(\sum_{K\in\mathcal{T}_h}\|\triangle w\|^2_{L^2(K)}\right)^{\frac{1}{2}}\left(\sum_{e\in\mathcal{E}_h}|e|^{-1}\left\|\left[\!\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\!\right]\right\|^2_{L^2(e)}\right)^{\frac{1}{2}}. \qquad (4.99)
\end{aligned}$$

Here $\mathcal{T}_e$ is the set of the elements in $\mathcal{T}_h$ that share the common edge e.

Next we show the coercivity of $\mathcal{A}_h$. Similar to (4.91), it holds that

$$\int_K m(\triangle+\tau)v(\triangle+\tau)v+\tau^2 vv\,\mathrm{d}x \geq \int_K C_1|\triangle v|^2+C_2|v|^2\,\mathrm{d}x,$$

for some positive constants C_1 and C_2 depending on $m(x)$ and τ. Using the inequality of arithmetic and geometric means and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
\mathcal{A}_h(v,v)\ \geq\ & C_1\sum_{K\in\mathcal{T}_h}\|\triangle v\|_{L^2(K)}+\sigma\sum_{e\in\mathcal{E}_h}|e|^{-1}\left\|\left[\!\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\!\right]\right\|^2_{L^2(e)} \\
& -C\left(\sum_{K\in\mathcal{T}_h}\|\triangle v\|^2_{L^2(K)}\right)^{\frac{1}{2}}\left(\sum_{e\in\mathcal{E}_h}|e|^{-1}\left\|\left[\!\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\!\right]\right\|^2_{L^2(e)}\right)^{\frac{1}{2}} \\
\geq\ & \frac{C_1}{2}\sum_{K\in\mathcal{T}_h}\|\triangle v\|^2_{L^2(K)} \\
& +\left(\sigma-\frac{C^2}{C_1}\right)\sum_{e\in\mathcal{E}_h}|e|^{-1}\left\|\left[\!\!\left[\frac{\partial v}{\partial \nu_e}\right]\!\!\right]\right\|^2_{L^2(e)}. \qquad (4.100)
\end{aligned}$$

Provided σ is large enough, one has that

$$\mathcal{A}_h(v,v)\geq C\|v\|_h^2 \quad \text{for all } v\in V_h.$$

The existence and uniqueness of the discrete problem follows immediately.

Let u be the exact solution and u_h be the discrete solution of the source problem. We have the consistency relation [49]

$$\mathcal{A}_h(u-u_h,v)=0 \quad \text{for all } v\in V_h.$$

Let $v \in V_h$ be arbitrary.

$$\begin{aligned}
\|u-u_h\|_h &\le \|u-v\|_h + \|v-u_h\|_h \\
&\le \|u-v\|_h + C \max_{w\in V_h\setminus\{0\}} \frac{\mathcal{A}_h(v-u_h,w)}{\|w\|_h} \\
&\le \|u-v\|_h + C \max_{w\in V_h\setminus\{0\}} \frac{\mathcal{A}_h(v-u,w)}{\|w\|_h} \\
&\le C\|u-v\|_h,
\end{aligned}$$

and hence

$$\|u-u_h\|_h \le C \inf_{v\in V_h} \|u-v\|_h.$$

Let $I_h : C^0(\overline{\Omega}) \to V_h$ be the Lagrange nodal interpolation operator. Then the following inequalities hold (Section 3.4 of [49])

$$\|u - I_h u\|_h \le Ch^\beta \|u\|_{H^{2+\beta}(\Omega)} \le Ch^\beta |f|_{H^1(\Omega)}, \tag{4.101}$$

where $\beta = \min\{\alpha, k-1\}$. Note that β is limited by the regularity of the solution and the orders of the Lagrange elements.

Let $V = H_0^2(\Omega)$. Summarizing the approximation property and the error estimate, we obtain the following lemma.

Lemma 4.6.3. *(Quasi-optimality) We have that*

$$\lim_{h\to 0} \inf_{v_h\in V_h} \|v - v_h\|_h = 0 \quad \textit{for all } v \in V.$$

The discrete problem (4.95) has a unique solution and

$$\|u - u_h\|_h \le Ch^\beta \|u\|_{H^{2+\beta}(\Omega)} \le Ch^\beta |f|_{H^1(\Omega)}, \tag{4.102}$$

where C is a constant independent of the mesh size.

4.6.2 The Eigenvalue Problem

The C^0IPG for the eigenvalue problem can be stated as follows. Find $\mu_h \in \mathbb{R}$ and $u_h \in V_h$ such that

$$\mathcal{A}_h(u_h, v) = \mu_h \mathcal{B}_h(u_h, v) \quad \text{for all } v \in V_h. \tag{4.103}$$

Following the abstract convergence theory in Section 1.4.1 (see also [114]) and the spirit of discontinuous Galerkin method for the Laplace eigenvalue problem [11], we would like to show that the C^0 IPG is *spectrally correct*, namely,

1. non-pollution of the spectrum: no discrete spurious eigenvalues;
2. completeness of the spectrum: all eigenvalues smaller than a fixed value are approximated when the mesh is fine enough;

3. non-pollution: there are no spurious eigenfunctions;
4. completeness of the eigenspaces: the eigenspace approximations have the right dimension.

To carry out subsequent discussions, we recall some classical results of spectral theory (see [165]). We define two solution operators

$$T : H^1(\Omega) \to V, \quad \mathcal{A}(Tf, v) = \mathcal{B}(f, v) \quad \text{for all } v \in V,$$

for the continuous problem (4.90) and

$$T_h : H^1(\Omega) \to V_h, \quad \mathcal{A}_h(T_h f, v) = \mathcal{B}_h(f, v) \quad \text{for all } v \in V_h,$$

for the discrete problem (4.95).

Since T is symmetric, positive definite, and compact due to the compact embedding of V into $H_0^1(\Omega)$, T has a sequence of positive eigenvalues $\{\lambda_j\}$ with zero being the only accumulation point. The inverse of $\{\lambda_j\}$, i.e., $\{\mu_j = 1/\lambda_j\}$, are the eigenvalues of (4.89) with ∞ being the only accumulation point.

Let $\sigma(T)$ and $\rho(T)$ be the spectrum and resolvent sets of T, respectively. Recall that the resolvent operator is defined as

$$R_z(T) = (z - T)^{-1} \quad z \in \rho(T).$$

Similarly, for T_h, we have $\sigma(T_h), \rho(T_h)$, and

$$R_z(T_h) = (z - T_h)^{-1} \quad z \in \rho(T_h).$$

Our goal is to show that the C^0 IPG (4.103) is spectrally correct and prove the optimal convergence rate.

For non-pollution of the spectrum, we can show that any open set containing $\sigma(T)$ also contains $\sigma(T_h)$ for h small enough. We first show that for z away from $\sigma(T)$, $z - T$ is bounded from below.

Lemma 4.6.4. *Let $z \in \rho(T), z \neq 0$. There exists a positive constant C only depending upon Ω and $|z|$ such that*

$$\|(z - T)f\|_h \geq C\|f\|_h \quad \textit{for all } f \in V(h).$$

Proof. Let $z \in \rho(T), z \neq 0$ be fixed and $f \in V(h)$. Set $g = (z - T)f$. Since $Tf \in V$, we have that $g \in V(h)$. Note that

$$T = ((\triangle + \tau)m(\triangle + \tau) + \tau^2)^{-1}\triangle = \tilde{T}^{-1}\triangle : H_0^1(\Omega) \to V$$

in the weak sense. Then $zf - g = Tf$ implies

$$\tilde{T}(zf - g) = \triangle f.$$

Hence $zf - g \in V$ is the solution of the following problem

$$\begin{aligned}
\tilde{T}(zf-g) - \frac{1}{z}\triangle(zf-g) &= \frac{\triangle g}{z} \quad \text{in } \Omega, \\
zf - g &= 0 \quad \text{on } \partial\Omega, \\
\frac{\partial}{\partial\nu}(zf-g) &= 0 \quad \text{on } \partial\Omega.
\end{aligned}$$

Since the above problem is a lower order perturbation of (4.87), we deduce that for some C [139]

$$\|zf - g\|_V \le \frac{C}{|z|}\|\nabla g\| \le \frac{C}{|z|}\|g\|_h.$$

Since $zf - g \in V$, we have that

$$\|zf - g\|_h \le C\|zf - g\|_V$$

and

$$\|zf - g\|_h \le \frac{C}{|z|}\|g\|_h.$$

Using the triangle inequality, we obtain the desired result

$$\begin{aligned}
\|f\|_h &\le \frac{1}{z}(\|zf-g\|_h + \|g\|_h) \\
&\le C(|z|)\|g\|_h \\
&= C(|z|)\|(z-T)f\|_h.
\end{aligned}$$

□

Next we show that a similar property holds for T_h as well.

Lemma 4.6.5. *For $z \in \rho(T), z \neq 0$, there exists a positive constant C only depending on Ω and $|z|$ such that, for h small enough,*

$$\|(z - T_h)f\|_h \ge C\|f\|_h \quad \textit{for all } f \in V(h). \tag{4.104}$$

Proof. By the triangle inequality,

$$\|(z - T_h)f\|_h \ge \|(z-T)f\|_h - \|(T - T_h)f\|_h.$$

By Lemma 4.6.4, Lemma 4.6.2, and Lemma 4.6.3, we have that

$$\|(z - T_h)f\|_h \ge C(|z|)\|f\|_h - Ch^\beta\|f\|_h,$$

where $C(|z|)$ is the constant in Lemma 4.6.4. Since $C(|z|)$ only depends on Ω and z, (4.104) is readily verified for h small enough. □

Lemma 4.6.6. *Let $F \subset \rho(T)$ be closed. There exists a positive constant C independent of h such that, for h small enough, we have*

$$\|R_z(T_h)\|_{\mathcal{L}(V(h),V(h))} \le C \quad \textit{for all } z \in F.$$

Proof. Let $z \in F$ be fixed. Since $z \in \rho(T)$, we have that

$$\|R_z(T_h)\|_{\mathcal{L}(V(h),V(h))} = \sup_{g \in V(h), \|g\|_h = 1} \|(z - T_h)^{-1} g\|_h.$$

Letting $\|g\|_h = 1$ and $(z - T_h)^{-1} g = f$, we obtain

$$\|(z - T_h) f\|_h = \|g\|_h = 1.$$

From Lemma 4.6.5, for h small enough, we get

$$C\|f\|_h \leq \|(z - T_h) f\|_h = 1$$

and the lemma follows immediately. □

Lemma 4.6.6 claims that, for any $z \in \rho(T)$ and h small enough, $(z - T_h)$ admits a bounded inverse from $V(h)$ to $V(h)$, i.e., $R_z(T_h)$ is well defined and continuous from $V(h)$ to $V(h)$. Thus we have shown the following theorem which implies non-pollution of the spectrum.

Theorem 4.6.7. *(Non-pollution of the spectrum) Let $A \subset \mathbb{C}$ be an open set containing $\sigma(T)$. Then, for h small enough, $\sigma(T_h) \subset A$.*

For fixed $z \in \rho(T)$ and $f \in V(h)$, we can write

$$\begin{aligned} \|zf - Tf\|_h &\leq |z|\|f\|_h + \|Tf\|_h \\ &\leq |z|\|f\|_h + C\|f\|_h \\ &\leq C(|z|)\|f\|_h, \end{aligned}$$

due to the stability estimate of the continuous problem and the Poincaré inequality of Lemma 4.6.2. Using Lemma 4.6.4, for all fixed $z \in \rho(T)$, $z - T : V(h) \to V(h)$ is a continuous invertible operator with continuous inverse. A direct consequence of this fact is an analogue of Lemma 4.6.4: let $F \subset \rho(T)$ be closed; then, there exists a positive constant C independent of h such that

$$\|R_z(T)\|_{\mathcal{L}(V(h),V(h))} \leq C$$

for all $z \in F$. From continuity of $T : H^1(\Omega) \to H^1(\Omega)$, if $F \subset \rho(T)$ is closed, there exists a positive constant C such that

$$\|R_z(T)\|_{\mathcal{L}(H^1(\Omega),H^1(\Omega))} \leq C \tag{4.105}$$

for all $z \in F$.

Let λ be an eigenvalue of T with algebraic multiplicity p. Denote by Γ a circle in the complex plane centered at λ such that no other eigenvalue lies inside Γ. Recall the spectral projections E from $H^1(\Omega)$ into V and E_h from $H^1(\Omega)$ into V_h by (see [157])

$$E := \frac{1}{2\pi i}\int_\Gamma R_z(T)\mathrm{d}z, \quad E_h := \frac{1}{2\pi i}\int_\Gamma R_z(T_h)\mathrm{d}z. \tag{4.106}$$

Let X and Y be closed subspaces of $V(h)$. We also recall the "distance" between X and Y as

$$d(X,Y) = \max\{\delta_h(X,Y), \delta_h(Y,X)\},$$

where

$$\delta_h(X,Y) := \sup_{x\in X, \|x\|=1} \inf_{y\in Y} \|x-y\|.$$

We first show that E_h converges to E in operator norm as $h \to 0$.

Theorem 4.6.8. *Let E and E_h be defined as in* (4.106). *Then*

$$\lim_{h\to 0} \|E - E_h\|_{\mathcal{L}(H^1(\Omega),V(h))} = 0. \tag{4.107}$$

Proof. It is easy to see that

$$(z-T)^{-1} - (z-T_h)^{-1} = (z-T_h)^{-1}(T-T_h)(z-T)^{-1},$$

i.e.,

$$R_z(T) - R_z(T_h) = R_z(T_h)(T-T_h)R_z(T).$$

Let $f \in H_0^1(\Omega)$. We have

$$\begin{aligned} & \|R_z(T_h)(T-T_h)R_z(T)f\|_h \\ \leq\ & \|R_z(T_h)\|_{\mathcal{L}(V(h),V(h))}\|T-T_h\|_{\mathcal{L}(H^1(\Omega),V(h))} \\ & \qquad \cdot\|R_z(T)\|_{\mathcal{L}(H^1(\Omega),H^1(\Omega))}\|f\|_{H^1(\Omega)}. \end{aligned}$$

From Lemma 4.6.3, Lemma 4.6.6, and (4.105), we obtain (4.107). □

Theorem 4.6.9. *(Non-pollution of the eigenspace)*

$$\lim_{h\to 0} \delta_h(E_h(V_h), E(V)) = 0.$$

Proof. With $E(H^1(\Omega)) = E(V)$ and $E_h y_h = y_h$ for all $y_h \in E_h(V_h)$, we have

$$\begin{aligned} & \sup_{y_h\in E_h(V_h), \|y_h\|_h=1} \inf_{x\in E(V)} \|y_h - x\|_h \\ =\ & \sup_{y_h\in E_h(V_h), \|y\|_h=1} \inf_{x\in E(H^1(\Omega))} \|y_h - x\|_h \\ =\ & \sup_{y_h\in E_h(V_h), \|y\|_h=1} \inf_{x\in H^1(\Omega)} \|E_h y_h - Ex\|_h. \end{aligned}$$

Letting $x = y_h$ and using the discrete Poincaré inequality, we obtain

$$\begin{aligned} & \sup_{y_h\in E_h(V_h), \|y\|_h=1} \inf_{x\in H^1(\Omega)} \|E_h y_h - Ex\|_h \\ \leq\ & \sup_{y_h\in E_h(V_h), \|y\|_h=1} \|E_h y_h - E y_h\|_h \\ \leq\ & \sup_{y_h\in E_h(V_h), \|y\|_h=1} \|E_h - E\|_{\mathcal{L}(H^1(\Omega),V(h))}\|y_h\|_h. \end{aligned}$$

Application of Theorem 4.6.8 completes the proof. □

Theorem 4.6.10. *(Completeness of the eigenspaces)*

$$\lim_{h\to 0} \delta_h(E(V), E_h(V_h)) = 0.$$

Proof.

$$\sup_{x\in E(V),\|x\|_h=1} \inf_{y_h\in E_h(V_h)} \|x - y_h\|_h = \sup_{x\in E(V),\|x\|_h=1} \inf_{y_h\in V_h} \|Ex - E_h y_h\|_h.$$

From quasi-optimality of V_h, there exists $x_h \in V_h$ such that

$$\lim_{h\to 0} \|x - x_h\|_h = 0.$$

So we have

$$\begin{aligned}
& \inf_{y_h\in V_h} \|Ex - E_h y_h\|_h \\
\le\ & \|Ex - E_h x_h\|_h \\
\le\ & \|E(x - x_h)\|_h + \|(E - E_h)x_h\|_h \\
\le\ & C\|E\|_{\mathcal{L}(V(h),V(h))}\|x - x_h\|_h + \|E - E_h\|_{\mathcal{L}(V(h),V(h))}\|x_h\|_h.
\end{aligned}$$

Since E is a projection, the first term goes to 0 as $h \to 0$. Using the fact that

$$\|E - E_h\|_{\mathcal{L}(V(h),V(h))} \le \|E - E_h\|_{\mathcal{L}(H^1(\Omega),V(h))},$$

and Theorem 4.6.8, we have that

$$\|E - E_h\|_{\mathcal{L}(V(h),V(h))} \to 0 \quad \text{as } h \to 0.$$

Note that $E(V)$ is finite dimensional. Pointwise convergence implies uniform convergence, which completes the proof. □

Completeness of the spectrum can be verified once we have completeness of the eigenspaces.

Theorem 4.6.11. *For all $z \in \sigma(T)$, there exists a family of $\{z_h\}$, $z_h \in \sigma(T_h)$ such that*

$$\lim_{h\to 0} z_h = z.$$

Proof. Theorem 4.6.9 and Theorem 4.6.10 imply that

$$d(E(V), E_h(V_h)) \to 0, \quad h \to 0.$$

Hence for h small enough, $E(V)$ and $E_h(V_h)$ have the same dimensions. Let D_Γ be the domain bounded by Γ. If $D_\Gamma \cap \sigma(T) \ne \emptyset$, then, for h small enough, $D_\Gamma \cap \sigma(T_h) \ne \emptyset$. Since T only has a point spectrum, without loss of generality, one can choose D_Γ a disk with radius $\epsilon > 0$ centered at z. Hence for h small enough, there must be an element in $\sigma(T_h)$ which is close enough to z (less than ϵ). The theorem follows consequently. □

Let $s = \dim E(H_0^1(\Omega))$. It has been shown that, for h small enough, there are s eigenvalues of T_h such that

$$\lim_{h\to 0} \sup_{1\le i\le s} |\lambda - \lambda_{i,h}| = 0.$$

Due to the approximation property of V_h (4.101), we have that

$$\delta_h(E(V), V_h) \le Ch^\beta.$$

Theorem 4.6.12. *For h small enough, we have that*

$$\sup_{1\le i\le n} |\lambda - \lambda_{i,h}| \le Ch^{2\beta}.$$

Proof. By (4.101), we have that

$$\begin{aligned}
\|E - E_h\|_{\mathcal{L}(E(V),V(h))} &\le C\|T - T_h\|_{\mathcal{L}(E(V),V(h))}\\
&\le C \sup_{x\in E(V),\|x\|_h=1} \|Tx - T_h x\|_h\\
&\le Ch^\beta.
\end{aligned}$$

Since E is a projection, for h small enough, $E_h|_{E(V)} : E(V) \to E_h(V_h)$ is an invertible mapping that we denote by $F_h = E_h|_{E(V)}$. Its inverse is uniformly bounded with respect to h.

Let $\tilde{T} = T|_{E(V)}$ and $\tilde{T}_h = F_h^{-1} T_h F_h : E(V) \to E(V)$. We obtain [11]

$$\sup_{1\le i\le n} |\lambda - \lambda_{i,h}| \le C\|\tilde{T} - \tilde{T}_h\|_{\mathcal{L}(E(V),V(h))}.$$

Let $S_h = F_h^{-1} E_h : H^1(\Omega) \to V(h)$, which is a continuous operator. For all $x \in E(V)$, $S_h T x = \tilde{T} x$ and $S_h T_h x = \tilde{T}_h x$. So we get

$$(\tilde{T} - \tilde{T}_h)x = S_h(T - T_h)x \quad \text{for all } x \in E(V),$$

and

$$\begin{aligned}
\|\tilde{T} - \tilde{T}_h\|_{\mathcal{L}(E(V),V(h))} &= \sup_{x\in E(V),\|x\|_h=1} \|\tilde{T}x - \tilde{T}_h x\|_h\\
&\le C \sup_{x\in E(V),\|x\|_h=1} \|Tx - T_h x\|_h\\
&\le Ch^\beta.
\end{aligned}$$

It is clear that the problem considered is self-adjoint. Since the C^0 IPG is symmetric, one actually has that

$$\sup_{1\le i\le n} |\lambda - \lambda_{i,h}| \le Ch^{2\beta}.$$

□

4.6.3 Numerical Examples

In this section, we present some preliminary examples using Lagrange elements. As usual, we choose two polygonal domains: the unit square given by

$$(-1/2, 1/2) \times (-1/2, 1/2),$$

and the L-shaped domain given by

$$(-1/2, 1/2) \times (-1/2, 1/2) \setminus [0, 1/2] \times [-1/2, 0].$$

We generate initial quasi-uniform meshes with $h \approx 0.1$ for the two domains and uniformly refine them three times. Again, we use the relative error defined as

$$\text{Rel. Err.} = \frac{|\lambda_{h_i} - \lambda_{h_{i+1}}|}{\lambda_{h_{i+1}}},$$

where λ_{h_i} is the computed smallest eigenvalue on the mesh with size h_i. We set the penalty parameter $\sigma = 20$ for all numerical examples according to the criteria in [163].

Let $m = 1/15$ and $\tau = 4$. We first choose quadratic Lagrange elements and compute the smallest 6 eigenvalues for the two domains. In Table. 4.20, we show the eigenvalues for the unit square. It is clear that all eigenvalues converge as the mesh size decreases. Similar behavior can be observed for the L-shaped domain (Table. 4.21).

h	1st	2nd	3rd	4th	5th	6th
1/10	3.8145	6.4631	6.4750	9.3590	11.3859	12.2733
1/20	3.6706	6.0516	6.0567	5.9342	10.3035	11.2301
1/40	3.6293	5.9342	5.9358	8.2867	10.0001	10.9324
1/80	3.6180	5.9022	5.9027	8.2194	9.9188	10.8506

Table 4.20: The first 6 eigenvalues of the unit square ($m = 1/15, k = 2$).

h	1st	2nd	3rd	4th	5th	6th
1/10	9.5893	10.8592	12.4876	15.1766	17.7363	22.6025
1/20	8.7479	9.9375	11.3513	13.3911	15.3455	19.4319
1/40	8.4692	9.6627	11.0076	12.8613	14.6383	18.4943
1/80	8.3748	9.5839	10.9096	12.7131	14.4357	18.2145

Table 4.21: The first 6 eigenvalues of the L-shaped domain ($m = 1/15, k = 2$).

In Fig. 4.8, we show the first and second eigenfunctions for the two domains. In Fig. 4.9, we plot relative errors for the first and the second eigenvalues against mesh

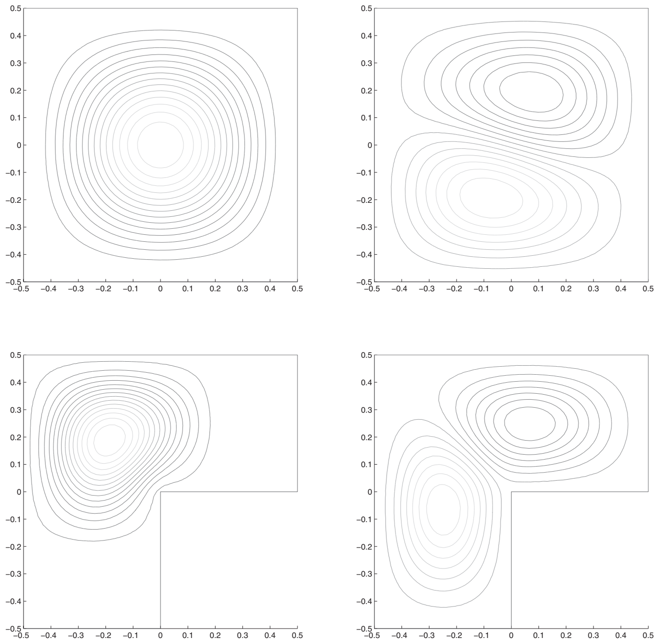


Figure 4.8: The first row: the first and the second eigenfunctions for the unit square. The second row: the first and the second eigenfunctions for the L-shaped domain.

sizes in log scale. For the unit square, we can see roughly the second order convergence is achieved for both eigenvalues. For the L-shaped domain, the convergence rate of the first eigenvalue is less than 2 due to the reentrant angle which leads to low regularity of the eigenfunction. The numerical result suggests that the convergence rate should be $O(h^{4/3})$ (dotted line). The convergence rate of the second eigenvalue is higher indicating the second eigenfunction is smoother than the first one.

In Fig. 4.10, we repeat the plot for cubic Lagrange elements. For the unit square, we see that the relative error is roughly of $O(h^4)$ for both eigenvalues. For the L-shaped domain, the convergence rate is less than $O(h^4)$ for both eigenvalues. Again, the numerical results suggest that the convergence rate should be $O(h^{4/3})$ for the first eigenvalue. However, the second eigenfunction has more regularity than the first eigenfunction which ends up with higher convergence rate. We note that the order of convergence rate is related to the regularities of the eigenfunctions. If the multiplicity of the eigenvalue is more than one, the convergence rate is related to the regularity of the eigenspace [23].

Next we set $m = 1/(7 + x + y)$ and $\tau = 4$. We first choose the quadratic Lagrange element and show the first 6 eigenvalues for the unit square in Table. 4.22. The second and third computed eigenvalues are the approximations of an eigenvalue

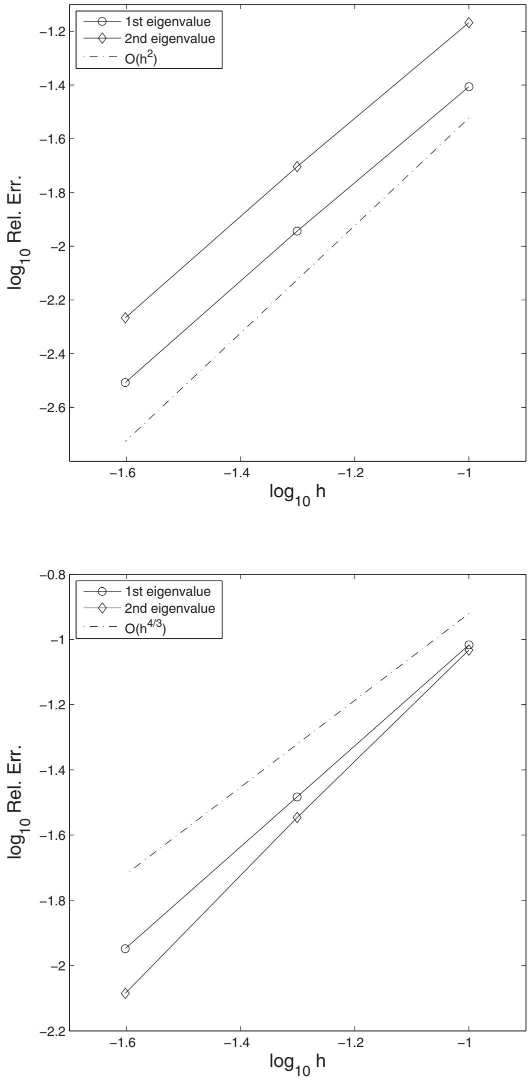


Figure 4.9: Convergence rates of the first and second eigenvalues by the quadratic Lagrange element ($k = 2$). Top: the unit square. Bottom: the L-shaped domain.

with multiplicity 2. The plot of these two eigenfunctions also supports our argument (see Fig. 4.11).

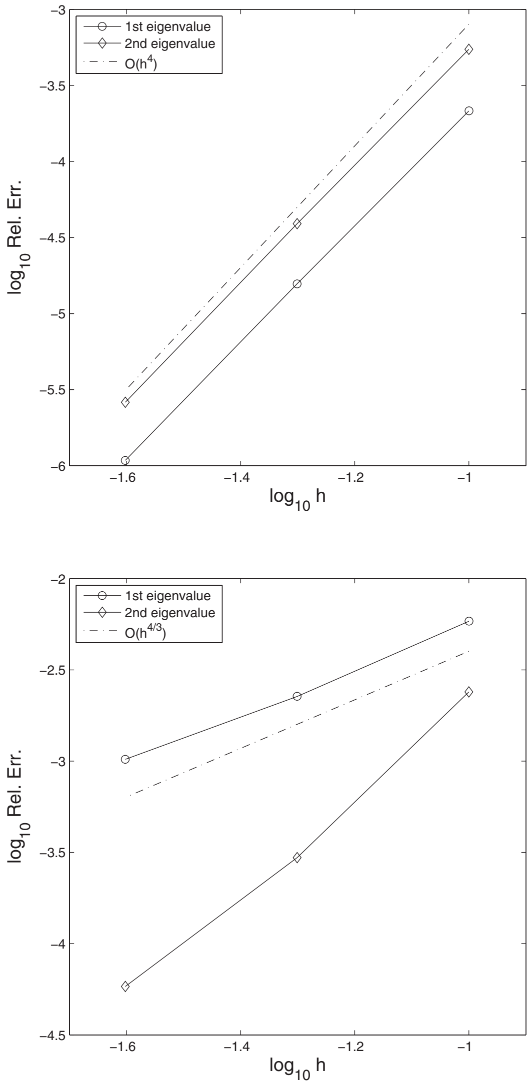


Figure 4.10: Convergence rates of the first and second eigenvalues by the cubic Lagrange element ($k = 3$). Top: the unit square. Bottom: the L-shaped domain.

Similar to the unit square, we show the results for the L-shaped domain in Table. 4.23. The plots for the first and second eigenfunctions are shown in Fig. 4.12.

h	1st	2nd	3rd	4th	5th	6th
1/10	7.4740	13.5351	13.5614	19.8779	24.2052	25.9899
1/20	7.1635	12.6567	12.6665	18.0989	21.8852	23.7695
1/40	7.0742	12.4054	12.4086	17.5744	21.2382	23.1319
1/80	7.0498	12.3370	12.3380	17.4295	21.0633	22.9562

Table 4.22: The first 6 eigenvalues of the unit square ($m = 1/(7 + x + y), k = 2$).

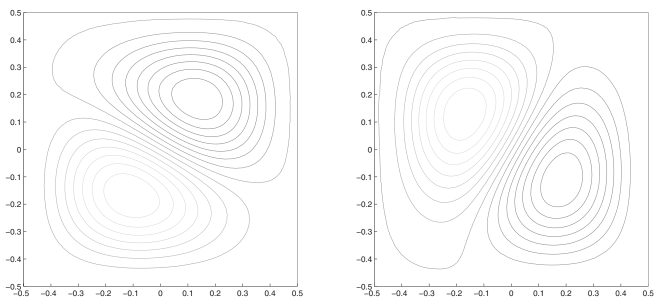


Figure 4.11: The second and third eigenfunctions of the unit square ($m = 1/(7 + x + y)$).

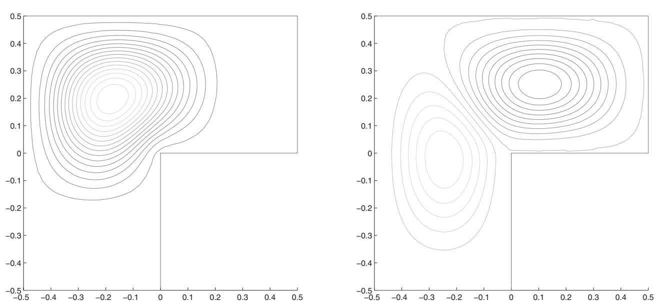


Figure 4.12: The first and second eigenfunctions of the L-shaped domain ($m = 1/(7 + x + y)$).

h	1st	2nd	3rd	4th	5th	6th
1/10	20.1694	22.7661	26.7383	32.2638	38.4223	47.6902
1/20	18.4038	20.8072	24.2673	28.4764	33.1298	40.7528
1/40	17.8190	20.2204	23.5332	27.3577	31.5403	38.7368
1/80	17.6199	20.0520	23.3250	27.0446	31.0834	38.1561

Table 4.23: The first 6 eigenvalues of the L-shaped domain ($m = 1/(7+x+y), k = 2$).

4.7 Appendix: MATLAB Code for the Mixed Method

In this section, we illustrate the implementation of the mixed method (4.26) described in Section 4.3 for the biharmonic eigenvalue problem using linear Lagrange elements. Using the subroutines in Chapter 3 for the Dirichlet eigenvalue problem, the MATLAB code of the mixed method for the biharmonic eigenvalue problem contains a few lines.

We assume a triangular mesh $\mathcal{T}$ for Ω is given. Let V_h be the linear Lagrange element space. Let $\{\phi_1, \phi_2, \dots, \phi_N\}$ be the basis functions associated with the interior nodes of $\mathcal{T}$ and $\{\phi_{N+1}, \dots, \phi_{N+M}\}$ be the basis functions associated with the boundary nodes of $\mathcal{T}$. In other words,

$$\text{span}\{\phi_1, \phi_2, \dots, \phi_N, \phi_{N+1}, \dots, \phi_{N+M}\} = V_h$$

and

$$\text{span}\{\phi_1, \phi_2, \dots, \phi_N\} = V_h \cap H_0^1(\Omega).$$

Let $u = \sum_{i=1}^{N} u_i\phi_i$ and $\sigma = \sum_{i=1}^{N+M} \sigma_i\phi_i$. Define two vectors

$$\mathbf{u} = (u_1, \dots, u_N)^T$$

and

$$\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_{N+M})^T.$$

The stiffness matrix S is given by

$$S = (\nabla\phi_j, \nabla\phi_i), \quad i, j = 1, \dots, N+M,$$

and the mass matrix M is given by

$$M = (\phi_j, \phi_i), \quad i, j = 1, \dots, N+M.$$

The discrete form for (4.26) is given by

$$M\boldsymbol{\sigma} - A\mathbf{u} = 0, \tag{4.108a}$$

$$A^T\boldsymbol{\sigma} = \lambda D\mathbf{u}, \tag{4.108b}$$

where

$$A = S(1 : N + M, 1 : N),$$
$$D = M(1 : N, 1 : N).$$

We can solve $\mathbf{u}$ using (4.108a)

$$\boldsymbol{\sigma} = M^{-1} A\mathbf{u}.$$

Substitution $\mathbf{u}$ in (4.108b) leads to the following generalized eigenvalue problem

$$A^T M^{-1} A\mathbf{u} = \lambda D\mathbf{u}.$$

Assume that $\mathcal{T}$ is given in MATLAB code, i.e., 'p' contains the vertices, 't' contains the triangles, and 'e' contains the boundary edges. The following MATLAB code computes 'num' smallest biharmonic eigenvalues.

```
1. function lambda = BiharmonicEig(p, t, e, num)
2. [S, M]=assemble(p,t);
3. N=length(p);
%-------Find boundary nodes-------------------------
4. bdnodeE = unique([e(1,:),e(2,:)]);
5. Inode = setdiff(linspace(1,N,N), bdnodeE);
6. A = S(:, Inode);
7. D = M(Inode, Inode);
% -------- clamped plate eigenvalues --------------
8. disp('clamped plate eigenvalues');
9. [V,D]=eigs(A'*inv(M)*A, D, num,'sm');
10 lambda = diag(D);
```

Line 2 calls the subroutine 'assemble' to construct the stiffness matrix S and the mass matrix M using the linear Lagrange element. The subroutine 'assemble.m' is given in Section 3.5.

Lines 4 and 5 find all interior vertices.

Lines 6 and 7 set two matrices A and D.

Line 9 computes 'num' smallest eigenvalues.

Chapter 5

The Maxwell's Eigenvalue Problem

5.1 Introduction

The classic equations describing the macroscopic electromagnetic field are called the Maxwell's equations. The finite element methods for the Maxwell's equations and the associated eigenvalue problem have been studied by many researchers [42, 164, 226, 43, 38, 113, 72, 147, 39, 202, 35]. For example, discontinuous Galerkin methods are considered in [58, 2, 59], non-conforming Maxwell's eigensolvers are discussed in [51], mixed methods are discussed in detail by Boffi in [38, 35], and nodal elements are used by electric engineers [77, 110]. Note that for non-convex polyhedra, nodal elements might lead to spurious eigenvalues [42, 176, 177] and a suitable remedy is needed [29, 99, 100, 101].

In this chapter, we focus on the mixed method by Kikuchi [168] and follow the analysis of Demkowicz and Monk [112].

We first introduce the Maxwell's equations following [202]. Let Ω be a bounded simply connected Lipschitz polyhedron. Let $\mathcal{E}$, $\mathcal{H}$, $\mathcal{D}$, and $\mathcal{B}$ be the electric field, the magnetic field, the electric displacement, and the magnetic induction, respectively. In addition, let ρ be the electric charge density and $\mathcal{J}$ be the current density function.

Maxwell's equations in Ω are given by

$$\frac{\partial \mathcal{B}}{\partial t} + \nabla \times \mathcal{E} = 0, \qquad \text{Faraday's law} \tag{5.1a}$$

$$\nabla \cdot \mathcal{D} = \rho, \qquad \text{Gauss's law} \tag{5.1b}$$

$$\frac{\partial \mathcal{D}}{\partial t} - \nabla \times \mathcal{H} = -\mathcal{J}, \qquad \text{Ampére's circuital law} \tag{5.1c}$$

$$\nabla \cdot \mathcal{B} = 0. \qquad \text{solenoidal} \tag{5.1d}$$

We consider the time-harmonic case, i.e.,

$$\mathcal{E}(\mathbf{x}, t) = \mathcal{R}\left(\exp(-i\omega t)\hat{\mathbf{E}}(\mathbf{x})\right), \tag{5.2}$$

$$\mathcal{D}(\mathbf{x}, t) = \mathcal{R}\left(\exp(-i\omega t)\hat{\mathbf{D}}(\mathbf{x})\right), \tag{5.3}$$

$$\mathcal{H}(\mathbf{x}, t) = \mathcal{R}\left(\exp(-i\omega t)\hat{\mathbf{H}}(\mathbf{x})\right), \tag{5.4}$$

$$\mathcal{B}(\mathbf{x}, t) = \mathcal{R}\left(\exp(-i\omega t)\hat{\mathbf{B}}(\mathbf{x})\right), \tag{5.5}$$

$$\mathcal{J}(\mathbf{x}, t) = \mathcal{R}\left(\exp(-i\omega t)\hat{\mathbf{J}}(\mathbf{x})\right), \tag{5.6}$$

$$\rho(\mathbf{x}, t) = \mathcal{R}\left(\exp(-i\omega t)\hat{\rho}(\mathbf{x})\right), \tag{5.7}$$

where $\omega > 0$ is the temporal frequency. Simple calculation gives the time-harmonic Maxwell's equations:

$$-i\omega\hat{\mathbf{B}} + \nabla \times \hat{\mathbf{E}} = 0, \tag{5.8a}$$

$$\nabla \cdot \hat{\mathbf{D}} = \hat{\rho}, \tag{5.8b}$$

$$-i\omega\hat{\mathbf{D}} - \nabla \times \hat{\mathbf{H}} = -\hat{\mathbf{J}}, \tag{5.8c}$$

$$\nabla \cdot \hat{\mathbf{B}} = 0. \tag{5.8d}$$

Furthermore, the following two constitutive laws hold:

$$\hat{\mathbf{D}} = \epsilon\hat{\mathbf{E}} \quad \text{and} \quad \hat{\mathbf{B}} = \mu\hat{\mathbf{H}},$$

where ϵ and μ are called the electric permittivity and magnetic permeability, respectively. In general, ϵ and μ are 3×3 positive-definite matrix functions of position. In free space, they are given by $\epsilon_0 I$ and $\mu_0 I$, respectively, where

$$\epsilon_0 \approx 8.845 \times 10^{-12}, \quad \mu_0 = 4\pi \times 10^{-7}.$$

In a conducting material, we have the Ohm's law:

$$\hat{\mathbf{J}} = \sigma\hat{\mathbf{E}} + \hat{\mathbf{J}}_a, \tag{5.9}$$

where σ is the conductivity, $\hat{J}_a$ is the applied current density. If $\sigma > 0$, the material is a conductor. If $\sigma = 0$ and $\epsilon \neq \epsilon_0$, the material is called a dielectric.

To further simplify the equations, we introduce new variables

$$\mathbf{E} = \epsilon^{1/2}\hat{\mathbf{E}} \quad \text{and} \quad \mathbf{H} = \mu_0^{1/2}\hat{\mathbf{H}}.$$

The relative permittivity and permeability are defined as

$$\epsilon_r = \frac{1}{\epsilon_0}\left(\epsilon + \frac{i\sigma}{\omega}\right)$$

and

$$\mu_r = \frac{\mu}{\mu_0},$$

respectively.

Substituting them into (5.8), we obtain the following system

$$-i\kappa\mu_r\mathbf{H} + \nabla \times \mathbf{E} = 0, \tag{5.10}$$

$$-i\kappa\epsilon_r\mathbf{E} - \nabla \times \mathbf{H} = -\frac{1}{i\kappa}\mathbf{F}, \tag{5.11}$$

$$\nabla \cdot (\epsilon_r\mathbf{E}) = -\frac{1}{\kappa^2}\nabla \cdot \mathbf{F}, \tag{5.12}$$

$$\nabla \cdot (\mu_r\mathbf{H}) = 0, \tag{5.13}$$

where $\kappa = \omega\sqrt{\epsilon_0\mu_0}$ is called the wavenumber and $\mathbf{F} = i\kappa\mu_0^{1/2}\hat{\mathbf{J}}_a$. Eliminating the magnetic field $\mathbf{H}$, we obtain the second-order Maxwell's system

$$\nabla \times (\mu_r^{-1}\nabla \times \mathbf{E}) - \kappa^2\epsilon_r\mathbf{E} = \mathbf{F} \tag{5.14}$$

with the divergence condition (5.12) and the perfect electrically conducting boundary condition for $\mathbf{E}$

$$\boldsymbol{\nu} \times \mathbf{E} = 0 \quad \text{on } \partial\Omega, \tag{5.15}$$

where $\boldsymbol{\nu}$ is the unit outward norm to $\partial\Omega$.

The values of κ such that (5.14) fails to have a unique solution are called Maxwell's eigenvalues or resonant frequencies of Ω. For simplicity, we will assume that $\mu_r = \epsilon_r = 1$ in the rest of this chapter. For $\mathbf{f}$ such that $\nabla \cdot \mathbf{f} = \mathbf{0}$, we consider the following curl-curl source problem of finding $\mathbf{E}$ such that

$$\nabla \times \nabla \times \mathbf{E} = \mathbf{f} \qquad \text{in } \Omega \tag{5.16a}$$

$$\nabla \cdot \mathbf{E} = 0 \qquad \text{in } \Omega \tag{5.16b}$$

$$\boldsymbol{\nu} \times \mathbf{E} = \mathbf{0} \qquad \text{on } \partial\Omega. \tag{5.16c}$$

The Maxwell's eigenvalue problem is to find $(\lambda, \mathbf{E})$ such that

$$\nabla \times \nabla \times \mathbf{E} = \lambda\mathbf{E} \qquad \text{in } \Omega \tag{5.17a}$$

$$\nabla \cdot \mathbf{E} = 0 \qquad \text{in } \Omega \tag{5.17b}$$

$$\boldsymbol{\nu} \times \mathbf{E} = \mathbf{0} \qquad \text{on } \partial\Omega. \tag{5.17c}$$

5.2 The Maxwell's Eigenvalue Problem

5.2.1 Preliminaries

To analyze the finite element method for Maxwell's equations, we need some functional analysis tools. The presentation of this part follows closely to the book by Monk [202]. We first define collectively compact operators.

Definition 5.2.1. *Let X be a Hilbert space. A set*

$$\mathcal{A} = \{T_n : X \to X, n = 0, 1, 2, \ldots\}$$

of bounded linear operators is called collectively compact if, for each bounded set $U \subset X$, the set

$$\mathcal{A} = \{T_n u | \text{ for all } u \in U \text{ and } T_n \in \mathcal{A}\}$$

is relatively compact.

Collectively compact operators have the following properties.

Lemma 5.2.1. *Let X be a Hilbert space. Assume that*

$$\mathcal{A} = \{T_n : X \to X, n = 0, 1, 2, \ldots\}$$

of bounded linear operators is collectively compact. Then the operators $\{T_n, n = 0, 1, 2, \ldots\}$ are uniformly bounded.

Proof. Let $U = \{u \in X | \|u\|_X = 1\}$. By the definition of collectively compact operators, the set $\mathcal{A}(U)$ is relatively compact and thus bounded. This implies a uniform bound on $\|T_n\|_{X \to X}, n = 0, 1, 2, \ldots$. □

Definition 5.2.2. *The operators $\{T_n, n = 0, 1, 2, \ldots\}$ are said to converge pointwise to an operator $T : X \to X$ if, for each $f \in X, T_n f \to Tf$ in X as $n \to \infty$.*

We quote two results on pointwise convergent operators in [202].

Lemma 5.2.2. *(Lemma 2.50 of [202]) Let X and Y be Hilbert spaces and let $T_n : X \to Y, n = 1, 2, \ldots$ be a family of bounded, linear, and pointwise convergent operators with limit operator $T : X \to Y$. Then the convergence is uniform on compact subsets U of X, i.e.,*

$$\sup_{\phi \in U} \|T_n \phi - T\phi\|_Y \to 0 \quad \text{as } n \to \infty.$$

Lemma 5.2.3. *Suppose $\{T_n : X \to X, n = 0, 1, 2, \ldots\}$ is a collectively compact set of bounded linear operators and that the operators are pointwise convergent to a compact operator $T : X \to X$. Then*

$$\|(T_n - T)T\|_{\mathcal{L}(X,X)} \to 0 \quad \text{and} \quad \|(T_n - T)T_n\|_{\mathcal{L}(X,X)} \to 0$$

as $n \to \infty$.

5.2.2 The Curl-curl Problem

Let $\Omega \subset \mathbb{R}^3$ be a bounded, simply connected Lipschitz polyhedral domain. The functional space for the Maxwell's equations is $H_0(\text{curl};\Omega)$ defined as

$$H(\text{curl};\Omega) := \left\{\mathbf{u} \in L^2(\Omega)^3 \mid \nabla \times \mathbf{u} \in L^2(\Omega)^3\right\}.$$

The inner product is defined as

$$(\mathbf{u},\mathbf{v})_{H(\text{curl};\Omega)} = (\mathbf{u},\mathbf{v}) + (\nabla \times \mathbf{u}, \nabla \times \mathbf{v}) \quad \mathbf{u},\mathbf{v} \in H(\text{curl};\Omega),$$

which induces a norm $\|\cdot\|_{H(\text{curl};\Omega)}$ on $H(\text{curl};\Omega)$. Next we define

$$H_0(\text{curl};\Omega) := \{\mathbf{u} \in H(\text{curl};\Omega) \mid \mathbf{u} \times \boldsymbol{\nu} = 0 \text{ on } \partial\Omega\}.$$

For $H_0(\text{curl};\Omega)$, the well-known Helmholtz decomposition holds.

Theorem 5.2.4. *Let $\nabla H_0^1(\Omega)$ be the set of gradients of functions in $H_0^1(\Omega)$. Then $\nabla H_0^1(\Omega)$ is a closed subspace of $H_0(\text{curl};\Omega)$ such that*

$$H_0(\text{curl};\Omega) = Y \oplus \nabla H_0^1(\Omega)$$

where

$$Y = \{\mathbf{u} \in H_0(\text{curl};\Omega) | (\mathbf{u}, \nabla p) = 0 \quad \textit{for all } p \in H_0^1(\Omega)\}. \tag{5.18}$$

We also need the space $H(\text{div};\Omega)$ of functions with square-integrable divergence defined as

$$H(\text{div};\Omega) = \{\mathbf{u} \in L^2(\Omega)^3 \mid \nabla \cdot \mathbf{u} \in L^2(\Omega)\},$$

equipped with the scalar product

$$(\mathbf{u},\mathbf{v})_{H(\text{div};\Omega)} = (\mathbf{u},\mathbf{v}) + (\nabla \cdot \mathbf{u}, \nabla \cdot \mathbf{v})$$

and the corresponding norm $\|\cdot\|_{H(\text{div};\Omega)}$.

Taking the divergence-free condition into account, we define

$$H(\text{div}^0;\Omega) = \{\mathbf{u} \in H(\text{div};\Omega) \,|\, \nabla \cdot \mathbf{u} = 0 \text{ in } \Omega\}. \tag{5.19}$$

The following compactness result is a simplified version of Theorem 4.7 of [202].

Theorem 5.2.5. *If Ω is a bounded Lipschitz domain, the space Y is compactly embedded in $L^2(\Omega)^3$.*

For functions in Y, the following Friedrichs inequality holds.

Theorem 5.2.6. *Suppose that Ω is a bounded Lipschitz domain. If Ω is simply connected, and has a connected boundary, there is a constant $C > 0$ such that for every $\mathbf{u} \in Y$*

$$\|\mathbf{u}\| \leq C\|\nabla \times \mathbf{u}\|. \tag{5.20}$$

Using the space Y, the weak formulation for (5.16) can be stated as follows. Given $\mathbf{f} \in H(\mathrm{div}^0;\Omega)$, find $\mathbf{u} \in Y$ such that

$$(\nabla \times \mathbf{u}, \nabla \times \boldsymbol{\phi}) = (\mathbf{f}, \boldsymbol{\phi}) \quad \text{for all } \boldsymbol{\phi} \in Y. \tag{5.21}$$

Theorem 5.2.7. *There exists a unique solution* $\mathbf{u} \in Y$ *for* (5.21).

Proof. It is clear that the sesqulinear form $(\nabla \times \mathbf{u}, \nabla \times \boldsymbol{\phi})$ on $Y \times Y$ is bounded. Due to the Friedrichs inequality (5.20), it is also coercive. Then the theorem follows the Lax-Milgram Lemma 1.3.1. □

Consequently, we can define a solution operator

$$T : H(\mathrm{div}^0;\Omega) \subset L^2(\Omega)^3 \to Y \subset L^2(\Omega)^3$$

such that

$$\mathbf{u} = T\mathbf{f}. \tag{5.22}$$

It is obvious that T is self-adjoint and compact due to Theorem 5.2.5.

The finite element method for the curl-curl problem uses a mixed formulation since a Y-conforming finite element space is difficult to construct. In the following, we collect some results for a mixed formulation for the curl-curl problem and refer the readers to [168, 147, 202] for details.

The mixed problem is stated as follows. Given $\mathbf{f} \in H(\mathrm{div}^0;\Omega)$, find $(\mathbf{u}, p) \in H_0(\mathrm{curl};\Omega) \times H_0^1(\Omega)$ such that

$$(\nabla \times \mathbf{u}, \nabla \times \boldsymbol{\phi}) + (\boldsymbol{\phi}, \nabla p) = (\mathbf{f}, \boldsymbol{\phi}) \qquad \text{for all } \boldsymbol{\phi} \in H_0(\mathrm{curl};\Omega), \tag{5.23a}$$

$$(\mathbf{u}, \nabla q) = 0 \qquad \text{for all } q \in H_0^1(\Omega). \tag{5.23b}$$

The results introduced in Chapter 1 can be employed to study the above variational formulation. To this end, we define the sesqulinear forms

$$a : H_0(\mathrm{curl};\Omega) \times H_0(\mathrm{curl};\Omega) \to \mathbb{C} \quad \text{and} \quad b : H_0(\mathrm{curl};\Omega) \times H_0^1(\Omega) \to \mathbb{C}$$

such that

$$a(\mathbf{u}, \mathbf{v}) := (\nabla \times \mathbf{u}, \nabla \times \mathbf{v}) \qquad \text{for all } \mathbf{u}, \mathbf{v} \in H_0(\mathrm{curl};\Omega), \tag{5.24a}$$

$$b(\mathbf{u}, \xi) := (\mathbf{u}, \nabla \xi) \qquad \text{for all } \mathbf{u} \in H_0(\mathrm{curl};\Omega), \xi \in H_0^1(\Omega). \tag{5.24b}$$

Then (5.23) can be written as follows. Find $(\mathbf{u}, p) \in H_0(\mathrm{curl};\Omega) \times H_0^1(\Omega)$ such that

$$a(\mathbf{u}, \boldsymbol{\phi}) + b(\boldsymbol{\phi}, p) = (\mathbf{f}, \boldsymbol{\phi}) \qquad \text{for all } \boldsymbol{\phi} \in H_0(\mathrm{curl};\Omega), \tag{5.25a}$$

$$b(\mathbf{u}, q) = 0 \qquad \text{for all } q \in H_0^1(\Omega). \tag{5.25b}$$

Theorem 5.2.8. *Given* $\mathbf{f} \in H(\mathrm{div}^0;\Omega)$*, the mixed problem* (5.25) *has a unique solution*

$$(\mathbf{u}, p) \in H_0(\mathrm{curl};\Omega) \times H_0^1(\Omega).$$

Furthermore, $p = 0$ *and* $\mathbf{u}$ *satisfies*

$$\|\mathbf{u}\|_{H(\mathrm{curl};\Omega)} \leq C\|\mathbf{f}\|.$$

Proof. The problem is in the form of Theorem 1.3.3. One only needs to check if the conditions of Theorem 1.3.3 are satisfied.

Define a subspace of $H(\text{curl};\Omega)$:

$$Z = \{\mathbf{u} \in H_0(\text{curl};\Omega) \,|\, b(\mathbf{u},\xi) = 0 \quad \text{for all } \xi \in H_0^1(\Omega)\}.$$

This space coincides with Y defined in (5.18). Due to the Friedrichs inequality (Theorem 5.2.6), for $\mathbf{u} \in Y$, we have that

$$\begin{aligned} a(\mathbf{u},\mathbf{u}) &= (\nabla \times \mathbf{u}, \nabla \times \mathbf{u}) \\ &\geq \frac{1}{2}\|\nabla \times \mathbf{u}\|^2 + C\|\mathbf{u}\|^2 \\ &\geq \alpha\|\mathbf{u}\|^2_{H(\text{curl};\Omega)} \end{aligned}$$

for some $\alpha > 0$. Thus $a(\cdot,\cdot)$ is coercive on Y.

For $b(\cdot,\cdot)$, by choosing $\mathbf{w} = \nabla p$ and using the Poincaré inequality for $H_0^1(\Omega)$, we have that

$$\sup_{\mathbf{w}\in H(\text{curl};\Omega)} \frac{|(\mathbf{w},\nabla p)|}{\|\mathbf{w}\|_{H(\text{curl};\Omega)}} \geq \frac{|(\nabla p,\nabla p)|}{\|\nabla p\|_{H(\text{curl};\Omega)}} \geq \|\nabla p\| \geq \beta\|p\|_{H^1}$$

for some $\beta > 0$. Then by Theorem 1.3.3, there exists a unique solution $(\mathbf{u}, p)$ of (5.23). Choosing $\mathbf{u} = \nabla p$ in (5.23) and using the fact that $\mathbf{f}$ is divergence free, we see that $(\nabla p, \nabla p) = 0$ and thus $p = 0$.

Since $\mathbf{f} \in H(\text{div}^0;\Omega)$, by the Cauchy-Shwarz inequality, we have that

$$(\mathbf{f},\mathbf{v}) \leq \|\mathbf{f}\| \cdot \|\mathbf{v}\| \leq \|\mathbf{f}\|\|\mathbf{v}\|_{H(\text{curl};\Omega)},$$

which implies $\|\mathbf{f}\|_{H(\text{curl};\Omega)'} \leq \|\mathbf{f}\|$. Then Theorem 1.3.3 leads to

$$\|\mathbf{u}\|_{H(\text{curl};\Omega)} \leq C\|\mathbf{f}\|.$$

The proof is complete. □

5.2.3 Divergence-conforming Elements

To treat three dimensional problems involving divergence operators, it is desirable to have divergence-conforming elements. Although we do not use it here, it is relevant to the Maxwell's equations and would be helpful to understand the edge element in the next section. The presentation here follows Section 5.4 of [202].

We define $\tilde{P}_k$ the space of homogeneous polynomials of total degree exactly k and

$$D_k = (P_{k-1})^3 \oplus \tilde{P}_{k-1}\mathbf{x},$$

where $\mathbf{x} = (x_1, x_2, x_3)^T$. It is easy to show the following properties of D_k:

(1) $\mathbf{u} \in D_1$ if and only if $\mathbf{u} = \mathbf{a} + b\mathbf{x}$, where $\mathbf{a} \in \mathbb{C}^3$ and $b \in \mathbb{C}$;

(2) $\dim(D_k) = \frac{1}{2}(k+3)(k+1)k$;

(3) $\nabla \cdot D_k = P_{k-1}$.

Now we define the divergence-conforming finite element due to Nédélec [208].

Definition 5.2.3. *Let $\hat{K}$ be the reference tetrahedron whose vertices are $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$. Let $\hat{\mathcal{P}} = D_k$ and $\mathbf{u} \in (H^{1/2+\delta}(\hat{K}))^3, \delta > 0$. Then the degrees of freedom $\hat{\mathcal{N}}$ are defined as*

$$N_{\hat{f}}(\hat{\mathbf{u}}) = \left\{ \int_{\hat{f}} \hat{\mathbf{u}} \cdot \boldsymbol{\nu} q \, ds \quad \text{for all } q \in P_{k-1}(\hat{f}) \text{ and } \hat{f} \right\}, \tag{5.26}$$

$$N_{\hat{K}}(\hat{\mathbf{u}}) = \left\{ \int_{\hat{K}} \hat{\mathbf{u}} \cdot \mathbf{q} \quad \text{for all } \mathbf{q} \in (P_{k-2})^3 \right\}, \tag{5.27}$$

where $\boldsymbol{\nu}$ is the unit outward normal to $\hat{f}$.

Note that it would be ideal if we only require $\mathbf{u} \in H(\text{div}; \hat{K})$. However, the traces of such functions might not have the regularity for the above degrees of freedom to be well defined.

Let $\mathcal{T}$ be a tetrahedral mesh for Ω and $K \in \mathcal{T}$. The affine mapping $F_K : \hat{K} \to K$ is given by

$$F_K \hat{\mathbf{x}} = B_K \hat{\mathbf{x}} + \hat{\mathbf{b}}.$$

We relate the vectorial function $\mathbf{u}$ on $K \in \mathcal{T}$ to $\hat{\mathbf{u}}$ on the reference tetrahedron $\hat{K}$ such that

$$\mathbf{u} \cdot F_K = \frac{1}{\det(B_k)} B_K \hat{\mathbf{u}}. \tag{5.28}$$

If $\boldsymbol{\nu}$ is a unit outward norm to $\hat{K}$, then $\boldsymbol{\nu}$ such that

$$\boldsymbol{\nu} \circ F_K = \frac{1}{|(B_K^{-1})^T \hat{\boldsymbol{\nu}}|} (B_K^{-1})^T \hat{\boldsymbol{\nu}}$$

is a unit (inward or outward depending on the sign of $\det(B_K)$) normal to K. It is shown in [202] that the degrees of freedom for $\hat{\mathbf{u}}$ on $\hat{K}$ and for $\mathbf{u}$ on K are identical provided $\det(B_K) > 0$.

If $\mathbf{u} \in (H^{1/2+\delta}(K))^3, \delta > 0$, then there exists a unique $\mathbf{u}_K \in D_k$ such that

$$M_f(\mathbf{u} - \mathbf{u}_K) = \{0\} \quad \text{and} \quad M_K(\mathbf{u} - \mathbf{u}_K) = \{0\}.$$

Let $\{\mathcal{T}_h, h > 0\}$ be a regular family of meshes of Ω. The global set of degrees of freedom is defined as

$$\mathcal{N} = \cup_{K \in \mathcal{T}_h} \mathcal{N}_K.$$

We have the following theorem.

Theorem 5.2.9. *A vector function $\mathbf{u} \in D_k$ defined on tetrahedron K is determined uniquely by the degree of freedom* (5.26) *and* (5.27). *Moreover, the space W_h of finite elements for the mesh $\mathcal{T}_h$ defined element-wise is divergence-conforming, i.e., $W_h \subset H(div; \Omega)$.*

The above theorem implies that there exist an interpolation operator

$$\mathbf{w}_h : H^{1/2+\delta}(\Omega)^3 \to W_h, \quad \mathbf{w}_h \mathbf{u}|_K = \mathbf{w}_k \mathbf{u} \quad \text{for each } K \in \mathcal{T}_h. \tag{5.29}$$

The interpolation error, which is fundamental for the error analysis, is proved in [202] (Theorem 5.25 therein).

Theorem 5.2.10. *Suppose $\{\mathcal{T}_h\}_{h>0}$ is a regular family of meshes on Ω and $0 < \delta < 1/2$. Then if $\mathbf{u} \in H^s(\Omega)^3$, $1/2 + \delta \le s \le k$, there exists a constant C independent of h and $\mathbf{u}$ such that*

$$\|\mathbf{u} - \mathbf{w}_h \mathbf{u}\|_{L^2(\Omega)^3} \le C h^s \|\mathbf{u}\|_{H^s(\Omega)^3}, \quad 1/2 + \delta \le s \le k, \tag{5.30}$$

and

$$\|\nabla \cdot (\mathbf{u} - \mathbf{w}_h \mathbf{u})\| \le C h^s \|\nabla \cdot \mathbf{u}\|_{H^s(\Omega)}, \quad 1/2 + \delta \le s \le k. \tag{5.31}$$

5.2.4 Curl-conforming Edge Elements

The lowest-order edge element first appeared in [244] by Whitney. Later Nédélec rigorously extended edge elements to higher orders [208]. We present a short introduction of edge elements following [202].

We assume that the domain Ω is covered by a regular tetrahedral mesh and denote the mesh by $\mathcal{T}_h$ where h is the maximum diameter of the elements in $\mathcal{T}_h$. Let P_k be the space of polynomials of maximum total degree k and $\tilde{P}_k$ the space of homogeneous polynomials of degree k. We define

$$R_k = (P_{k-1})^3 \oplus \{\mathbf{p} \in (\tilde{P}_k)^3 \mid \mathbf{x} \cdot \mathbf{p}(\mathbf{x}) = 0 \text{ for all } \mathbf{x} \in \mathbb{R}^3\}.$$

Then one has

$$\dim(R_k) = \frac{1}{2} k(k+2)(k+3).$$

In addition, the following two properties hold:

1. $(P_k)^3 = R_k \oplus \nabla \tilde{P}_{k+1}$,
2. If $\mathbf{u} \in R_k$ such that $\nabla \times \mathbf{u} = 0$ then $\mathbf{u} = \nabla p$ for some $p \in P_k$.

Let $\hat{K}$ be the reference tetrahedron. The degrees of freedom of edge elements are associated with the edges $\hat{e}$, faces $\hat{f}$, and the volume of $\hat{K}$. Let $\hat{\boldsymbol{\tau}}$ denote a unit vector parallel to $\hat{e}$ and $\hat{\boldsymbol{\nu}}$ denote the unit outward normal to $\hat{f}$. The degrees of freedom are given by

$$M_{\hat{e}}(\hat{\mathbf{u}}) = \left\{ \int_{\hat{e}} \hat{\mathbf{u}} \cdot \hat{\boldsymbol{\tau}} \hat{q} \, \mathrm{d}s \quad \text{for all } \hat{q} \in P_{k-1}(\hat{e}) \text{ for } \hat{e} \text{ of } \hat{K} \right\},$$

$$M_{\hat{f}}(\hat{\mathbf{u}}) = \left\{ \frac{1}{|\hat{f}|} \int_{\hat{f}} \hat{\mathbf{u}} \cdot \hat{\mathbf{q}} \, \mathrm{d}\hat{A} \quad \text{for all } \hat{\mathbf{q}} \in (P_{k-2}(\hat{f}))^3, \hat{\mathbf{q}} \cdot \hat{\boldsymbol{\nu}} \text{ for } \hat{f} \text{ of } \hat{K} \right\},$$

$$M_{\hat{K}}(\hat{\mathbf{u}}) = \left\{ \int_{\hat{K}} \hat{\mathbf{u}} \cdot \hat{\mathbf{q}} \, \mathrm{d}\hat{V} \quad \text{for all } \hat{\mathbf{q}} \in (P_{k-3}(\hat{K}))^3 \right\},$$

where $|\hat{f}|$ denotes the area of $\hat{f}$.

To program edge elements, if one wishes to use the reference tetrahedron $\hat{K}$, we need to define the mapping

$$F_K : \hat{K} \to K,$$

which is a continuously differentiable bijective and $\det(\mathrm{d}F_K)$ is one sign on $\hat{K}$. In particular, if we assume the vertices of $\hat{K}$ are given by

$$\hat{\mathbf{a}}_1 = (0,0,0)^T, \quad \hat{\mathbf{a}}_2 = (1,0,0)^T, \quad \hat{\mathbf{a}}_3 = (0,1,0)^T, \quad \hat{\mathbf{a}}_4 = (0,0,1)^T,$$

and the vertices of an element K are given by

$$\mathbf{a}_1 = (x_1,y_1,z_1)^T, \quad \mathbf{a}_2 = (x_2,y_2,z_2)^T, \quad \mathbf{a}_3 = (x_3,y_3,z_3)^T, \mathbf{a}_4 = (x_4,y_4,z_4)^T,$$

we have that

$$F_K\hat{\mathbf{x}} := B_K\hat{\mathbf{x}} + \mathbf{b}_k,$$

where

$$B_K = \begin{pmatrix} x_2-x_1 & x_3-x_1 & x_4-x_1 \\ y_2-y_1 & y_3-y_1 & y_4-y_1 \\ z_2-z_1 & z_3-z_1 & z_4-z_1 \end{pmatrix}, \quad \mathbf{b}_K = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix}.$$

The vectors in R_k need to be transformed in a special way (see Sec. 5.5 of [202])

$$\mathbf{u} \circ F_K = (B_K^T)^{-1}\hat{\mathbf{u}} \tag{5.32}$$

such that

$$\nabla \times \mathbf{u} = \frac{1}{\det(\mathrm{B_K})} B_K \hat{\nabla} \times \hat{\mathbf{u}}.$$

Then the vector

$$\boldsymbol{\tau} = \frac{B_K\hat{\boldsymbol{\tau}}}{|B_K\hat{\boldsymbol{\tau}}|} \tag{5.33}$$

is a unit tangent vector to the edge e of K. Under the transformation (5.32), R_K is invariant.

Now we are ready to define the curl-conforming edge element.

Definition 5.2.4. *The curl-conforming edge element $(K, \mathcal{P}, \mathcal{N})$ is defined as follows.*

- *K is a tetrahedron,*
- *$\mathcal{P} = R_k$,*
- *The degrees of freedom are associated with the edges e, faces f, and the volume of an element $K \in \mathcal{T}_h$. Letting $\boldsymbol{\tau}$ denote a unit vector in the direction of e and $\boldsymbol{\nu}$ denote the unit outward normal to f, the degrees of freedom of edge*

elements are given by

$$M_e(\mathbf{u}) = \left\{ \int_e \mathbf{u} \cdot \boldsymbol{\tau} q \, ds \quad \textit{for all } q \in P_{k-1}(e) \right. \tag{5.34}$$
$$\left. \textit{for each edge } e \textit{ of } K \right\},$$

$$M_f(\mathbf{u}) = \left\{ \frac{1}{|f|} \int_f \mathbf{u} \cdot \mathbf{q} \, dA \quad \textit{for all } \mathbf{q} = B_K \hat{\mathbf{q}}, \right. \tag{5.35}$$
$$\left. \hat{\mathbf{q}} \in (P_{k-2}(\hat{f}))^2, \hat{\mathbf{q}} \cdot \hat{\boldsymbol{\nu}} = 0 \, \textit{for each face } f \textit{ of } K \right\},$$

$$M_K(\mathbf{u}) = \left\{ \int_K \mathbf{u} \cdot \mathbf{q} \, dV \quad \textit{for all } \mathbf{q} \, \textit{such that} \right. \tag{5.36}$$
$$\left. \mathbf{q} \circ F_K = \frac{1}{\det(\mathrm{B_K})} B_K \hat{\mathbf{q}}, \hat{\mathbf{q}} \in (P_{k-3}(K))^3 \right\}.$$

Then $\mathcal{N} = M_e(\mathbf{u}) \cup M_f(\mathbf{u}) \cup M_K(\mathbf{u})$.

The edge element has the following properties.

1. Suppose $\det(B_K) > 0$ and the tangent vectors $\boldsymbol{\tau}$ on the edges of K are related to those of $\hat{K}$ by (5.33). Then each of the sets of degrees of freedom (5.34)–(5.36) are identical to the degrees of freedom for $\hat{\mathbf{u}}$ on $\hat{K}$ under the transformation (5.32).
2. If $\mathbf{u} \in R_K$ is such that the degrees of freedom (5.34) and (5.35) vanish, then $\mathbf{u} \times \boldsymbol{\nu} = \mathbf{0}$ on f.
3. If all degrees of freedom of $\mathbf{u} \in R_K$ vanish, then $\mathbf{u} = \mathbf{0}$.

The global curl-conforming edge element space is then given by

$$U_h = \{\mathbf{v} \in H(\text{curl}; \Omega) \mid \mathbf{v}|_K \in R_k \text{ for all } K \in \mathcal{T}_h\}.$$

The $H_0(\text{curl}; \Omega)$ conforming edge element space is simply

$$U_{0,h} = \{\mathbf{u}_h \in U_h \mid \boldsymbol{\nu} \times \mathbf{u}_h = 0 \quad \text{on } \partial\Omega\}, \tag{5.37}$$

which can be obtained by setting the degrees of freedom associated with edges and faces on $\partial\Omega$ to zero.

Assuming $\mathbf{u}$ is smooth enough, the element-wise interpolant $\mathbf{r}_K \mathbf{u} \in R_K$ satisfies

$$M_e(\mathbf{u} - \mathbf{r}_K \mathbf{u}) = M_f(\mathbf{u} - \mathbf{r}_K \mathbf{u}) = M_K(\mathbf{u} - \mathbf{r}_K \mathbf{u}) = \{0\}.$$

Then the global interpolant $\mathbf{r}_h \mathbf{u} \in U_h$ is such that

$$\mathbf{r}_h \mathbf{u}|_K = \mathbf{r}_K \mathbf{u} \quad \text{for all } K \in \mathcal{T}_h.$$

The following result holds for $\mathbf{r}_h \mathbf{u}$.

Lemma 5.2.11. *(Lemma 5.38 of [202]) Suppose there are constants* $\delta > 0$ *and* $p > 2$ *such that* $\mathbf{u} \in H^{1/2+\delta}(K)^3$ *and* $\operatorname{curl} \mathbf{u} \in L^p(K)^3$ *for each* $K \in \mathcal{T}_h$*. Then* $\mathbf{r}_h\mathbf{u}$ *is well defined and bounded.*

The following result provides error estimates for the interpolant.

Lemma 5.2.12. *(Theorem 5.41 of [202]) Let* $\mathcal{T}_h$ *be a regular mesh on* Ω*. Then*

(1) If $\mathbf{u} \in H^s(\Omega)^3$ *and* $\nabla \times \mathbf{u} \in H^s(\Omega)^3$ *for* $1/2 + \delta \leq s \leq k$ *for* $\delta > 0$ *then*

$$\begin{aligned}\|\mathbf{u} - \mathbf{r}_h\mathbf{u}\|_{L^2(\Omega)^3} &+ \|\nabla \times (\mathbf{u} - \mathbf{r}_h\mathbf{u})\|_{L^2(\Omega)^3} \\ &\leq Ch^s \left(\|\mathbf{u}\|_{H^s(\Omega)^3} + \|\nabla \times \mathbf{u}\|_{H^s(\Omega)^3}\right). \quad (5.38)\end{aligned}$$

(2) If $\mathbf{u} \in H^{1/2+\delta}(K)^3$*,* $0 < \delta \leq 1/2$ *and* $\operatorname{curl} \mathbf{u}|_K \in P_k$*, then*

$$\|\mathbf{u} - \mathbf{r}_h\mathbf{u}\|_{L^2(\Omega)^3} \leq C\left(h_K^{1/2+\delta}\|\mathbf{u}\|_{H^{1/2+\delta}(K)^3} + h_K\|\nabla \times \mathbf{u}\|_{L^2(K)^3}\right).$$

(3) If $\mathbf{u} \in H^s(\Omega)^3$ *and* $\nabla \times \mathbf{u} \in H^s(\Omega)^3$ *for* $1/2 + \delta \leq s \leq k$ *and* $\delta > 0$*, the following result holds*

$$\|\nabla \times (\mathbf{u} - \mathbf{r}_h\mathbf{u})\|_{L^2(\Omega)^3} \leq Ch^s\|\nabla \times \mathbf{u}\|_{H^s(\Omega)^3}.$$

For linear edge element, one has that

$$R_1 = \left\{\mathbf{u}(\mathbf{x}) = \mathbf{a} + \mathbf{b} \times \mathbf{x}, \quad \text{where } \mathbf{a}, \mathbf{b} \in \mathbb{C}^3\right\}.$$

The six degrees of freedom are determined from the moments $\int_e \mathbf{u} \cdot \boldsymbol{\tau} \mathrm{d}s$ on the six edges of K. This explains why they are called the edge elements. In fact, the basis function with unit integral on edge $e_{i,j}$, where i, j denote the vertex indices, is given by

$$\boldsymbol{\psi}_{i,j} = \lambda_i \nabla \lambda_j - \lambda_j \nabla \lambda_i,$$

where λ_i is the barycentric coordinate function associated with vertex $\mathbf{a}_i$. In addition, we have that

$$\nabla \times \boldsymbol{\psi}_{i,j} = 2\nabla \lambda_i \times \nabla \lambda_j.$$

This is also called the Whitney's representation of the basis function [244].

Remark 5.2.1. *The edge elements discussed above are sometimes referred to as the first family of edge elements on tetrahedra. There are also the second family of edge elements on tetrahedra (see Chapter 8 of [202]).*

The following inverse inequality for edge elements will be useful in the forthcoming error analysis (see Section 3.6 of [147]).

Lemma 5.2.13. *Let* $\mathcal{T}_h$ *be a regular mesh for* Ω*. Then for* $\mathbf{u}_h \in U_h$*,*

$$\|\mathbf{u}_h\|_{H(\mathrm{curl};\Omega)} \leq Ch^{-1}\|\mathbf{u}_h\|$$

for some constant C *independent of* $\mathbf{u}_h$ *and* h*.*

5.2.5 Convergence Analysis

Let the finite element space for $H_0^1(\Omega)$ be given by

$$S_h = \left\{ p_h \in H_0^1(\Omega) \mid p_h|_K \in P_k \text{ for all } K \in \mathcal{T}_h \right\}.$$

It follows that $\nabla S_h \subset U_{0,h}$. The discrete Helmholtz decomposition can be defined via

$$U_{0,h} = Y_h \oplus \nabla S_h,$$

where Y_h is given by

$$Y_h = \{ \mathbf{u}_h \in U_{0,h} \mid (\mathbf{u}_h, \nabla \xi_h) = 0 \quad \text{for all } \xi_h \in S_h \}. \tag{5.39}$$

Then the discrete problem for (5.23) is to find $(\mathbf{u}_h, p_h) \in U_{0,h} \times S_h$ such that

$$(\nabla \times \mathbf{u}_h, \nabla \times \boldsymbol{\phi}_h) + (\nabla p_h, \boldsymbol{\phi}_h) = (\mathbf{f}, \boldsymbol{\phi}_h) \qquad \text{for all } \boldsymbol{\phi}_h \in U_{0,h}, \tag{5.40a}$$

$$(\mathbf{u}_h, \nabla q_h) = 0 \qquad \text{for all } q_h \in S_h. \tag{5.40b}$$

To prove the well-posedness of the discrete problem, we need an important property of the edge element: the discrete compactness. The property was first studied by Kikuchi for the lowest-order edge element [169]. Demkowicz and Monk extended the result to all orders of edge elements [112]. We will follow the version in [112].

We start with some regularity result for the source problem. For given $\mathbf{f} \in Y'$ with $\nabla \cdot \mathbf{f} = 0$ in Ω, we consider the problem of finding $\mathbf{u} \in Y$ such that

$$\nabla \times \nabla \times \mathbf{u} = \mathbf{f}. \tag{5.41}$$

There exists a constant $\sigma_0 > 0$ such that for all $\sigma, 0 \leq \sigma < \sigma_0$ and $\mathbf{f} \in H^{\sigma-1}(\Omega)^3$, one can write

$$\mathbf{u} = \mathbf{u}^* + \nabla \chi,$$

where $\mathbf{u}^* \in H^{\sigma+1}(\Omega)^3$ and $\chi \in H_0^1(\Omega)$ with $\triangle \chi \in H^\sigma$. Furthermore, one has that

$$\|\mathbf{u}^*\|_{H^{\sigma+1}(\Omega)} + \|\chi\|_{H^1(\Omega)} \leq C \|\mathbf{f}\|_{H^{\sigma-1}(\Omega)}, \tag{5.42a}$$

$$\|\triangle \chi\| \leq C \|\mathbf{f}\|_{H^{-1}(\Omega)}. \tag{5.42b}$$

Let $\Lambda = \{h_n\}_{n=1}^\infty$ be a countable set of mesh sizes such that $h_n \to 0$ as $n \to \infty$.

Definition 5.2.5. *We say the spaces $\{V_h \subset V, h \in \Lambda\}$ have the discrete compactness property if for every sequence $\{\mathbf{u}_h\}_{h\in\Lambda}$ such that*

1. $\mathbf{u}_h \in V_h$ *for each* $h \in \Lambda$*, and*
2. *there is a constant C independent of $\mathbf{u}_h$ such that $\|\mathbf{u}_h\|_V \leq C$ independent of $h \in \Lambda$,*

there exists a subsequence, still denoted $\{\mathbf{u}_h\}$, and a function $\mathbf{u} \in V$ such that

$$\mathbf{u}_h \to \mathbf{u} \text{ strongly in } L_2(\Omega)^3 \text{ as } h \to 0 \text{ in } \Lambda.$$

We would like to show that $Y_h, h \in \Lambda$ has the discrete compactness property. We need the following regularity result.

Lemma 5.2.14. *(Lemma 7.15 of [202]) Let Ω be a bounded simply connected Lipschitz polyhedron. Let $\mathcal{T}_h$ be a regular, quasi-uniform mesh on $\partial\Omega$. Let $\mathbf{u}_h \in Y_h$ and suppose $\mathbf{u} \in Y$ satisfies*

$$\begin{aligned} \nabla \times \mathbf{u} &= \nabla \times \mathbf{u}_h \quad \text{in } \Omega, \\ \boldsymbol{\nu} \times \mathbf{u} &= \boldsymbol{\nu} \times \mathbf{u}_h \quad \text{on } \partial\Omega. \end{aligned}$$

Then there is a $\delta > 0$ with $\delta \leq 1/3$ such that $\mathbf{u} \in H^{1/2+s}(\Omega)^3$ for $0 \leq s < \delta$. Furthermore,

$$\|\mathbf{u}\|_{H^{1/2+s}(\Omega)^3} \leq C\left(\|\nabla \times \mathbf{u}\|_{L^2(\Omega)^3} + \|\boldsymbol{\nu} \times \mathbf{u}\|_{H^s(\partial\Omega)^3}\right).$$

Using the above regularity results, we can show that $\{Y_h, h \in \Lambda\}$ possesses the discrete compactness property. The following theorem is again taken from [202].

Lemma 5.2.15. *Let Ω be a bounded simply connected Lipschitz domain. Assume the meshes $\mathcal{T}_h$ be regular and quasi-uniform. Then Y_h possesses the discrete compactness property.*

The proofs of the above lemmas are rather technical. We refer the readers to [202] for details.

As a consequence, we have the following discrete version of the Friedrichs inequality.

Theorem 5.2.16. *(Lemma 7.20 of [202]) Let Ω be a bounded simply connected Lipschitz domain. Assume that the mesh is regular and quasi-uniform. There exists a positive constant C independent of $h \in \Lambda$ such that if $u_h \in Y_h$, for $h \in \Lambda$ small enough, then*

$$\|\mathbf{u}_h\|_{L^2(\Omega)^3} \leq C\left(\|\nabla \times \mathbf{u}_h\|_{L^2(\Omega)^3} + \|\nu \times \mathbf{u}_h\|_{L^2(\partial\Omega)^3}\right). \tag{5.43}$$

It is clear that, if $\mathbf{u}_h \in Y_h$, i.e., $\nu \times \mathbf{u}_h = \mathbf{0}$, one has that

$$\|\mathbf{u}_h\|_{L^2(\Omega)^3} \leq C\|\nabla \times \mathbf{u}_h\|_{L^2(\Omega)^3}.$$

With Theorem 5.2.16, we can prove the existence and uniqueness of a solution to the discrete problem.

Theorem 5.2.17. *The discrete problem* (5.40) *has a unique solution $(\mathbf{u}_h, p_h) \in U_{0,h} \times S_h$ with $p_h = 0$. In addition, if $(\mathbf{u}, p) \in H_0(\mathrm{curl}; \Omega) \times H_0^1(\Omega)$ is the solution of* (5.25) *with $p = 0$, there exists a constant C independent of h, $\mathbf{u}$, and $\mathbf{u}_h$ such that*

$$\|\mathbf{u} - \mathbf{u}_h\|_{H(\mathrm{curl};\Omega)} \leq C \inf_{\mathbf{v}_h \in U_{0,h}} \|\mathbf{u} - \mathbf{v}_h\|_{H(\mathrm{curl};\Omega)}.$$

Proof. Similar to the continuous case, we need to check the conditions of Theorem 1.3.3. We have the same seqsquilinear forms but on different spaces:

$$a(\mathbf{u}_h, \mathbf{v}_h) := (\nabla \times \mathbf{u}_h, \nabla \times \mathbf{v}_h) \qquad \text{for all } \mathbf{u}, \mathbf{v} \in U_{0,h}, \tag{5.44a}$$

$$b(\mathbf{u}_h, \xi_h) := (\mathbf{u}_h, \nabla \xi_h) \qquad \text{for all } \mathbf{u} \in U_{0,h}, \xi \in S_h. \tag{5.44b}$$

Define a space

$$\{\mathbf{u}_h \in U_{0,h} \,|\, b(\mathbf{u}_h, \xi_h) = 0 \quad \text{for all } \xi_h \in S_h\}.$$

It is obvious that the above space is Y_h. From Theorem 5.2.16 and the boundary condition of $\mathbf{u}_h$, we have that, for $\mathbf{u}_h \in Y_h$,

$$a(\mathbf{u}_h, \mathbf{u}_h) = \|\nabla \times \mathbf{u}_h\|_{L^2(\Omega)^3} \geq C\left(\|\nabla \times \mathbf{u}_h\|_{L^2(\Omega)^3} + \|\mathbf{u}_h\|_{L^2(\Omega)^3}\right)$$

for some constant $C > 0$. Thus $a(\cdot, \cdot)$ is coercive on Y_h.

Let $\boldsymbol{\phi}_h = \nabla p_h$. The discrete Babuška-Brezzi condition holds since

$$\sup_{\boldsymbol{\phi}_h \in U_{0,h}} \frac{|b(\boldsymbol{\phi}_h, p_h)|}{\|\boldsymbol{\phi}_h\|_{H(\mathrm{curl};\Omega)}} \geq \frac{|(\nabla p_h, \nabla p_h)|}{\|\nabla p_h\|_{H(\mathrm{curl};\Omega)}} = \|\nabla p_h\| \geq \beta \|p_h\|_{H^1(\Omega)}.$$

Then there exists a unique solution to (5.44). Letting $\mathbf{u}_h = \nabla p_h$ in (5.40b), we have that $\|\nabla p_h\| = 0$ and thus $p_h = 0$ by the Poincaré inequality for $H_0^1(\Omega)$. From Theorem 2.3.6, we have that

$$\|\mathbf{u} - \mathbf{u}_h\|_{H(\mathrm{curl};\Omega)} \leq C \inf_{\mathbf{v}_h \in U_h} \|\mathbf{u} - \mathbf{v}_h\|_{H(\mathrm{curl};\Omega)}.$$

□

Consequently, there exists a discrete solution operator

$$T_h : L^2(\Omega)^3 \to Y_h \quad \text{such that} \quad \mathbf{u}_h = T_h \mathbf{f}.$$

Using the result from Lemma 5.2.11, for $\mathbf{u} \in H^s(\Omega)^3$ and $\mathrm{curl}\,\mathbf{u} \in H^s(\Omega)^3$, we have that

$$\|(T - T_h)\mathbf{f}\|_{H(\mathrm{curl};\Omega)} \leq Ch^s\left(\|\mathbf{u}\|_{H^s(\Omega)^3} + \|\nabla \times \mathbf{u}\|_{H^s(\Omega)^3}\right). \tag{5.45}$$

5.2.6 The Eigenvalue Problem

The Maxwell's eigenvalue problem is to find $\lambda \in \mathbb{R}$ and $(\mathbf{u}, p) \in H_0(\mathrm{curl};\Omega) \times H_0^1(\Omega)$ such that

$$a(\mathbf{u}, \boldsymbol{\phi}) + b(\boldsymbol{\phi}, p) = \lambda(\mathbf{u}, \boldsymbol{\phi}) \qquad \text{for all } \boldsymbol{\phi} \in H_0(\mathrm{curl};\Omega), \tag{5.46a}$$

$$b(\mathbf{u}, q) = 0 \qquad \text{for all } q \in H_0^1(\Omega). \tag{5.46b}$$

Using the space Y, the eigenvalue problem can be written as finding $\lambda \in \mathbb{R}$ and $\mathbf{u} \in Y$ such that

$$(\nabla \times \mathbf{u}, \nabla \times \boldsymbol{\phi}) = \lambda(\mathbf{u}, \boldsymbol{\phi}) \quad \text{for all } \boldsymbol{\phi} \in Y. \tag{5.47}$$

We can write the above equation as an operator equation: Find $\lambda \in \mathbb{R}$ and $\mathbf{u} \in Y$ such that

$$T\mathbf{u} = \mu\mathbf{u}, \tag{5.48}$$

where $\mu = 1/\lambda$.

Similar to the continuous case, the discrete eigenvalue problem is to find $\lambda_h \in \mathbb{R}$ and $(u_h, p_h) \in U_{0,h} \times S_h$ such that

$$(\nabla \times \mathbf{u}_h, \nabla \times \boldsymbol{\phi}_h) + (\nabla p_h, \boldsymbol{\phi}_h) = \lambda_h(\mathbf{u}_h, \boldsymbol{\phi}_h) \qquad \text{for all } \boldsymbol{\phi}_h \in U_{0,h}, \tag{5.49a}$$

$$(\mathbf{u}_h, \nabla q_h) = 0 \qquad \text{for all } q_h \in S_h. \tag{5.49b}$$

Using the operator T_h, the eigenvalue problem is to find $\mathbf{u}_h \in Y_h$ and $\mu_h \in \mathbb{R}$ such that

$$T_h\mathbf{u}_h = \mu_h\mathbf{u}_h \tag{5.50}$$

where $\mu_h = 1/\lambda_h$.

Now we are ready to prove the error estimate for the eigenvalue value problem. We define the set

$$W = \cup_{h\in\Lambda} Y_h \subset H_0(\text{curl}; \Omega). \tag{5.51}$$

The following theorem claims that the embedding of W into $L^2(\Omega)^3$ is compact.

Theorem 5.2.18. *(Lemma 4.3 of [112]) Suppose that $\{Y_h\}_{h\in\Lambda}$ has the discrete compactness property. Then the embedding of W into $L^2(\Omega)^3$ is compact, i.e.,*

$$W \hookrightarrow L^2(\Omega)^3.$$

Proof. Let $\{\mathbf{u}_n\}$ be a bounded sequence in $W \subset H(\text{curl}; \Omega)$. For each n, $\mathbf{u}_n \in Y_{h_n}$. If $h_n \geq \delta > 0$, the fact that $h_n \in \Lambda$ implies there are only finitely many h_n used. Hence $\{\mathbf{u}_n\}$ is contained in a finite dimensional space and the result is trivial. Otherwise, we may assume that $h_n \to 0$ as $n \to \infty$. Then $\{\mathbf{u}_n\}_{n=1}^{\infty}$ satisfies the discrete compactness property. Thus there is a convergent subsequence in $L^2(\Omega)^3$ and the proof is complete. □

Let $\mathcal{A}$ be the collection of operators

$$\mathcal{A} = \{T_h : L^2(\Omega)^3 \to L^2(\Omega)^3, h \in \Lambda\}. \tag{5.52}$$

Theorem 5.2.19. *Suppose $\{Y_h\}_{h\in\Lambda}$ has the discrete compactness property. Then $\mathcal{A}$ is collectively compact.*

Proof. Let U be a bounded set in $L^2(\Omega)^3$. If $\mathbf{u} \in U$, $T_h\mathbf{u} \in Y_h$ such that

$$(\nabla \times T_h\mathbf{u}, \nabla \times \mathbf{v}_h) = (\mathbf{u}, \mathbf{v}_h) \quad \text{for all } \mathbf{v}_h \in Y_h.$$

Consequently, $\|\nabla \times T_h \mathbf{u}\| \leq C\|\mathbf{u}\|$. By the discrete Friedrichs inequality, we have that

$$\|T_h \mathbf{u}\| + \|\nabla \times T_h \mathbf{u}\| \leq C\|\mathbf{u}\|.$$

Thus $\{T_h \mathbf{u} \subset W, h \in \Lambda\}$ is bounded in $H(\text{curl}; \Omega)$. Since $W \hookrightarrow L^2(\Omega)^3$, there is a subsequence that converges strongly in $L^2(\Omega)^3$. Hence $\mathcal{A}(U)$ is precompact. □

Let μ be an eigenvalue and Γ be a simple closed curve with only eigenvalue μ inside. Denote by $R(E)$ the range of the spectral projection E. With suitable regularity, we have the following theorem.

Theorem 5.2.20. *Let $\{Y_h\}_{h\in\Lambda}$ have the discrete compactness property. Let μ be an eigenvalue of T with multiplicity m. Then there are exactly m discrete eigenvalues $\mu_{h,j}, j = 1, 2, \ldots, m$, of T_h such that*

$$|\mu - \mu_{h,j}| \to 0, \quad 1 \leq j \leq m \quad \text{as } h \to 0.$$

Furthermore, if all the eigenfunctions $\boldsymbol{\phi} \in E(\mu)$ are such that $\boldsymbol{\phi} \in H^p(\Omega)^3$ and $\nabla \times \boldsymbol{\phi} \in H^p(\Omega)^3$, then

$$|\mu - \mu_{j,h}| = O(h^{2p}) \quad 1 \leq j \leq m. \tag{5.53}$$

Proof. It is obvious that T and T_h are self-adjoint. Let $\boldsymbol{\phi}_i, i = 1, \ldots m$ be a basis for $E(\mu)$. Then we have that

$$\begin{aligned} ((T - T_h)\boldsymbol{\phi}_i, \boldsymbol{\phi}_j) &= (\nabla \times (T - T_h)\boldsymbol{\phi}_i, \nabla \times A\boldsymbol{\phi}_j) \\ &= (\nabla \times (T - T_h)\boldsymbol{\phi}_i, \nabla \times (T - T_h)\boldsymbol{\phi}_j). \end{aligned}$$

By Theorem 1.4.4, we have

$$|\mu - \mu_{h,j}| \leq C \left\{ \max_i \|\nabla \times (T - T_h)\boldsymbol{\phi}_i\| + \|(T - T_h)|_{E(\mu)}\|^2 \right\}.$$

Since $E(\mu)$ is finite dimensional, the pointwise convergence of T_h to T in $H(\text{curl}; \Omega)$ shows that both terms on the right-hand side go to zero as $h \to 0$.

Let $\boldsymbol{\phi} \in E(\mu)$. We have that

$$\begin{aligned} \|(T - T_h)\boldsymbol{\phi}\|_{H(\text{curl};\Omega)} &\leq \inf_{\mathbf{v}_h \in Y_h} \|T\boldsymbol{\phi} - \mathbf{v}_h\|_{H(\text{curl};\Omega)} \\ &\leq Ch^p \left(\|T\boldsymbol{\phi}\|_{H^p(\Omega)^3} + \|\nabla \times T\boldsymbol{\phi}\|_{H^p(\Omega)^3} \right) \\ &= \frac{Ch^p}{\mu} \left(\|\boldsymbol{\phi}\|_{H^p(\Omega)^3} + \|\nabla \times \boldsymbol{\phi}\|_{H^p(\Omega)^3} \right). \end{aligned}$$

Again, due to the fact that $E(\mu)$ is finite dimensional, we obtain

$$\|(T - T_h)|_{E(\mu)}\|_{H(\text{curl};\Omega)} \leq C_\mu h^p,$$

which implies (5.53). □

5.2.7 An Equivalent Eigenvalue Problem

The actual computation of the Maxwell's eigenvalue problem is complicated using the mixed formulation. In practice, one can ignore the divergence-free condition and keep the non-zero eigenvalues of a simpler problem by working with $H_0(\text{curl};\Omega)$. The following argument is based on Section 4.7 of [202].

We consider the eigenvalue problem of finding non-trivial pairs $\mathbf{u} \in H_0(\text{curl};\Omega)$ and $\lambda \in \mathbb{R}$ such that

$$(\nabla \times \mathbf{u}, \nabla \times \mathbf{v}) = \lambda(\mathbf{u}, \mathbf{v}) \quad \text{for all } \mathbf{v} \in H_0(\text{curl};\Omega). \tag{5.54}$$

Using the Helmholtz decomposition, we can write $\mathbf{u}$ as

$$\mathbf{u} = \mathbf{u}_0 + \nabla p \quad \text{where } \mathbf{u}_0 \in Y, p \in H_0^1(\Omega). \tag{5.55}$$

Taking $\mathbf{v} = \nabla\xi$ for some $\xi \in H_0^1(\Omega)$ in (5.54), we have that

$$\lambda(\nabla p, \nabla\xi) = 0 \quad \text{for all } \xi \in H_0^1(\Omega).$$

Thus one has either $\lambda = 0$ or $(\nabla p, \nabla\xi) = 0$. If $\lambda \neq 0$, choosing $\xi = p$, we have that $\nabla p = 0$. Since $p \in H_0^1(\Omega)$, $p = 0$. From (5.55), we obtain $\mathbf{u} = \mathbf{u}_0 \in Y$. Hence eigenfunctions of (5.54) corresponding to non-zero eigenvalues are eigenfunctions of the Maxwell's eigenvalue problem (5.46).

When $\lambda = 0$, we have that $\mathbf{u}_0 \in Y$ satisfies

$$(\nabla \times \mathbf{u}_0, \nabla \times \mathbf{v}) = 0 \quad \text{for all } \mathbf{v} \in Y.$$

The Friedrichs inequality implies that $\mathbf{u}_0 = \mathbf{0}$. Again from (5.55), $\mathbf{u} = \nabla p$ for some $p \in H_0^1(\Omega)$. Thus $\lambda = 0$ is an eigenvalue of infinite multiplicity of (5.54). The corresponding eigenspace is $\nabla H_0^1(\Omega)$.

Note that $\lambda = 0$ is not an eigenvalue of (5.46). Since if we enforce the divergence-free condition, i.e., $\nabla \cdot \nabla p = 0$, we obtain $p = 0$. Then the eigenfunction is trivial. Nonetheless, we can compute non-zero eigenvalues of (5.54) to obtain eigenvalues of (5.46).

The same argument can be carried out for the discrete case. Let $U_{0,h}$ be the edge finite element space. The discrete eigenvalue problem for (5.54) is to find $\mathbf{u}_h \in U_{0,h}$, $\mathbf{u}_h \neq \mathbf{0}$ and λ_h such that

$$(\nabla \times \mathbf{u}_h, \nabla \times \mathbf{v}_h) = \lambda(\mathbf{u}_h, \mathbf{v}_h) \quad \text{for all } \mathbf{v}_h \in U_{0,h}. \tag{5.56}$$

It is obvious that $\mathbf{u}_h = \nabla p_h$ for any $p_h \in S_h$ is an eigenfunction corresponding to the eigenvalue $\lambda_h = 0$. The discrete Friedrichs inequality implies that only eigenfunctions corresponding to $\lambda_h = 0$ belong to ∇S_h.

For $\lambda_h \neq 0$, we choose $\mathbf{v}_h = \nabla\xi_h$ for any $\xi_h \in S_h$ in (5.56) to obtain

$$\lambda_h(\mathbf{u}_h, \nabla\xi_h) = 0 \quad \text{for all } \xi_h \in S_h.$$

Hence $\mathbf{u}_h$ is discrete divergence-free, i.e., $\mathbf{u}_h \in Y_h$. Similar to the continuous case, one computes non-zero eigenvalues of (5.56) which coincides with the eigenvalues of the mixed problem (5.49).

5.2.8 Numerical Examples

We first derive exact Maxwell's eigenvalues for certain domains. The following result is classical and can be found in many references, e.g., [45]. For the unit cube $(0,1)^3$, the eigenfunctions are tensor products of trigonometric functions. Eigenvalues are of the form $\{k\pi^2\}$ where

$$k = k_1^2 + k_2^2 + k_3^2$$

are non-negative integers satisfying

$$k_1k_2 + k_2k_3 + k_3k_1 > 0.$$

For example, the following

$$\begin{pmatrix} \cos(k_1\pi x_1)\sin(k_2\pi x_2)\sin(k_3\pi x_3) \\ \sin(k_1\pi x_1)\cos(k_2\pi x_2)\sin(k_3\pi x_3) \\ \sin(k_1\pi x_1)\sin(k_2\pi x_2)\cos(k_3\pi x_3) \end{pmatrix}$$

is an eigenfunction. A few smallest eigenvalues with their multiplicities are listed in Table 5.1.

eigenvalues	multiplicity
$2\pi^2$	3
$3\pi^2$	2
$5\pi^2$	6
$6\pi^2$	6
$8\pi^2$	3
$9\pi^2$	6
$10\pi^2$	6

Table 5.1: Maxwell's eigenvalues of the unit cube.

For the unit ball, the eigenvalues are given by

$$\{\omega_{mn}^2, \hat{\omega}_{mn}^2, n = 1, 2, \ldots, m = -n, \ldots, -1, 0, 1, \ldots, n\}.$$

The eigenvalues are split into two groups:

Transverse Electric (TE), which satisfy

$$j_m(\omega_{mn}) = 0;$$

Transverse Magnetic (TM), which satisfy

$$j_m(\hat{\omega}_{mn}) + \hat{\omega}_{mn} j'_m(\hat{\omega}_{mn}) = 0.$$

	ω_i^2	mode	multiplicity
1	7.5279e+00	TM($\hat{\omega}_{11}^2$)	3
2	1.4979e+01	TM($\hat{\omega}_{21}^2$)	5
3	2.0191e+01	TE(ω_{11}^2)	3
4	2.4735e+01	TM($\hat{\omega}_{31}^2$)	7
5	3.3217e+01	TE(ω_{21}^2)	5
6	3.6747e+01	TM($\hat{\omega}_{41}^2$)	9
7	3.7415e+01	TM($\hat{\omega}_{12}^2$)	11

Table 5.2: Maxwell's eigenvalues of the unit ball.

Here j_m is the mth order spherical Bessel function and j_m' is its derivative. A few smallest eigenvalues are given in Table 5.2.

We partition the unit cube and obtain four levels of tetrahedral meshes. In Table 5.3, we present the numerical results using the linear edge element. The first three discrete eigenvalues are the approximation of the exact eigenvalue $2\pi^2$ with multiplicity 3. The last column is the error given by

$$\text{Err.} = \frac{\lambda_1 - \frac{1}{3}\sum_{j=1}^{3} \lambda_{1,j,h}}{\lambda_1}. \tag{5.57}$$

h	1st	2nd	3rd	Err.
0.3933	18.225383	18.903474	19.072456	0.050936
0.2153	19.603864	19.566819	19.573561	0.007994
0.1193	19.710621	19.707713	19.708828	0.001528
0.0585	19.733369	19.733043	19.732653	0.000313

Table 5.3: The first three Maxwell's eigenvalues for the unit cube using the linear edge element.

The second domain is the unit ball. Again, the first eigenvalue has multiplicity 3. We show the first three discrete eigenvalues and compute the relative error in (5.57).

h	1st	2nd	3rd	Err.
0.5313	7.739196	7.754744	7.777255	0.030442
0.3608	7.625388	7.628210	7.631304	0.013337
0.1934	7.552370	7.552471	7.552831	0.003275
0.0958	7.533982	7.534158	7.534202	8.2546e-04

Table 5.4: The first three Maxwell's eigenvalues for the unit ball using the linear edge element.

Finally, we consider the L-shaped domain given by

$$(0,1)^3 \setminus (0.5,1) \times (0.5,1) \times (0.5,1).$$

The numerical results indicate that the first eigenvalue is simple. Since there is no exact value available, we compute the relative error as follows

$$\text{Rel. Err.} = \frac{|\lambda_{1,h_1} - \lambda_{1,h_2}|}{\lambda_{1,h_2}}.$$

We show the results in Table 5.5.

h	1st eigenvalue	Rel. Err.
0.4099	10.458863	-
0.2262	11.982769	0.145704
0.1111	12.448686	0.038882
0.0591	12.738332	0.023267

Table 5.5: The first Maxwell's eigenvalue for the L-shaped domain using the linear edge element.

In Fig. 5.1, we show the convergence rates of the first eigenvalues for three domains. It can be seen that, for the unit cube and the unit ball, the convergence rates are $O(h^2)$. For the L-shaped domain, it seems that the second order convergence rate cannot be obtained, indicating that the reentrant corner leads to lower regularity of the eigenfunction.

5.3 The Quad-curl Eigenvalue Problem

The quad-curl problem arises in the inverse electromagnetic scattering theory [203] and magnetohydrodynamics (MHD) equations [254]. Unlike the Maxwell's eigenvalue problem, which has been studied extensively in the literature (see, for example, [101] and [35]), there are few results on the quad-curl eigenvalue problem. Construction of conforming finite elements with suitable regularity for the quad-curl problem can be extremely technical and prohibitively expensive, even if such finite elements exist.

In this section, we present a mixed finite element method for the quad-curl eigenvalue problem by Sun [231]. The major advantage of this approach lies in the fact that only curl-conforming edge elements are needed [208]. Similar to the Maxwell's eigenvalue problem, the divergence-free condition, which is usually treated using Lagrange multipliers, can be ignored for the quad-curl eigenvalue problem.

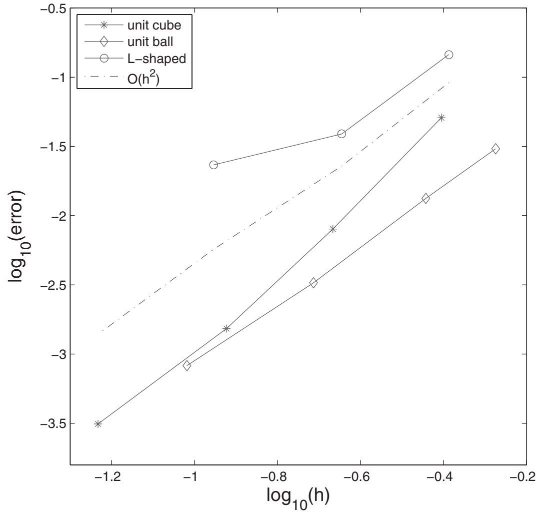


Figure 5.1: Convergence rates of the first Maxwell's eigenvalue using the linear edge element.

5.3.1 The Quad-curl Problem

The quad-curl problem is stated as follows. For $\mathbf{f} \in H(\mathrm{div}^0;\Omega)$, find $\mathbf{u}$ such that

$$\nabla\times\nabla\times\nabla\times\nabla\times\, \mathbf{u} = \mathbf{f} \quad \text{in } \Omega, \tag{5.58a}$$

$$\nabla\cdot\mathbf{u} = 0 \quad \text{in } \Omega, \tag{5.58b}$$

$$\mathbf{u}\times\boldsymbol{\nu} = 0 \quad \text{on } \partial\Omega, \tag{5.58c}$$

$$(\nabla\times\mathbf{u})\times\boldsymbol{\nu} = 0 \quad \text{on } \partial\Omega. \tag{5.58d}$$

The variational approach we will describe for the quad-curl problem requires several Hilbert spaces. We define

$$H^s(\mathrm{curl};\Omega) := \left\{\mathbf{u}\in L^2(\Omega)^3 \mid (\nabla\times)^j\,\mathbf{u}\in L^2(\Omega)^3,\, 1\le j\le s\right\}$$

equipped with the scalar product

$$(\mathbf{u},\mathbf{v})_{H^s(\mathrm{curl};\Omega)} = (\mathbf{u},\mathbf{v}) + \sum_{j=1}^{s}((\nabla\times)^j\,\mathbf{u}, (\nabla\times)^j\,\mathbf{v})$$

and the corresponding norm $\|\cdot\|_{H^s(\mathrm{curl};\Omega)}$. We will use the standard notation $H(\mathrm{curl};\Omega)$ when $s=1$.

We define

$$H_0^2(\text{curl};\Omega) := \left\{\mathbf{u} \in H^2(\text{curl};\Omega) \mid \mathbf{u} \times \boldsymbol{\nu} = 0 \text{ and } (\nabla \times \mathbf{u}) \times \boldsymbol{\nu} = 0 \text{ on } \partial\Omega\right\}.$$

We start with the weak formulation of the quad-curl problem. Let V and W be given by

$$V := \left\{\mathbf{u} \in H_0^2(\text{curl};\Omega) \cap H(\text{div};\Omega) \mid \nabla \cdot \mathbf{u} = 0\right\}, \tag{5.59}$$

$$W := \{\mathbf{u} \in H^2(\text{curl};\Omega) \cap H(\text{div};\Omega) \mid \nabla \cdot \mathbf{u} = 0\}. \tag{5.60}$$

We define a bilinear form $\mathcal{C} : V \times V \to \mathbb{R}$ by

$$\mathcal{C}(\mathbf{u},\mathbf{v}) := (\nabla \times \nabla \times \mathbf{u}, \nabla \times \nabla \times \mathbf{v}) \quad \text{for all } \mathbf{u},\mathbf{v} \in V. \tag{5.61}$$

Let $\mathbf{f} \in H(\text{div}^0;\Omega)$. The weak formulation for the quad-curl problem is to find $\mathbf{u} \in V$ such that

$$\mathcal{C}(\mathbf{u},\mathbf{v}) = (\mathbf{f},\mathbf{v}) \quad \text{for all } \mathbf{v} \in V. \tag{5.62}$$

Theorem 5.3.1. *There exists a unique solution* $\mathbf{u} \in V$ *to* (5.62).

Proof. Due to the fact that functions in V are divergence-free, using the Friedrichs inequality twice, we see that the bilinear form $\mathcal{C}$ is elliptic on V. Then Lax-Milgram Lemma 1.3.1 implies that there exists a unique solution $\mathbf{u}$ of (5.62) in V. □

To the authors' knowledge, there are no regularity results for the quad-curl problem in the literature. For Maxwell's equations, it is well known that non-convexity leads to singularities; see [101] and [99]. For the biharmonic equation with clamped plate boundary conditions, convexity is sufficient for the solution to be in $H^3(\Omega)$ [139]. Therefore the mixed finite element method given in [89] for the corresponding biharmonic eigenvalue problem does not produce spurious modes. However, whether convexity is sufficient for the quad-curl solution to be in $H^3(\text{curl};\Omega)$ is a non-trivial open problem. On the other hand, for biharmonic eigenvalue problems on non-convex domains, we have seen that mixed finite methods compute spurious modes (see Section 4.5.4 and [52]). Thus non-convexity might lead to the failure of the mixed method for the quad-curl eigenvalue problem. In the rest of this chapter, we assume that the solution $\mathbf{u}$ of (5.62) belongs to $H^3(\text{curl};\Omega)$.

Let $\boldsymbol{\phi} = \nabla \times \nabla \times \mathbf{u}$. We define

$$X := \{\mathbf{u} \in H(\text{curl};\Omega) \cap H(\text{div};\Omega) | \nabla \times \mathbf{u} = 0 \text{ in } \Omega\},$$

and

$$a(\mathbf{u},\mathbf{v}) = (\mathbf{u},\mathbf{v}), \quad b(\mathbf{u},\mathbf{v}) = -(\nabla \times \mathbf{u}, \nabla \times \mathbf{v}).$$

Let Y be defined as (5.18). The mixed formulation for the quad-curl problem can be stated as follows. For $\mathbf{f} \in H(\text{div}^0;\Omega)$, find $(\mathbf{u},\boldsymbol{\phi}) \in Y \times X$ such that

$$a(\mathbf{f},\mathbf{v}) + b(\boldsymbol{\phi},\mathbf{v}) = 0 \qquad \text{for all } \mathbf{v} \in Y, \tag{5.63a}$$

$$b(\mathbf{u},\boldsymbol{\psi}) = -(\boldsymbol{\phi},\boldsymbol{\psi}) \qquad \text{for all } \boldsymbol{\psi} \in X. \tag{5.63b}$$

In the following, we derive the equivalence of the above mixed formulation to the quad-curl problem. We employ a technique similar to that in Section 4.3 for the biharmonic equation (see also Section 7.1 of [88]).

The solution of the quad-curl problem is the solution of the following unconstrained minimization problem: Find $\mathbf{u}$ such that

$$J(\mathbf{u}) = \inf_{\mathbf{v}\in V} J(\mathbf{v}), \tag{5.64}$$

where

$$J(\mathbf{v}) = \frac{1}{2}\int_\Omega |\nabla\times\nabla\times\mathbf{v}|^2\,\mathrm{d}x - \int_\Omega \mathbf{f}\cdot\mathbf{v}\,\mathrm{d}x. \tag{5.65}$$

This is due to the fact that (5.62) is the Euler-Lagrange equation for the minimization problem.

Equivalently we consider the constrained minimization problem associated with the quadratic form

$$\mathcal{J}(\mathbf{v},\boldsymbol{\psi}) = \frac{1}{2}\int_\Omega |\boldsymbol{\psi}|^2\,\mathrm{d}x - \int_\Omega \mathbf{f}\cdot\mathbf{v}\,\mathrm{d}x \tag{5.66}$$

for $(\mathbf{v},\boldsymbol{\psi}) \in V\times L^2(\Omega)^3$ such that $\nabla\times\nabla\times\mathbf{v} = \boldsymbol{\psi}$.

We define

$$\mathcal{V} := \left\{(\mathbf{v},\boldsymbol{\psi}) \in Y\times L^2(\Omega)^3 \mid \beta((\mathbf{v},\boldsymbol{\psi}),\boldsymbol{\mu}) = 0 \text{ for all } \boldsymbol{\mu}\in X\right\},$$

where

$$\beta((\mathbf{v},\boldsymbol{\psi}),\boldsymbol{\mu}) = \int_\Omega \nabla\times\mathbf{v}\cdot\nabla\times\boldsymbol{\mu}\,\mathrm{d}x - \int_\Omega \boldsymbol{\psi}\cdot\boldsymbol{\mu}\,\mathrm{d}x. \tag{5.67}$$

Thus the problem can be stated as: Find $(\mathbf{u},\boldsymbol{\phi}) \in \mathcal{V}$ such that

$$\int_\Omega \boldsymbol{\phi}\cdot\boldsymbol{\psi}\,\mathrm{d}x = \int_\Omega \mathbf{f}\cdot\mathbf{v}\,\mathrm{d}x \quad \text{for all } (\mathbf{v},\boldsymbol{\psi}) \in \mathcal{V}.$$

Lemma 5.3.2. *The mapping*

$$(\mathbf{v},\boldsymbol{\psi}) \in \mathcal{V} \to \|\boldsymbol{\psi}\|$$

is a norm over $\mathcal{V}$. Furthermore,

$$\mathcal{V} := \left\{(\mathbf{v},\boldsymbol{\psi}) \in V\times L^2(\Omega)^3 \mid \nabla\times\nabla\times\mathbf{v} = \boldsymbol{\psi}\right\}.$$

Proof. The lemma follows directly from the Friedrichs inequality. □

Theorem 5.3.3. *If $\mathbf{u}\in V$ is the solution of* (5.64)*, we have that*

$$\mathcal{J}(\mathbf{u},\nabla\times\nabla\times\,\mathbf{u}) = \inf_{(\mathbf{v},\boldsymbol{\psi})\in\mathcal{V}} \mathcal{J}(\mathbf{v},\boldsymbol{\psi}) \tag{5.68}$$

and $(\mathbf{u},\nabla\times\nabla\times\mathbf{u}) \in \mathcal{V}$ is the unique solution of (5.68).

Proof. Since the mapping

$$((\mathbf{u}, \boldsymbol{\phi}), (\mathbf{v}, \boldsymbol{\psi})) \in \mathcal{V} \times \mathcal{V} \to \int_\Omega \boldsymbol{\phi} \cdot \boldsymbol{\psi} \,\mathrm{d}x$$

is continuous and $\mathcal{V}$-elliptic,

$$(\mathbf{v}, \boldsymbol{\psi}) \in \mathcal{V} \to \int_\Omega \mathbf{f} \cdot \mathbf{v} \,\mathrm{d}x$$

is continuous, the minimization problem of finding $(\mathbf{u}^*, \boldsymbol{\phi}) \in \mathcal{V}$ such that

$$\mathcal{J}(\mathbf{u}^*, \boldsymbol{\phi}) = \inf_{(\mathbf{v}, \boldsymbol{\psi}) \in \mathcal{V}} \mathcal{J}(\mathbf{v}, \boldsymbol{\psi})$$

has a unique solution that satisfies

$$\int_\Omega \boldsymbol{\phi} \cdot \boldsymbol{\psi} \,\mathrm{d}x = \int_\Omega \mathbf{f} \cdot \mathbf{v} \,\mathrm{d}x \quad \text{for all } (\mathbf{v}, \boldsymbol{\psi}) \in \mathcal{V}.$$

From Lemma 5.3.2, we see that $\mathbf{u}^* \in V$ and that $\nabla \times \nabla \times \mathbf{u}^* = \boldsymbol{\phi}$. We have

$$\int_\Omega \nabla \times \nabla \times \mathbf{u} \cdot \nabla \times \nabla \times \mathbf{v} \,\mathrm{d}x = \int_\Omega \mathbf{f} \cdot \mathbf{v} \,\mathrm{d}x.$$

Consequently, $\mathbf{u}^*$ is the solution $\mathbf{u}$ of (5.64).

$\square$

Based on the above theorem, we define two solution operators

$$A : H(\mathrm{div}^0; \Omega) \to X$$

such that

$$A\mathbf{f} = \boldsymbol{\phi}$$

and

$$B : H(\mathrm{div}^0; \Omega) \to Y$$

for (5.63) such that

$$B\mathbf{f} = \mathbf{u}.$$

We can write (5.63) as

$$a(A\mathbf{f}, \mathbf{v}) + b(\mathbf{v}, B\mathbf{f}) = 0 \qquad \text{for all } \mathbf{v} \in X, \tag{5.69a}$$

$$b(A\mathbf{f}, \mathbf{q}) = -(\mathbf{f}, \mathbf{q}) \qquad \text{for all } \mathbf{q} \in Y. \tag{5.69b}$$

Next we consider the edge element method for the minimization problem. Let

$$\mathcal{V}_h = \{(\mathbf{v}_h, \boldsymbol{\psi}_h) \in Y_h \times X_h \mid \beta((\mathbf{v}_h, \boldsymbol{\psi}_h), \boldsymbol{\mu}_h) = 0 \quad \text{for all } \boldsymbol{\mu}_h \in X_h\},$$

where Y_h is defined in (5.39) and X_h is such that

$$U_h = X_h \oplus \nabla S_h.$$

Note that $Y_h \not\subset Y$. The discrete problem corresponding to (5.66) is to find $(\mathbf{u}_h, \boldsymbol{\phi}_h) \in \mathcal{V}_h$ such that

$$\mathcal{J}(\mathbf{u}_h, \boldsymbol{\phi}_h) = \inf_{(\mathbf{v}_h, \boldsymbol{\psi}_h) \in \mathcal{V}_h} \mathcal{J}(\mathbf{v}_h, \boldsymbol{\psi}_h). \tag{5.70}$$

It is easy to see that the discrete problem (5.70) has a unique solution and $(\mathbf{u}_h, \boldsymbol{\phi}_h) \in \mathcal{V}_h$ satisfies

$$\int_\Omega \boldsymbol{\phi}_h \cdot \boldsymbol{\psi}_h \,\mathrm{d}x = \int_\Omega \mathbf{f} \cdot \mathbf{v}_h \,\mathrm{d}x \quad \text{for all } (\mathbf{v}_h, \boldsymbol{\psi}_h) \in \mathcal{V}_h. \tag{5.71}$$

Theorem 5.3.4. *Let* $(\mathbf{u}, \phi)$ *and* $(\mathbf{u}_h, \boldsymbol{\phi}_h)$ *be the solutions of* (5.68) *and* (5.70)*, respectively, and assume that* $\mathbf{u} \in H^3(\text{curl}; \Omega)$. *There exists a constant* C *independent of* h *such that*

$$\begin{aligned} &\|\nabla \times \mathbf{u} - \nabla \times \mathbf{u}_h\| + \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\phi}_h\| \\ &\qquad \le C \left(\inf_{(\mathbf{v}_h, \boldsymbol{\psi}_h) \in \mathcal{V}_h} (\|\nabla \times \mathbf{u} - \nabla \times \mathbf{v}_h\| + \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\psi}_h\|) \right. \\ &\qquad\qquad \left. + \inf_{\boldsymbol{\mu}_h \in X_h} \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\mu}_h\|_{H(\text{curl};\Omega)} \right). \end{aligned} \tag{5.72}$$

Proof. Assuming that $\mathbf{u} \in H^3(\text{curl}; \Omega)$, it holds that

$$\begin{aligned} & \int_\Omega \nabla \times (\nabla \times \nabla \times \mathbf{u}) \cdot \nabla \times \mathbf{v} \,\mathrm{d}x \\ = & \int_\Omega \nabla \times \nabla \times \mathbf{u} \cdot \nabla \times \nabla \times \mathbf{v} \,\mathrm{d}x \\ = & \int_\Omega \mathbf{f} \cdot \mathbf{v} \,\mathrm{d}x \end{aligned}$$

for all $\mathbf{v} \in \mathcal{C}_0^\infty(\Omega)^3$, the space of smooth functions with compact support in Ω. Hence for all $\mathbf{v} \in H_0(\text{curl}; \Omega)$, the following holds

$$\int_\Omega \nabla \times (\nabla \times \nabla \times \mathbf{u}) \cdot \nabla \times \mathbf{v} \,\mathrm{d}x = \int_\Omega \mathbf{f} \cdot \mathbf{v} \,\mathrm{d}x. \tag{5.73}$$

Thus for any $\mathbf{v} \in H_0(\text{curl}; \Omega)$ and $\boldsymbol{\psi} \in L^2(\Omega)^3$, we obtain

$$\beta\left((\mathbf{v}, \boldsymbol{\psi}), \nabla \times \nabla \times \mathbf{u}\right) = \int_\Omega \mathbf{f} \cdot \mathbf{v} \,\mathrm{d}x - \int_\Omega \boldsymbol{\psi} \cdot \nabla \times \nabla \times \mathbf{u} \,\mathrm{d}x.$$

For any $(\mathbf{v}_h, \boldsymbol{\psi}_h) \in \mathcal{V}_h$ and $\boldsymbol{\mu}_h \in X_h$, using the fact that

$$\beta((\mathbf{v}_h, \boldsymbol{\psi}_h), \boldsymbol{\mu}_h) = 0,$$

(5.73), and (5.71), we have

$$\begin{aligned}
& \beta\left((\mathbf{u}_h-\mathbf{v}_h,\boldsymbol{\phi}_h-\boldsymbol{\psi}_h),\nabla\times\nabla\times\mathbf{u}-\boldsymbol{\mu}_h\right) \\
= & \int_\Omega \nabla\times(\mathbf{u}_h-\mathbf{v}_h)\cdot\nabla\times(\nabla\times\nabla\times\mathbf{u}-\boldsymbol{\mu}_h)\,\mathrm{d}x \\
& \qquad -\int_\Omega(\boldsymbol{\phi}_h-\boldsymbol{\psi}_h)\cdot(\nabla\times\nabla\times\mathbf{u}-\boldsymbol{\mu}_h)\,\mathrm{d}x \\
= & \int_\Omega \nabla\times(\mathbf{u}_h-\mathbf{v}_h)\cdot\nabla\times\nabla\times\nabla\times\mathbf{u}\,\mathrm{d}x-\int_\Omega\nabla\times(\mathbf{u}_h-\mathbf{v}_h)\cdot\nabla\times\boldsymbol{\mu}_h\,\mathrm{d}x \\
& \qquad -\int_\Omega(\boldsymbol{\phi}_h-\boldsymbol{\psi}_h)\cdot\nabla\times\nabla\times\mathbf{u}\,\mathrm{d}x+\int_\Omega(\boldsymbol{\phi}_h-\boldsymbol{\psi}_h)\cdot\boldsymbol{\mu}_h\,\mathrm{d}x \\
= & \int_\Omega \nabla\times(\mathbf{u}_h-\mathbf{v}_h)\cdot\nabla\times\nabla\times\nabla\times\mathbf{u}\,\mathrm{d}x-\int_\Omega(\boldsymbol{\phi}_h-\boldsymbol{\psi}_h)\cdot\nabla\times\nabla\times\mathbf{u}\,\mathrm{d}x \\
= & \int_\Omega \mathbf{f}\cdot(\mathbf{u}_h-\mathbf{v}_h)\,\mathrm{d}x-\int_\Omega(\boldsymbol{\phi}_h-\boldsymbol{\psi}_h)\cdot\nabla\times\nabla\times\mathbf{u}\,\mathrm{d}x \\
= & \int_\Omega \boldsymbol{\phi}_h\cdot(\boldsymbol{\phi}_h-\boldsymbol{\psi}_h)\,\mathrm{d}x-\int_\Omega(\boldsymbol{\phi}_h-\boldsymbol{\psi}_h)\cdot\nabla\times\nabla\times\mathbf{u}\,\mathrm{d}x \\
= & -\int_\Omega(\nabla\times\nabla\times\mathbf{u}-\boldsymbol{\phi}_h)\cdot(\boldsymbol{\phi}_h-\boldsymbol{\psi}_h)\,\mathrm{d}x.
\end{aligned} \tag{5.74}$$

On the other hand, for all $\boldsymbol{\mu}_h \in X_h$, one has

$$\int_\Omega \nabla\times\mathbf{u}_h\cdot\nabla\times\boldsymbol{\mu}_h\mathrm{d}\,x=\int_\Omega\boldsymbol{\phi}_h\cdot\boldsymbol{\mu}_h\mathrm{d}x,$$
$$\int_\Omega \nabla\times\mathbf{v}_h\cdot\nabla\times\boldsymbol{\mu}_h\mathrm{d}\,x=\int_\Omega\boldsymbol{\psi}_h\cdot\boldsymbol{\mu}_h\mathrm{d}x.$$

Taking the difference and letting $\boldsymbol{\mu}_h=\mathbf{u}_h-\mathbf{v}_h$,

$$\int_\Omega \nabla\times(\mathbf{u}_h-\mathbf{v}_h)\cdot\nabla\times(\mathbf{u}_h-\mathbf{v}_h)\mathrm{d}x=\int_\Omega(\boldsymbol{\phi}_h-\boldsymbol{\psi}_h)\cdot(\mathbf{u}_h-\mathbf{v}_h)\mathrm{d}x,$$

which implies

$$\|\nabla\times(\mathbf{u}_h-\mathbf{v}_h)\|\le C\|\boldsymbol{\phi}_h-\boldsymbol{\psi}_h\|, \tag{5.75}$$

where C is the constant in the discrete Friedrichs inequality.

Using the above inequality and (5.74), we get

$$\begin{aligned}
& \left|\int_\Omega(\nabla\times\nabla\times\mathbf{u}-\boldsymbol{\phi}_h)\cdot(\boldsymbol{\phi}_h-\boldsymbol{\psi}_h)\,\mathrm{d}x\right| \\
= & \left|\beta\left((\mathbf{u}_h-\mathbf{v}_h,\boldsymbol{\phi}_h-\boldsymbol{\psi}_h),\nabla\times\nabla\times\mathbf{u}-\boldsymbol{\mu}_h\right)\right| \\
\le & \|\nabla\times(\mathbf{u}_h-\mathbf{v}_h)\|\,\|\nabla\times(\nabla\times\nabla\times\mathbf{u}-\boldsymbol{\mu}_h)\| \\
& \qquad +\|\boldsymbol{\phi}_h-\boldsymbol{\psi}_h\|\,\|\nabla\times\nabla\times\mathbf{u}-\boldsymbol{\mu}_h\| \\
\le & C\|\boldsymbol{\phi}_h-\boldsymbol{\psi}_h\|\,\|\nabla\times(\nabla\times\nabla\times\mathbf{u}-\boldsymbol{\mu}_h)\| \\
& \qquad +\|\boldsymbol{\phi}_h-\boldsymbol{\psi}_h\|\,\|\nabla\times\nabla\times\mathbf{u}-\boldsymbol{\mu}_h\| \\
\le & C_1\|\boldsymbol{\phi}_h-\boldsymbol{\psi}_h\|\,\|\nabla\times\nabla\times\mathbf{u}-\boldsymbol{\mu}_h\|_{H(\mathrm{curl};\Omega)},
\end{aligned}$$

where $C_1 = \max\{C, 1\}$. Consequently,

$$
\begin{aligned}
&\|\boldsymbol{\phi}_h - \boldsymbol{\psi}_h\|^2 \\
= \quad & -\int_\Omega (\boldsymbol{\phi}_h - \boldsymbol{\psi}_h) \cdot (\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\phi}_h)\,\mathrm{d}x \\
& \qquad\qquad + \int_\Omega (\boldsymbol{\phi}_h - \boldsymbol{\psi}_h) \cdot (\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\psi}_h)\,\mathrm{d}x \\
\leq \quad & C_1 \|\boldsymbol{\phi}_h - \boldsymbol{\psi}_h\| \, \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\mu}_h\|_{H(\mathrm{curl};\Omega)} \\
& \qquad\qquad + \|\boldsymbol{\phi}_h - \boldsymbol{\psi}_h\| \, \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\psi}_h\|
\end{aligned}
$$

and hence

$$
\|\boldsymbol{\phi}_h - \boldsymbol{\psi}_h\| \leq C_1 \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\mu}_h\|_{H(\mathrm{curl};\Omega)} + \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\psi}_h\|.
$$

Moreover, we have that

$$
\begin{aligned}
& \|\nabla \times \mathbf{u} - \nabla \times \mathbf{u}_h\| + \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\phi}_h\| \\
\leq \quad & \|\nabla \times \mathbf{u} - \nabla \times \mathbf{v}_h\| + \|\nabla \times \mathbf{v}_h - \nabla \times \mathbf{u}_h\| \\
& \qquad\qquad + \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\psi}_h\| + \|\boldsymbol{\psi}_h - \boldsymbol{\phi}_h\| \\
\leq \quad & \|\nabla \times \mathbf{u} - \mathrm{curl}\,\mathbf{v}_h\| + \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\psi}_h\| + (1 + C)\|\boldsymbol{\psi}_h - \boldsymbol{\phi}_h\|,
\end{aligned}
$$

where (5.75) is used. Combining the above inequalities, we obtain

$$
\begin{aligned}
& \|\nabla \times \mathbf{u} - \nabla \times \mathbf{u}_h\| + \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\phi}_h\| \\
\leq \quad & \|\nabla \times \mathbf{u} - \nabla \times \mathbf{v}_h\| + \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\psi}_h\| \\
& \quad + (1 + C)\left(C_1 \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\mu}_h\|_{H(\mathrm{curl};\Omega)} + \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\psi}_h)\|\right) \\
\leq \quad & \|\nabla \times \mathbf{u} - \nabla \times \mathbf{v}_h\| + (2 + C)\|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\psi}_h\| \\
& \qquad\qquad + (1 + C)C_1 \|\nabla \times \nabla \times \, \mathbf{u} - \boldsymbol{\mu}_h\|_{H(\mathrm{curl};\Omega)}.
\end{aligned}
$$

The proof is complete by taking the infimum over all $(\mathbf{v}_h, \boldsymbol{\psi}_h) \in \mathcal{V}_h$ and $\boldsymbol{\mu}_h \in X_h$. $\square$

Theorem 5.3.5. *Let* $(\mathbf{u}, \phi)$ *and* $(\mathbf{u}_h, \phi_h)$ *solve* (5.68) *and* (5.70)*, respectively. Let* $\alpha(h) = C_1/h$ *where* C_1 *is the constant in Lemma 5.2.13. Then there exists a constant* C *independent of the mesh size* h *such that*

$$
\begin{aligned}
& \|\nabla \times \mathbf{u} - \nabla \times \mathbf{u}_h\| + \|\nabla \times \nabla \times \, \mathbf{u} - \boldsymbol{\phi}_h\| \\
\leq \quad & C \left\{ (1 + \alpha(h)) \inf_{\mathbf{v}_h \in Y_h} \|\nabla \times \mathbf{u} - \nabla \times \mathbf{v}_h\| \right. \\
& \qquad\qquad \left. + \inf_{\boldsymbol{\mu}_h \in X_h} \|\nabla \times \nabla \times \, \mathbf{u} + \boldsymbol{\mu}_h\|_{H(\mathrm{curl};\Omega)} \right\}.
\end{aligned}
$$

Proof. Let $(\mathbf{v}_h, \boldsymbol{\psi}_h) \in \mathcal{V}_h$ and $\boldsymbol{\mu}_h \in X_h$. Writing $\mathbf{w}_h = \boldsymbol{\mu}_h + \boldsymbol{\psi}_h$, one has that $\beta((\mathbf{v}_h, \boldsymbol{\psi}_h), \mathbf{w}_h) = 0$, i.e.,

$$
\int_\Omega \nabla \times \mathbf{v}_h \cdot \nabla \times \mathbf{w}_h \,\mathrm{d}x - \int_\Omega \boldsymbol{\psi}_h \cdot \mathbf{w}_h \,\mathrm{d}x = 0.
$$

Using the fact that $\boldsymbol{\nu} \times (\nabla \times \mathbf{u}) = 0$ on $\partial\Omega$, we obtain

$$\int_\Omega \nabla \times \nabla \times \mathbf{u} \cdot \mathbf{w}_h \, \mathrm{d}x = \int_\Omega \nabla \times \mathbf{u} \cdot \nabla \times \mathbf{w}_h \, \mathrm{d}x.$$

Combination of the above two equations leads to

$$\int_\Omega (\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\psi}_h) \cdot \mathbf{w}_h \, \mathrm{d}x = \int_\Omega \nabla \times (\mathbf{u} - \mathbf{v}_h) \cdot \nabla \times \mathbf{w}_h \, \mathrm{d}x.$$

Therefore,

$$\begin{aligned} \left| \int_\Omega (\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\psi}_h) \cdot \mathbf{w}_h \, \mathrm{d}x \right| &\le \|\nabla \times \mathbf{u} - \nabla \times \mathbf{v}_h\| \, \|\nabla \times \mathbf{w}_h\| \\ &\le \alpha(h) \|\nabla \times \mathbf{u} - \nabla \times \mathbf{v}_h\| \, \|\mathbf{w}_h\| \end{aligned}$$

and

$$\begin{aligned} \|\mathbf{w}_h\|^2 &= \int_\Omega (\boldsymbol{\mu}_h + \nabla \times \nabla \times \mathbf{u}) \cdot \mathbf{w}_h \, \mathrm{d}x + \int_\Omega (\boldsymbol{\psi}_h - \nabla \times \nabla \times \mathbf{u}) \cdot \mathbf{w}_h \, \mathrm{d}x \\ &\le \|\boldsymbol{\mu}_h + \nabla \times \nabla \times \mathbf{u}\| \, \|\mathbf{w}_h\| + \alpha(h) \|\nabla \times \mathbf{u} - \nabla \times \mathbf{v}_h\| \, \|\mathbf{w}_h\|. \end{aligned}$$

From this inequality, we deduce that

$$\begin{aligned} \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\psi}_h\| &\le \|\nabla \times \nabla \times \mathbf{u} + \boldsymbol{\mu}_h\| + \|\mathbf{w}_h\| \\ &\le 2\|\nabla \times \nabla \times \mathbf{u} + \boldsymbol{\mu}_h\| + \alpha(h) \|\nabla \times \mathbf{u} - \nabla \times \mathbf{v}_h\|, \end{aligned}$$

and thus,

$$\begin{aligned} &\inf_{(\mathbf{v}_h, \boldsymbol{\psi}_h) \in \mathcal{V}_h} \left(\|\nabla \times \mathbf{u} - \nabla \times \mathbf{v}_h\| + \|\nabla \times \nabla \times \mathbf{u} - \boldsymbol{\psi}_h\| \right) \\ &\quad \le (1 + \alpha(h)) \inf_{\mathbf{v}_h \in Y_h} \|\nabla \times \mathbf{u} - \nabla \times \mathbf{v}_h\| + 2 \inf_{\boldsymbol{\mu}_h \in X_h} \|\nabla \times \nabla \times \mathbf{u} + \boldsymbol{\mu}_h\|. \end{aligned}$$

Combination of this inequality and (5.72) completes the proof. □

Theorem 5.3.6. *Let* $(\mathbf{u}, \phi)$ *and* $(\mathbf{u}_h, \boldsymbol{\phi}_h)$ *be the solutions of* (5.68) *and* (5.70)*, respectively. Furthermore, assume that* $(\nabla \times)^i \mathbf{u} \in H^s(\Omega)^3, i = 1, 2, 3$ *and* s *is the same as in Lemma 5.2.12. Then there exists a constant* C *independent of the mesh size* h *such that*

$$\begin{aligned} &\|\nabla \times \mathbf{u} - \nabla \times \mathbf{u}_h\| + \|\nabla \times \nabla \times \ \mathbf{u} - \boldsymbol{\phi}_h\| \\ &\qquad \le C h^{s-1} \left(\|\mathbf{u}\|_{H^s(\Omega)^3} + \|\nabla \times \mathbf{u}\|_{H^s(\Omega)^3} \right). \end{aligned} \tag{5.76}$$

Proof. We define the Fortin operator $\Pi_h : Y \to Y_h$ such that $\Pi_h \mathbf{u}$ is the first component $\mathbf{u}_h$ of (5.40) with $(\mathbf{f}, \boldsymbol{\phi}_h)$ replaced by $(\nabla \times \mathbf{u}, \nabla \times \boldsymbol{\phi}_h)$ (see Sec. 3 of [34]). According to Lemma 5.2.17, we have that

$$\|\mathbf{u} - \Pi_h \mathbf{u}\|_{H(\mathrm{curl};\Omega)} \le C \inf_{\mathbf{v}_h \in U_{0,h}} \|\mathbf{u} - \mathbf{v}_h\|_{H(\mathrm{curl};\Omega)}.$$

Using Lemma 5.2.12, the following inequality holds

$$\|\nabla\times\mathbf{u}-\nabla\times\Pi_h\mathbf{u}\|\le Ch^s\left(\|\mathbf{u}\|_{H^s(\Omega)^3}+\|\nabla\times\mathbf{u}\|_{H^s(\Omega)^3}\right).$$

For $\mathbf{w}=\nabla\times\nabla\times\mathbf{u}$, we define the $H(\mathrm{curl};\Omega)$ orthogonal projection

$$P_h: H(\mathrm{curl};\Omega)\to U_h$$

such that

$$(\nabla\times(\mathbf{w}-P_h\mathbf{w}),\nabla\times\boldsymbol{\phi}_h)+(\mathbf{w}-P_h\mathbf{w},\boldsymbol{\phi}_h)=0 \quad \text{for all } \boldsymbol{\phi}_h\in U_h.$$

Then Cea's Lemma leads to the following estimate (see Sec. 7.2 of [202])

$$\|\mathbf{w}-P_h\mathbf{w}\|_{H(\mathrm{curl};\Omega)}=\inf_{\boldsymbol{\mu}_h\in U_h}\|\mathbf{w}-\boldsymbol{\mu}_h\|_{H(\mathrm{curl};\Omega)}.$$

Letting $\boldsymbol{\phi}_h=\nabla\xi_h$ for $\xi_h\in S_h$, we find that $P_h\mathbf{w}$ is discrete divergence-free, i.e., $P_h\mathbf{w}\in X_h$. Thus

$$\inf_{\boldsymbol{\mu}_h\in X_h}\|\nabla\times\nabla\times\mathbf{u}-\boldsymbol{\mu}_h\|_{H(\mathrm{curl};\Omega)}\le\inf_{\boldsymbol{\mu}_h\in U_h}\|\nabla\times\nabla\times\ \mathbf{u}-\boldsymbol{\mu}_h\|_{H(\mathrm{curl};\Omega)}.$$

From Lemma 5.2.12, we have that

$$\begin{aligned}&\inf_{\boldsymbol{\mu}_h\in U_h}\|\nabla\times\nabla\times\mathbf{u}-\boldsymbol{\mu}_h\|_{H(\mathrm{curl};\Omega)}\\ \le\ & Ch^s\left(\|\nabla\times\nabla\times\ \mathbf{u}\|_{H^s(\Omega)^3}+\|(\nabla\times)^3\,\mathbf{u}\|_{H^s(\Omega)^3}\right)\end{aligned}$$

for some constants C independent of h. Using Theorem 5.3.5, we obtain that

$$\begin{aligned}&\|\nabla\times\mathbf{u}-\nabla\times\mathbf{u}_h\|+\|\nabla\times\nabla\times\ \mathbf{u}-\boldsymbol{\phi}_h\|\\ \le\ & C\left(1+\frac{C_1}{h}\right)h^s\left(\|\mathbf{u}\|_{H^s(\Omega)^3}+\|\nabla\times\mathbf{u}\|_{H^s(\Omega)^3}\right)\\ &\qquad +Ch^s\left(\|\nabla\times\nabla\times\ \mathbf{u}\|_{H^s(\Omega)^3}+\|(\nabla\times)^3\,\mathbf{u}\|_{H^s(\Omega)^3}\right)\\ \le\ & Ch^{s-1}\left(\|\mathbf{u}\|_{H^s(\Omega)^3}+\|\nabla\times\mathbf{u}\|_{H^s(\Omega)^3}\right)\\ &\qquad +Ch^s\left(\|\nabla\times\nabla\times\mathbf{u}\|_{H^s(\Omega)^3}+\|(\nabla\times)^3\,\mathbf{u}\|_{H^s(\Omega)^3}\right)\\ \le\ & Ch^{s-1}\left(\|\mathbf{u}\|_{H^s(\Omega)^3}+\|\nabla\times\mathbf{u}\|_{H^s(\Omega)^3}\right).\end{aligned}$$

□

We now use the theory from Section 5.2.2 to prove an L^2-norm convergence result for $\mathbf{u}-\mathbf{u}_h$. Of course, since we are using edge elements of the first kind [208], the convergence rate in $L^2(\Omega)$ cannot be better than the convergence rate in $H(\mathrm{curl};\Omega)$. So nothing would be gained from a duality argument.

Theorem 5.3.7. *Under the conditions of Theorem 5.3.6, there exists a constant C independent of $\mathbf{u}$, $\mathbf{u}_h$, and h such that*

$$\|\mathbf{u}-\mathbf{u}_h\|\le Ch^{s-1}\left(\|\mathbf{u}\|_{H^s(\Omega)^3}+\|\nabla\times\mathbf{u}\|_{H^s(\Omega)^3}\right).$$

Proof. Let $\mathbf{v}_h \in Y_h$ be the first component of the solution of (5.40) with

$$\mathbf{f} = \nabla \times \nabla \times \mathbf{u}$$

so that $\mathbf{u}$ is the exact solution. By Lemma 5.2.17 and Lemma 5.2.12, we have that

$$\|\mathbf{u} - \mathbf{v}_h\|_{H(\text{curl};\Omega)} \leq Ch^s \left(\|\mathbf{u}\|_{H^s(\Omega)^3} + \|\nabla \times \mathbf{u}\|_{H^s(\Omega)^3}\right). \tag{5.77}$$

Then, using the triangle inequality and the discrete Friedrichs inequality in Lemma 5.2.16, we have that

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_h\| &\leq \|\mathbf{u} - \mathbf{v}_h\| + \|\mathbf{v}_h - \mathbf{u}_h\| \\ &\leq \|\mathbf{u} - \mathbf{v}_h\| + C\|\nabla \times (\mathbf{v}_h - \mathbf{u}_h)\| \\ &\leq C(\|\mathbf{u} - \mathbf{v}_h\|_{H(\text{curl};\Omega)} + \|\nabla \times (\mathbf{u} - \mathbf{u}_h)\|). \end{aligned}$$

Combination of Theorem 5.3.6 and (5.77) completes the proof. $\square$

5.3.2 The Quad-curl Eigenvalue Problem

The quad-curl eigenvalue problem is to find λ and $\mathbf{u}$ such that

$$\nabla \times \nabla \times \nabla \times \nabla \times \mathbf{u} = \lambda \mathbf{u} \quad \text{in } \Omega, \tag{5.78a}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \tag{5.78b}$$

$$\mathbf{u} \times \boldsymbol{\nu} = 0 \quad \text{on } \partial\Omega, \tag{5.78c}$$

$$(\nabla \times \mathbf{u}) \times \boldsymbol{\nu} = 0 \quad \text{on } \partial\Omega. \tag{5.78d}$$

We call λ a quad-curl eigenvalue and $\mathbf{u}$ the associated eigenfunction. Due to the well-posedness of the quad-curl problem, we can define an operator

$$T : L^2(\Omega)^3 \to L^2(\Omega)^3$$

such that $T\mathbf{f} = \mathbf{u}$ for (5.62). It is obvious that T is self-adjoint. Furthermore, because of the compact imbedding of V into $L^2(\Omega)^3$, T is a compact operator.

The weak formulation for the quad-curl eigenvalue problem is to find $(\lambda, \mathbf{u}) \in \mathbb{R} \times V$ such that

$$\mathcal{C}(\mathbf{u}, \mathbf{q}) = \lambda(\mathbf{u}, \mathbf{q}) \quad \text{for all } \mathbf{q} \in V. \tag{5.79}$$

It is clear that λ is an eigenvalue satisfying (5.79) if and only if $\mu = 1/\lambda$ is an eigenvalue of T.

Lemma 5.3.8. *There is an infinite discrete set of quad-curl eigenvalues $\lambda_j > 0, j = 1, 2, \ldots$ and corresponding eigenfunctions $\mathbf{u}_j \in V$, $\mathbf{u}_j \neq \mathbf{0}$ such that* (5.79) *is satisfied and $0 < \lambda_1 \leq \lambda_2 \leq \ldots$. Furthermore*

$$\lim_{j \to \infty} \lambda_j = \infty.$$

The eigenfunctions satisfy $(\mathbf{u}_j, \mathbf{u}_l)_{L^2(\Omega)^3} = 0$ if $j \neq l$.

Proof. Applying the Hilbert-Schmidt theory (Theorem 1.1.13, see also, for example, Theorem 2.36 of [202]), we immediately have the above result. □

Using the Helmholtz decomposition, we can easily obtain the following result. Thus we omit its proof.

Lemma 5.3.9. *The quad-curl eigenvalues coincide with the non-zero eigenvalues of the following problem. Find* $(\lambda, \mathbf{u}) \in \mathbb{R} \times H_0^2(curl;\Omega)$ *such that*

$$\mathcal{C}(\mathbf{u}, \mathbf{q}) = \lambda(\mathbf{u}, \mathbf{q}) \quad \textit{for all } \mathbf{q} \in H_0^2(curl;\Omega). \tag{5.80}$$

Then the quad-curl eigenvalue problem in mixed form can be written as: Find $\lambda \in \mathbb{R}$, $(\mathbf{0}, \mathbf{0}) \neq (\mathbf{u}, \boldsymbol{\phi}) \in W \times X$ satisfying

$$a(\boldsymbol{\phi}, \mathbf{v}) + b(\mathbf{v}, \mathbf{u}) = 0 \qquad \text{for all } \mathbf{v} \in X, \tag{5.81a}$$

$$b(\boldsymbol{\phi}, \mathbf{q}) = -\lambda(\mathbf{u}, \mathbf{q}) \qquad \text{for all } \mathbf{q} \in Y. \tag{5.81b}$$

It is easy to see that if $(\lambda, (\mathbf{u}, \boldsymbol{\phi}))$ is an eigenpair of (5.81), then $\lambda B\mathbf{u} = \mathbf{u}, \mathbf{u} \neq \mathbf{0}$, i.e., $(\lambda, \mathbf{u})$ is a quad-curl eigenpair. If $\lambda B\mathbf{u} = \mathbf{u}, \mathbf{u} \neq \mathbf{0}$, then there exists $\boldsymbol{\phi} \in X$ such that $(\lambda, (\mathbf{u}, \boldsymbol{\phi}))$ is an eigenpair of (5.81).

Recall that the mixed finite element method for the quad-curl problem is as follows. For $\mathbf{f} \in H(\mathrm{div}^0;\Omega)$, find $A_h\mathbf{f} \in Y_h$, $B_h\mathbf{f} \in X_h$ such that

$$a(A_h\mathbf{f}, \mathbf{v}_h) + b(\mathbf{v}_h, B_h\mathbf{f}) = 0 \qquad \text{for all } \mathbf{v}_h \in X_h, \tag{5.82a}$$

$$b(A_h\mathbf{f}, \mathbf{q}_h) = -(\mathbf{f}, \mathbf{q}_h) \qquad \text{for all } \mathbf{q}_h \in Y_h. \tag{5.82b}$$

From Theorems 5.3.6 and 5.3.7, we have that

$$\|(B - B_h)\mathbf{f}\| \leq Ch^{s-1}\left(\|B\mathbf{f}\|_{H^s(\Omega)^3} + \|\nabla \times B\mathbf{f}\|_{H^s(\Omega)^3}\right), \tag{5.83}$$

$$\|(A - A_h)\mathbf{f}\| \leq Ch^{s-1}\left(\|B\mathbf{f}\|_{H^s(\Omega)^3} + \|\nabla \times B\mathbf{f}\|_{H^s(\Omega)^3}\right). \tag{5.84}$$

In the following, we assume that

$$\|\mathbf{u}\|_{H^s(\Omega)^3} \leq C\|\mathbf{f}\|$$

and

$$\|\nabla \times \mathbf{u}\|_{H^s(\Omega)^3} \leq C\|\mathbf{f}\|$$

hold for some constant C. Note that when $s = 2$, the above regularity result is a consequence of Theorem 5.2.16 and the fact that $\mathbf{u}$ is the solution of the quad-curl problem. Thus we have the norm convergence

$$\lim_{h\to 0} \|B - B_h\| = 0$$

and

$$\lim_{h\to 0} \|A - A_h\| = 0.$$

The discrete eigenvalue problem is to find $\lambda_h \in \mathbb{R}$, $(\mathbf{u}_h, \boldsymbol{\phi}_h) \in Y_h \times X_h$ such that

$$a(\boldsymbol{\phi}_h, \mathbf{v}_h) + b(\mathbf{v}_h, \mathbf{u}_h) = 0 \qquad \text{for all } \mathbf{v}_h \in X_h, \tag{5.85a}$$

$$b(\boldsymbol{\phi}_h, \mathbf{q}_h) = -\lambda_h(\mathbf{u}_h, \mathbf{q}_h) \qquad \text{for all } \mathbf{q}_h \in Y_h. \tag{5.85b}$$

Theorem 5.3.10. *The discrete quad-curl eigenvalues of* (5.85) *coincide with the non-zero eigenvalues of the following problem. Find* $\lambda_h \in \mathbb{R}$ *and* $\mathbf{u}_h \in U_{0,h}, \boldsymbol{\phi}_h \in U_h$ *such that*

$$(\boldsymbol{\phi}_h, \mathbf{v}_h) - (\nabla \times \mathbf{v}_h, \nabla \times \mathbf{u}_h) = 0 \qquad \textit{for all } \mathbf{v}_h \in U_h, \tag{5.86a}$$

$$(\nabla \times \boldsymbol{\phi}_h, \nabla \times \mathbf{q}_h) = -\lambda_h(\mathbf{u}_h, \mathbf{q}_h) \qquad \textit{for all } \mathbf{q}_h \in U_{0,h}. \tag{5.86b}$$

Proof. We write

$$\mathbf{u}_h = \mathbf{u}_h^0 + \nabla\varphi_h, \quad \mathbf{u}_h^0 \in Y_h, \varphi_h \in S_h.$$

Letting $\mathbf{q}_h = \nabla\xi_h$ in (5.86b), we have that

$$0 = (\boldsymbol{\phi}_h, \nabla \times \nabla\xi_h) = -\lambda_h(\mathbf{u}_h, \nabla\xi_h) = \lambda_h(\nabla\varphi_h, \nabla\xi_h) \quad \text{for all } \xi_h \in S_h.$$

Then either $\lambda_h = 0$ or $(\nabla\varphi_h, \nabla\xi_h) = 0$ for all $\xi_h \in S_h$. It is clear that if $\lambda_h \neq 0$, we have $(\nabla\varphi_h, \nabla\xi_h) = 0$ for all $\xi_h \in S_h$, which implies $\nabla\varphi_h = 0$. Thus $\mathbf{u}_h = \mathbf{u}_h^0$, which is discrete divergence-free. □

Let μ be a non-zero eigenvalue of B. Recall that the ascent r of $\mu - B$ is defined as the smallest integer such that

$$N((\mu - B)^r) = N((\mu - B)^{r+1}),$$

where N denotes the null space. Let $m = \dim N((\mu - B)^r)$ be the algebraic multiplicity of μ. The geometric multiplicity of μ is $\dim N(\mu - B)$. Note that since B is self-adjoint, the two multiplicities are the same. Then there are m eigenvalues of B_h, $\mu_1(h), \dots, \mu_m(h)$ such that

$$\lim_{h\to 0} \mu_j(h) = \mu, \quad \text{for } j = 1, \dots, m. \tag{5.87}$$

Theorem 5.3.11. *Let* $\lambda = 1/\mu$ *be an exact quad-curl eigenvalue with multiplicity* m *and* $\lambda_{j,h}, j = 1, \dots, m$ *be the corresponding computed eigenvalues. Then we have that*

$$|\lambda - \lambda_{j,h}| \leq Ch^{2s-2} \tag{5.88}$$

for some constant C.

Proof. From (5.83) and (5.84), we obtain that

$$\|(B - B_h)\| \leq Ch^{s-1} \quad \text{and} \quad \|(A - A_h)\| \leq Ch^{s-1}.$$

Then the theorem is proved using Theorem 11.1 of [23]. □

5.3.3 Numerical Examples

In this section, we show two preliminary examples. Due to the restriction of computation power, we can only compute three mesh levels for the model problem. However, the result seems to verify the convergence of the mixed method.

The first one is for the quad-curl source problem. It is well known that the divergence of the curl of a smooth function is zero. So, to obtain a test solution, we just need to take $\mathbf{u}$ as the curl of an appropriate function.

Let $\Omega = [0,1]^3$. To satisfy the boundary condition, one can simply make the H^2 trace of $\mathbf{u}$ zero. Hence, we set

$$\mathbf{w} = (\sin^3 \pi x \sin^3 \pi y \sin^3 \pi z, 0, 0)$$

and

$$\mathbf{u} = \nabla \times \mathbf{w} = \begin{pmatrix} 0 \\ 3\pi \cos \pi z \sin^3 \pi x \sin^3 \pi y \sin^2 \pi z \\ -3\pi \cos \pi y \sin^3 \pi x \sin^2 \pi y \sin^3 \pi z \end{pmatrix}$$

satisfying

$$\nabla \cdot \mathbf{u} = 0, \quad \boldsymbol{\nu} \times \mathbf{u} = \boldsymbol{\nu} \times (\nabla \times \mathbf{u}) = 0$$

and

$$\begin{aligned} \mathbf{f} &= \nabla \times \nabla \times \nabla \times \nabla \times \mathbf{u} \\ &= \pi^5 \begin{pmatrix} 0 \\ 216 \cos^2 \pi x \cos^2 \pi y \cos \pi z \sin \pi x \sin \pi y \sin^2 \pi z \\ \quad +72 \cos^2 \pi x \cos^3 \pi z \sin \pi x \sin^3 \pi y \\ \quad -540 \cos^2 \pi x \cos \pi z \sin \pi x \sin^3 \pi y \sin^2 \pi z \\ \quad +72 \cos^2 \pi y \cos^3 \pi z \sin^3 \pi x \sin \pi y \\ \quad -540 \cos^2 \pi y \cos \pi z \sin^3 \pi x \sin \pi y \sin^2 \pi z \\ \quad -132 \cos^3 \pi z \sin^3 \pi x \sin^3 \pi y \\ \quad +615 \cos \pi z \sin^3 \pi x \sin^3 \pi y \sin^2 \pi z \\ -72 \cos^2 \pi x \cos^3 \pi y \sin \pi x \sin^3 \pi z \\ \quad -216 \cos^2 \pi x \cos \pi y \cos^2 \pi z \sin \pi x \sin^2 \pi y \sin \pi z \\ \quad +540 \cos^2 \pi x \cos \pi y \sin \pi x \sin^2 \pi y \sin^3 \pi z \\ \quad -72 \cos^3 \pi y \cos^2 \pi z \sin^3 \pi x \sin \pi z \\ \quad +132 \cos^3 \pi y \sin^3 \pi x \sin^3 \pi z \\ \quad +540 \cos \pi y \cos^2 \pi z \sin^3 \pi x \sin^2 \pi y \sin \pi z \\ \quad -615 \cos \pi y \sin^3 \pi x \sin^2 \pi y \sin^3 \pi z \end{pmatrix}. \end{aligned}$$

We employ the linear edge element R_1. Then the problem of approximating the solution of a quad-curl problem is reduced to two discrete curl-curl problems as described at the end of the previous section.

Table 5.6 gives the L^2 error in terms of the number of degrees of freedom. If N is the number of degrees of freedom, we expect that the error decreases $O(N^{-1/3})$ which corresponds to a convergence rate of $O(h)$ as expected.

mesh	vertices	edges	tetrahedra	$(\text{vertices})^{-1/3}$	L_2 error	Order
A	3403	21462	16999	0.0655	0.128	
B	12049	74345	60333	0.0436	0.094	0.76
C	195757	1332110	1119033	0.0172	0.032	1.16

Table 5.6: Convergence rate of the mixed method.

The second example is for the quad-curl eigenvalue problem. As we see previously, we can ignore the divergence-free condition when we compute the quad-curl eigenvalues. This enables us to work with the edge element space directly. Namely, we only need to solve the following problem. Find $\lambda \in \mathbb{R}$, $(\mathbf{u}_h, \mathbf{w}_h) \in U_{0,h} \times U_h$ such that

$$a(\mathbf{w}_h, \mathbf{v}_h) + b(\mathbf{v}_h, \mathbf{u}_h) = 0 \qquad \text{for all } \mathbf{v}_h \in U_h, \tag{5.89a}$$

$$b(\mathbf{w}_h, \mathbf{q}_h) = -\lambda_h(\mathbf{u}_h, \mathbf{q}_h) \qquad \text{for all } \mathbf{q}_h \in U_{0,h}. \tag{5.89b}$$

Let

$$\{\boldsymbol{\phi}_i, i = 1, \ldots, N\}$$

be a basis for $U_{0,h}$ and

$$\{\boldsymbol{\phi}_i, i = 1, \ldots, N, N+1, \ldots, M\}$$

be a basis for U_h. The matrix form corresponding to the above equations is given by

$$\begin{pmatrix} \mathbf{0}_{N\times N} & \mathcal{C}_{N\times M} \\ -\mathcal{C}_{M\times N} & \mathcal{M}_{M\times M} \end{pmatrix} = \lambda \begin{pmatrix} \mathcal{M}_{N\times N} & \mathbf{0}_{N\times M} \\ \mathbf{0}_{M\times N} & \mathbf{0}_{M\times M} \end{pmatrix} \tag{5.90}$$

where

$$\begin{aligned} \mathcal{C}_{N\times M}(i,j) &= (\nabla \times \boldsymbol{\phi}_j, \nabla \times \boldsymbol{\phi}_i), i = 1, \ldots N, j = 1, \ldots, M, \\ \mathcal{C}_{N\times M}(i,j) &= (\nabla \times \boldsymbol{\phi}_j, \nabla \times \boldsymbol{\phi}_i), i = 1, \ldots M, j = 1, \ldots, N, \\ \mathcal{M}_{N\times N}(i,j) &= (\boldsymbol{\phi}_j, \boldsymbol{\phi}_i), i = 1, \ldots N, j = 1, \ldots, N, \\ \mathcal{M}_{M\times M}(i,j) &= (\boldsymbol{\phi}_j, \boldsymbol{\phi}_i), i = 1, \ldots M, j = 1, \ldots, M. \end{aligned}$$

The resulting algebraic eigenvalue problem is solved by Matlab '*eigs*' on a desktop computer.

We consider two domains: the unit ball and the unit cube. Due to the restriction of the computational power available, the largest matrices we can compute are obtained using a rather coarse mesh ($h \approx 0.1$). This is why we are not able to show the convergence order. However, the eigenvalues seem to converge for all examples. Of course, a better eigenvalue solver on a more powerful machine is very much desired.

We show the results on a few meshes on both domains in Tables 5.7 and 5.8. The degrees of freedom are denoted by DoF in the table. For the unit ball, three meshes, corresponding to the mesh sizes $h \approx 0.3$, $h \approx 0.2$, $h \approx 0.15$, are used. For both the linear and quadratic edge elements, we see some numerical evidence of the convergence in Table 5.7.

	$h \approx 0.3$	$h \approx 0.2$	$h \approx 0.15$	$h \approx 0.1$
linear edge element	201.6299	199.3129	197.8822	196.6903
DoF	6580	21282	49792	917576
quadratic edge element	206.0821	200.7491	198.7244	-
DoF	35608	115348	270072	-

Table 5.7: The first quad-curl eigenvalues for the unit ball on a few meshes using the linear and quadratic edge elements. Besides the computed eigenvalue, we also show the degrees of freedom (DoF) of the discrete problems which equal the dimension of the matrices defined in (5.90).

For the unit cube, we use three meshes corresponding to the mesh sizes $h \approx 0.4$, $h \approx 0.2$, $h \approx 0.1$. The results are shown in Table 5.8 for both the linear and quadratic edge elements.

	$h \approx 0.4$	$h \approx 0.2$	$h \approx 0.1$	$h \approx 0.05$
linear edge element	1.5209e+03	1.6210e+03	1.6922e+03	1.7072e+03
DoF	734	5121	40359	325911
quadratic edge element	1.8240e+03	1.7486e+03	1.7236e+03	-
DoF	3924	27698	218854	-

Table 5.8: The first quad-curl eigenvalues for the unit cube on a few meshes using the linear and quadratic edge elements. Besides the computed eigenvalue, we also show the degrees of freedom (DoF) of the discrete problems which equal the dimension of the matrices defined in (5.90).

Chapter 6

The Transmission Eigenvalue Problem

6.1 Introduction

The transmission eigenvalue (TE) problem first appeared in the analysis of inverse problems in Kirsch [170] and in more generality in Colton and Monk [94]. The main goal at that time was to show that transmission eigenvalues can be easily avoided such that qualitative methods, e.g., the linear sampling method, can be used to reconstruct the unknown target.

Rynne and Sleeman [219] showed that there is at most a countable set of real transmission eigenvalues with the only possible accumulation point being infinity. Later, for spherically homogeneous media, it is proved that the transmission eigenvalues form at most a discrete set with infinity as the only possible accumulation point by the analytic Fredholm theory [96]. However, little was known about the existence of the transmission eigenvalues except the spherically stratified medium.

Recently, Païvarinta and Sylvester [213] proved the existence of at least one eigenvalue, and soon thereafter Cakoni, Gintides, and Haddar [64] proved the existence of infinitely many real transmission eigenvalues together with estimates, which started the program of research on using transmission eigenvalues to infer properties of the scatterer. Cakoni et al. [60, 144] have shown that transmission eigenvalues can be recovered from measurements of the scattered far field data. The recovery of transmission eigenvalues from near field data is studied by Sun in [227]. Later, Kirsch and Lechleiter [172] (see also Lechleiter and Rennoch [183], Lechleiter and Peters [182, 181]) studied the problem along the line of an inside-outside duality. Inverse spectral problems for transmission eigenvalues are also considered in [5, 130, 241].

The interior transmission eigenvalue problem is neither elliptic nor self-adjoint. It is not covered in any standard theory of partial differential equations. Furthermore, the problem is a system of two second order equations. A reformulation leads to a fourth order nonlinear eigenvalue problem. These properties make the computation of transmission eigenvalues very challenging.

Since 2010, significant efforts have been devoted to develop effective numerical methods for transmission eigenvalues [95, 228, 203, 161, 246, 232, 6, 160, 173, 7, 162, 129, 68, 187, 156, 8, 125]. The first numerical treatment appeared in [95], where three finite element methods were proposed. Later, a mixed finite element method using Lagrange elements was developed in [161]. However, error analysis was not addressed. An and Shen [6] proposed an efficient spectral-element based numerical method for transmission eigenvalues of two-dimensional, radially-stratified media. The first numerical method supported by a rigorous convergence analysis was introduced by Sun in [228], in which transmission eigenvalues are computed as roots of a nonlinear function whose values are generalized eigenvalues of a related positive definite fourth order problem. The method has two drawbacks: 1) only real transmission eigenvalues can be obtained, and 2) many fourth order eigenvalue problems need to be solved. In [98] boundary integral equations are used to compute real transmission eigenvalues in the special case when the index of refraction is constant. Recently, Cakoni et al. [68] reformulated the problem and proved convergence (based on Osborn's compact operator theory [211]) of a mixed finite element method. Li et al. [187] developed a finite element method based on writing the transmission eigenvalue problem as a quadratic eigenvalue problem.

Some non-traditional methods, including the linear sampling method in the inverse scattering theory [229] and the inside-out duality [183], were proposed to search transmission eigenvalues using scattering data. However, these methods seem to be computationally expensive since they rely on solving tremendous numbers of direct problems. Other methods [129, 160, 162] and the related source problem [149, 246] have been studied in the literature as well.

The transmission eigenvalues are related to the scattering of acoustic waves by a bounded simply connected inhomogeneous medium $\Omega \subset \mathbb{R}^2$. Let $n(x) \in L^\infty(\Omega)$, the index of refraction, be a bounded function and k be the wave number. The scattering problem for an incident u^i by an inhomogeneous medium is to find the total field $u := u^i + u^s$ such that

$$\Delta u + k^2 u = 0, \qquad \text{in } \mathbb{R}^2 \setminus \Omega, \tag{6.1a}$$

$$\Delta u + k^2 n(x) u = 0, \qquad \text{in } \Omega, \tag{6.1b}$$

$$u^+ - u^- = 0, \qquad \text{on } \partial\Omega, \tag{6.1c}$$

$$\left(\frac{\partial u}{\partial \nu}\right)^+ - \left(\frac{\partial u}{\partial \nu}\right)^- = 0, \qquad \text{on } \partial\Omega, \tag{6.1d}$$

$$\lim_{r\to\infty} r^{1/2}\left(\frac{\partial u^s}{\partial r} - iku^s\right) = 0, \tag{6.1e}$$

where $r = |x|$ and $\pm$ denote the values approaching from inside and outside of Ω, respectively. The Sommerfeld radiation condition (6.1e) holds uniformly in $\hat{x} = x/|x|$. This models the scattering of time harmonic acoustic waves by an inhomogeneous medium. It has a unique solution $u \in H^1_{loc}(\mathbb{R}^2)$ under suitable assumptions on $n(x)$ [93].

The transmission eigenvalue problem is related to the above scattering problem: find $k \in \mathbb{C}$, $w, v \in L^2(\Omega)$, $w - v \in H^2(\Omega)$ such that

$$\Delta w + k^2 n(x) w = 0, \qquad \text{in } \Omega, \tag{6.2a}$$

$$\Delta v + k^2 v = 0, \qquad \text{in } \Omega, \tag{6.2b}$$

$$w - v = 0, \qquad \text{on } \partial\Omega, \tag{6.2c}$$

$$\frac{\partial w}{\partial \nu} - \frac{\partial v}{\partial \nu} = 0, \qquad \text{on } \partial\Omega, \tag{6.2d}$$

where ν is the unit outward normal to $\partial\Omega$ and the index of refraction $n(x)$ is positive. Values of $k \neq 0$ such that there exists a nontrivial solution (w, v) to (6.2) are called the transmission eigenvalues (see [95]).

It is helpful to discuss the physical meaning of transmission eigenvalues. The transmission eigenvalue problem is related to the non-scattering of an incident wave. Note that, if u^i is such that $u^s = 0$, then $w := u|_\Omega$ and $v := u^i|_\Omega$ satisfy (6.2), However, even when k is a transmission eigenvalue, the scattered field does not vanish in general. This is due to the fact that it is impossible to extend v outside Ω such that v satisfies the Helmholtz equation in $\mathbb{R}^2$. Nevertheless, it is known that the solutions to the Helmholtz equation in Ω can be approximated by entire solutions in appropriate norms.

Define the Herglotz wave function by

$$v_g(x) := \int_{\mathbb{S}} g(d) e^{ikx\cdot d} \mathrm{d}s(d), \quad g \in L^2(\mathbb{S}) \tag{6.3}$$

where $\mathbb{S} := \{x \in \mathbb{R}^2 : |x| = 1\}$. Let k be a transmission eigenvalue with the nontrivial (w, v) satisfying (6.2). Then for a given $\epsilon > 0$, there is a v_{g_ϵ} such that

$$\|v_{g_\epsilon} - v\| < \epsilon$$

and the scattered field u^s corresponding to the incident field v_{g_ϵ} is $O(\epsilon)$, i.e., the scattered field u^s can be arbitrarily small by a suitable choice of the incident field.

6.2 Existence of Transmission Eigenvalues

We present some existence results for transmission eigenvalues in $\mathbb{R}^3$. Similar results hold in $\mathbb{R}^2$. Note that the theoretical results are still partial. For example, the existence of complex transmission eigenvalues for general domains with the index of refraction $n(x)$ being a function is still open.

6.2.1 Spherically Stratified Media

The early study of transmission eigenvalues focused on the simpler case of the spherically stratified media [96]. Consider the transmission eigenvalue problem (6.2) when $n(x) = n(r)$ is spherically stratified. Let Ω be a ball $\{x : |x| < a\}$ and $n \in C^2[0, a]$.

We can expand v and w in a series of spherical harmonics

$$v(x) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_l^m j_l(kr) Y_l^m(\hat{x}), \tag{6.4a}$$

$$w(x) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} b_l^m y_l(r) Y_l^m(\hat{x}), \tag{6.4b}$$

where $r = |x|$, $\hat{x} = x/|x|$, j_l is a spherical Bessel function of order l, and y_l is a real valued solution of

$$y'' + \frac{2}{r} y' + \left(k^2 n(r) - \frac{l(l+1)}{r^2} \right) y = 0 \tag{6.5}$$

normalized such that $y_l(r)$ behaves like $j_l(kr)$ as $r \to 0$.

From [92] we can represent this solution in the form

$$y_l(r) = j_l(kr) + k^2 \int_0^r G(r, s, k) j_l(ks) \, \mathrm{d}s, \tag{6.6}$$

where G is real valued and twice continuously differentiable for $0 \leq s \leq r$ and is an even entire function of k of finite exponential type. Setting $f_l(r) = ry_l(r)$ we see from (6.5) that f_l satisfies

$$f'' + \left(k^2 n(r) - \frac{l(l+1)}{r^2}\right) f = 0 \tag{6.7}$$

and from [93] we can deduce that for fixed $r > 0$ f_l is a bounded function of k as $k \to \infty$. Hence for fixed $r > 0$, y_l is an entire function of k of finite exponential type that is bounded for k on the positive real axis. The following existence result is from [95].

Theorem 6.2.1. *Assume that $n(x) = n(r)$ is spherically stratified, Ω is the ball $\{x : |x| < a\}$, and $n \in C^2[0, a]$. Then if $n(r)$ is not identically equal to one there exist a countably infinite number of transmission eigenvalues for* (6.2).

In some special cases, we can find the transmission eigenvalue exactly. Let $\Omega \subset \mathbb{R}^2$ be a disk of radius a and let the index of refraction n be a positive real constant. Solutions of the Helmholtz equation $\triangle v + k^2 v = 0$ in Ω are

$$J_m(kr)\cos m\theta, \quad J_m(kr)\sin m\theta, \quad m \geq 0, \tag{6.8}$$

where J_m is the first kind Bessel function of order m. Solutions of the Helmholtz equation $\triangle w + k^2 n w = 0$ in Ω are

$$J_m(k\sqrt{n}r)\cos m\theta, \quad J_m(k\sqrt{n}r)\sin m\theta, \quad m \geq 0. \tag{6.9}$$

For a fixed m, in order to make $v - w$ vanish on $\partial\Omega$, one can choose

$$v = J_m(kr)\cos m\theta, \quad m \geq 0$$

and

$$w = \frac{J_m(ka)}{J_m(k\sqrt{n}a)} J_m(k\sqrt{n}r)\cos m\theta, \quad m \geq 0.$$

The transmission eigenvalues are k's such that

$$\frac{\partial v}{\partial r} = \frac{\partial w}{\partial r} \quad \text{on} \quad \partial\Omega.$$

Using the recursive formula for the derivatives of Bessel's functions, one has that

$$\begin{aligned}
\frac{\partial J_m(kr)}{\partial r} &= k\left(J_{m-1}(kr) - \frac{m}{kr}J_m(kr)\right), \\
\frac{\partial J_m(k\sqrt{n}r)}{\partial r} &= k\sqrt{n}\left(J_{m-1}(k\sqrt{n}r) - \frac{m}{k\sqrt{n}r}J_m(k\sqrt{n}r)\right).
\end{aligned}$$

Then the eigenvalues are k's such that

$$J_1(ka)J_0(k\sqrt{n}a) = \sqrt{n}J_0(ka)J_1(k\sqrt{n}a), \quad m = 0, \tag{6.10}$$

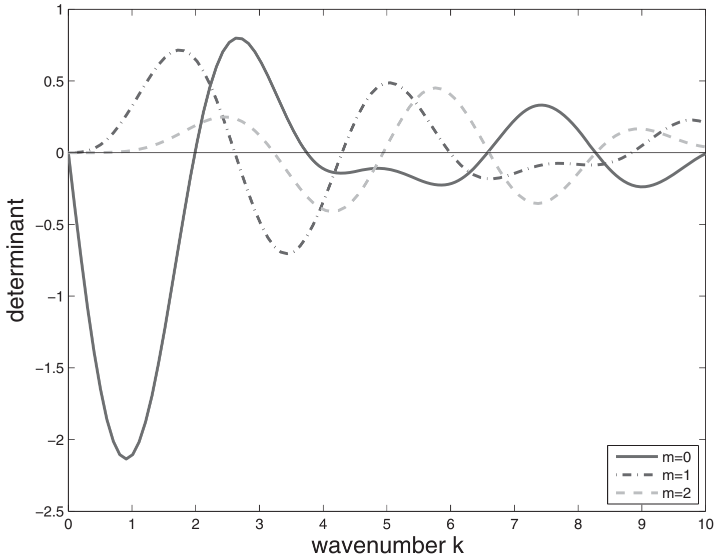


Figure 6.1: The plot of d_m against k for $m = 0, 1, 2$. The transmission eigenvalues are the intersections of the curves and the x-axis.

or

$$J_{m-1}(ka)J_m(k\sqrt{n}a) = \sqrt{n}J_m(ka)J_{m-1}(k\sqrt{n}a), \quad m \geq 1.$$

The case of $m = 0$ corresponds to the spherically stratified media when the index of refraction is a constant [96].

Considering a simple case when $a = 1/2$ and $n = 16$, we have that

$$J_1(k/2)J_0(2k) = 4J_0(k/2)J_1(2k), \quad m = 0,$$

or

$$J_{m-1}(k/2)J_m(2k) = 4J_m(k/2)J_{m-1}(2k), \quad m \geq 1.$$

In Fig. 6.1, we plot the value d_m against the wave number k where

$$d_0 = J_1(k/2)J_0(2k) - 4J_0(k/2)J_1(2k), \tag{6.11}$$

$$d_m = J_{m-1}(k/2)J_m(2k) - 4J_m(k/2)J_{m-1}(2k),\ m = 1, 2. \tag{6.12}$$

The transmission eigenvalues are those k's where $d_m = 0$. From Fig. 6.1, we see that the distribution of the (real) transmission eigenvalues is quite complicated. In Table 6.1, we show some transmission eigenvalues. The eigenvalues for $m > 0$ have multiplicity 2 since the above derivation works for both $\cos m\theta$ and $\sin m\theta$, $m > 0$ in (6.8) and (6.9). Note that similar derivation holds in $\mathbb{R}^3$.

m	eigenvalues		
0	1.9880	3.7594	6.5810
1	2.6129	4.2954	5.9875
2	3.2240	4.9462	6.6083
3	3.8248	5.5870	7.2591
4	4.4556	6.2278	7.9099

Table 6.1: Transmission eigenvalues corresponding to different m's of a disk with $a = 1/2$ and $n = 16$. These values are computed from (6.11) and (6.12).

6.2.2 General Media

In order for transmission eigenvalues to form a discrete set, it is clearly necessary that $n(x)$ is not identically equal to one. For general $n \in L^\infty(\Omega)$ the sharpest conditions to date on $n(x)$ for transmission eigenvalues to exist and form a discrete set are that $n(x)$ is either greater than or less than one in $\overline{\Omega}$ [64, 63]. This is clearly not optimal since for the case when $n(x) = n(r)$ depends only on $r = |x|$, it can be shown [96] that transmission eigenvalues exist and form a discrete set provided

$$\int_0^a \sqrt{n(r)}\,\mathrm{d}r \neq a, \tag{6.13}$$

where Ω is the ball $\{x : |x| < a\}$.

In two recent papers [64, 63], for a general domain Ω, Cakoni et al. obtained upper and lower bounds on $n(x)$ in terms of transmission eigenvalues for balls with constant index of refraction. In particular, they proved the following theorem.

Theorem 6.2.2. *Let $n(x) \in L^\infty(\Omega)$ and let B_1 be the largest ball such that $B_1 \subset \Omega$ and B_2 the smallest ball such that $\Omega \subset B_2$. Let $\gamma, \beta > 0$. Then*

1) If $1 + \gamma \le n_* \le n(x) \le n^* < \infty$ then

$$0 < k_{1,B_2,n^*} \le k_{1,D,n(x)} \le k_{1,B_1,n_*}.$$

2) If $0 < n_* \le n(x) \le n^* < 1 - \beta$ then

$$0 < k_{1,B_2,n_*} \le k_{1,D,n(x)} \le k_{1,B_1,n^*}.$$

Here k_{1,B_i,n_*} and k_{1,B_i,n^*}, $i = 1, 2$ are the first (real) transmission eigenvalues corresponding to the ball B_i with constant index of refraction n_* and n^*, respectively. $k_{1,\Omega,n(x)}$ is the first transmission eigenvalue of Ω with index of refraction $n(x)$.

Previously, Colton et al. obtained a Faber-Krahn type inequality [96]

$$k_1^2(\Omega) \ge \frac{\lambda_0(\Omega)}{\sup_\Omega n(x)}, \tag{6.14}$$

where k_1 is the smallest real transmission eigenvalue and $\lambda_0(\Omega)$ is the first Dirichlet eigenvalue. In addition, Theorem 6.2.2 shows that for constant index of refraction the first transmission eigenvalue depends monotonically on the index of refraction. Thus from a knowledge of the first transmission eigenvalue for Ω and $n(x)$ and the balls B_1 and B_2 we can obtain (in Case 1 of Theorem 6.2.2) a lower bound for $\sup n$ and an upper bound for $\inf n$. Similar estimates hold in Case 2 of Theorem 6.2.2.

6.2.3 Non-existence of Imaginary Transmission Eigenvalues

Having shown that transmission eigenvalues exist in the case of a spherically stratified medium, it is desirable to determine where they are located in the complex plane. We will show later that numerical evidence suggests that complex eigenvalues exist in the case of a spherically stratified medium [185]. However, the existence of complex transmission eigenvalues for a general medium is still open. Nevertheless, it can be shown that, if $n(x)$ is never equal to 1, there do not exist purely imaginary transmission eigenvalues. The following theorem is from [95].

Theorem 6.2.3. *Assume $n(x) > 1$ for $x \in \overline{\Omega}$ or $n(x) < 1$ for $x \in \overline{\Omega}$. Then there are no purely imaginary transmission eigenvalues.*

Proof. We first rewrite (6.2) as a fourth order problem. Let us recall the Sobolev space

$$H_0^2(\Omega) = \left\{ u \in H^2(\Omega) : u = 0 \text{ and } \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial\Omega \right\}.$$

Let $u = w - v \in H_0^2(\Omega)$. Subtracting (6.2b) from (6.2a), we obtain

$$(\triangle + k^2)u = -k^2(n(x) - 1)w.$$

Dividing $n(x) - 1$ and applying $(\triangle + k^2 n(x))$ to both sides of the above equation, we obtain

$$(\triangle + k^2 n(x)) \frac{1}{n(x) - 1} (\triangle + k^2)u = 0.$$

Then the weak formulation is to find a nontrivial solution $u \in H_0^2(\Omega)$ and $k \in \mathbb{C}$ such that

$$\int_\Omega \frac{1}{n-1} (\triangle u + k^2 u)(\triangle \bar{v} + k^2 n \bar{v})\, \mathrm{d}x = 0 \tag{6.15}$$

for all $v \in H_0^2(\Omega)$.

Let $n(x) > 1$ for $x \in \overline{\Omega}$ and assume, contrary to the statement of the theorem, that there exist purely imaginary transmission eigenvalues. Then $\frac{1}{n-1} \geq \sigma > 0$ and we define

$$\mathcal{A}_\tau(u, v) = \left(\frac{1}{n-1} (\triangle u + \tau u), (\triangle v + \tau v) \right) + \tau^2 (u, v), \tag{6.16}$$

$$\mathcal{B}(u, v) = (\nabla u, \nabla v), \tag{6.17}$$

where $\tau = k^2$. Then (6.15) can be written as

$$\mathcal{A}_\tau(u, v) - \tau \mathcal{B}(u, v) = 0 \quad \text{for all } v \in H_0^2(\Omega).$$

If k is purely imaginary, $\tau = -\sigma < 0$ with $\sigma > 0$. Setting $v = u$, we have

$$\begin{aligned} 0 &= \mathcal{A}_\tau(u,u) + \sigma\mathcal{B}(u,u) \\ &\geq \sigma^2(u,u) + \sigma(\nabla u, \nabla u) \end{aligned}$$

and this implies $u = 0$, which leads to a contradiction.

Similarly, if $n < 1$, then $\frac{n}{1-n} \geq \sigma > 0$. Let

$$\begin{aligned} \tilde{\mathcal{A}}_\tau(u,v) &= \left(\frac{1}{n-1}(\triangle u + \tau n u), (\triangle v + \tau n v)\right) + \tau^2(nu, v) \\ &= \left(\frac{n}{1-n}(\triangle u + \tau u), (\triangle v + \tau v)\right) + (\triangle u, \triangle v). \end{aligned} \tag{6.18}$$

Then

$$\begin{aligned} 0 &= \tilde{\mathcal{A}}_\tau(u,u) + \sigma\mathcal{B}(u,u) \\ &\geq (\triangle u, \triangle u) + \sigma(\nabla u, \nabla u). \end{aligned}$$

By Poincaré's inequality this again implies $u = 0$ and the proof is complete. □

6.2.4 Complex Transmission Eigenvalues

So far we only discuss the existence of real transmission eigenvalues for spherically stratified media. Since the problem is not self-adjoint, we can not exclude the possibility of complex transmission eigenvalues. In fact, early numerical experiments indicate the existence of complex eigenvalues [95]. Using (6.10), it is possible to search for transmission eigenvalues in the whole complex plane $\mathbb{C}$.

Define

$$Z_0(k) = J_1(ka)J_0(k\sqrt{n}a) - \sqrt{n}J_0(ka)J_1(k\sqrt{n}a).$$

Then the zeros of $Z_0(k)$, if they exist, are transmission eigenvalues, including the real k's given above. Again let $a = 1/2$ and $n = 16$. Using a contour plot for $|Z_0|$, we find that there exist a pair of complex transmission eigenvalues around $k = 4.901 \pm 0.5781i$ along with other real and complex eigenvalues. In Fig. 6.2 we plot $|Z_0|$ in a neighborhood of the origin. It can be seen that in addition to real transmission eigenvalues, there are also complex transmission eigenvalues. Note that since we require that $n(x)$ is real, the complex transmission eigenvalues must appear in complex conjugate pairs.

The first theoretical study of the existence of complex eigenvalues for spherically stratified media appeared in [185]. Using the analytic function theory, tt is shown that there possibly exist infinitely many complex transmission eigenvalues. We quote the following result from [185] without proof.

Theorem 6.2.4. *Consider the transmission eigenvalue problem* (6.2) *where the domain Ω is the unit disc in $\mathbb{R}^2$ or the unit ball in $\mathbb{R}^3$ and $n = n(r) > 0$ is a positive constant. Then*

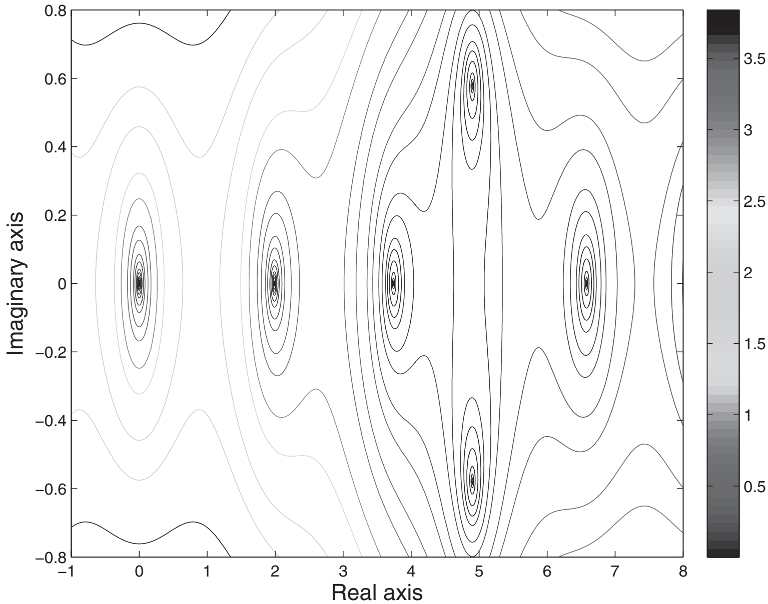


Figure 6.2: The contour plot of $|Z_0(k)|$ suggests the existence of complex transmission eigenvalues around $4.901 \pm 0.5781i$.

(i) In $\mathbb{R}^2$, if $n \neq 1$, then there exists an infinite number of complex eigenvalues.

(ii) In $\mathbb{R}^3$, if n is a positive integer not equal to one, then all transmission eigenvalues corresponding to spherically symmetric eigenfunctions are real. On the other hand if n is a rational positive number $n = p/q$ such that either $q < p < 2q$ or $p < q < 2p$, then there exists an infinite number of complex eigenvalues.

The theory of transmission eigenvalues is a rapid expanding field in the inverse scattering theory. There are many open problems which are of importance for theories and applications, for examples, the existence of complex transmission eigenvalues for general non-absorbing media, the existence of real transmission eigenvalues for absorbing media, and the conditions which guarantee the discreteness of transmission eigenvalues. We refer the readers to the reviewer paper [66] and the *Special Issue on Transmission Eigenvalues, Inverse Problem, Vol. 29, no. 10, Oct. 2013*.

6.3 Argyris Element for Real Transmission Eigenvalues

Starting from this section, we will present several finite element methods for transmission eigenvalues. We first introduce a conforming finite element using the Argyris element proposed in [228]. The method depends on a fourth order reformulation of the problem. Some functions are constructed involving an associated generalized fourth order eigenvalue problem. The roots of these functions are shown to be the transmission eigenvalues. Then iterative methods are applied to search the roots of these functions. The associated generalized eigenvalue problems are computed by the Argyris element. The convergence of the iterative methods is proved using the derivative of generalized eigenvalues [132, 107, 10].

6.3.1 A Fourth Order Reformulation

From last section, we see that the weak formulation for the transmission eigenvalue problem can be stated as follows. Find $(k^2 \neq 0, u) \in \mathbb{C} \times H_0^2(\Omega)$ such that

$$\left(\frac{1}{n(x)-1}(\triangle u + k^2 u), \triangle v + k^2 n(x) v\right) = 0 \quad \text{for all } v \in H_0^2(\Omega). \tag{6.19}$$

Recall the bilinear forms

$$\mathcal{A}_\tau(u,v) = \left(\frac{1}{n(x)-1}(\triangle u + \tau u), (\triangle v + \tau v)\right) + \tau^2(u,v), \tag{6.20a}$$

$$\begin{aligned}\tilde{\mathcal{A}}_\tau(u,v) &= \left(\frac{1}{1-n(x)}(\triangle u + \tau n(x) u), (\triangle v + \tau n(x) v)\right) + \tau^2(n(x)u,v)\\ &= \left(\frac{n(x)}{1-n(x)}(\triangle u + \tau u), (\triangle v + \tau v)\right) + (\Delta u, \Delta v),\end{aligned} \tag{6.20b}$$

$$\mathcal{B}(u,v) = (\nabla u, \nabla v), \tag{6.20c}$$

where $\tau := k^2$. For simplicity, we also call τ a transmission eigenvalue if k is. From (6.19), the transmission eigenvalues are τ's such that

$$\mathcal{A}_\tau(u,v) - \tau\mathcal{B}(u,v) = 0 \quad \text{for all } v \in H_0^2(\Omega), \tag{6.21}$$

when $n(x) > 1$ and

$$\tilde{\mathcal{A}}_\tau(u,v) - \tau\mathcal{B}(u,v) = 0 \quad \text{for all } v \in H_0^2(\Omega), \tag{6.22}$$

when $n(x) < 1$.

The following lemma provides useful properties of the generalized eigenvalue problems. Because of its importance for the iterative methods we will introduce shortly, we sketch its proof and refer the readers to [65] for more details.

Lemma 6.3.1. *(Lemma 2.1 of [228]) Let the index of refraction* $n(x)$ *satisfy*

$$\frac{1}{n(x)-1} > \gamma > 0, \quad \textit{a.e. in } \Omega, \tag{6.23}$$

or

$$\frac{n(x)}{1-n(x)} > \gamma > 0, \quad \textit{a.e. in } \Omega. \tag{6.24}$$

Then $\mathcal{A}_\tau$ *or* $\tilde{\mathcal{A}}_\tau$ *is a coercive sesquilinear form on* $H_0^2(\Omega) \times H_0^2(\Omega)$*. Moreover,* $\mathcal{B}$ *is symmetric and non-negative on* $H_0^2(\Omega)$*.*

Proof. Assuming that $n(x)$ satisfies (6.23), we have

$$\begin{aligned}
\mathcal{A}_\tau(u,u) &\geq \gamma\|\Delta u + \tau u\|^2 + \tau^2\|u\|^2 \\
&\geq \gamma\|\Delta u\|^2 - 2\gamma\tau\|\Delta u\|\|u\| + (\gamma+1)\tau^2\|u\|^2 \\
&= \epsilon\left(\tau\|u\| - \frac{\gamma}{\epsilon}\|\Delta u\|\right)^2 + \left(\gamma - \frac{\gamma^2}{\epsilon}\right)\|\Delta u\|^2 + (1+\gamma-\epsilon)\tau^2\|u\|^2 \\
&\geq \left(\gamma - \frac{\gamma^2}{\epsilon}\right)\|\Delta u\|^2 + (1+\gamma-\epsilon)\tau^2\|u\|^2
\end{aligned}$$

for $\gamma < \epsilon < \gamma + 1$. Moreover, letting $\lambda_0(\Omega)$ be the first Dirichlet eigenvalue of $-\Delta$ in Ω and using the Poincaré inequality, we get

$$\|\nabla u\|^2 \leq \frac{1}{\lambda_0(\Omega)}\|\Delta u\|^2$$

since $\nabla u \in H_0^1(\Omega)^2$. Thus $\mathcal{A}_\tau$ is a coercive sesquilinear form on $H_0^2(\Omega) \times H_0^2(\Omega)$, i.e.,

$$\mathcal{A}_\tau(u,u) \geq C_\tau\|u\|^2_{H^2(\Omega)} \tag{6.25}$$

for some positive constant C_τ.

Similarly, it can be shown that $\tilde{\mathcal{A}}_\tau$ is a coercive sesquilinear form on $H_0^2(\Omega) \times H_0^2(\Omega)$ provided (6.24) is satisfied. The conclusion on $\mathcal{B}$ is obvious. □

Hence we can define the following bounded self-adjoint linear operators

$$A_\tau : H_0^2(\Omega) \to H_0^2(\Omega), \qquad (A_\tau u, v) = \mathcal{A}_\tau(u,v), \tag{6.26a}$$

$$\tilde{A}_\tau : H_0^2(\Omega) \to H_0^2(\Omega), \qquad (\tilde{A}_\tau u, v) = \tilde{\mathcal{A}}_\tau(u,v), \tag{6.26b}$$

$$B : H_0^2(\Omega) \to H_0^2(\Omega), \qquad (Bu, v) = \mathcal{B}(u,v). \tag{6.26c}$$

Lemma 6.3.1 shows that B is a non-negative operator, A_τ is a positive definite operator if $\frac{1}{n(x)-1} > \gamma > 0$, and $\tilde{A}_\tau$ is a positive definite operator if $\frac{n(x)}{1-n(x)} > \gamma > 0$. Since $H_0^1(\Omega)^2$ is compactly embedded in $L^2(\Omega)^2$, B is a compact operator. In addition, A_τ and $\tilde{A}_\tau$ depend continuously on $\tau \in (0,\infty)$.

Now we consider the following generalized eigenvalue problems of finding $\lambda(\tau) \in \mathbb{R}$ and $u \in H_0^2(\Omega)$ such that

$$\mathcal{A}_\tau(u,v) - \lambda(\tau)\mathcal{B}(u,v) = 0 \quad \text{for all } v \in H_0^2(\Omega) \tag{6.27}$$

for $\frac{1}{n(x)-1} > \gamma > 0$ and finding $\lambda(\tau) \in \mathbb{R}$ and $u \in H_0^2(\Omega)$ such that

$$\tilde{\mathcal{A}}_\tau(u,v) - \lambda(\tau)\mathcal{B}(u,v) = 0 \quad \text{for all } v \in H_0^2(\Omega) \tag{6.28}$$

for $\frac{n(x)}{1-n(x)} > \gamma > 0$. It is obvious that $\lambda(\tau)$ is a continuous function of τ. From (6.21) and (6.22), a transmission eigenvalue is a root of

$$f(\tau) := \lambda(\tau) - \tau. \tag{6.29}$$

We will show the existence of an interval containing at least one root of (6.29). It can be obtained using the analytic results on bounds for transmission eigenvalues. We introduce an abstract theorem in [65] which provides the conditions for the existence of solutions of (6.29).

Theorem 6.3.2. *Let $\tau \to A_\tau$ be a continuous mapping from $(0,\infty)$ to the set of self-adjoint and positive definite bounded linear operators on a Hilbert space U, and let B be a self-adjoint and non-negative compact bounded linear operator on U. We assume that there exists two positive constants $\tau_0 > 0$ and $\tau_1 > 0$ such that*

1. *$A_{\tau_0} - \tau_0 B$ is positive on U,*
2. *$A_{\tau_1} - \tau_1 B$ is non-positive on a k-dimensional subspace W_k of U.*

Then each of the equations $\lambda_j(\tau) = \tau$ for $j = 1, \dots, k$ has at least one solution in $[\tau_0, \tau_1]$ where $\lambda_j(\tau)$ is the jth eigenvalue (counting multiplicity) of A_τ with respect to B, i.e., ker$(A_\tau - \lambda_j(\tau)B) \neq \{0\}$.

Under suitable assumptions on $n(x)$, the operators A_τ or $\tilde{A}_\tau$ with B satisfy the conditions of the above theorem with $U = H_0^2(\Omega)$. Let $n_* = \inf_\Omega(n)$, $n^* = \sup_\Omega(n)$, and $\mu_p(\Omega) > 0$ be the $(p+1)$th biharmonic eigenvalue with clamped plate boundary condition (counting the multiplicity) on Ω. Set

$$\theta_p(\Omega) := 4\frac{\mu_p(\Omega)^{1/2}}{\lambda_0(\Omega)} + 4\frac{\mu_p(\Omega)}{\lambda_0(\Omega)^2}. \tag{6.30}$$

The following theorem in [228] is a modification of Theorem 3.1 in [65], which provides conditions on $n(x)$ and gives intervals containing transmission eigenvalues.

Theorem 6.3.3. *(Theorem 2.3 of [228]) Let $n(x) \in L^\infty(\Omega)$ satisfy either one of the following assumptions*

$$1) \quad 1 + \theta_p(\Omega) \le n_* \le n(x) \le n^* < \infty \tag{6.31}$$

and

$$2) \quad 0 < n_* \le n(x) \le n^* < \frac{1}{1+\theta_p(\Omega)}. \tag{6.32}$$

Then, there exist $p+1$ transmission eigenvalues (counting multiplicity) in the interval $[\tau_0, \tau_1]$ where

$$\tau_0 = \frac{\lambda_0(\Omega)}{\sup_\Omega(n)} - \epsilon, \quad \tau_1 = \frac{\lambda_0(\Omega) - 2M\mu_p(\Omega)^{1/2}}{2+2M}, \quad M = \frac{1}{n_*-1} \tag{6.33}$$

for Case 1) and

$$\tau_0 = \lambda_0(\Omega) - \epsilon, \quad \tau_1 = \frac{\lambda_0(\Omega) - 2M\mu_p(\Omega)^{1/2}}{2M}, \quad M = \frac{n^*}{1-n^*} \tag{6.34}$$

for Case 2) with any $\epsilon > 0$.

The assumptions of Theorem 6.3.3 are restrictive and the estimates are crude. As a refined version of Theorem 6.2.2, the following result in [64] can also be used to determine an interval containing transmission eigenvalues. Let $\epsilon > 0$ such that Ω contains $m = m(\epsilon)$ disjoint disks B_ϵ of radius ϵ. Also let B_{r_1} be the largest ball of radius r_1 such that $B_{r_1} \subset \Omega$ and B_{r_2} be the smallest ball of radius r_2 such that $\Omega \subset B_{r_2}$.

Theorem 6.3.4. *Assume that $n(x) \in L^\infty(\Omega)$ and α, β are positive constants. Let k_{1,n_*} and k_{1,n^*} be the first transmission eigenvalues corresponding to the ball B_1 of radius one with the index of refraction n_* and n^*, respectively. Let $k_{1,\Omega}$ be the first transmission eigenvalue of Ω with index of refraction $n(x)$.*

1) If $1 + \alpha \le n_* \le n(x) \le n^* < \infty$, then

$$0 < \frac{k_{1,n^*}}{r_2} \le k_{1,\Omega} \le \frac{k_{1,n_*}}{r_1}. \tag{6.35}$$

There are at least $m(\epsilon)$ transmission eigenvalues in the interval $[\frac{k_{1,n^*}}{r_2}, \frac{k_{1,n_*}}{\epsilon}]$.

2) If $0 \le n_* \le n(x) \le n^* < 1 - \beta$, then

$$0 < \frac{k_{1,n_*}}{r_2} \le k_{1,\Omega} \le \frac{k_{1,n^*}}{r_1}. \tag{6.36}$$

There are at least $m(\epsilon)$ transmission eigenvalues in the interval $[\frac{k_{1,n_*}}{r_2}, \frac{k_{1,n^*}}{\epsilon}]$.

In general, the values in (6.35) and (6.36) provide better bounds for the transmission eigenvalues under milder conditions than Theorem 6.3.3. However, to use Theorem 6.3.4, we need to know the transmission eigenvalues of disks with constant index of refraction which is no easier than a general domain. The bounds in Theorem 6.3.3 can be obtained easily using finite element methods for Dirichlet eigenvalues and biharmonic eigenvalues.

The numerical methods are based on finding the root of a discrete version of (6.29). Since $\lambda(\tau)$ is the generalized eigenvalue of operator A_τ or $\tilde{A}_\tau$ with respect to B, we need to compute an approximation $\lambda_h(\tau)$ for $\lambda(\tau)$. This is done by using finite element methods for the generalized eigenvalue problems (6.27) and (6.28). In particular, we use the H^2-conforming Argyris elements [88], denoted by S_h.

Let $\mathcal{T}$ be a triangular mesh for Ω and assume that $\lambda_{j,h}(\tau)$ is the jth eigenvalue of the discrete eigenvalue problem

$$A_{\tau,h}\mathbf{x} = \lambda_{j,h}(\tau) B_h \mathbf{x}, \tag{6.37}$$

where $A_{\tau,h}$ (or $\tilde{A}_{\tau,h}$) and B_h are the finite element matrices for (6.27) (or (6.28)). Note that $\lambda_{j,h}(\tau)$ depends on τ continuously. To compute the jth transmission eigenvalue, we fix an index j and compute the jth eigenvalues $\lambda_{j,h}(\tau)$ of (6.37). These values are then used to compute the roots of (6.29). For simplicity, we drop the index j in the following except j needs to be specified otherwise. The following result is from [228].

Theorem 6.3.5. *Assume that we apply the Argyris finite element method for* (6.27) *or* (6.28) *on a Lipschitz domain* Ω *and the index of refraction* $n(x)$ *satisfies the condition of Lemma 6.3.1. Let* $\lambda_h(\tau)$ *be the finite element approximation of a generalized eigenvalue* $\lambda(\tau)$ *on a triangular mesh* $\mathcal{T}$ *with mesh size* h. *Then for any* $\epsilon > 0$, *there exists an* h_0 *such that if* $h \leq h_0$ *then*

$$|\lambda_h(\tau) - \lambda(\tau)| \leq \epsilon.$$

In the following, we present two iterative methods to compute the roots of

$$f_h(\tau) := \lambda_h(\tau) - \tau. \tag{6.38}$$

6.3.2 Bisection Method

We start with finding τ_0 and τ_1 such that the desired transmission eigenvalues are in $[\tau_0, \tau_1]$. According to the discussion in the previous section, we can either compute τ_0 and τ_1 using the smallest Dirichlet eigenvalue and the clamp plate eigenvalues ((6.33) or (6.34) of Theorem 6.3.3) for Ω or the transmission eigenvalues for disks containing or contained Ω ((6.35) or (6.36) of Theorem 6.3.4).

The bisection algorithm to compute N smallest transmission eigenvalues is as follows. The tolerance is denoted by tol.

Bisection Method:

Input:

- the index of refraction,
- the tolerance tol, and
- the number of transmission eigenvalues N to compute

Output:

- N real transmission eigenvalues

1. generate a regular triangular mesh for Ω
2. compute τ_0 and τ_1 and construct matrix B_h
3. for each $i, 1 \leq i \leq N$

   while $\text{abs}(\tau_0 - \tau_1) > tol$

- $\tau = (\tau_0 + \tau_1)/2$
- construct matrix $A_{\tau,h}$ depending on τ
- compute ith eigenvalue $\lambda_{i,h}$ of $A_{\tau,h}\mathbf{x} = \lambda B_h \mathbf{x}$
- if $\lambda_{i,h} - \tau > 0$
- $\quad \tau_0 = \tau$
- elseif $\lambda_{i,h} - \tau < 0$
- $\quad \tau_1 = \tau$
- else
- $\quad$ break
- end

end

In the following, we will establish the convergence of the above method using the derivatives of eigenvalues [132, 107, 10]. Let λ_h be a generalized eigenvalue of (6.37) and X be a matrix of eigenvectors associated with λ_h such that $X^T B_h X = I$. Thus we have

$$A_{\tau,h} X = B_h X \Lambda_h,$$

where $\Lambda_h = \lambda_h I$. In general, the repeated eigenvalue λ_h will separate as τ changes and the derivative of the eigenvalue λ_h with multiplicity m is not a scalar. We will denote it by

$$\Lambda_h' = \text{diag}(\lambda_{1,h}', \ldots, \lambda_{m,h}').$$

It is well known that the choice of X is not unique [107] and there exists a suitable matrix $\Gamma \in \mathbb{R}^{m \times m}$ such that $\Gamma^T \Gamma = I$ and the columns of orthogonal transformation $Z = X\Gamma$ are the eigenvectors for which a derivative can be defined.

Differentiating $A_{\tau,h} Z = B_h Z \Lambda_h$, we obtain

$$A_{\tau,h}' Z + A_{\tau,h} Z' = B_h' Z \Lambda_h + B_h Z' \Lambda_h + B_h Z \Lambda_h'.$$

Collecting similar terms, we obtain

$$(A_{\tau,h} - \lambda_h B_h) Z' = (\lambda_h B_h' - A_{\tau,h}') Z + B_h Z \Lambda_h'.$$

Multiplying the above equation by X^T, substituting $Z = X\Gamma$, and using the fact that

$$X^T (A_{\tau,h} - \lambda_h B_h) = 0,$$

we have

$$X^T (A_{\tau,h}' - \lambda_h B_h') X \Gamma = \Gamma \Lambda_h'.$$

Note that B_h does not depend on τ; we have $B_h' = 0$ and thus

$$\Lambda_h' = (X\Gamma)^T (A_{\tau,h}')(X\Gamma). \tag{6.39}$$

If λ_h is a distinct eigenvalue, we have

$$\lambda_h' = \mathbf{x}^T A_{\tau,h}' \mathbf{x}$$

where $\mathbf{x}$ is the associated eigenvector such that $\mathbf{x}^T B_h \mathbf{x} = 1$.

Next we show that $f_h'(\tau)$ is negative on an interval right to τ_0. Let $\lambda_h(\tau)$ be a generalized eigenvalue and X be the associated matrix of eigenvectors such that $X^T B_h X = I$. In addition, let $Z = X\Gamma$ be the transformation whose columns are the eigenvectors for which a derivative can be defined. This is true since we have generalized Hermitian eigenvalue problems.

Theorem 6.3.6. *(Lemma 3.2 of [228]) Let $\mathcal{A}'_{\tau,h}$ and $\tilde{\mathcal{A}}'_{\tau,h}$ represent the derivatives of $\mathcal{A}_{\tau,h}$ and $\tilde{\mathcal{A}}_{\tau,h}$, respectively. If $|\nabla \frac{1}{n(x)-1}| < c_g$ for some constant c_g for $n(x) > 1$ or $|\nabla \frac{1}{n(x)-1}| < c_g$ for some constant c_g for $n(x) < 1$, we have $f_h'(\tau) < 0$ when*

$$\tau < \frac{\left(1 + \frac{2}{n^*-1} - c_g - \frac{c_g}{\lambda_0(D)}\right)\lambda_0(\Omega)}{2\left(\frac{1}{n_*-1} + 1\right)} \tag{6.40}$$

for $n(x) > 1$ and

$$\tau < \frac{\left(1 + \frac{2}{1-n_*} - c_g - \frac{c_g}{\lambda_0(D)}\right)\lambda_0(\Omega)}{\frac{2n^*}{1-n^*}} \tag{6.41}$$

for $n(x) < 1$.

Proof. Assume that the index of refraction $n(x) > 1$ and $|\nabla \frac{1}{n(x)-1}| < c_g$ for some constant c_g. By simple calculations, we have

$$\begin{aligned} \mathcal{A}'_{\tau,h}(u,v) &= -\left(\nabla \frac{u}{n(x)-1}, \nabla v\right) - \left(\nabla u, \nabla \frac{v}{n(x)-1}\right) \\ &\quad + 2\tau\left(\frac{1}{n(x)-1}u, v\right) + 2\tau(u,v). \end{aligned}$$

Letting $v = u$, we have

$$\begin{aligned} \mathcal{A}'_{\tau,h}(u,u) &= -2\left(\left(\nabla \frac{1}{n(x)-1}\right)u, \nabla u\right) - 2\left(\frac{1}{n(x)-1}\nabla u, \nabla u\right) \\ &\quad + 2\tau\left(\frac{1}{n(x)-1}u, u\right) + 2\tau(u,u) \\ &\le c_g(\|u\|^2 + \|\nabla u\|^2) - \frac{2}{n^*-1}\|\nabla u\|^2 + \frac{2\tau}{n_*-1}\|u\|^2 + 2\tau\|u\|^2 \\ &\le c_g(\|u\|^2 + \|\nabla u\|^2) - \frac{2}{n^*-1}\|\nabla u\|^2 + 2\tau\left(\frac{1}{n_*-1} + 1\right)\|u\|^2. \end{aligned}$$

Let $\mathbf{x}$ be a column of Z and u be the corresponding function of $\mathbf{x}$ in S_h. Note that $(\nabla u, \nabla u) = 1$. Let $A'_{\tau,h}$ be the matrix corresponding to $\mathcal{A}'_{\tau,h}$. Then we have

$$\begin{aligned} \lambda_h'(\tau) &= Z^T A'_{\tau,h} Z \\ &\le c_g\|u\|^2 + c_g - \frac{2}{n^*-1} + 2\tau\left(\frac{1}{n_*-1} + 1\right)\|u\|^2 \\ &\le c_g\frac{1}{\lambda_0(\Omega)} + c_g - \frac{2}{n^*-1} + 2\tau\left(\frac{1}{n_*-1} + 1\right)\frac{1}{\lambda_0(\Omega)}, \end{aligned}$$

where we have applied the Poincaré inequality. Thus if

$$c_g \frac{1}{\lambda_0(\Omega)} + c_g - \frac{2}{n^* - 1} + 2\tau \left(\frac{1}{n_* - 1} + 1 \right) \frac{1}{\lambda_0(\Omega)} < 1,$$

i.e.,

$$\tau < \frac{\left(1 + \frac{2}{n^*-1} - c_g - \frac{c_g}{\lambda_0(\Omega)}\right) \lambda_0(\Omega)}{2\left(\frac{1}{n_*-1} + 1\right)}, \tag{6.42}$$

then

$$f_h'(\tau) = \lambda_h'(\tau) - 1 < 0,$$

which implies $f(\tau)$ is monotonically decreasing.

Similarly, let $\tilde{\mathcal{A}}'_{\tau,h}$ represent the derivative of $\tilde{\mathcal{A}}_{\tau,h}$ with respect to τ. Assume that the index of refraction $n(x) < 1$ and $|\nabla \frac{n(x)}{1-n(x)}| < c_g$ for some constant c_g. We have

$$\tilde{\mathcal{A}}'_{\tau,h}(u, v) = -\left(\nabla u, \nabla \frac{n(x)v}{1 - n(x)}\right) - \left(\nabla \frac{n(x)u}{1 - n(x)}, \nabla v\right) + 2\tau \left(\frac{n(x)}{1 - n(x)} u, v\right).$$

Letting $v = u$, we get

$$\begin{aligned} \tilde{\mathcal{A}}'_{\tau,h}(u, u) &= -2\left(\left(\nabla \frac{n(x)}{1 - n(x)}\right) u, \nabla u\right) - 2\left(\frac{n(x)}{1 - n(x)} \nabla u, \nabla u\right) \\ &\quad + 2\tau \left(\frac{n(x)}{1 - n(x)} u, u\right) \\ &\leq c_g(\|u\|^2 + \|\nabla u\|^2) - \frac{2}{1 - n_*} \|\nabla u\|^2 + \frac{2\tau n^*}{1 - n^*} \|u\|^2. \end{aligned}$$

Hence

$$\begin{aligned} \lambda_h'(\tau) &= Z^T \tilde{A}'_{\tau,h} Z \\ &= c_g \|u\|^2 + c_g - \frac{2}{1 - n_*} + \frac{2\tau n^*}{1 - n^*} \|u\|^2 \\ &\leq c_g \frac{1}{\lambda_0(\Omega)} + c_g - \frac{2}{1 - n_*} + \frac{2\tau n^*}{1 - n^*} \frac{1}{\lambda_0(\Omega)}, \end{aligned}$$

where again we applied the Poincaré inequality. Thus if

$$c_g \frac{1}{\lambda_0(\Omega)} + c_g - \frac{2}{1 - n_*} + \frac{2\tau n^*}{1 - n^*} \frac{1}{\lambda_0(\Omega)} < 1,$$

i.e.,

$$\tau < \frac{\left(1 + \frac{2}{1-n_*} - c_g - \frac{c_g}{\lambda_0(\Omega)}\right) \lambda_0(\Omega)}{\frac{2n^*}{1-n^*}}, \tag{6.43}$$

we also obtain

$$f_h'(\tau) = \lambda_h'(\tau) - 1 < 0.$$

Note that the above derivation does not depend on the mesh size h. □

In the case of constant index of refraction for a simple eigenvalue, the results can be simplified. Assuming $n > 1$ is constant and using integration by part, we obtain

$$A_{\tau,h} = H - \frac{2\tau}{n-1}G + \tau^2 \frac{n}{n-1}M, \tag{6.44}$$

where H, G, and M are matrices corresponding to $\frac{1}{n-1}(\triangle u, \triangle v)$, $(\nabla u, \nabla v)$, and (u, v), respectively. Then we obtain

$$A'_{\tau,h} = -\frac{2}{n-1}G + 2\tau \frac{n}{n-1}M.$$

Assume that $\mathbf{x}$ is an eigenvector associated with λ_h. Letting u be the corresponding function of $\mathbf{x}$ in S_h, we have $(\nabla u, \nabla u) = 1$. Hence

$$\begin{aligned}
\lambda'_h(\tau) &= \mathbf{x}^T A'_{\tau,h}\mathbf{x} \\
&= -\frac{2}{n-1}(\nabla u, \nabla u) + 2\tau \frac{n}{n-1}(u, u) \\
&\leq -\frac{2}{n-1} + 2\tau \frac{n}{n-1}\frac{1}{\lambda_0(\Omega)}.
\end{aligned}$$

Thus we get $f'_h(\tau) < 0$ if

$$-\frac{2}{n-1} + 2\tau \frac{n}{n-1}\frac{1}{\lambda_0(\Omega)} < 1,$$

i.e.,

$$\tau < \frac{n+1}{2n}\lambda_0(\Omega). \tag{6.45}$$

For the case of $\tilde{A}_{\tau,h}$, assuming the index of refraction $0 < n < 1$ is a constant, we have

$$\tilde{A}'_{\tau,h} = -\frac{2n}{1-n}G + 2\tau\left(\frac{n^2}{1-n} + n\right)M.$$

Hence

$$\begin{aligned}
\lambda'_h(\tau) &= \mathbf{x}^T \tilde{A}'_{\tau,h}\mathbf{x} \\
&= -\frac{2n}{1-n}(\nabla u, \nabla u) + 2\tau\left(\frac{n^2}{1-n} + n\right)(u, u) \\
&\leq -\frac{2n}{1-n} + 2\tau\left(\frac{n^2}{1-n} + n\right)\frac{1}{\lambda_0(\Omega)}.
\end{aligned}$$

Thus we have $f'_h(\tau) < 0$ if

$$-\frac{2n}{1-n} + 2\tau\left(\frac{n^2}{1-n} + n\right)\frac{1}{\lambda_0(\Omega)} < 1,$$

i.e.,

$$\tau < \frac{n+1}{2n}\lambda_0(\Omega).$$

Now we show some numerical study of function $f_h(\tau)$ as a verification of the above results. We consider the case when Ω is a disk with radius $1/2$. The computation is done on a mesh $\mathcal{T}$ for Ω whose size $h \approx 0.05$. In Fig. 6.3, we plot $f_{1,h} = \lambda_{1,h}(\tau) - \tau$ with $n = 24, 16, 8, 4$. We see that $f_{1,h}$ is positive for small positive τ and monotonically decreasing in an interval right to zero. From (6.45), we

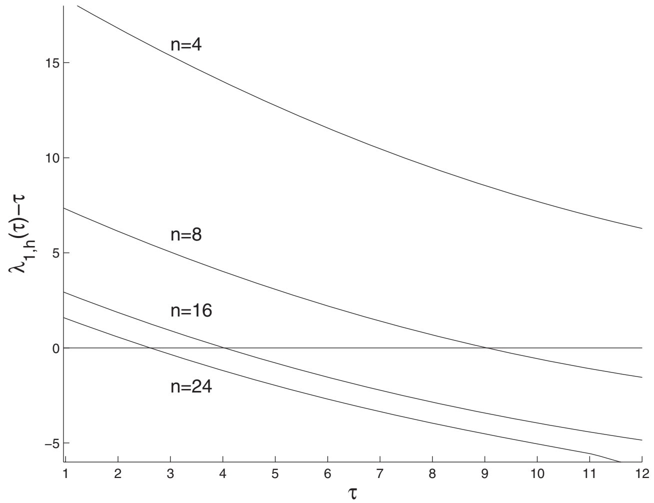


Figure 6.3: $\lambda_{1,h}(\tau) - \tau$ versus τ for $n = 24, 16, 8, 4$ when Ω is a disk of radius $1/2$.

have

$$\tau_2 := \frac{n+1}{2n}\lambda_0(\Omega) \approx 12.2311.$$

According to Theorem 6.3.6, for each j,

$$f_{j,h}(\tau) = \lambda_{j,h}(\tau) - \tau$$

is monotonically decreasing on (τ_0, τ_2). This conclusion is verified in Fig. 6.4.

The following lemma states that the root of (6.38) approximates the root of (6.29) well if the mesh size is small enough.

Theorem 6.3.7. *Let $f(\tau)$ and $f_h(\tau)$ be two continuous functions. For a small enough $\epsilon > 0$, we assume that*

$$f_h'(\tau) \leq -\delta < 0 \quad \text{for some } \delta > 0$$

and $|f(\tau) - f_h(\tau)| < \epsilon$ on an interval $[a - \epsilon/\delta, b + \epsilon/\delta]$ for some $0 < a < b$. If $f_h(\tau_0) = 0$ for some $\tau_0 \in [a, b]$, then there exists a τ_* such that $f(\tau_*) = 0$ and

$$|\tau_* - \tau_0| < \epsilon/\delta.$$

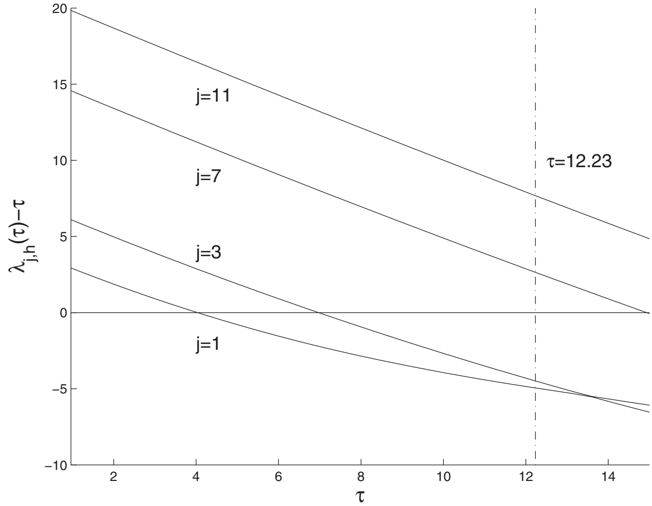


Figure 6.4: $\lambda_{j,h}(\tau) - \tau$ versus τ for $j = 1, 3, 7, 11$ when Ω is a disk of radius $1/2$ and $n = 16$.

Proof. Since $f_h'(\tau) \leq -\delta < 0$, if ϵ is small enough, there must exist τ_1 and τ_2 such that $f_h(\tau_1) > \epsilon$ and $f_h(\tau_2) < -\epsilon$. Furthermore, $|f(\tau) - f_h(\tau)| < \epsilon$ for all τ implies that $f(\tau_1) > 0$ and $f(\tau_2) < 0$. The existence of τ_* such that $f(\tau_*) = 0$ follows immediately since $f(\tau)$ is continuous.

Assume that $|\tau_* - \tau_0| \geq \epsilon/\delta$. Since $f_h(\tau_0) = 0$, we have $f_h(\tau_*) = f_h'(\xi)(\tau_* - \tau_0)$ for ξ between τ_0 and τ_*. Thus we have either $f_h(\tau_*) > \epsilon$ or $f_h(\tau_*) < -\epsilon$. Both contradict the fact that $|f_h(\tau_*) - f(\tau_*)| < \epsilon$. This completes the proof. $\square$

Combining Theorems 6.3.5, 6.3.6, and 6.3.7 and assuming we carry out the bisection method using the tolerance tol, we have the following convergence result.

Theorem 6.3.8. *(Theorem 3.4 of [228]) Assume that we apply the conforming finite element method for* (6.27) *or* (6.28) *using the Argyris element on a regular mesh* $\mathcal{T}$ *with mesh size* h *and the conditions in Theorems 6.3.5, 6.3.6, and 6.3.7 are satisfied. Let* τ_* *be the root of* (6.29) *and* τ_h *be the approximation of* τ_* *computed by the bisection method. Assuming that* τ *satisfies* (6.40) *and* (6.41)*, then for any* $\epsilon > 0$*, there exists* h_0 *such that for* $h < h_0$ *we have*

$$|\tau_h - \tau_*| \leq \epsilon/\delta + tol$$

for some fixed $\delta > 0$ *not depending on* ϵ.

Proof. Let $\lambda_h(\tau)$ be the finite element approximation of $\lambda(\tau)$ for the generalized eigenproblems (6.21) or (6.22). Then by Theorem 6.3.5, for any $\epsilon > 0$, there exist h_0 such that for a regular mesh with $h < h_0$, we have

$$|\lambda_h(\tau) - \lambda(\tau)| < \epsilon.$$

Assume τ_0 is the root of $f_h(\tau)$, i.e., $\lambda_h(\tau_0) - \tau_0 = 0$. It is obvious that $|\tau_h - \tau_0| < tol$. If τ satisfies (6.40) and (6.41), from the derivation of Theorem 6.3.6, there exist $\delta > 0$ such that $f_h'(\tau) < -\delta$ in a neighborhood of τ_0. Using Theorem 6.3.7, we have

$$|\tau_* - \tau_0| < \epsilon/\delta.$$

Then an application of the triangle inequality completes the proof. □

6.3.3 Secant Method

To use the above bisection method, we need to decide an interval $[\tau_0, \tau_1]$ which contains the desired transmission eigenvalues. However, computation of τ_0 and τ_1 using Theorem 6.3.3 would require the Dirichlet and the clamped plate eigenvalues. Conditions (6.31) and (6.32) of Theorem 6.3.3 also put a strict condition on the index of refraction $n(x)$. Theorem 6.3.4 provides an alternative way to decide τ_0 and τ_1 under mild restriction on $n(x)$. However, it requires the computation of the transmission eigenvalues of disks with constant index of refraction. To overcome these difficulties, we propose the following secant method to search the roots of $f_h(\tau)$. The method turns out to be very efficient in general.

Secant Method:

Input:

- x_0, x_1 - two initial values
- $n(x)$ - index of refraction, tol - tolerance
- N - number of transmission eigenvalues to be computed
- $maxit$ - maximum number of iteration

Output:

- N smallest real transmission eigenvalues

1. generate a regular triangular mesh for Ω
2. construct matrix B_h
3. for each $i, 1 \leq i \leq N$ do the following
   a. set $it = 0$ and $\delta =$abs$(x_1 - x_0)$
   b. $it = it + 1$
   c. $t = x_0$

d. construct matrix corresponding to $A_{t,h}$

e. compute the ith smallest generalized eigenvalue λ_A of $A_{t,h}\mathbf{x} = \lambda B_h \mathbf{x}$

f. $t = x_1$

g. construct matrices corresponding to $A_{t,h}$

h. compute the ith generalized eigenvalue λ_B of $A_{t,h}\mathbf{x} = \lambda B_h \mathbf{x}$

i. while $\delta > tol$ and $it < maxit$

- $t = x_1 - \lambda_B \frac{x_1 - x_0}{\lambda_B - \lambda_A}$
- construct the matrix corresponding to $A_{t,h}$
- compute the ith smallest eigenvalue λ_t of $A_{t,h}\mathbf{x} = \lambda B_h \mathbf{x}$
- $\delta = \text{abs}(\lambda_t - t)$
- $x_0 = x_1, x_1 = t, \lambda_A = \lambda_B, \lambda_B = \lambda_t, it = it + 1.$

end

Here x_0 and x_1 are initial values which are chosen close to zero and $x_0 < x_1$. This is due to the fact that $f_h(\tau)$ is monotonically decreasing in an interval I right to zero. The $maxit$ is the maximum number of iterations. Similar to the bisection method, we have the following convergence theorem whose proof is straightforward (see [17]).

Theorem 6.3.9. *(Theorem 3.5 of [228]) Assume we apply the conforming finite element method for* (6.21) *or* (6.22) *using the Argyris element on a regular mesh $\mathcal{T}$ with mesh size h. Let*

$$f_h'(\tau) < -\delta < 0 \quad \textit{for } \delta > 0$$

on some interval $[a, b]$ where $a = \tau_0$ is given by (6.33) *and b is given by* (6.42) *for $n(x) > 1$ (*(6.34) *and* (6.43) *for $n(x) < 1$). Let ϵ be an arbitrary positive number. Assume that τ_* is the root of $f(\tau)$ such that $\tau_* \in [a + \epsilon/\delta, b - \epsilon/\delta]$. Let τ_0 be the root of $f_h(\tau)$ computed by the secant method. Then there exist an h_0 such that for $h < h_0$ we have*

$$|\tau_0 - \tau_*| \leq \epsilon/\delta + tol.$$

In the following we present some numerical examples. We use the linear Lagrange finite element to compute the smallest Dirichlet eigenvalue and the Argyris element to compute the clamped plate eigenvalues and the generalized eigenvalue problems. In all examples, we use a regular mesh with mesh size $h \approx 0.05$ and $tol = 10^{-6}$. The transmission eigenvalues computed here are consistent with the results in [95].

The major advantages of the proposed iterative methods over the finite element methods in [95] are the accuracy and speed. For example, it is impossible to use the Argyris method in [95] on a mesh with mesh size $h < 0.05$ for a disk with radius $1/2$ since solving the non-Hermitian eigenvalue problem using Matlab's *eig* leads to *Out of memory*. Instead of *eig*, one might use *sptarn* which is more efficient and needs less memory. However, a search interval needs to be specified precisely otherwise *sptarn* might not converge for our problem. In addition, there are no convergence results for the non-Hermitian iterative solvers up to date [27].

We compute the first transmission eigenvalue for three different domains: a disk Ω_1 of radius $R = 1/2$ centered at the origin, the unit square Ω_2 centered at the origin and a triangle Ω_3 whose vertices are given by $(-\frac{\sqrt{3}}{2}, -\frac{1}{2})$, $(\frac{\sqrt{3}}{2}, -\frac{1}{2})$, and $(0, 1)$. The mesh size $h \approx 0.05$ for all three domains. Table 6.2 shows the results of the bisection method when $n = 24$ and $n = \frac{1}{24}$. The sizes of the matrices are also shown in the table.

Domain	Size of $A_{\tau,h}$	n	τ_0	τ_1	$k_1^2(\Omega)$
Ω_1	17846×17846	24	0.9640	9.3825	2.5872
Ω_2	5630×5630	24	0.8225	7.9462	2.3275
Ω_3	2183×2183	24	0.7360	7.0675	2.1712
Ω_1	17846×17846	$\frac{1}{24}$	23.1373	225.1791	55.8562
Ω_2	5630×5630	$\frac{1}{24}$	19.7392	190.7097	62.0928
Ω_3	2183×2183	$\frac{1}{24}$	17.6641	169.6204	52.1111

Table 6.2: The first transmission eigenvalue computed by the bisection method using Theorem 6.3.3 for three domains: a disk Ω_1 of radius $R = 1/2$, the unit square Ω_2, and a triangle Ω_3 whose vertices are given by $(-\frac{\sqrt{3}}{2}, -\frac{1}{2})$, $(\frac{\sqrt{3}}{2}, -\frac{1}{2})$, and $(0, 1)$.

Next we consider the case when the index of refraction is not constant. We choose two domains: a disk Ω_1 of radius $R = 1/2$ and the unit square Ω_2 centered at the origin. The indices of refraction are given by $8 + 4|x|$ and $8 + x_1 - x_2$, respectively. We make these choices because the first transmission eigenvalues of both cases are obtained in [227] via the inverse scattering scheme and they can be used to verify the numerical results. The computed eigenvalues are shown in Table 6.3. We can see that the values computed by the bisection method (direct way) and by the inverse scattering scheme agree very well. Since [227] uses k_1 instead of k_1^2, we also show k_1 in Table 6.3.

Domain	$n(x)$	$k_1(\Omega)$ (inverse scattering)	$k_1(\Omega)$ (bisection method)
Ω_1	$8 + 4\|x\|$	2.78	2.8292
Ω_2	$8 + x_1 - x_2$	2.90	2.8834

Table 6.3: The first transmission eigenvalue when index of refraction is not constant for two domains: a disk Ω_1 of radius $R = 1/2$ and the unit square Ω_2 centered at the origin. The third column contains the values from [227] reconstructed by the inverse scattering scheme. The fourth column is computed by the bisection method.

6.3.4 Some Discussions

A major drawback of using Theorem 6.3.3 is the restriction on the index of refraction. It becomes severe if we want to compute several transmission eigenvalues. For example, suppose we want to compute five smallest transmission eigenvalues. Since $\mu_5(\Omega) \approx 25,337.6304$, we obtain $\theta_5(\Omega) \approx 216.8401$. This would require

$$n(x) > 1 + \theta_5(\Omega) \approx 217.8401$$

for the condition in Theorem 6.3.3 to be satisfied. As an alternative we can use the bounds given in Theorem 6.3.4 which requires the transmission eigenvalues of disks with constant index of refraction. We refer the readers to [95] for some discussion on how to obtain these transmission eigenvalues.

Let $n(x) = 16$ and Ω be the unit square. Then $B_1 = \{x; |x| < 1/2\}$ is the largest disk such that $B_1 \subset \Omega$ and $B_2 = \{x; |x| < 0.8\}$ is a disk such that $\Omega \subset B_2$. Note that the condition in Theorem 6.3.3 is not satisfied since $16 < 1 + \theta_0(\Omega) \approx 21.8749$. Let $k_{1,\Omega}$, k_{1,B_1}, and k_{1,B_2} be the first transmission eigenvalues of the above domains, respectively. From [95] we have

$$k_{1,B_1} \approx 1.9912, k_{1,B_2} \approx 1.2443.$$

Using these bounds in the bisection method, we obtain $k_{1,\Omega} \approx 1.8651$.

Next let Ω be the triangle whose vertices are given by $(-\frac{\sqrt{3}}{2}, -\frac{1}{2})$, $(\frac{\sqrt{3}}{2}, -\frac{1}{2})$, and $(0,1)$. Then $B_1 = \{x; |x| < 1/2\}$ satisfies $B_1 \subset \Omega$ and $B_2 = \{x; |x| < 1\}$ satisfies $\Omega \subset B_2$. Again $n(x) = 16$ violates the condition of Theorem 6.3.3. We have

$$k_{1,B_1} \approx 1.9912, k_{1,B_2} \approx 0.9956.$$

Using these bounds in the bisection method, we obtain $k_{1,\Omega} \approx 1.7885$.

The secant method only needs the value of the function and converges quickly for the smallest a few transmission eigenvalues. In Table 6.4, we show six smallest transmission eigenvalues computed by the secant method for a disk with radius $1/2$ and $n = 24$. The secant method converges much faster than the bisection method. For example, for the smallest transmission eigenvalue, the bisection method uses 27 iterations compared to 4 iterations by the secant method.

j	k_j^2	number of iterations
1	2.5872	4
2	4.5364	4
3	4.5389	4
4	6.9483	4
5	6.9525	4
6	8.7960	5

Table 6.4: Secant method: smallest 6 transmission eigenvalues for a disk with radius $1/2$ and $n = 24$.

6.4 A Mixed Method Using The Argyris Element

The method in the previous section only computes real transmission eigenvalue. In this section, we present a method that can also compute complex transmission eigenvalues due to Cakoni, Monk, and Sun [68]. It reformulates the problem into a mixed system involving a fourth order problem and a second order problem. The Argyris element and Lagrange element are then employed to discretize the system. Finally, the convergence of the eigenvalue problem is proved using a theorem by Osborn [211].

6.4.1 The Mixed Formulation

We first recall the fourth order formulation of the transmission eigenvalue problem. Find an eigenvalue $k \in \mathbb{C}$ and the corresponding nontrivial transmission eigenfunction $u \in H_0^2(\Omega)$ such that

$$\left(\frac{1}{n-1}(\Delta u + k^2 u), \Delta \overline{v} + k^2 \overline{nv}\right) = 0 \quad \text{for all } v \in H_0^2(\Omega), \tag{6.46}$$

where $\overline{v}$ denotes the complex conjugate of v. In this section, we assume $n > n_0 > 1$ almost everywhere where n_0 is constant, although, with obvious changes, the theory also holds for n strictly less than 1.

Remark 6.4.1. *In [66], the proof of the discreteness of eigenvalues of (6.46) uses fractional powers of certain compact operators to convert the problem to an eigenvalue problem for a system of compact operators. These operators are not convenient for numerical computation since computing fractional powers of inverses of solution operators is time consuming. The formulation below uses operators involving just the Laplacian that are easy to implement.*

Expanding (6.46) we obtain the problem of finding non-trivial $u \in H_0^2(\Omega)$ and $k \in \mathbb{C}$ such that

$$(\Delta u, \Delta v)_{n-1} + k^2 (u, \Delta v)_{n-1} + k^2 (\Delta u, nv)_{n-1} + k^4 (nu, v)_{n-1} = 0,$$

where

$$(u, v)_{n-1} = \int_\Omega \frac{1}{n-1} u \overline{v} \, \mathrm{d}x.$$

Obviously $k = 0$ is not an eigenvalue of this problem since the sesquilinear form $(\Delta u, \Delta v)_{n-1}$ is coercive on $H_0^2(\Omega)$. Define $\tau = k^2$ and let $w \in H_0^1(\Omega)$ satisfy

$$\Delta w = \tau \frac{n}{n-1} u \quad \text{in } \Omega.$$

Then we may rewrite the above transmission eigenvalue problem as the problem of finding $\tau \in \mathbb{C}$ and a nontrivial pair of functions $(u, w) \in H_0^2(\Omega) \times H_0^1(\Omega)$ such that

$$\begin{aligned} (\Delta u, \Delta v)_{n-1} &= -\tau \left((u, \Delta v)_{n-1} + (\Delta u, nv)_{n-1} - (\nabla w, \nabla v) \right), \\ (\nabla w, \nabla z) &= -\tau (nu, z)_{n-1}, \end{aligned}$$

for all $v \in H_0^2(\Omega)$ and $z \in H_0^1(\Omega)$. This is a new nonself-adjoint eigenvalue problem.

To analyze the problem, define the following sesquilinear forms where $u, v \in H_0^2(\Omega)$ and $z, w \in H_0^1(\Omega)$:

$$\begin{aligned} a(u, v) &= (\Delta u, \Delta v)_{n-1}, \\ b_1(u, v) &= (u, \Delta v)_{n-1} + (\Delta u, nv)_{n-1}, \\ b_2(w, v) &= -(\nabla w, \nabla v), \\ c(u, z) &= (nu, z)_{n-1}, \\ d(w, z) &= (\nabla w, \nabla z). \end{aligned}$$

Then we define the sesquilinear form A on

$$(H_0^2(\Omega) \times H_0^1(\Omega)) \times (H_0^2(\Omega) \times H_0^1(\Omega))$$

by

$$A((u, w), (v, z)) = a(u, v) + d(w, z).$$

It is clear that A is an inner product on $H_0^2(\Omega) \times H_0^1(\Omega)$.

The eigenvalue problem is then to find $\lambda \in \mathbb{C}$ and non-trivial $(u, w) \in H_0^2(\Omega) \times H_0^1(\Omega)$ such that

$$\lambda A((u, w), (v, z)) = b_1(u, v) + b_2(w, v) + c(u, z)$$

for all $(v, z) \in H_0^2(\Omega) \times H_0^1(\Omega)$, where $\lambda = -1/\tau$. Recall that $\tau = k^2 = 0$ is not a transmission eigenvalue for the fourth order formulation.

Now we define the operator

$$T : H_0^2(\Omega) \times H_0^1(\Omega) \to H_0^2(\Omega) \times H_0^1(\Omega)$$

by

$$A(T(u, w), (v, z)) = b_1(u, v) + b_2(w, v) + c(w, z)$$

for all $(v, z) \in H_0^2(\Omega) \times H_0^1(\Omega)$. Then, in operator notation, we seek $\lambda \in \mathbb{C}$ and non-trivial $(u, w) \in H_0^2(\Omega) \times H_0^1(\Omega)$ such that

$$\lambda(u, w) = T(u, w).$$

Note that if $\lambda \neq 0$, $(0, w)$, $w \in H_0^1(\Omega)$, is not an eigenfunction of this system, so we have not introduced spurious eigenvalues into the problem.

Assume we use conforming finite element spaces $X_h \subset H_0^2(\Omega)$ and $Y_h \subset H_0^1(\Omega)$ to compute a finite dimensional eigenvalue problem. We cover Ω with a shape-regular triangulation $\mathcal{T}_h$ consisting of triangles K with maximum diameter h. In this case an obvious choice is to use Argyris elements to build X_h, and this is the choice we will use later in the numerical tests. To build Y_h we could use simple continuous piecewise polynomials, and in our code we use piecewise linear or piecewise quadratic Lagrange elements.

The finite element problem is to seek $\lambda_h \in \mathbb{C}$ and non-trivial $(u_h, v_h) \in X_h \times Y_h$ such that

$$\lambda_h A((u_h, w_h), (v_h, z_h)) = b_1(u_h, v_h) + b_2(w_h, v_h) + c(w_h, z_h)$$

for all $(v_h, z_h) \in X_h \times Y_h$. We next define an approximation to the operator T denoted by

$$T_h : H_0^2(\Omega) \times H_0^1(\Omega) \to X_h \times Y_h$$

such that for $(p, q) \in H_0^2(\Omega) \times H_0^1(\Omega)$, $T_h(p, q) \in X_h \times Y_h$ satisfies

$$A(T_h(p, q), (v_h, z_h)) = b_1(p, v_h) + b_2(q, v_h) + c(q, z_h)$$

for all $(v_h, z_h) \in X_h \times Y_h$.

6.4.2 Convergence Analysis

The discrete eigenvalue problem is to find approximate transmission eigenvalues $\lambda_h \in \mathbb{C}$ and non-trivial eigenfunctions $(u_h, w_h) \in X_h \times Y_h$ satisfying

$$\lambda_h(u_h, w_h) = T_h(u_h, w_h).$$

To prove convergence we will apply a theorem due to Osborn [211, Theorem 3] (stated here in terms of Hilbert spaces rather than Banach spaces as in Osborn's paper).

Let X denote a complex Hilbert space with $S : X \to X$ a compact operator. For a non-zero eigenvalue λ of S with algebraic multiplicity m, let Γ be a circle centered at λ containing no other eigenvalues. Recall the spectral projection

$$E = \frac{1}{2\pi i} \int_\Gamma (z - S)^{-1} \,\mathrm{d}z$$

and $R(E)$ the range of E (the dimension of $R(E)$ is m). Similarly, let $R(E^*)$ denote the range of the spectral projection E^* for the Hilbert adjoint S^* of S where now the eigenvalue is $\overline{\lambda}$.

Let $T_h : X \to X$ denote a sequence of compact operators for $h > 0$ (in fact constructed by finite elements). Osborn [211, Theorem 2] gives conditions under which the eigenvalues of S_h converge to those of S. If λ is an eigenvalue of S with multiplicity m, suppose $\lambda_{h,1}, \cdots, \lambda_{h,m}$ converge to λ then define

$$\hat{\lambda}_h = \frac{1}{m} \sum_{j=1}^{m} \lambda_{h,j}.$$

Theorem 6.4.1. *(Theorem 3 of [211]) Suppose $S_h \to S$ in norm and $S_h^* \to S^*$ in* *norm. Let* $\phi_1, \cdots, \phi_m$ *be a basis for* $R(E)$ *and let* $\phi_1^*, \cdots, \phi_m^*$ *be the dual basis. Then there is a constant* C *such that*

$$|\lambda - \hat{\lambda}_h| \leq \frac{1}{m}\sum_{j=1}^{m} |[(S - S_h)\phi_j, \phi_j^*]| + C\|(S - S_h)|_{R(E)}\| \, \|(S^* - S_h^*)|_{R(E^*)}\|,$$

where $[(S - S_h)\phi_j, \phi_j^*]$ *denotes the Hilbert space duality pairing.*

Remark 6.4.2. *This is actually Theorem 1.4.5 for Hilbert spaces.*

The following lemma states that the norm convergence of T_h to T and T_h^* to T^*:

Lemma 6.4.2. *Under the standing conditions on the domain and finite element spaces and provided* n *is smooth and* $n - 1 > 0$ *in* Ω*,* $T_h \to T$ *as* $h \to 0$ *in norm. In particular*

$$\|T - T_h\|_{\mathcal{L}(H^2(\Omega)\times H^1(\Omega), H^2(\Omega)\times H^1(\Omega))} \leq Ch^{\min(\alpha, 2s)},$$

where $\min(\alpha, 2s) > 0$ *and depends on the interior angles of the Lipschitz polygon as described in the proof. Similarly* $T_h^* \to T^*$ *in norm, and the same estimate holds for* $\|T^* - T_h^*\|_{\mathcal{L}(H^2(\Omega)\times H^1(\Omega), H^2(\Omega)\times H^1(\Omega))}$*.*

Remark 6.4.3. *If the domain is convex, we have at least first order convergence.*

Proof. It is clear that we have Galerkin orthogonality:

$$A((T - T_h)(u, w), (v_h, z_h)) = 0 \quad \text{for all } (v_h, z_h) \in X_h \times Y_h.$$

Then as usual

$$A((T - T_h)(u, w), (T - T_h)(u, w)) = A((T - T_h)(u, w), T(u, w) - (v_h, z_h))$$

for any $v_h, z_h \in X_h \times Y_h$. Hence

$$\|(T - T_h)(u, w)\|_{H^2(\Omega)\times H^1(\Omega)} \leq \|T(u, w) - (v_h, z_h))\|_{H^2(\Omega)\times H^1(\Omega)}. \tag{6.47}$$

We can now complete the estimate using the regularity of u and v and standard finite element error estimates. First let $T(u, w) = (k_1, k_2) \in H_0^2(\Omega) \times H_0^1(\Omega)$. Then $k_2 \in H_0^1(\Omega)$ satisfies

$$(\nabla k_2, \nabla z) = (nu, z)_{n-1}.$$

Since $n/(n - 1) \in L^\infty(\Omega)$ and Ω is a Lipschitz polygon, there is an $\alpha_0 > 0$ such that

$$\|k_2\|_{H^{1+\alpha}(\Omega)} \leq C\|nu/(n - 1)\|_{H^{-1+\alpha}(\Omega)},$$

where $\alpha_0 > \alpha \geq 1/2$ and where α_0 depends on the interior angles of the polygon. In particular, $\alpha_0 > 1/2$ and if the domain is convex $\alpha_0 = 1$ [139]. Choosing $z_h = P_{1,h}k_2$ where $P_{1,h}$ is the $H_0^1(\Omega)$ projection into Y_h we have

$$\begin{aligned} \|k_2 - z_h\|_{H^1(\Omega)} &\leq Ch^\alpha \|k_2\|_{H^{1+\alpha}(\Omega)} \\ &\leq Ch^\alpha \|nu/(n - 1)\|_{H^{-1+\alpha}(\Omega)} \\ &\leq Ch^\alpha \|u\| \end{aligned} \tag{6.48}$$

for $1/2 < \alpha < \min(\alpha_0, 1)$, provided Y_h contains polynomials of degree at least one (which must hold since Y_h is H^1-conforming).

Now $k_1 \in H_0^2(\Omega)$ satisfies

$$(\Delta k_1, \Delta v)_{n-1} = (u, \Delta v)_{n-1} + (\Delta u, nv)_{n-1} - (\nabla w, \nabla v)$$

for all $v \in H_0^2(\Omega)$.

In strong form $k_1 \in H_0^2(\Omega)$ satisfies

$$\Delta\left(\frac{1}{n-1}\Delta k_1\right) = \Delta\left(\frac{u}{n-1}\right) + \frac{n}{n-1}\Delta u + \Delta w := F.$$

If n is smooth, the right-hand side is in $H^{-1}(\Omega)$. Furthermore,

$$\|k_1\|_{H^{2+2s}(\Omega)} \le C\|F\|_{H^{-2+2s}(\Omega)} \le C\|F\|_{H^{-1}(\Omega)}$$

for $0 < s < \min(1/2, s_0/2)$. Here $s_0 > 0$ is the regularity limit given by [24, Section 4]. If Ω is convex, $s = 1/2$. So $k_1 \in H^{2+2s}(\Omega)$ where s depends on the interior angles of the domain.

Choosing $v_h = P_{2,h}k_1$ where $P_{2,h}$ is the $H^2(\Omega)$ projection into X_h we have

$$\begin{aligned} \|k_1 - P_{2,h}k_1\|_{H^2(\Omega)} &\le Ch^{2s}\|k_1\|_{H^{2+2s}(\Omega)} \le Ch^{2s}\|F\|_{H^{-1}(\Omega)} \\ &\le Ch^{2s}\left(\|u\|_{H^2(\Omega)} + \|w\|_{H^{1+\alpha}(\Omega)}\right). \end{aligned} \tag{6.49}$$

Putting together the estimates from (6.48) and (6.49) we have proved that

$$\begin{aligned} &\inf_{v_h, z_h \in X_h \times Y_h} \|T(u, w) - (v_h, z_h))\|_{H^2(\Omega)\times H^1(\Omega)} \\ &\le Ch^{\min(\alpha, 2s)}\left((\|u\|_{H^2(\Omega)} + \|w\|_{H^{1+\alpha}(\Omega)}\right). \end{aligned}$$

Using this in (6.47) proves the first estimate of the lemma.

Now consider the adjoint operator

$$T^* : H_0^2(\Omega) \times H_0^1(\Omega) \to H_0^2(\Omega) \times H_0^1(\Omega).$$

For $(v, z) \in H_0^2(\Omega) \times H_0^1(\Omega)$, it is defined by

$$A((u, w), T^*(v, z)) = b_1(u, v) + b_2(w, v) + c(w, z)$$

for all $(u, w)) \in H_0^2(\Omega) \times H_0^1(\Omega)$. Letting $T^*(v, z) = (t_1^*, t_2^*)$, the strong form of this equation is

$$\Delta\left(\frac{1}{n-1}\Delta t_1^*\right) = \frac{1}{n-1}\Delta v + \Delta\frac{n}{n-1}v + \frac{n}{n-1}z := G, \tag{6.50}$$

$$\Delta t_2^* = \Delta v. \tag{6.51}$$

In the same way as before, since $v \in H^2(\Omega)$, we have that $t_2^* \in H^{1+\alpha}(\Omega)$ and so choosing $z_h = P_{1,h}t_2^*$ gives

$$\begin{aligned} \|t_2^* - z_h\|_{H^1(\Omega)} &\le Ch^{\alpha}\|t_2^*\|_{H^{1+\alpha}(\Omega)} \\ &\le Ch^{\alpha}\|v\|_{H^2(\Omega)}. \end{aligned}$$

In addition, since $n/(n-1)$ is smooth, the right-hand side of (6.50) has the regularity $G \in L^2(\Omega)$, and again

$$\begin{aligned} \|t_1^* - P_{2,h}t_1^*\|_{H^2(\Omega)} &\le Ch^{2s}\|t_1^*\|_{H^{2+2s}(\Omega)} \\ &\le Ch^{2s}\|G\| \\ &\le Ch^{2s}\left(\|v\|_{H^2(\Omega)} + \|z\|\right). \end{aligned}$$

The proof is now complete. □

Theorem 6.4.3. *Under the assumptions of Lemma 6.4.2,*

$$|\lambda - \hat{\lambda}_h| = O(h^{2\min(\alpha,2s)}),$$

where α and s are the exponents in Lemma 6.4.2.

Remark 6.4.4. *From [24, Figure 1] we expect that s can be chosen so that $s > 1/2$ so the theorem predicts at least $O(h)$ convergence for the eigenvalues. If the domain is convex we predict quadratic convergence.*

Proof. Suppose we have m eigenfunctions

$$T(u_j, v_j) = \lambda(u_j, v_j)$$

together with a dual basis for $R(E)$ denoted $(u_j^*, v_j^*) \in H_0^2(\Omega) \times H_0^1(\Omega)$ such that

$$A((u_j, v_j), (u_\ell^*, v_\ell^*)) = \delta_{j,\ell}.$$

We apply Theorem 6.4.1 using $\phi = (u, v) \in H_0^2(\Omega) \times H_0^1(\Omega)$ and $S\phi = T(u, v)$ (similarly for T^*). By Lemma 6.4.2 we have the norm convergence of the operators. It remains to estimate the term $[(S - S_h)\phi_j, \phi_j^*]$. In our case

$$\begin{aligned} [(S - S_h)\phi_j, \phi_j^*] &= A((T - T_h)(u_j, v_j), (u_j^*, v_j^*)) \\ &= A((T - T_h)(u_j, v_j), T^*(u_j^*, v_j^*)). \end{aligned}$$

By Galerkin orthogonality this implies that

$$[(S - S_h)\phi_j, \phi_j^*] = A((T - T_h)(u_j, v_j), (T^* - T_h^*)(u_j^*, v_j^*)).$$

Using the error estimate from Lemma 6.4.2 completes the proof. □

6.4.3 Numerical Examples

Now we show some simple examples. Let V_h be the finite element space generated by the Argyris elements on a regular triangular mesh of Ω. Let $X_h \subset V_h \cap H_0^2(\Omega)$. We choose Y_h to be the standard continuous piecewise linear Lagrange element such that $Y_h \subset H_0^1(\Omega)$. Let $\{\phi_i\}_{i=1}^{N_h}$ be the basis for X_h and $\{\psi_i\}_{i=1}^{M_h}$ be the basis for Y_h. We define the following matrices

$$\begin{aligned}
A_{ij} &= (\triangle\phi_j, \triangle\phi_i)_{n-1},\\
S^1_{ij} &= (\triangle\phi_j, \phi_i)_{n-1},\\
S^2_{ij} &= (n\phi_j, \triangle\phi_i),\\
S_{ij} &= (\nabla\psi_j, \nabla\phi_i),\\
S'_{ij} &= (\nabla\psi_j, \nabla\psi_i),\\
M_{ij} &= (\psi_j, n\phi_i)_{n-1},
\end{aligned}$$

where $\mathbf{u} = (u_1, \ldots, u_{N_h})^T$ such that $u_h = \sum_{i=1}^{N_h} u_i\phi_i$ and $\mathbf{w} = (w_1, \ldots, w_{M_h})^T$ such that $w_h = \sum_{i=1}^{M_h} w_i\psi_i$. The matrix eigenvalue problem is given by

$$\mathcal{A}\mathbf{x} = \tau\mathcal{B}\mathbf{x},$$

where

$$\begin{aligned}
\mathcal{A} &= \begin{pmatrix} A & 0 \\ 0 & S' \end{pmatrix},\\
\mathcal{B} &= -\begin{pmatrix} S^1 + S^2 & -S \\ M & 0 \end{pmatrix},\\
\mathbf{x} &= \begin{pmatrix} \mathbf{u} \\ \mathbf{w} \end{pmatrix}.
\end{aligned}$$

We choose three test domains: the unit square, an L-shaped domain, and the disk with radius $1/2$ centered at the origin. The unit square and the L-shaped domain are given by

$$(-1/2, 1/2) \times (-1/2, 1/2)$$

and

$$(-1/2, 1/2) \times (-1/2, 1/2) \setminus ([0, 1/2] \times [-1/2, 0]),$$

respectively.

For simplicity, we choose the index of refraction $n(x) = 16$ since we can then compare the results computed here to those in the previous sections (see also [95, 228]). For each domain we generate a coarse triangular mesh and then uniformly refine the mesh to perform a convergence study. In the case of the circle each refinement gives a better polygonal approximation of the curved boundary. So we do not use curved elements for the circular domain, and this may have a major effect on the convergence rates in that case.

The computed transmission eigenvalues are shown in Table 6.5. They are consistent with the values in [95, 228].

In Fig. 6.5, we plot the relative error for the first real transmission eigenvalue against the mesh size h when the linear Lagrange element is used to discretize $H_0^1(\Omega)$. For the circle we can compute the true relative error using precise estimates

Shape	Base mesh	1 refinement	2 refinements	3 refinements
unit square	1.877313	1.879039	1.879455	1.879557
Number of DoFs	1587	6407	25767	103367
L-shaped	2.971278	2.964095	2.958426	2.955279
Number of DoFs	1187	4807	19367	77767
circle	1.989962	1.988407	1.988088	1.988017
Number of DoFs	1245	5023	20199	81031

Table 6.5: The first (real) transmission eigenvalues for the test domains on a series of uniformly refined meshes. The index of refraction is $n = 16$. DoFs refer to the total number of degree of freedoms ($M_h + N_h$).

of the transmission eigenvalue computed via special functions (see Table 6.1). For the other domains we compare the difference between the eigenvalues on successive meshes. The results indicate convergence rates for each domain. The convergence orders for the unit square and the circle are 2. The convergence order for the L-shaped domain is less than $1/2$. This is to be expected since even for smooth eigenfunctions the order of convergence is limited by the piecewise linear space. An interesting observation is that the eigenvalues converge from below for the unit square while from above for the L-shaped domain and the circle.

Using the linear Lagrange element for $H_0^1(\Omega)$ and the Argyris element for the biharmonic terms limits the maximum possible convergence rate to that of the lower order space. In Fig. 6.6 we show results using piecewise quadratic elements to discretize $H_0^1(\Omega)$. As expected, the convergence rate for the first eigenvalue on the L-shaped domain does not change compared to that in Fig. 6.5 because the eigenfunction is singular near the reentrant corner. For the square the convergence rate is now fourth order since the first eigenfunction is smooth. The convergence rate for eigenfunctions on the circular domain does not increase to fourth order despite the fact that the eigenfunctions are smooth. This is likely because we approximate the circular domain with a mesh of triangles so there is a geometric error that pollutes the eigenvalue calculation. This example suggests that using a higher order space to discretize $H_0^1(\Omega)$ improves the convergence rate for smooth eigenfunctions.

6.5 A Mixed Method using Lagrange Elements

The previous mixed method uses the Argyris element which is H^2-conforming. In this section, we present a simpler mixed finite element using Lagrange elements due to Ji, Sun, and Turner [161]. The formulation is similar to the mixed method for the biharmonic equation in Section 4.3 (see also [89]).

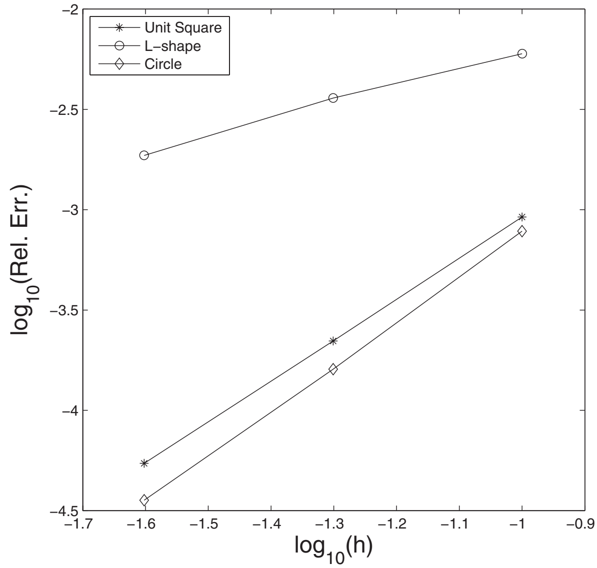


Figure 6.5: Convergence rate of the first real transmission eigenvalue using piecewise linear elements to discretize $H_0^1(\Omega)$. As expected the convergence rate for the circle and square is second order, while for the L-shaped domain it is lower.

6.5.1 Another Mixed Formulation

Let Ω be a convex Lipschitz domain. For simplicity, we assume that $n(x) - 1 \geq \delta > 0$ on Ω.

The starting point is the fourth order equation of the transmission eigenvalue problem as well. We recall the weak form of the fourth problem of finding $(k^2, u) \in \mathbb{C} \times H_0^2(\Omega)$ such that

$$\left(\frac{1}{n-1}(\triangle + k^2 n)u, (\triangle + k^2)\phi\right) = 0 \quad \text{for all } \phi \in H_0^2(\Omega). \tag{6.52}$$

Let

$$v = \frac{1}{n-1}(\Delta + k^2 n)u.$$

We have

$$\begin{aligned} (\Delta + k^2)v &= 0, \\ \frac{1}{n-1}(\Delta + k^2 n)u &= v. \end{aligned}$$

Following the mixed method approach [89], we obtain the following weak problem.

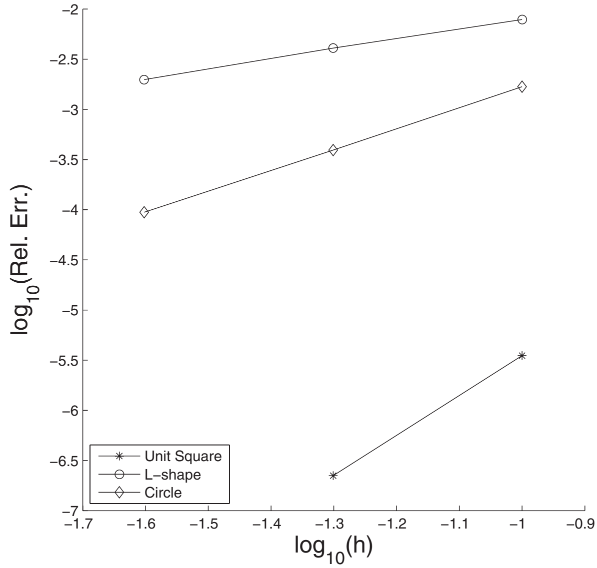


Figure 6.6: Convergence rate of the first real transmission eigenvalue using piecewise quadratic elements to discretize $H_0^1(\Omega)$. Compared with Fig. 6.5 the convergence rate for the L-shaped domain is unchanged reflecting the low regularity of the eigenfunction in that case. For the square domain the convergence rate increases to $O(h^4)$. For the circle a corresponding increase in the convergence rate is not seen (see the text for more discussion).

Find $(k^2, u, v) \in \mathbb{C} \times H_0^1(\Omega) \times H^1(\Omega)$ such that

$$\begin{aligned} (\nabla v, \nabla \phi) &= k^2(v, \phi) \quad \text{for all } \phi \in H_0^1(\Omega), \\ (\nabla u, \nabla \varphi) + ((n-1)v, \varphi) &= k^2(nu, \varphi) \quad \text{for all } \varphi \in H^1(\Omega). \end{aligned}$$

Given finite dimensional spaces $S_h \subset H^1(\Omega)$ and $S_h^0 \subset H_0^1(\Omega)$ such that $S_h^0 \subset S_h$, the discrete problem is to find $(k_h^2, u_h, v_h) \in \mathbb{C} \times S_h^0 \times S_h$ such that

$$\begin{aligned} (\nabla v_h, \nabla \phi_h) &= k_h^2(v_h, \phi_h) \quad \text{for all } \phi_h \in S_h^0, \\ (\nabla u_h, \nabla \varphi_h) + ((n-1)v_h, \varphi_h) &= k_h^2(nu_h, \varphi_h) \quad \text{for all } \varphi_h \in S_h. \end{aligned}$$

Matrix	Dimension	Definition
$S_{K\times T}$	$K\times T$	$S_{K\times T}^{i,j}=(\nabla\psi_i,\nabla\psi_j), 1\le i\le K, 1\le j\le T$
$S_{T\times K}$	$T\times K$	$S_{T\times K}^{i,j}=(\nabla\psi_i,\nabla\psi_j), 1\le i\le T, 1\le j\le K$
$M_{K\times T}$	$K\times T$	$M_{K\times T}^{i,j}=(\psi_i,\psi_j), 1\le i\le K, 1\le j\le T$
$M_{T\times K}^{n}$	$T\times K$	$(M_{T\times K}^{n})^{i,j}=(n\psi_i,\psi_j), 1\le i\le T, 1\le j\le K$
$M_{T\times T}^{n-1}$	$T\times T$	$(M_{T\times T}^{n-1})^{i,j}=((n-1)\psi_i,\psi_j), 1\le i\le T, 1\le j\le T$

Table 6.6: Definition of various matrices for the mixed method using the linear Lagrange element.

6.5.2 The Discrete Problem

We use standard piecewise linear finite elements to discretize the problem. Let

$$\begin{aligned} S_h &= \text{the space of continuous piecewise linear finite elements on } \Omega, \\ S_h^0 &= S_h \cap H_0^1(\Omega) \\ &= \text{the subspace of functions in } S_h \text{ that have vanishing DoF on } \partial\Omega, \end{aligned}$$

where DoF stands for degree of freedom. Let $\psi_1,\dots,\psi_K$ be a basis for S_h^0 and

$$\psi_1,\dots,\psi_K,\psi_{K+1},\dots,\psi_T$$

be a basis for S_h. Let $u_h=\sum_{i=1}^K u_i\psi_i$ and $v_h=\sum_{i=1}^T v_i\psi_i$. Furthermore, let $\mathbf{u}=(u_1,\dots,u_K)^T$ and $\mathbf{v}=(v_1,\dots,v_T)^T$.

The matrix problem corresponding to the above problem is

$$\begin{aligned} S_{K\times T}\mathbf{v} &= k_h^2 M_{K\times T}\mathbf{v}, \\ S_{T\times K}\mathbf{u}+M_{T\times T}^{n-1}\mathbf{v} &= k_h^2 M_{T\times K}^{n}\mathbf{u}, \end{aligned}$$

where the matrices are defined in Table 6.6.

For convenience, we write the generalized eigenvalue problem as

$$\begin{pmatrix} S_{K\times T} & 0_{K\times K} \\ M_{T\times T}^{n-1} & S_{T\times K} \end{pmatrix}\begin{pmatrix} \mathbf{v} \\ \mathbf{u} \end{pmatrix} = k_h^2 \begin{pmatrix} M_{K\times T} & 0_{K\times K} \\ 0_{T\times T} & M_{T\times K}^{n} \end{pmatrix}\begin{pmatrix} \mathbf{v} \\ \mathbf{u} \end{pmatrix}. \tag{6.53}$$

In contrast to some mixed methods for the biharmonic eigenvalue problems and Dirichlet eigenvalue problem which needs the inversion of a certain matrix, here we have the general eigenvalue problem directly. This is certainly an advantage thanks to the property of the original problem.

For simplicity, we rewrite the above problem as

$$A\mathbf{x}=\lambda B\mathbf{x} \tag{6.54}$$

where

$$A = \begin{pmatrix} S_{K\times T} & 0_{K\times K} \\ M^{n-1}_{T\times T} & S_{T\times K} \end{pmatrix},$$
$$B = \begin{pmatrix} M_{K\times T} & 0_{K\times K} \\ 0_{T\times T} & M^{n}_{T\times K} \end{pmatrix},$$
$$\mathbf{x} = \begin{pmatrix} \mathbf{v} \\ \mathbf{u} \end{pmatrix}.$$

The generalized eigenvalue problem obtained above is large, sparse, and non-Hermitian. Use of a direct method is prohibitively expensive even on a rather coarse mesh [95]. If we only need a few smallest real transmission eigenvalues in inverse scattering theory, iterative methods are the obvious choice. For this purpose we will devise an adaptive algorithm using the Arnoldi method [132, 220] to compute the transmission eigenvalues. This choice was influenced by the fact that Matlab has an implemented Arnoldi solver named 'sptarn' which could be integrated easily into the finite element code.

Mablab command *sptarn* uses Arnoldi iteration (see Section 9.3) with spectral transformation. To guarantee efficiency, we need to specify a small search interval, i.e., to estimate accurately an interval containing the desired transmission eigenvalues. Using the Faber-Krahn type inequality (6.14), we have a lower bound for transmission eigenvalues as long as we have the first Dirichlet eigenvalue. In fact, this can be done easily since we have the necessary matrices for the mixed finite element already. The discrete Dirichlet eigenvalue problem is simply the following generalized eigenvalue problem

$$S_{K\times K}\mathbf{x} = \lambda M_{K\times K}\mathbf{x}, \tag{6.55}$$

where $S_{K\times K}$ and $M_{K\times K}$ are the stiffness matrix and the mass matrix, respectively.

Since *sptarn* might compute complex transmission eigenvalues, we need to exclude them as well. Assuming a triangular mesh $\mathcal{T}$ is already generated for D, the following adaptive algorithm computes several smallest transmission eigenvalues efficiently. The algorithm is implemented using Matlab. The inputs are a triangular mesh $\mathcal{T}$ for domain D, the supremum of the index of refraction $n(x)$, and the number of transmission eigenvalues to be computed. The outputs are the desired transmission eigenvalues.

Matrices A and B are then sent to the adaptive Arnoldi method to search for the required transmission eigenvalues. This is done by first computing the left bound lb of an interval using (6.14) and setting the right bound of the search interval $rb = lb + 1$. It is necessary to keep the interval small since a larger interval might contain many transmission eigenvalues and it would keep *sptarn* searching forever. In fact, the distribution of real transmission eigenvalue is quite complicated [95]. In our algorithm the search interval is moved to the right by one unit until all desired transmission eigenvalues are found.

Adaptive Mixed FEM:

Input:

- a regular triangular mesh for Ω
- the index of refraction $n(x)$ and $n^*(\sup_\Omega(n(x)))$
- the number of transmission eigenvalues noe to be computed

Output:

- noe smallest transmission eigenvalues

1. construct matrices S, M, M_n
2. construct matrices A, B from S, M, M_n
3. compute λ_0 from S and M
4. set $TE = \emptyset, lb = \frac{\lambda_0}{\sup_D n}, rb = lb + 1$
5. while length(TE) < noe
   - $it = it + 1$
   - $[V, D] = sptarn(A, B, lb, rb)$
   - delete complex values in D
   - $TE = TE \cup D$
   - $lb = rb, rb = lb + it + 1$

6.5.3 Numerical Examples

Now we provide some numerical examples to show the effectiveness of our algorithm. We first consider the case when the index of refraction is constant; here we choose $n(x) = 16$. We choose two geometries for Ω: a disk centered at $(0, 0)$ with radius $1/2$ and a unit square given by $[-1/2, 1/2] \times [-1/2, 1/2]$.

domain	index of refraction n	1st	2nd	3rd	4th
disk ($r = 1/2$)	16	1.9986	2.6334	2.6343	3.2641
unit square	16	1.8873	2.4596	2.4599	2.8928

Table 6.7: Computed transmission eigenvalues by the mixed method using the linear Lagrange element.

The computed transmission eigenvalues, using a quasi-uniform triangular mesh $\mathcal{T}$ for Ω with $h \approx 0.05$, are shown in Table 6.7. These are consistent with the values given by Colton et al. [95].

In Fig. 6.7, we show the eigenfunctions associated with the first and second (real) transmission eigenvalues for the disk ($n = 16$). In Fig. 6.8, we show the eigenfunctions associated with the first and second (real) transmission eigenvalues for the unit

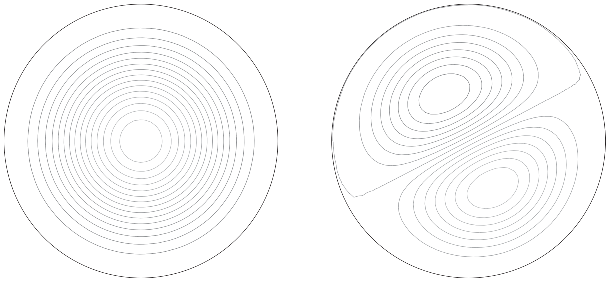


Figure 6.7: The eigenfunctions associated with the first and second (real) transmission eigenvalues for the disk ($n = 16$). Left: the first eigenfunction. Right: the second eigenfunction.

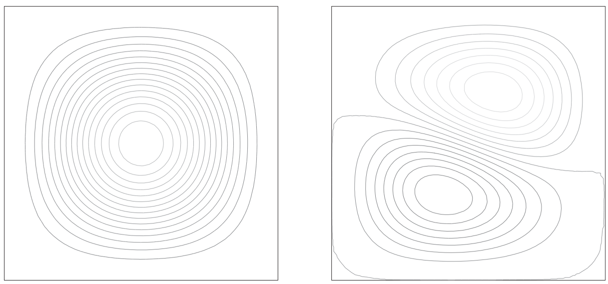


Figure 6.8: The eigenfunctions associated with the first and second (real) transmission eigenvalues for the unit square ($n = 16$). Left: the first eigenfunction. Right: the second eigenfunction.

square ($n = 16$). Note that since we used the fourth order formulation (6.52), the plots are the differences of w and v in (6.2).

Next we check the convergence numerically. We start with a quasi-uniform triangular mesh $\mathcal{T}$ for Ω with $h \approx 0.1$. Then we uniformly refine the mesh a couple of times. In Figure 6.9, we plot the convergence of the relative error of the smallest transmission eigenvalue, which is defined as

$$\text{Rel. Err.} = \frac{|\lambda_1(h) - \lambda_1(\frac{h}{2})|}{\lambda_1(\frac{h}{2})}. \tag{6.56}$$

Here $\lambda_1(h)$ is the smallest transmission eigenvalue with mesh size h. It is clearly we obtain a second order convergence.

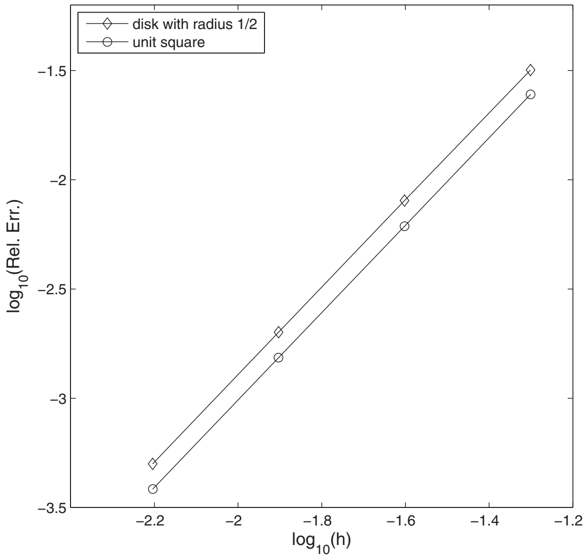


Figure 6.9: The plot of $\log_{10}(\text{Rel. Err.})$ against $\log_{10}(h)$ for the smallest transmission eigenvalue.

domain	index of refraction n	1st	2nd	3rd	4th
disk ($r = 1/2$)	$8 + 4\|x\|$	2.7770	3.5571	3.5584	4.3605
unit square	$8 + x_1 - x_2$	2.8373	3.5632	3.5642	4.1582

Table 6.8: Computed transmission eigenvalues for non-constant indices of refraction.

Finally, we compute the transmission eigenvalues when the index of refraction is a function. We set $n(x) = 8 + 4|x|$ for the disk and $8 + x_1 - x_2$ for the unit square. Several smallest transmission eigenvalues are shown in Table 6.8. The computed values are consistent with the results in [228] and [95]. In particular, the smallest transmission eigenvalues are consistent with the values in [227], which are computed from the near field data using an inverse scattering algorithm.

6.6 The Maxwell's Transmission Eigenvalues

We studied the interior transmission eigenvalues associated with the Helmholtz in previous sections. The goal of this section is to disscus finite element methods for transmission eigenvalues for the vector case, i.e., Maxwell's transmission eigenvalues (MTEs). In particular, we present two finite element methods proposed in Monk and Sun [203]. The first approach is a curl-conforming finite element method based on a formulation first given by Kirsch [171] (see also [95] for a similar derivation for the Helmholtz transmission eigenvalue problem). The second approach is a mixed finite element method for the fourth order reformulation of the transmission eigenvalue problem. Both methods enjoy simplicity and easy implementation using the edge elements. These methods are strongly related. The resulting non-Hermitian matrix eigenvalue problem is then computed by an adaptive Arnoldi method.

We first introduce the transmission eigenvalue problem for the Maxwell's equations and recall some existence results in the literature. Then we derive the Maxwell's transmission eigenvalues for balls with constant index of refraction. This serves as a benchmark problem and is used to verify the computational results. We present two finite element methods, i.e., the curl-conforming finite element method and the mixed finite element method. To solve the fully discretized matrix eigenvalue problem, an adaptive Arnoldi method similar to that in Section 4.3 is employed. The method employs a Faber-Krahn type inequality for the Maxwell's transmission eigenvalues and adaptively updates search intervals as in the previous section. Various numerical examples are provided to show the effectiveness of the proposed methods and test the tightness of several inequalities for transmission eigenvalues.

Let $\Omega \subset \mathbb{R}^3$ be a bounded Lipschitz polyhedron. Let ν be the unit outward normal to $\partial\Omega$. We recall the Hilbert space $H(\text{curl}, \Omega)$ defined in Section 5.1 as

$$H(\text{curl}; \Omega) = \{u \in L^2(\Omega)^3 : \nabla \times u \in L^2(\Omega)^3\}$$

and

$$H_0(\text{curl}; \Omega) = \{u \in H(\text{curl}, \Omega) : \nu \times u = 0 \text{ on } \partial\Omega\}.$$

We need two more spaces

$$\begin{aligned}\mathcal{U}(\Omega) &= \{u \in H(\text{curl}; \Omega) : \nabla \times u \in H(\text{curl}; \Omega)\},\\ \mathcal{U}_0(\Omega) &= \{u \in H_0(\text{curl}; \Omega) : \nabla \times u \in H_0(\text{curl}; \Omega)\}.\end{aligned}$$

Let N be a 3×3 matrix-valued index of refraction defined on Ω such that $N \in L^\infty(\Omega, \mathbb{R}^{3\times 3})$.

Definition 6.6.1. *A real matrix field N is said to be bounded positive definite on Ω if $N \in L^\infty(\Omega, \mathbb{R}^{3\times 3})$ and there exists a constant $\gamma > 0$ such that*

$$\bar{\xi} \cdot N\xi \geq \gamma|\xi|^2, \quad \textit{for all } \xi \in \mathbb{C}^3 \textit{ a.e. in } \Omega.$$

We assume that N, N^{-1} and either $(N-I)^{-1}$ or $(I-N)^{-1}$ are bounded positive-definite real matrix fields on Ω.

We consider the time-harmonic electromagnetic incident plane wave given by

$$E^i(x,d,p) = \frac{i}{k}\,\mathrm{curl}\,\mathrm{curl}\, p\, e^{ikx\cdot d}$$

and

$$H^i(x,d,p) = \mathrm{curl}\, p\, e^{ikx\cdot d},$$

where $d \in \mathbb{R}^3$ is a unit vector giving the direction of propagation of the wave, and the vector p is called the polarization.

The scattering by an anisotropic medium leads to the following problem for the interior electric and magnetic fields E, H and the scattered electric and magnetic fields E^s, H^s satisfying

$$\mathrm{curl}\, E^s - ikH^s = 0, \qquad \text{in } \mathbb{R}^3 \setminus \Omega, \tag{6.57a}$$

$$\mathrm{curl}\, H^s + ikE^s = 0, \qquad \text{in } \mathbb{R}^3 \setminus \Omega, \tag{6.57b}$$

$$\mathrm{curl}\, E - ikH = 0, \qquad \text{in } \Omega, \tag{6.57c}$$

$$\mathrm{curl}\, H + ikN(x)H = 0, \qquad \text{in } \Omega, \tag{6.57d}$$

$$\nu \times (E^s + E^i) - \nu \times E = 0, \qquad \text{on } \partial\Omega, \tag{6.57e}$$

$$\nu \times (H^s + H^i) - \nu \times H = 0, \qquad \text{on } \partial\Omega, \tag{6.57f}$$

and the Silver-Müller radiation condition

$$\lim_{r\to\infty} (H^s \times x - rE^s) = 0, \tag{6.58}$$

where $r = |x|$ and k is the wave number. Under suitable conditions on N and Ω, the well-posedness of the above problem is known (Theorem 4.2 of [61]) and the scattered fields have the following asymptotic behavior

$$E^s(x,d,p) = \frac{e^{ikr}}{r} E_\infty(\hat{x},d,p) + O\left(\frac{1}{r^2}\right), \qquad r \to \infty, \tag{6.59a}$$

$$H^s(x,d,p) = \frac{e^{ikr}}{r} \hat{x} \times E_\infty(\hat{x},d,p) + O\left(\frac{1}{r^2}\right), \qquad r \to \infty, \tag{6.59b}$$

where $\hat{x} = x/r$ and E_∞ is the electric far field pattern [93]. Given E_∞, one can define the far field operator $F : L^2_t(\mathbb{S}) \to L^2_t(\mathbb{S})$ by

$$(Fg)(\hat{x}) := \int_\Omega E_\infty(\hat{x}, d, g(d))\, \mathrm{d}s, \tag{6.60}$$

where $\mathbb{S} = \{\hat{x} \in \mathbb{R}^3; |\hat{x}| = 1\}$ and

$$L^2_t(\mathbb{S}) := \left\{u \in L^2(\mathbb{S})^3 : \nu \cdot u = 0 \quad \text{on } \mathbb{S}\right\}.$$

The far field operator F has fundamental importance in the study of qualitative methods, for example, the linear sampling method (see Section 3.3 of [61]). For the case of anisotropic media, F has dense range provided k is not a transmission eigenvalue which we define next. We refer the readers to [61, 96, 93] for the mathematical derivation and interpretation of the above scattering problem.

In terms of electric fields, the transmission eigenvalue problem for the anisotropic Maxwell's equations can be formulated as the following.

Definition 6.6.2. *A value of $k^2 \neq 0$ is called a transmission eigenvalue if there exist real-valued fields $E, E_0 \in L^2(\Omega)^3$ with $E - E_0 \in \mathcal{U}_0(\Omega)$ such that*

$$\nabla \times \nabla \times E - k^2 N E = 0, \qquad \textit{in } \Omega, \tag{6.61a}$$

$$\nabla \times \nabla \times E_0 - k^2 E_0 = 0, \qquad \textit{in } \Omega, \tag{6.61b}$$

$$\nu \times E = \nu \times E_0, \qquad \textit{on } \partial\Omega, \tag{6.61c}$$

$$\nu \times \nabla \times E = \nu \times \nabla \times E_0, \qquad \textit{on } \partial\Omega. \tag{6.61d}$$

Similar to the scalar case, the above problem can be rewritten as a fourth order problem. Let $u = E - E_0$ and $v = NE - E_0$. Then we have that

$$E = (N - I)^{-1}(v - u),$$
$$E_0 = (I - N)^{-1}(Nu - v).$$

Subtracting (6.61b) from (6.61a), we obtain

$$\nabla \times \nabla \times u = k^2 v,$$

and therefore

$$E = (N - I)^{-1}\left(\frac{1}{k^2}\nabla \times \nabla \times u - u\right). \tag{6.62}$$

Substituting for E in (6.61a) and taking the boundary conditions (6.61c) and (6.61d) into account, we end up with a fourth order differential equation for $u \in \mathcal{U}_0(\Omega)$ satisfying

$$(\nabla \times \nabla \times -k^2 N)(N - I)^{-1}(\nabla \times \nabla \times u - k^2 u) = 0. \tag{6.63}$$

Therefore the variational formulation for the transmission eigenvalue problem can be stated as follows. Find $k^2 \neq 0$ and $u \in \mathcal{U}_0(\Omega)$ such that

$$\left((N - I)^{-1}\left(\nabla \times \nabla \times -k^2 I\right) u, (\nabla \times \nabla \times -k^2 N)\phi\right) = 0 \tag{6.64}$$

for all $\phi \in \mathcal{U}_0(\Omega)$.

The following theorem shows that, under certain conditions on the index of refraction, there exists an infinite countable set of Maxwell's transmission eigenvalues.

Theorem 6.6.1. *(Theorem 2.10 of [64]) Assume that $N \in L^\infty(\Omega, \mathbb{R}^{3\times 3})$ satisfies either one of the following assumptions:*

$$\begin{aligned} &1) \qquad 1+\alpha \le n_* \le \bar{\xi}\cdot N(x)\xi \le n^* < \infty, \\ &2) \qquad 0 < n_* \le \bar{\xi}\cdot N(x)\xi \le n^* < 1-\beta, \end{aligned}$$

for every $\xi \in \mathbb{C}^3$ such that $|\xi| = 1$ and some constants $\alpha, \beta > 0$. Then there exists an infinite countable set of transmission eigenvalues with $+\infty$ as the only accumulation point.

Next we recall theorems which provide lower and upper bounds for transmission eigenvalues. We will test the efficiency of these inequalities later.

Theorem 6.6.2. *(Theorem 4.33 in [61]) Let $k_{1,\Omega,N(x)}$ be the first transmission eigenvalue and let α and β be positive constants. Denote by k_{1,Ω,n_*} and k_{1,D,n^*} the first transmission eigenvalue for $N = n_* I$ and $N = n^* I$, respectively.*

1. *If $\|N(x)\|_2 \ge \alpha > 1$, then $0 < k_{1,\Omega,n^*} \le k_{1,\Omega,N(x)} \le k_{1,\Omega,n_*}$ for all $x \in \Omega$.*
2. *If $0 < \|N(x)\|_2 \le 1-\beta$, then $0 < k_{1,\Omega,n_*} \le k_{1,\Omega,N(x)} \le k_{1,\Omega,n^*}$ for all $x \in \Omega$.*

The bounds in terms of transmission eigenvalues on balls are obtained in [64]. Let B_{r_1} be the largest ball of radius r_1 such that $B_{r_1} \subset \Omega$ and B_{r_2} the smallest ball of radius r_2 such that $\Omega \subset B_{r_2}$. We denote by k_{1,n_*} and k_{1,n^*} the first transmission eigenvalue for the unit ball with index of refraction n_* and n^*, respectively. For a given $0 < \epsilon \le r_1$ let $m(\epsilon) \in \mathbb{N}$ be the number of balls B_ϵ of radius ϵ that are contained in Ω. Then we have the following estimate.

Theorem 6.6.3. *(Corollary 2.11 in [64]) Assume that $N \in L^\infty(\Omega, \mathbb{R}^{d\times d})$, $d = 2, 3$, and let $k_{1,D,N(x)}$ be the first transmission eigenvalue.*

1. *If $1+\alpha \le n_* \le \bar{\xi}\cdot N(x)\xi \le n^* < \infty$ for every $\xi \in \mathbb{C}^d$ such that $\|\xi\| = 1$, and some constant $\alpha > 0$, then*
$$0 < \frac{k_{1,n^*}}{r_2} \le k_{1,\Omega,N(x)} \le \frac{k_{1,n_*}}{r_1}. \tag{6.65}$$
Furthermore, there exist at least $m(\epsilon)$ transmission eigenvalues in the interval $[k_{1,n^*}/r_2, k_{1,n_*}/\epsilon]$.
2. *If $0 < n_* \le \bar{\xi}\cdot N(x)\xi \le n^* < 1-\beta$ for every $\xi \in \mathbb{C}^d$ such that $\|\xi\| = 1$ and some constant $\beta > 0$, then*
$$0 < \frac{k_{1,n_*}}{r_2} \le k_{1,\Omega,N(x)} \le \frac{k_{1,n^*}}{r_1}. \tag{6.66}$$
Furthermore, there exist at least $m(\epsilon)$ transmission eigenvalues in the interval $[k_{1,n_*}/r_2, k_{1,n^*}/\epsilon]$.

Note that the above two theorems hold for arbitrarily small positive numbers α and β which is a rather mild requirement on the index of refraction. We refer the readers to [64, 67, 253] for more results on the existence of the transmission eigenvalues.

6.6.1 Transmission Eigenvalues of Balls

In this section we derive the transmission eigenvalues on balls with constant index of refraction. The eigenvalue problem on a ball has theoretical importance (see Theorem 6.6.3) and will also serve as a benchmark problem in Section 6.6.5.

We assume that the index of refraction $N = N_0 I$ where N_0 is a scalar constant. Let $u = j_n(k\rho)Y_n^m(\hat{x})$ and $v = j_n(k\sqrt{N_0}\rho)Y_n^m(\hat{x})$ where j_n is the spherical Bessel's function of order n and Y_n^m is the spherical harmonic (see, e.g., [93]). Here $\hat{x} = x/|x|$ and $\rho = |x|$. Note that u and v are solutions of the Helmholtz equation (see p. 235 of [202]). Then the following are solutions to the Maxwell's equations (6.61a) and (6.61b), respectively,

$$\tilde{M}_u = \nabla \times \{xu\}, \quad \tilde{N}_u = \frac{1}{ik}\nabla \times \{\tilde{M}_u\}, \quad n > 1,$$
$$\tilde{M}_v = \nabla \times \{xv\}, \quad \tilde{N}_v = \frac{1}{ik}\nabla \times \{\tilde{M}_v\}, \quad n > 1.$$

Using the curl in spherical coordinates (ρ, θ, ϕ), we have solutions for Maxwell's equations of TE (transverse electric) modes

$$\begin{aligned}\tilde{M}_u &= -\frac{\partial u}{\partial \theta}e_\phi + \frac{1}{\sin\theta}\frac{\partial u}{\partial \phi}e_\theta \\ &= -j_n(k\rho)\frac{\partial Y_n^m(\hat{x})}{\partial \theta}e_\phi + \frac{1}{\sin\theta}j_n(k\rho)\frac{\partial Y_n^m(\hat{\mathbf{x}})}{\partial \phi}e_\theta.\end{aligned}$$

Taking the curl of the above equation and dropping the constant $1/ik$, we have solutions of TM (transverse magnetic) modes

$$\begin{aligned}&(\nabla \times \tilde{M}_u)_\rho \\ = \;& \frac{1}{\rho\sin\theta}\left\{\frac{\partial}{\partial\theta}\left(\sin\theta\left[-j_n(k\rho)\frac{\partial Y_n^m(\hat{x})}{\partial\theta}\right]\right) - \frac{1}{\sin\theta}j_n(k\rho)\frac{\partial^2 Y_n^m(\hat{x})}{\partial\phi^2}\right\} \\ = \;& \frac{j_n(k\rho)}{\rho\sin\theta}\left\{-\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y_n^m(\hat{\mathbf{x}})}{\partial\theta}\right) - \frac{1}{\sin\theta}\frac{\partial^2 Y_n^m(\hat{x})}{\partial\phi^2}\right\}, \\ &(\nabla \times \tilde{M}_u)_\theta \\ = \;& \frac{1}{\rho}\left\{-\frac{\partial}{\partial\rho}\left(-\rho j_n(k\rho)\frac{\partial Y_n^m(\hat{x})}{\partial\theta}\right)\right\} \\ = \;& \frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho j_n(k\rho)\right)\frac{\partial Y_n^m(\hat{x})}{\partial\theta}\end{aligned}$$

and

$$(\nabla \times \tilde{M}_u)_\phi = \frac{1}{\rho\sin\theta}\frac{\partial}{\partial\rho}\left(\rho j_n(k\rho)\right)\frac{\partial Y_n^m(\hat{x})}{\partial\phi}.$$

Note that similar results hold for $\tilde{M}_v$.

For TE mode solutions, in order to satisfy the boundary conditions (6.61c) and (6.61d), the wave number k^2's need to satisfy

$$\begin{vmatrix} j_n(k\rho) & j_n(k\sqrt{N_0}\rho) \\ \frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho j_n(k\rho)\right) & \frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho j_n(k\sqrt{N_0}\rho)\right) \end{vmatrix} = 0, \quad n \geq 1. \tag{6.67}$$

The zeros of (6.67) provide the first group, i.e., TE modes, of the Maxwell's transmission eigenvalues. We refer the readers to Example 3.2 of [67] for a detailed derivation of (6.67).

Next we consider the TM modes. Let

$$E_u = \nabla \times \tilde{M}_u$$

and

$$E_v = \nabla \times \tilde{M}_v.$$

Simple calculation shows that

$$\begin{aligned}(\nabla \times E_u)_\rho &= 0,\\ (\nabla \times E_u)_\theta &= \frac{k^2}{\sin\theta} j_n(k\rho)\frac{\partial Y_n^m(\hat{x})}{\partial \phi},\\ (\nabla \times E_u)_\phi &= k^2 j_n(k\rho)\frac{\partial Y_n^m(\hat{x})}{\partial \theta},\end{aligned}$$

and

$$\begin{aligned}(\nabla \times E_v)_\rho &= 0,\\ (\nabla \times E_v)_\theta &= \frac{k^2 N_0}{\sin\theta} j_n(k\sqrt{N_0}\rho)\frac{\partial Y_n^m(\hat{x})}{\partial \phi},\\ (\nabla \times E_v)_\phi &= k^2 N_0 j_n(k\sqrt{N_0}\rho)\frac{\partial Y_n^m(\hat{x})}{\partial \theta}.\end{aligned}$$

Similar to the TE modes, the transmission eigenvalues for TM modes are k^2's such that

$$\left|\begin{array}{cc} \frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho j_n(k\rho)\right) & \frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho j_n(k\sqrt{N_0}\rho)\right) \\ k^2 j_n(k\rho) & k^2 N_0 j_n(k\sqrt{N_0}\rho) \end{array}\right| = 0, \quad n \geq 1. \tag{6.68}$$

This set of transmission eigenvalues gives the second group of the Maxwell's transmission eigenvalues. Note that the multiplicities of the transmission eigenvalues for TE and TM modes are $3, 5, 7, \ldots$, which correspond to the number of spherical harmonics of order $n = 1, 2, 3, \ldots$

Let Ω be the unit ball and set $N_0 = 16$. In Fig. 6.10 and Fig. 6.11 we show the plots of the determinants corresponding to TE and TM modes, respectively. By searching for the zeros of the determinants in (6.67) and (6.68), we obtain the Maxwell's transmission eigenvalues for the unit ball which are shown in Table 6.9.

Note that the smallest transmission eigenvalue belongs to the TM modes. This is similar to the standard Maxwell's eigenvalue problem. The smallest Maxwell's eigenvalue for the unit ball belongs to the TM mode [45].

One can also search in the complex plane of zeros of the determinants defined in (6.67) and (6.68). In Fig. 6.12, we plot the absolute values of the two determinants for the first TE and TM modes. The zeros on the real axis coincide with the values in Table 6.9. The plots also indicate the likely existence of complex Maxwell's transmission eigenvalues.

i	Transmission eigenvalue (k^2)	Type	Multiplicity
1	1.1654	TM	3
2	1.4608	TE	3
3	1.4751	TM	5
4	1.7640	TE	5
5	1.7775	TM	7
6	2.0611	TE	7

Table 6.9: Maxwell transmission eigenvalues (real) for the unit ball with $N = 16I$ determined by locating the zeros of the determinants in (6.67) and (6.68).

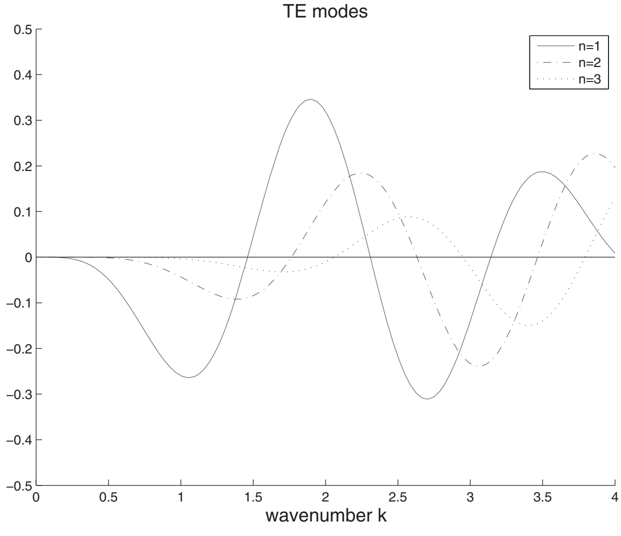


Figure 6.10: The determinant in (6.67) as a function of wave number k for $n = 1, 2, 3$. Zeros of the determinants are transmission eigenvalues for the unit ball with $N_0 = 16$ (TE modes).

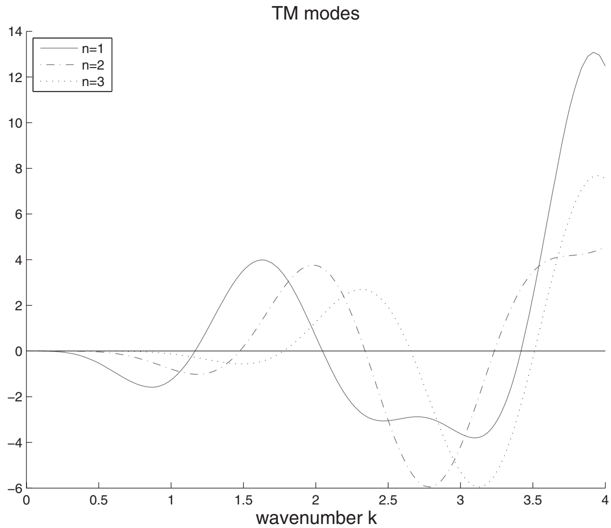


Figure 6.11: Graphs of the determinant in (6.68) as a function of wave number k for $n = 1, 2, 3$. Zeros of the determinants are transmission eigenvalues for the unit ball with $N_0 = 16$ (TM modes).

6.6.2 A Curl-conforming Edge Element Method

The first method is a curl-conforming finite element method based on the equations (6.61a)–(6.61d) directly (see also [171, 95]). Multiplying by suitable test functions and integrating by parts, a variational formulation of (6.61a)–(6.61d) can be stated as follows. Find $k^2 \neq 0$, $E_0 \in H(\text{curl}; \Omega)$ satisfying

$$(\nabla \times E_0, \nabla \times \phi) - k^2(E_0, \phi) = 0 \quad \text{for all } \phi \in H_0(\text{curl}; \Omega), \tag{6.69}$$

and $E \in H(\text{curl}; \Omega)$ satisfying

$$(\nabla \times E, \nabla \times \gamma) - k^2(NE, \gamma) = (\nabla \times E_0, \nabla \times \gamma) - k^2(E_0, \gamma), \tag{6.70}$$

for all $\gamma \in H(\text{curl}; \Omega)$ together with the essential boundary condition

$$E = E_0 \quad \text{on } \partial\Omega.$$

Note that, in (6.70), we have enforced the boundary condition (6.61d) weakly.

The following curl-conforming finite element method is based on this formulation. Let $\mathcal{T}_h$ be a regular tetrahedral mesh for Ω. Let S_h denote the smallest-order edge element space of Nédélec [208, 202] (see also Chapter 5).

We recall a subspace of S_h given by

$$S_h^0 = \{\xi_h \in S_h, \nu \times \xi_h = 0 \text{ on } \partial\Omega\} \subset H_0(\text{curl}; \Omega). \tag{6.71}$$

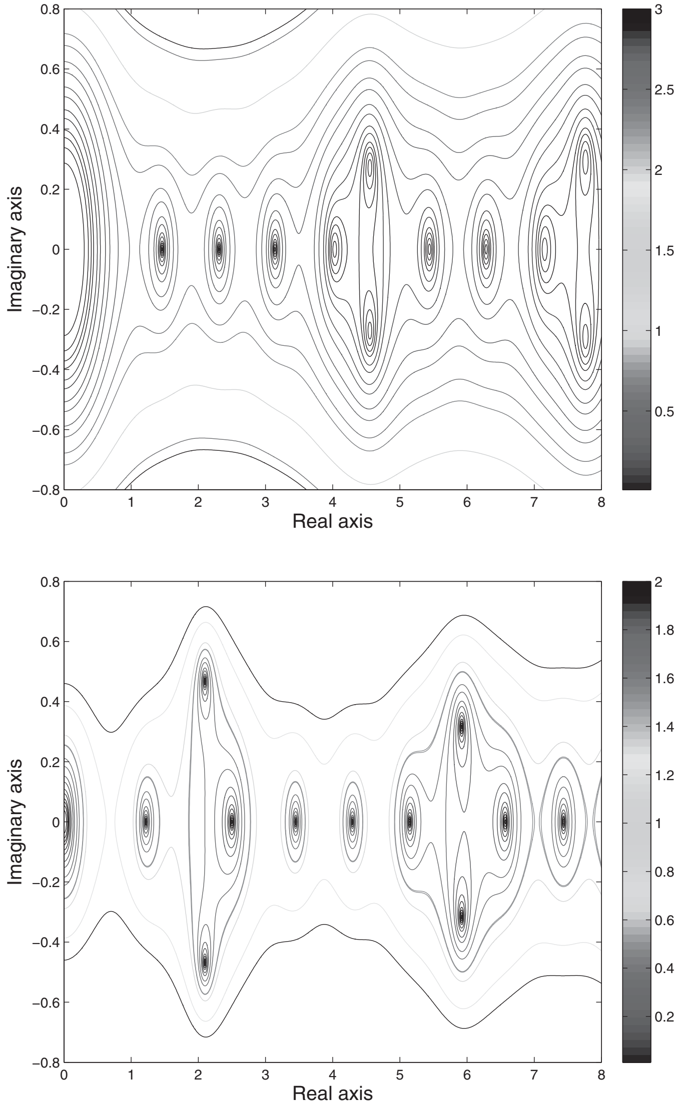


Figure 6.12: Contour plots of absolute values of the determinants for the first modes. The centers of the circles are the locations of transmission eigenvalues. We see that the plots also indicate the likely existence of complex Maxwell's transmission eigenvalues. Top: TE mode. Bottom: TM mode.

Let $T = \dim S_h$, $K = \dim S_h^0$, and $P = T - K$. Let $\xi_1, \ldots, \xi_T$ be a basis for S^h and $\xi_1, \ldots, \xi_K$ be a basis for S_h^0. Thus we have $S_h = \mathrm{span}\{\xi_j\}_{j=1}^T$ and $S_h^0 = \mathrm{span}\{\xi_j\}_{j=1}^K$. In addition, we define $S_h^B = \mathrm{span}\{\xi_j\}_{j=K+1}^T$.

Let w_h and v_h be the discrete approximations for E and E_0, respectively. We can write

$$\begin{aligned} w_h &= w_{0,h} + w_{B,h} \text{ where } w_{0,h} \in S_h^0 \text{ and } w_{B,h} \in S_h^B, \\ v_h &= v_{0,h} + w_{B,h} \text{ where } v_{0,h} \in S_h^0. \end{aligned}$$

First we choose a test function $\xi_h \in S_h^0$ and obtain

$$(\nabla \times (v_{0,h} + w_{B,h}), \nabla \times \xi_h) - k^2(v_{0,h} + w_{B,h}, \xi_h) = 0, \tag{6.72}$$

for all $\xi_h \in S_h^0$. In the same way, we have

$$(\nabla \times (w_{0,h} + w_{B,h}), \nabla \times \xi_h) - k^2(N(w_{0,h} + w_{B,h}), \xi_h) = 0, \tag{6.73}$$

for all $\xi_h \in S_h^0$. Rearranging terms in (6.70), we obtain

$$(\nabla \times (E - E_0), \nabla \times \gamma) - k^2(NE - E_0, \gamma) = 0,$$

for all $\gamma \in H(\text{curl}; \Omega)$. In the discrete case, for all $\gamma_h \in S_h^B$, we have

$$(\nabla \times (w_{0,h} - v_{0,h}), \nabla \times \gamma_h) - k^2(N(w_{0,h} + w_{B,h}) - (v_{0,h} + w_{B,h}), \gamma_h) = 0. \tag{6.74}$$

Rearranging the terms in (6.72), (6.73), and (6.74), we obtain

$$\begin{aligned} (\nabla \times (v_{0,h} + w_{B,h}), \nabla \times \xi_h) &= k^2(v_{0,h} + w_{B,h}, \xi_h), \\ (\nabla \times (w_{0,h} + w_{B,h}), \nabla \times \xi_h) &= k^2(N(w_{0,h} + w_{B,h}), \xi_h), \\ (\nabla \times (w_{0,h} - v_{0,h}), \nabla \times \gamma_h) &= k^2(N(w_{0,h} + w_{B,h}) - (v_{0,h} + w_{B,h}), \gamma_h), \end{aligned}$$

for all $\xi_h \in S_h^0$ and $\gamma_h \in S_h^B$. The definitions of the matrices are listed in Table 6.10.

Matrix	Size	Definition
A	$K \times K$	interior space stiffness matrix, $A_{j,\ell} = (\nabla \times \xi_j, \nabla \times \xi_\ell)$
B_N	$K \times P$	boundary mass matrices, $(B_N)_{j,\ell} = (N\xi_j, \xi_\ell)$,
B_1	$K \times P$	interior mass matrices, $(B_1)_{j,\ell} = (\xi_j, \gamma_\ell)$
C_N	$P \times P$	boundary space mass matrices, $(C_N)_{j,\ell} = (N\xi_j, \xi_\ell)$
C_1	$P \times P$	boundary space mass matrices, $(C_1)_{j,\ell} = (\xi_j, \xi_\ell)$
D	$K \times P$	interior stiffness matrix, $D_{j,\ell} = (\nabla \times \xi_j, \nabla \times \xi_\ell)$
M_N	$K \times K$	interior space mass matrices, $(M_N)_{j,\ell} = (N\xi_j, \xi_\ell)$
M_1	$K \times K$	interior space mass matrices, $(M_1)_{j,\ell} = (\xi_j, \xi_\ell)$

Table 6.10: Definition of matrices of the edge element method for the Maxwell's transmission eigenvalue problem.

The discrete problem we now need to solve is the following generalized eigenvalue problem

$$\mathcal{A}\vec{x} = k^2\mathcal{B}\vec{x} \tag{6.75}$$

where $\vec{x}$ has dimension $2K+P$ corresponding to $w_{0,h}$, $v_{0,h}$, and $w_{B,h}$. The matrices $\mathcal{A}$ and $\mathcal{B}$ are given blockwise by

$$\mathcal{A} = \begin{pmatrix} A & 0 & D \\ 0 & A & D \\ D^T & -D^T & 0 \end{pmatrix}$$

and

$$\mathcal{B} = \begin{pmatrix} M_N & 0 & B_N \\ 0 & M_1 & B_1 \\ B_N^T & -B_1^T & C_N - C_1 \end{pmatrix},$$

respectively.

Remark 6.6.1. *While it is possible to change variables to make $\mathcal{A}$ and $\mathcal{B}$ symmetric [232], neither would be positive definite. So* (6.75) *is not a standard positive-definite generalized eigenproblem.*

6.6.3 A Mixed Finite Element Method

The second method is based on a mixed formulation for the fourth order problem (6.63) which is similar to the mixed finite element approach for the quad-curl eigenvalue problem in Section 5.3.1 (see also [89, 201]). Recalling that $u = E - E_0$, we showed in (6.62) that $E = (N-I)^{-1}(\frac{1}{k^2}\nabla \times \nabla \times -I)u$. Hence we have that

$$\begin{aligned} (\nabla \times \nabla \times -k^2 N)E &= 0, \\ (\nabla \times \nabla \times -k^2 I)u &= (N-I)E. \end{aligned}$$

The mixed formulation can be stated as: find $(k^2, u, E) \in \mathbb{C} \times H_0(\text{curl};\Omega) \times H(\text{curl};\Omega)$ such that

$$\begin{aligned} (\nabla \times E, \nabla \times \phi) &= k^2(NE, \phi) \quad \text{for all } \phi \in H_0(\text{curl};\Omega), \\ (\nabla \times u, \nabla \times \varphi) - ((N-I)E, \varphi) &= k^2(u, \varphi) \quad \text{for all } \varphi \in H(\text{curl};\Omega). \end{aligned}$$

Given finite dimensional spaces $S_h \subset H(\text{curl};\Omega)$ and $S_h^0 \subset H_0(\text{curl};\Omega)$ such that $S_h^0 \subset S_h$, the discrete problem is to find $(k_h^2, u_h, E_h) \in \mathbb{C} \times S_h^0 \times S_h$ such that

$$\begin{aligned} (\nabla \times E_h, \nabla \times \phi_h) &= k_h^2(NE_h, \phi_h) \quad \text{for all } \phi_h \in S_h^0, \\ (\nabla \times u_h, \nabla \times \varphi_h) - ((N-I)E_h, \varphi_h) &= k_h^2(u_h, \varphi_h) \quad \text{for all } \varphi \in S_h. \end{aligned}$$

In the numerical tests, we again use the linear curl-conforming edge elements. Let $u_h = \sum_{i=1}^K u_i \xi_i$ and $E_h = \sum_{i=1}^T E_i \xi_i$. Then the corresponding matrix problem is

$$\begin{aligned} S_{K\times T} E_h &= k_h^2 M_{K\times T}^N E_h, \\ S_{T\times K} u_h - M_{T\times T}^{N-I} E_h &= k_h^2 M_{T\times K} u_h, \end{aligned}$$

where the matrices are defined in Table 6.11.

Matrix	Dimension	Definition
$S_{K\times T}$	$K \times T$	stiffness matrix $S^{i,j}_{K\times T} = (\nabla \times \xi_i, \nabla \times \xi_j)$
$S_{T\times T}$	$T \times T$	stiffness matrix $S^{i,j}_{T\times T} = (\nabla \times \xi_i, \nabla \times \xi_j)$
$M_{K\times T}$	$K \times T$	mass matrix $M^{i,j}_{K\times T} = (\xi_i, \xi_j)$
$M^{N}_{T\times K}$	$T \times K$	mass matrix $(M^{N}_{T\times K})^{i,j} = (N\xi_i, \xi_j)$
$M^{N-I}_{T\times T}$	$T \times T$	mass matrix $(M^{N-I}_{T\times T})^{i,j} = ((N-I)\xi_i, \xi_j)$

Table 6.11: Definition of matrices of the mixed method for the Maxwell's transmission eigenvalue problem.

We end up with the generalized eigenvalue problem

$$\mathcal{A}\vec{x} = k^2 \mathcal{B}\vec{x}, \tag{6.76}$$

where $\vec{x} = (E_h, u_h)^T$ and the matrices $\mathcal{A}$ and $\mathcal{B}$ are given by

$$\mathcal{A} = \begin{pmatrix} S_{K\times T} & 0_{K\times K} \\ -M^{N-I}_{T\times T} & S_{T\times K} \end{pmatrix}$$

and

$$\mathcal{B} = \begin{pmatrix} M^{N}_{K\times T} & 0_{K\times K} \\ 0_{T\times T} & M_{T\times K} \end{pmatrix},$$

respectively.

At the continuous level the fourth order problem (6.64) provides a weak form of the transmission eigenvalue problem that exactly respects the regularity requirements of the definition of the transmission eigenvalues. If we assume that E and E_0 are in $H(\mathrm{curl}, D)$, then at the continuous level the curl-conforming method and the mixed method we have outlined have, of course, the same spectrum in $(0, \infty)$. This equivalence carries over to the discrete problems (one discrete system can easily be derived from the other). As for the Maxwell's eigenvalue problem, $k_h = 0$ is an eigenvalue of large multiplicity for the discrete problem (see Section 4.7 of [202]). These eigenvalues are not physically relevant and should be excluded. Experimentally we find that the eigenspaces for $k_h = 0$ and $k_h = \infty$ differ between the two finite element methods.

The mixed method is easier to describe and implement since we have no need to impose the essential boundary condition on the difference of two fields. Both finite element methods have the advantage of using the standard linear edge elements. We find that the curl-conforming method performs slightly better in the Arnoldi process described in the next section, but this observation does not yet have any theoretical underpinning.

The generalized eigenvalue problem is non-Hermitian and the associated matrices are large and sparse. Direct methods are expensive even on a coarse mesh for two dimensional problems [95]. Therefore efficient computation of a few smallest transmission eigenvalues, which are important in algorithms to estimate material property in inverse scattering [62, 227], is a challenging problem.

6.6.4 An Adaptive Arnoldi Method

In this section, we apply an adaptive technique based on the Arnoldi method [132, 220] for the generalized eigenvalue problem obtained in the last section which is large, sparse and non-Hermitian. The process is similar to the adaptive Arnoldi method in Section 6.5, i.e., we employ the Matlab Arnoldi solver *sptarn* which can be integrated into our finite element code easily. *sptarn* uses the Arnoldi iteration with spectral transformation and requires an interval in which to search for the eigenvalues. On one hand, this is a rather appealing feature since we only need a few smallest transmission eigenvalues. Moreover, it avoids computing the smallest eigenvalue of the generalized systems (6.75) and (6.76) corresponding to the non-physical case of $k = 0$. On the other hand, the interval needs to be kept rather small in order to guarantee efficiency. Otherwise, *sptarn* will not return within a reasonable amount of time. Fortunately we are able to overcome this difficulty by coupling an iterative scheme with an estimation of the transmission eigenvalues.

To this end we first recall a Faber-Krahn type inequality for the Maxwell's transmission eigenvalues from [61].

Theorem 6.6.4. *(Theorem 4.29 of [61])*

1. *Assume that the imaginary part $\mathcal{I}(N(x)) = 0$ and $\|N(x)\|_2 \geq \delta > 1$ for all $x \in \Omega$ and some constant δ. Then,*

$$\sup_{\Omega} \|N\|_2 \geq \frac{\lambda_1(\Omega)}{k^2}, \tag{6.77}$$

where k is a transmission eigenvalue and $\lambda_1(\Omega)$ is the first Dirichlet eigenvalue of $-\Delta$ on Ω.

2. *Assume that the imaginary part $\mathcal{I}(N(x)) = 0$ and $0 \leq \beta \leq \|N(x)\|_2 \leq \delta < 1$ for all $x \in \Omega$ and some constant β. Then, if k is a transmission eigenvalue,*

$$k^2 \geq \lambda_1(\Omega), \tag{6.78}$$

where $\lambda_1(\Omega)$ is the first Dirichlet eigenvalue of $-\Delta$ on Ω.

The above theorem provides a lower bound for transmission eigenvalues in terms of the first Dirichlet eigenvalue and $\sup_\Omega \|N\|_2$. In the following we will only consider Case 1 of the above theorem, i.e., $\|N(x)\|_2 \geq \delta > 1$. The other case can be treated in exactly the same way. From (6.77), we have

$$k_1^2 \geq \frac{\lambda_1(\Omega)}{\sup_\Omega \|N\|_2}. \tag{6.79}$$

In fact, $\sup_\Omega \|N\|_2$ can be obtained from the given data easily and we can compute the first Dirichlet eigenvalue using standard linear finite elements. In particular, the discrete Dirichlet eigenvalue problem is simply the following generalized eigenvalue problem

$$S\vec{x} = \lambda M\vec{x}, \tag{6.80}$$

where S and M are the stiffness matrix and the mass matrix, respectively.

To compute a few smallest non-zero transmission eigenvalues, we start searching the transmission eigenvalues with a small interval to the right of the lower bound given in (6.79) (or (6.78)). If we successfully find transmission eigenvalues, we stop the process. Otherwise, we shift to the right, double the size of the interval and start a new search using 'sptarn' and continue this process until we find the desired transmission eigenvalues. The size of the search interval, denoted by s, should be rather small at the beginning, say, $1.0e-3$ and we can slowly increase it. From our experience, to maintain efficiency the interval cannot be too large. This is due to the fact that if there are too many eigenvalues in the interval, the efficiency of 'sptarn' will be significantly downgraded. Note that 'sptarn' also computes complex eigenvalues. Since only real transmission eigenvalues are of interest, we simply discard the complex ones. This process implicitly assumes that the Faber-Krahn lower bound is also a lower bound for the first non-zero discrete transmission eigenvalue, a fact we have not yet verified although our numerical experiments suggest it is true.

Assuming a tetrahedral mesh $\mathcal{T}$ is already generated for Ω, the following adaptive algorithm computes N_e smallest transmission eigenvalues.

Algorithm for MTEs:

Input:

- a tetrahedral mesh for Ω and the initial size of the search interval s
- the index of refraction $N(x)$ and $\sup_\Omega \|N(x)\|_2$
- the number of transmission eigenvalues N_e to be computed

Output:

- N_e Maxwell's transmission eigenvalues

1. construct matrices $\mathcal{A}$ and $\mathcal{B}$
2. compute $\lambda_1(\Omega)$
3. set $TE = \emptyset, lb = \frac{\lambda_1(\Omega)}{\sup_\Omega \|N(x)\|_2}, rb = lb + s$
4. while length(TE) $< N_e$
   - $[V, \Omega] = sptarn(A, B, lb, rb)$
   - delete complex values in Ω
   - $TE = TE \cup \Omega$
   - $lb = rb, s = \min(2s, 1), rb = lb + s$

Remark 6.6.2. *It is also possible to use the bounds in Theorem 6.6.3 to estimate a search interval. However, one needs to find balls inside and outside Ω and devise an effective way to compute transmission eigenvalues for balls with constant index of refraction. Since we use finite elements to compute transmission eigenvalues, it is easier to compute the Dirichlet eigenvalues using the same mesh.*

6.6.5 Numerical Examples

In this section we provide some numerical examples to show the viability of the proposed methods and test the efficiency of the inequalities at the beginning of this section. We choose two domains: Ω_1 the unit ball centered at the origin and Ω_2 the unit cube given by $[0,1] \times [0,1] \times [0,1]$ (see Fig. 6.13). We only consider when

$$\|N(x)\|_2 \geq \alpha > 1$$

since the case of

$$0 < \|N(x)\|_2 \leq 1 - \beta$$

is similar. We test three different cases for the index of refraction $N(x)$ corresponding to isotropic medium with constant index of refraction

$$1) \begin{pmatrix} 16 & 0 & 0 \\ 0 & 16 & 0 \\ 0 & 0 & 16 \end{pmatrix}, \tag{6.81}$$

anisotropic medium with constant index of refraction

$$2) \begin{pmatrix} 16 & 1 & 0 \\ 1 & 16 & 0 \\ 0 & 0 & 14 \end{pmatrix}, \tag{6.82}$$

and anisotropic medium with variable index of refraction

$$3) \begin{pmatrix} 16 & x & y \\ x & 16 & z \\ y & z & 14 \end{pmatrix}, \tag{6.83}$$

respectively.

Note that due to the 3D nature of the problem and the desktop computer available for the numerical tests, we have to restrict the mesh size h to be larger than roughly 0.2.

We compare the two finite element methods using the same meshes. The first example is the unit ball with index of refraction $N = 16I$. We use a mesh with $h \approx 0.4$. To make comparison, we compute the full spectrum of the generalized eigenvalue problems using Matlab's *eig*. Both methods end up with the same degree of freedom (DoF) 2566. The curl-conforming method computes 708 zero eigenvalues and the mixed method computes 228 zero eigenvalues. Note that the curl-conforming method has a large eigenspace corresponding to $k_h = 0$ is unsurprising since $k = 0$ is a non-trivial transmission eigenvalue for (6.61) with infinite dimensional eigenspace. Unlike for the Helmholtz equation, the fourth order problem (6.63) also has $k = 0$ as an eigenvalue. The mixed method computes $k_h = 0$ as an eigenvalue, but also computes many eigenvalues $k_h = inf$ since $\mathcal{B}$ in (6.76) is singular. The rest of the spectrum in $(0, \infty)$ coincides, even for complex eigenvalues, as we claimed earlier.

If we use the Arnoldi method in the interval $[1, 3]$ which contains 11 eigenvalues,

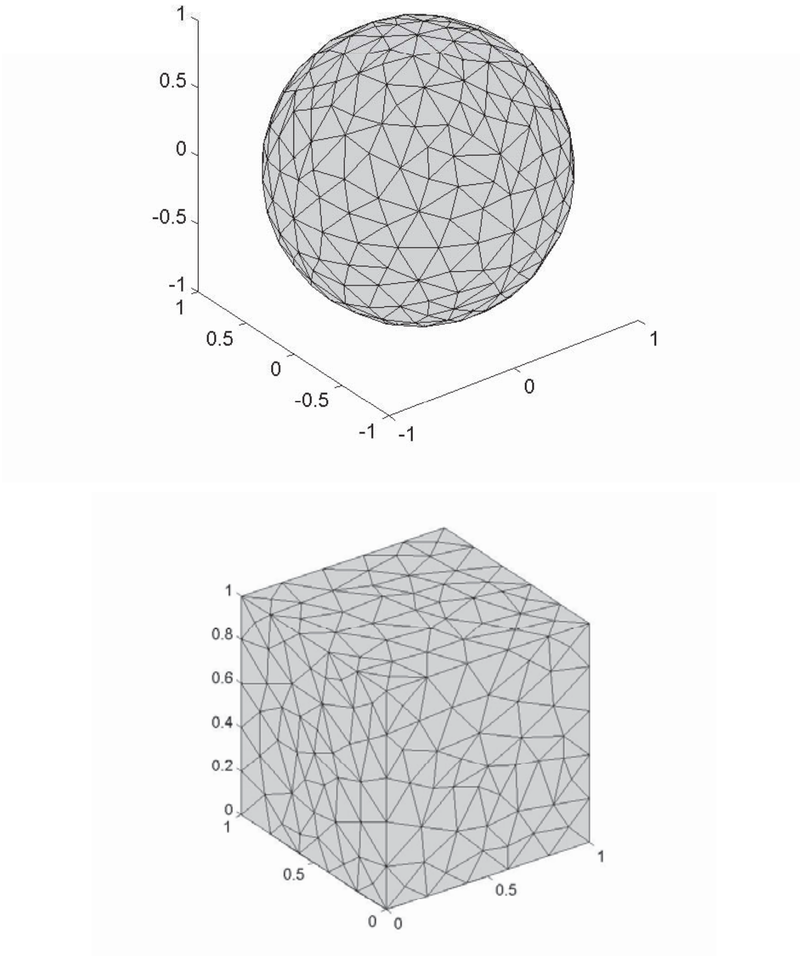


Figure 6.13: Two domains used for numerical examples and sample tetrahedra meshes. Top: the unit ball centered at the origin. Bottom: the unit cube given by $[0,1] \times [0,1] \times [0,1]$.

the curl-conforming method uses 4.38s (CPU time) which is slightly shorter than the mixed method with 4.98s (CPU time), providing a slight reason for preferring the curl-conforming method.

We have repeated this experiment for the unit cube with the index of refraction

$N = 16I$. We use a mesh $h \approx 0.3$. Both methods end up with the same number of DoF 1376. The curl-conforming method computes 444 zero eigenvalues and mixed method computes 102 zero eigenvalues (as well as some infinite eigenvalues). The rest of the spectrum in $(0, \infty)$ also coincides. If we use the Arnoldi method in the interval $[3, 5]$ which contains 3 eigenvalues, the curl-conforming method uses 0.74s (CPU time) which is slightly shorter than the mixed method with 0.91s (CPU time). We summarize the result in Table 6.12. Note that for both examples, the number of zero eigenvalues of the mixed method is twice the number of boundary nodes. In addition, the difference between the number of zero eigenvalues computed by the two methods coincides with the number of edges on the domain boundary. Since the two methods compute the same non-zero spectrum and the curl-conforming method is slightly more efficient, we will use the curl-conforming method in the subsequent examples.

domain	unit ball			unit cube		
	DoF	# of zero	CPU time	DoF	# of zero	CPU time
curl-conforming	2566	708	4.38s	1376	444	0.74s
mixed method	2566	228	4.98s	1376	102	0.91s

Table 6.12: Comparison of the curl-conforming method and the mixed method ($N = 16I$).

Next we show a few transmission eigenvalues for the unit ball with constant index of refraction $N = 16I$ (6.81) (see Table 6.13) computed using $h \approx 0.2$. These values coincide rather well with the exact transmission eigenvalues shown in Table 6.9 and have correct multiplicities.

	multiplicity	computed values
1.1654	3	1.1741, 1.1717, 1.1721
1.4608	3	1.4665, 1.4667, 1.4671
1.4751	5	1.4824, 1.4828, 1.4828, 1.4830, 1.4836
1.7640	5	1.7690, 1.7690, 1.7698, 1.7700, 1.7705
1.7775	7	1.7857, 1.7859, 1.7862, 1.7865, 1.7867, 1.7868, 1.7872

Table 6.13: Computed Maxwell's transmission eigenvalues for the unit ball with $N = 16I$. The mesh size $h \approx 0.2$. The first column is the transmission eigenvalues from Table 6.9. The second column is the multiplicities of the respective eigenvalues. The third column is the computed eigenvalues. The computed eigenvalues have the correct multiplicities.

Since we have exact values for this case, we can look at the convergence rate of the smallest transmission eigenvalue. This is done by carrying out the computation on a series of meshes with decreasing mesh size h. We plot the errors against the mesh size h in log scale in Fig. 6.14 where second order convergence can be seen clearly (see Table 6.14 for the actual h and the errors).

mesh size	computed eigenvalue	exact eigenvalue	error
$h \approx 0.66$	1.2145	1.1654	0.0491
$h \approx 0.50$	1.1970	1.1654	0.0316
$h \approx 0.40$	1.1837	1.1654	0.0183
$h \approx 0.29$	1.1761	1.1654	0.0107
$h \approx 0.22$	1.1720	1.1654	0.0066

Table 6.14: The errors of the smallest Maxwell's transmission eigenvalues for the unit ball with $N = 16I$. The exact values are from Table. 6.9.

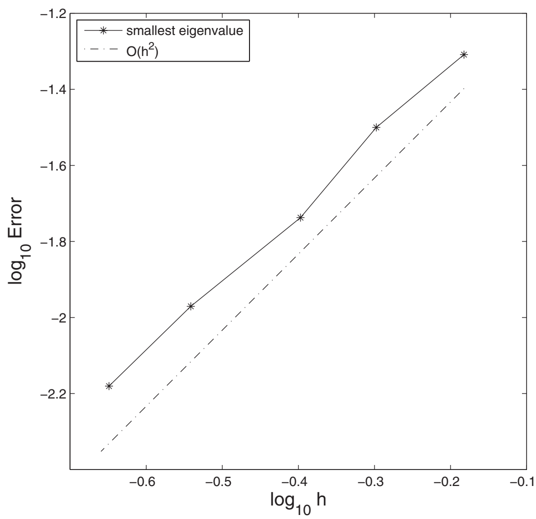


Figure 6.14: Convergence rate of the smallest transmission eigenvalue for the unit ball with $N = 16I$. Here h denotes the mesh size. Second order convergence is observed.

Next we check Theorem 6.6.2 for the index of refraction given in (6.83). Straightforward calculation shows that

$$n_* \approx 13.5697 \leq \hat{\xi} \cdot N(x)\xi \leq 17.0000 \approx n^*, \quad \text{for all } x \in \Omega_1. \tag{6.84}$$

Using a mesh with mesh size $h \approx 0.2$, we find that

$$k_{1,\Omega_1,n^*} \approx 1.1381 < k_{1,\Omega_1,N(x)} \approx 1.1857 < k_{1,\Omega_1,n_*} \approx 1.2877,$$

i.e., Theorem 6.6.2 gives a reasonable estimate for $k_{1,\Omega_1,N(x)}$ for this example.

Now we consider the unit cube, i.e., $\Omega_2 = [0,1] \times [0,1] \times [0,1]$. First, we check Theorem 6.6.2 for the index of refraction given in (6.83). Again, straightforward calculation shows that

$$n_* \approx 13.2679 \leq \hat{\xi} \cdot N(x)\xi \leq 17.5616 \approx n^*, \quad \text{for all } x \in \Omega_2. \tag{6.85}$$

Using a mesh with $h \approx 0.2$, we compute $k_{1,\Omega_2,N(x)} \approx 2.0527$ and have that

$$k_{1,\Omega_2,n^*} \approx 1.9920 < k_{1,\Omega_2,N(x)} \approx 2.0527 < k_{1,\Omega_2,n_*} \approx 2.2187.$$

Next, we check Theorem 6.6.3. It is obvious that the ball B_1 with radius $r_1 = 1/2$ is the largest ball such that $B_{r_1} \subset \Omega_2$ and the ball B_2 with radius $r_2 = \sqrt{2}$ is the smallest ball such that $\Omega_2 \subset B_{r_2}$. When the index of refraction is given by (6.82), we have that

$$n_* = 14, \quad n^* = 17.$$

Using the result of Section 6.6.1, we have that

$$k_{1,n^*} = 1.1277, \quad k_{1,n_*} = 1.2539.$$

The finite element method gives that $k_{1,\Omega_2,N(x)} \approx 2.0411$ and we have that

$$\frac{k_{1,n^*}}{\sqrt{2}} \approx 0.7974 \leq k_{1,\Omega,N(x)} \approx 2.0411 \leq \frac{k_{1,n_*}}{r_1} \approx 2.5078.$$

Now let $\epsilon = 1/4$. Then we can put $m(1/4) = 4$ balls $B_{1/4}$ with radius $1/4$ in Ω_2. According to Theorem 6.6.3, there are *at least* $m(1/4) = 4$ transmission eigenvalues in the interval

$$\left[\frac{k_{1,n^*}}{r_2}, \frac{k_{1,n_*}}{\epsilon}\right] \approx [0.7974, 5.0156].$$

The numerical method computes 16 transmission eigenvalues in $[0.7974, 3.1623]$.

When the index of refraction is given by (6.83), we have that

$$n_* = 13.2679, \quad n^* = 17.5616.$$

Once again using the result of Section 6.6.1, we have that

$$k_{1,n^*} = 1.1081, \quad k_{1,n_*} = 1.2918.$$

The finite element method gives $k_{1,\Omega_2,N(x)} \approx 2.0527$ and we have that

$$\frac{k_{1,n^*}}{\sqrt{2}} \approx 0.7835 \leq k_{1,\Omega,N(x)} \approx 2.0527 \leq \frac{k_{1,n_*}}{r_1} \approx 2.5836.$$

Similarly, according to Theorem 6.6.3, there are *at least* $m(1/4) = 4$ transmission eigenvalues in the interval

$$\left[\frac{k_{1,n^*}}{r_2}, \frac{k_{1,n_*}}{\epsilon}\right] \approx [0.7835, 5.1672].$$

The numerical method computes 19 transmission eigenvalues in $[0.7835, 3.1623]$.

6.7 Appendix: Code for the Mixed Method

Using the subroutine 'assemble' in Chapter 3, it is rather simple to implement the mixed method for transmission eigenvalues described in Section 6.5 when the index of refraction n is a constant. For simplicity, we use *eigs* instead of *sptarn*.

Suppose that a triangular mesh $\mathcal{T}$ for Ω is given. Let V_h be the linear Lagrange element space associated with $\mathcal{T}$. Let $\{\phi_1, \phi_2, \ldots, \phi_K\}$ be the basis functions associated with the interior nodes of $\mathcal{T}$ and $\{\phi_{K+1}, \ldots, \phi_T\}$ be the basis functions associated with the boundary nodes of $\mathcal{T}$. In other words,

$$\text{span}\{\phi_1, \phi_2, \ldots, \phi_N, \phi_{N+1}, \ldots, \phi_{N+M}\} = V_h$$

and

$$\text{span}\{\phi_1, \phi_2, \ldots, \phi_N\} = V_h \cap H_0^1(\Omega).$$

Let S be the stiffness matrix given by

$$S = (\nabla\phi_j, \nabla\phi_i), \quad i, j = 1, \ldots, N+M,$$

and M be the mass matrix given by

$$M = (\phi_j, \phi_i), \quad i, j = 1, \ldots, N+M.$$

According to Table 6.6 and (6.53), the respective matrices are

$$\begin{aligned}
S_{K\times T} &= S(1:K, 1:T), \\
S_{T\times K} &= S(1:T, 1:K), \\
M_{K\times T} &= M(1:K, 1:T), \\
M^n_{T\times K} &= nM(1:T, 1:K), \\
M^{n-1}_{T\times T} &= (n-1)M(1:T, 1:T).
\end{aligned}$$

A simple Matlab code "MixedFEMTE" is given below.

```
1. function lambda = MixedFEMTE(mesh, n, num)
% Input:
%   mesh - a triangular mesh
%   n    - index of refraction
%   num  - number of eigenvalues to compute
% Output:
%   lambda  - a vector of 'num' TEs
% ------- call assemble.m to construct matrices ---
2. [S, M]=assemble(mesh.p, mesh.t);
% ------- number of vertices in the mesh ----------
3. N=length(mesh.p);
%-------- Find boundary nodes ---------------------
```

```
4. bdnodeE = unique([mesh.e(1,:),mesh.e(2,:)]);
5. Inode = setdiff(linspace(1,N,N), bdnodeE);
%-------- Construct matrices -----------------------
6. Skt = S(Inode, :);
7. Stk = S(:, Inode);
8. Mkt = M(Inode, :);
9. Mtk = M(:, Inode);
%-------- assemble matrices ------------------------
10. L = length(Inode);
11. A = [Skt, sparse(L, L); (n-1)*M, Stk];
12. B = [Mkt, sparse(L, L); sparse(N, N), n*Mtk];
%---- call ’eigs’ to compute ’num’ eigenvalues
13. [V,D]=eigs(A, B, num, ’sm’);
14. lambda = sqrt(diag(D));
```

The following are some comments on the code.

a. Line 1: input "mesh" — a triangular mesh, "n" — index of refraction, and "num" — number of transmission eigenvalues to be computed,

b. Line 2: call the subroutine "assemble" (the same one in Chapter 3) to construct the stiffness and mass matrices,

c. Line 3: find the number of vertices in the mesh,

d. Lines 4–5: find the interior vertices and boundary vertices,

e. Lines 6–9: assign various matrices,

f. Line 10: find the number of interior matrices,

g. Lines 11–12: construction "A" and "B" according to (6.53),

h. Line 13: call *eigs* to compute the eigenvalues of the generalized eigenvalue problem,

i. Line 14: compute the transmission eigenvalues.

Chapter 7

The Schrödinger Eigenvalue Problem

7.1 Introduction

The Schrödinger eigenvalue problem

$$-\frac{1}{2}\triangle u + \mathcal{V}u = \lambda u \quad \text{in } \mathbb{R}^n \tag{7.1}$$

models the stationary state of N particles moving in an external potential $\mathcal{V}(x_1, \cdots, x_n)$ with $n = 3N$. Mathematically, we may assume that $\mathcal{V} \in L^{n/2}(\mathbb{R}^n) + L^\infty(\mathbb{R}^n)$.

There may or may not be a solution of (7.1), and if there is one it may not be unique [188]. Since (7.1) is intractable, reduced or equivalent models that are tractable are then introduced. We see that Hartree-Fock equations and Kohn-Sham equations are the most widely used models in electronic structure calculations. Note that electrons are fermions. One reduced model for a type of bosons is the so-called Gross-Pitaevskii equation, which is used to model a Bose-Einstein condensation (BEC) of ultracold dilute gas with N identical bosons confined in an external trap.

Such kind of models are nonlinear eigenvalue problems in $\mathbb{R}^3$, which usually require a so-called self-consistent field (SCF) iteration to linearize in computation. Consequently, the central computation in the application is the solution of the following linear Schrödinger equation

$$-\frac{1}{2}\triangle u + \mathcal{V}u = \lambda u \quad \text{in } \mathbb{R}^3, \tag{7.2}$$

where $\mathcal{V}$ is a potential and may be assumed to be a function in $L^{3/2}(\mathbb{R}^3) + L^\infty(\mathbb{R}^3)$ or

$$\lim_{|x|\to\infty} V(x) = \infty,$$

which is measurable and locally bounded.

Remark 7.1.1. *The above assumption is reasonable in solving Hartree-Fock equations, Kohn-Sham equations, and Gross-Pitaevskii equations, etc.*

Note that the Coulomb potential is singular at cores and the physical properties of solids depend essentially on valence electrons rather than the core electrons. As a result, a pseudopotential approximation is then proposed. With the pseudopotential approximation, the corresponding potential $\mathcal{V}$ and eigenfunction in (7.1) then have better regularity.

Remark 7.1.2. *In the pseudopotential setting, the resulting model is indeed a differential-integral equation. With the SCF iteration, we will solve (7.2) with $\mathcal{V}$ being a function operator.*

We observe that for finite atomic or molecular systems or BEC, the ground state decays exponentially, and the restrictions to bounded domains and homogeneous Dirichlet conditions are reasonable for the Schrödinger equation. While for crystals, for instance, we may pose the periodic boundary conditions. In application, we need to solve (7.1).

A variational problem associated with (7.1) is:

$$\inf\left\{\mathcal{E}(u) : u \in H^1(\mathbb{R}^n), \int_{\mathbb{R}^n} |u|^2 \,\mathrm{d}x = 1\right\}, \tag{7.3}$$

where

$$\mathcal{E}(u) = \frac{1}{2}\int_{\mathbb{R}^n} |\nabla u|^2 + \mathcal{V}|u|^2 \,\mathrm{d}x.$$

Here and hereafter in this section, we assume $n \geq 3$ for convenience.

If there exists $u_0 \in H^1(\mathbb{R}^n)$ such that

$$\lambda_0 \equiv \inf\left\{\mathcal{E}(u) : u \in H^1(\mathbb{R}^n), \int_{\mathbb{R}^n} |u|^2 \,\mathrm{d}x = 1\right\} = \mathcal{E}(u_0),$$

then u_0 is called the ground state and λ_0 the ground state energy. The variational problem (7.3) determines not only u_0 but also the corresponding eigenvalue λ_0, the smallest eigenvalue of (7.1).

We see that there may not exist a minimizer u_0 such that

$$\inf\left\{\mathcal{E}(u) : u \in H^1(\mathbb{R}^n), \int_{\mathbb{R}^n} |u|^2 \,\mathrm{d}x = 1\right\} = \mathcal{E}(u_0),$$

for instance, when $\mathcal{V} \equiv 0$.

By a standard argument, we have

Theorem 7.1.1. *Assume that $\mathcal{V} \in L^{n/2}(\mathbb{R}^n) + L^\infty(\mathbb{R}^n)$ and*

$$\lim_{|x|\to\infty} \mathcal{V}(x) = 0.$$

If

$$\lambda_0 = \inf\left\{\mathcal{E}(u) : u \in H^1(\mathbb{R}^n), \int_{\mathbb{R}^n} |u|^2\,dx = 1\right\} < 0,$$

then there exists a solution $u_0 \in H^1(\mathbb{R}^n)$ *such that* $\|u\| = 1$ *and* $\mathcal{E}(u_0) = \lambda_0$. *The minimizer* u_0 *also satisfies (7.1) in the sense of distribution.*

We refer to Section 11.5 of Lieb and Loss [189] for the proof. Furthermore, we have the following uniqueness of the minimizer; see Section 11.8 of Lieb and Loss [189].

Theorem 7.1.2. *Assume that* $u_0 \in H^1(\mathbb{R}^n)$ *is a minimizer of* $\mathcal{E}(u)$ *in* $H^1(\mathbb{R}^n)$, *namely,* $\mathcal{E}(u_0) = \lambda_0 > -\infty$ *and* $\|u_0\| = 1$. *If* $\mathcal{V} \in L^1_{loc}(\mathbb{R}^n)$ *and* $\mathcal{V}$ *is locally bounded from the above, then* u_0 *satisfying (7.1) with* $\lambda = \lambda_0$ *and* u_0 *can be chosen as a strictly positive function; such kind of positive minimizer is unique.*

Theorem 7.1.3. *Assume that* $\mathcal{V} \in L^1_{loc}(\mathbb{R}^n)$ *and* $\mathcal{V}$ *is bounded from the above and* $\lambda_0 > -\infty$. *If* $0 \le u \in H^1(\mathbb{R}^n)$ *satisfying (7.1) and* $\|u\| = 1$, *then* $\lambda = \lambda_0$ *and* u *is the unique positive minimizer* u_0.

Proof. Since u_0 is the solution of

$$-\frac{1}{2}\triangle u_0 + \mathcal{V}u_0 = \lambda_0 u_0 \quad \text{in } \mathbb{R}^n,$$

we have

$$\frac{1}{2}(\nabla u_0, \nabla u) + (\mathcal{V}u_0, u) = \lambda_0(u_0, u). \tag{7.4}$$

We also obtain from (7.1) that

$$\frac{1}{2}(\nabla u, \nabla u_0) + (\mathcal{V}u, u_0) = \lambda(u, u_0). \tag{7.5}$$

If $\lambda \neq \lambda_0$, then we get from (7.4) and (7.5) that

$$(u, u_0) = 0,$$

which is a contradiction since $u, u_0 \ge 0$. This completes the proof. □

In the rest of this chapter, we study finite element approximations to the following Schrödinger eigenvalue problem: Find $(\lambda, u) \in \mathbb{R} \times H^1_0(\Omega)$ such that

$$\begin{cases} -\frac{1}{2}\Delta u + \mathcal{V}u &= \lambda u \quad \text{in } \Omega, \\ \|u\| &= 1, \end{cases} \tag{7.6}$$

where $\Omega \subset \mathbb{R}^3$ is a polyhedral domain and $\mathcal{V}$ is the effective potential. We first investigate the finite element approximations when $\mathcal{V}$ is a nonlinear operator. Then we introduce a two-scale finite element discretization to (7.6) when $\mathcal{V}$ is a function, in particular, with Coulomb-type singularity.

7.2 Approximation to Gross–Pitaevskii Equation

In this section, we analyze the finite element approximation to the Gross-Pitaevskii equation, a nonlinear Schrödinger equation (7.6) with

$$\mathcal{V} = \mathcal{V}_{ext} + \beta |u|^2,$$

where β is constant and $\mathcal{V}_{ext} \geq 0$.

Define

$$V = \left\{ v \in H_0^1(\Omega) : \int_\Omega |v|^2 \mathcal{V}_{ext} \,\mathrm{d}x < \infty \right\},$$
$$\|v\|_V = \left(\|v\|^2_{H^1(\Omega)} + \|v\|^2_{\mathcal{V}} \right)^{1/2},$$

where

$$\|v\|_{\mathcal{V}} = \left(\int_\Omega |v|^2 \mathcal{V}_{ext} \,\mathrm{d}x \right)^{1/2}.$$

We see that $(V, \|\cdot\|_V)$ is a Hilbert space. Note that the Sobolev embedding theorem implies that $V \subset L^2(\Omega) \cap L^4(\Omega)$ and $\|\cdot\|_V$ is equivalent to $\|\cdot\|_V'$ in V, where

$$\|v\|_V' = \|v\|_{H^1(\Omega)} + \|v\|_{L^4(\Omega)} + \|v\|_{\mathcal{V}}.$$

For convenience, we use the norm $\|\cdot\|_V$ in our analysis.

Any eigenvalue λ of (7.6) can also be computed from its corresponding eigenfunction u as follows

$$\lambda = \int_\Omega \frac{1}{2}|\nabla u|^2 + \mathcal{V}_{ext}|u|^2 + \beta|u|^4 \,\mathrm{d}x = \mathcal{E}(u) + \frac{\beta}{2}\int_\Omega |u|^4 \,\mathrm{d}x,$$

where

$$\mathcal{E}(u) = \int_\Omega \frac{1}{2}|\nabla u|^2 + \mathcal{V}_{ext}|u|^2 + \frac{\beta}{2}|u|^4 \,\mathrm{d}x$$

is the energy. With a repulsive interaction, the BEC ground state solution u is the unique real non-negative function found by minimizing the energy $\mathcal{E}(v)$ under the constraint $\|v\| = 1$. Namely, if $\beta \geq 0$, then

$$u = \arg\min\{E(v) : v \in V, \|v\| = 1\} \geq 0 \tag{7.7}$$

is unique and solves (7.6)(see [190]).

The weak form of (7.6) is: Find $(\lambda, u) \in \mathbb{R} \times V$ such that

$$\left\{ \begin{array}{rcl} \frac{1}{2}(\nabla u, \nabla v) + (\mathcal{V}_{ext} u + \beta |u|^2 u, v) & = & \lambda(u, v) \ \text{ for all } \ v \in V, \\ u \geq 0, \quad \|u\| & = & 1. \end{array} \right. \tag{7.8}$$

In this section, we will study and analyze the finite element approximations to (7.8) or (7.6).

Let $V_h \subset V$ be a sequence of finite element subspaces such that

$$\lim_{h\to 0} \inf_{\chi\in V_h} \|v-\chi\|_V = 0 \qquad \text{for all } v \in V. \tag{7.9}$$

It is shown that for any $h \ll 1$, there exists a unique $u_h \in V_h$ such that $u_h \geq 0$ and

$$\mathcal{E}(u_h) = \min\{\mathcal{E}(v) : v \in V_h, \|v\| = 1\}, \tag{7.10}$$

which satisfies

$$\begin{cases} (\nabla u_h, \nabla v) + (\mathcal{V}_{ext} u_h + \beta|u_h^2|u_h, v) & = & \lambda_h(u_h, v) \text{ for all } v \in V_h, \\ u_h \geq 0, \|u_h\| & = & 1 \end{cases} \tag{7.11}$$

with

$$\lambda_h = \mathcal{E}(u_h) + \frac{\beta}{2}\int_\Omega u_h^4 \,\mathrm{d}x. \tag{7.12}$$

In fact, the existence of $u_h \in V_h$ is obvious and the uniqueness is quite difficult to prove and can be found in [84].

We assume here that $\{(u_h, \lambda_h)\}$ are approximations to (u, λ), namely, they satisfy (7.11) and (7.8), respectively. As a result, we have

$$\sup_{h\ll 1} \left(\mathcal{E}(u_h) + \frac{\beta}{2}\int_\Omega u_h^4 \,\mathrm{d}x \right) < \infty. \tag{7.13}$$

We will mention that assumption (7.9) is satisfied by most of the finite element spaces used in practice.

The following materials come from Zhou [255], where more general finite dimensional approximations have been investigated.

7.2.1 Convergence

The basic convergence of the finite element approximations is stated as follows.

Theorem 7.2.1. *There hold*

$$\lim_{h\to 0} \|u - u_h\| = 0, \tag{7.14}$$

$$\lim_{h\to 0} \lambda_h = \lambda, \tag{7.15}$$

$$\lim_{h\to 0} \mathcal{E}(u_h) = \mathcal{E}(u). \tag{7.16}$$

Proof. It is sufficient to prove that for any sequence $\{h_k\}$, there exists a subsequence $\{h_{k_j}\} \subset \{h_k\}$ such that

$$\lim_{j\to\infty} \|u - u_{h_{k_j}}\| = 0, \tag{7.17}$$

$$\lim_{j\to\infty} \lambda_{h_{k_j}} = \lambda, \tag{7.18}$$

and

$$\lim_{j\to\infty} \mathcal{E}(u_{h_{k_j}}) = \mathcal{E}(u). \tag{7.19}$$

Note that if $\{u_{h_k}\}$ are minimizers of (7.10) with $u_{h_k} \geq 0$ or equivalently $(\lambda_{h_k}, u_{h_k})(k = 1, 2, \cdots)$ satisfy (7.11), then (7.13) yields that there exists a convergent subsequence $\{\lambda_{h_{k_j}}\}$, a weakly convergent subsequence $\{u_{h_{k_j}}\}$, such that

$$u_{h_{k_j}} \rightharpoonup \tilde{u} \text{ in } V, \tag{7.20}$$

$$u_{h_{k_j}} \rightharpoonup \tilde{u} \text{ in } L^4(\Omega), \tag{7.21}$$

$$u_{h_{k_j}} \rightharpoonup \tilde{u} \text{ in } L^2(\Omega), \tag{7.22}$$

$$\lambda_{h_{k_j}} \to \tilde{\lambda}, \tag{7.23}$$

$$\mathcal{E}(u_{h_{k_j}}) \to \nu \tag{7.24}$$

for some $\tilde{\lambda} > 0, \nu > 0$, and $\tilde{u} \in V$ with $\tilde{u} \geq 0$.

Because $\|\cdot\|_V$ and $L^4(\Omega)$-norm are weakly lower semicontinuous, we have

$$\liminf_{j\to\infty} \mathcal{E}(u_{h_{k_j}}) \geq E(\tilde{u}). \tag{7.25}$$

Noting that $|u_{h_{k_j}}|^2$ converges to $|\tilde{u}|^2$ in $L^1(\Omega)$, we get that $\|\tilde{u}\| = 1$.

It is easy to see that (7.9) leads to that $\{u_{h_{k_j}}\}$ is a minimizing sequence for the functional $\mathcal{E}(v)$. Thus we obtain that $\tilde{u} \in V$ is a minimizer of $E(v)$ with $\tilde{u} \geq 0$ and $\|\tilde{u}\| = 1$. The uniqueness of the minimizer then yields $\tilde{u} = u$ and hence $\tilde{\lambda} = \lambda$ and $\nu = \mathcal{E}(u)$. Therefore we get (7.18) and (7.19).

Noting that (7.22) implies

$$\lim_{j\to\infty} (u_{h_{k_j}}, u) = (u, u),$$

we derive (7.17) from $\|u_{h_{k_j}}\| = \|u\| = 1$ and the identity

$$\|u_{h_{k_j}} - u\|^2 = \|u_{h_{k_j}}\|^2 - 2(u_{h_{k_j}}, u) + \|u\|^2.$$

This completes the proof. $\square$

7.2.2 Error Estimate

Now we turn to estimate the errors. In the following analysis, we need a useful identity.

Lemma 7.2.2. *If* $(\lambda, u) \in \mathbb{R} \times V$ *satisfies (7.8), then*

$$\begin{aligned}
&\frac{(\nabla v, \nabla v)/2 + (\mathcal{V}_{ext} v + \beta |v|^2 v, v)}{(v,v)} - \lambda \\
= \quad & \frac{(\nabla(v-u), \nabla(v-u))/2 + (\mathcal{V}_{ext}(v-u), v-u) + \beta(u^2(v-u), v-u)}{(v,v)} \\
&+ \frac{\beta((|v|^2 - |u|^2)v, v)}{(v,v)} - \lambda \frac{(v-u, v-u)}{(v,v)} \quad \textit{for all } v \in V.
\end{aligned} \tag{7.26}$$

Proof. Let $\mathcal{V} = \mathcal{V}_{ext} + \beta |u|^2$. We rewrite (7.8) as

$$\left\{ \begin{array}{rcl} \frac{1}{2}(\nabla u, \nabla v) + (\mathcal{V} u, v) & = & \lambda(u, v) \ \text{ for all } v \in V, \\ u \geq 0, \|u\| & = & 1. \end{array} \right. \tag{7.27}$$

Note that

$$\begin{aligned}
& \frac{1}{2}(\nabla(v-u), \nabla(v-u)) + (\mathcal{V}(v-u), v-u) \\
= \quad & \frac{1}{2}(\nabla v, \nabla v) + (\mathcal{V} v, v) + \left(\frac{1}{2}(\nabla u, \nabla(u - 2v)) + (\mathcal{V} u, u - 2v) \right)
\end{aligned}$$

for all $v \in V$. We have

$$\begin{aligned}
& \frac{1}{2}(\nabla(v-u), \nabla(v-u)) + (\mathcal{V}(v-u), v-u) \\
= \quad & \frac{1}{2}(\nabla v, \nabla v) + (\mathcal{V} v, v) + \lambda(u, u - 2v) \ \text{ for all } v \in V,
\end{aligned}$$

which implies

$$\begin{aligned}
& \frac{1}{2}(\nabla v, \nabla v) + (\mathcal{V}_{ext} v + \beta |v|^2 v, v) \\
= \quad & \frac{1}{2}(\nabla(v-u), \nabla(v-u)) + ((\mathcal{V}_{ext} + \beta |u|^2)(v-u), v-u) \\
& - \lambda(u, u - 2v) + (\beta(|v|^2 - |u|^2)v, v) \ \text{ for all } v \in V.
\end{aligned}$$

Using the identity

$$\lambda(v,v) = \lambda(v-u, v-u) - \lambda(u, u-2v) \ \text{ for all } v \in V,$$

we obtain for any $v \in V$ that

$$\begin{aligned} & \frac{1}{2}(\nabla v, \nabla v) + (\mathcal{V}_{ext} v + \beta |v|^2 v, v) - \lambda(v, v) \\ = \ & \frac{1}{2}(\nabla(v-u), \nabla(v-u)) + ((\mathcal{V}_{ext} + \beta |u|^2)(v-u), v-u) \\ & + (\beta(|v|^2 - |u|^2)v, v) - \lambda(v-u, v-u), \end{aligned}$$

which is nothing but (7.26). This completes the proof. □

Applying (7.26), we are able to give some upper bounds.

Theorem 7.2.3. *If* $h \ll 0$*, then*

$$| \lambda_h - \lambda | \leq C(\|u_h - u\| + \|u_h - u\|^2 + \inf_{v \in V_h} \|v - u\|_V^2), \tag{7.28}$$

$$\|u_h - u\|_V \leq C(\|u_h - u\| + \|u_h - u\|^2 + \inf_{v \in V_h} \|v - u\|_V). \tag{7.29}$$

Proof. We divide the proof into four steps. First, we give two basic estimations. Note that the Sobolev embedding theorem implies

$$\|v\|_{L^6(\Omega)} \leq C \|v\|_{H^1(\Omega)} \text{ for all } v \in H_0^1(\Omega) \tag{7.30}$$

and

$$\|u\|_{L^6(\Omega)} + \|u_h\|_{L^6(\Omega)} \leq C(\|u\|_{H^1(\Omega)} + \|u_h\|_{H^1(\Omega)}) \text{ for } h \ll 1. \tag{7.31}$$

Using the Hölder's inequality, we have

$$\begin{aligned} & |\beta((u_h - u)(u^2 + u_h u + u_h^2), v)| \\ \leq \ & C \|u_h - u\| (\|u\|_{L^6(\Omega)}^2 + \|u_h\|_{L^6(\Omega)}^2) \|v\|_{L^6(\Omega)} \text{ for all } v \in L^6(\Omega), \end{aligned}$$

which together with (7.30), (7.31), and (7.13) leads to

$$\begin{aligned} & |\beta((u_h - u)(u^2 + u_h u + u_h^2), v)| \\ \leq \ & C \|u_h - u\| \|v\|_V \text{ for all } v \in V_h. \end{aligned} \tag{7.32}$$

Obviously, the estimate (7.13) yields

$$\begin{aligned} & |(\lambda_h - \lambda)(u, v) + \lambda_h(u_h - u, v)| \\ \leq \ & C(| \lambda_h - \lambda | \, \|v\| + \|u_h - u\| \|v\|_V) \text{ for all } v \in V_h. \end{aligned} \tag{7.33}$$

Second, we establish an estimation for $u_h - u$. Note that (7.8) and (7.11) imply that for any $v \in V_h$,

$$\begin{aligned} & \frac{1}{2}(\nabla(u_h - u), \nabla v) + (\mathcal{V}_{ext}(u_h - u), v) \\ = \ & \lambda_h(u_h, v) - \lambda(u, v) - \beta(u_h^3 - u^3, v), \end{aligned}$$

where the facts that $u_h \geq 0$ and $u \geq 0$ are also used. Hence we have

$$\begin{aligned} & \frac{1}{2}(\nabla(u_h - u), \nabla v) + (\mathcal{V}_{ext}(u_h - u), v) \\ = \quad & (\lambda_h - \lambda)(u, v) + \lambda_h(u_h - u, v) \\ & -\beta((u_h - u)(u^2 + u_h u + u_h^2), v) \ \text{ for all } v \in V_h. \end{aligned} \tag{7.34}$$

Combining (7.32), (7.33), and (7.34), we obtain

$$\begin{aligned} & |\frac{1}{2}(\nabla(u_h - u), \nabla v) + (\mathcal{V}_{ext}(u_h - u), v)| \\ \leq \quad & C(| \lambda_h - \lambda | \, \|v\| + \|u_h - u\| \|v\|_V) \ \text{ for all } v \in V_h. \end{aligned} \tag{7.35}$$

Letting $P_h u \in V_h$ satisfying

$$\frac{1}{2}(\nabla(P_h u - u), \nabla v) + (\mathcal{V}_{ext}(P_h u - u), v) = 0 \ \text{ for all } v \in V_h \tag{7.36}$$

and setting $v = u_h - P_h u$ in (7.35), we arrive at

$$\|u_h - P_h u\|_V^2 \leq C(\|u_h - u\|^2 + | \lambda_h - \lambda | \, \|P_h u - u_h\|_V).$$

Thus from

$$\|u_h - u\|_V \leq \|u_h - P_h u\|_V + \|P_h u - u\|_V,$$

we have

$$\begin{aligned} & \|u_h - u\|_V^2 \\ & \quad \leq C(\|u_h - u\|^2 + | \lambda_h - \lambda | \, \|P_h u - u_h\| + \|P_h u - u\|_V^2). \end{aligned} \tag{7.37}$$

To complete the proof, third, we need also an estimation for $\lambda_h - \lambda$. By definition, there holds

$$\lambda_h = \frac{1}{2}(\nabla u_h, \nabla u_h) + (\mathcal{V}_{ext} u_h + \beta u_h^3, u_h).$$

Hence $\|u_h\| = 1$ and Lemma 7.2.2 yields

$$\begin{aligned} & \lambda_h - \lambda \\ = \quad & \frac{1}{2}(\nabla(u_h - u), \nabla(u_h - u)) + (\mathcal{V}_{ext}(u_h - u), u_h - u) \\ & +\beta(u^2(u_h - u), u_h - u) + \beta((u_h^2 - u^2)u_h, u_h) - \lambda(u_h - u, u_h - u). \end{aligned}$$

Since similar arguments show that

$$|\beta(u^2(u_h - u), u_h - u) + \beta((u_h^2 - u^2)u_h, u_h)|$$

can be bounded by

$$C(\|u_h - u\| \|u_h - u\|_V + \|u_h - u\|),$$

we obtain

$$| \lambda_h - \lambda | \leq C(\|u_h - u\| + \|u_h - u\|^2 + \|u_h - u\|_V^2). \tag{7.38}$$

Inserting (7.37) into (7.38), we get

$$\begin{aligned} | \lambda_h - \lambda | &\leq C(\|u_h - u\| + \|u_h - u\|^2 + \|P_h u - u\|_V^2) \\ &\quad + C | \lambda_h - \lambda | \, \|P_h u - u_h\|. \end{aligned}$$

Because of (7.13), we have the estimate

$$| \lambda_h - \lambda | \, \|P_h u - u_h\| \leq C\|u_h - u\| + | \lambda_h - \lambda | \, \|P_h u - u\|.$$

Hence

$$\begin{aligned} | \lambda_h - \lambda | &\leq C(\|u_h - u\| + \|u_h - u\|^2 + \|P_h u - u\|_V^2) \\ &\quad + C | \lambda_h - \lambda | \, \|P_h u - u\|. \end{aligned} \tag{7.39}$$

Finally, taking (7.9),

$$\|P_h u - u\| \leq C\|P_h u - u\|_V,$$

and Cea's Lemma (Lemma 2.3.1) that

$$\|P_h u - u\|_V \leq C \inf_{v \in V_h} \|u - v\|_V \tag{7.40}$$

into account, we then obtain

$$\|P_h u - u\| \ll 1 \text{ if } h \ll 1,$$

which together with (7.39) and (7.40) produces (7.28).

Note that a direct estimation of (7.37) shows

$$\|u_h - u\|_V^2 \leq C(\|u_h - u\|^2 + (\lambda_h - \lambda)^2 + \|P_h u - u\|_V^2),$$

or

$$\|u_h - u\|_V \leq C(\|u_h - u\| + | \lambda_h - \lambda | + \|P_h u - u\|_V). \tag{7.41}$$

Thus the estimate (7.29) is derived from (7.28), (7.40), and (7.41). This completes the proof. □

From (7.14), (7.28), and (7.29), we immediately obtain

Theorem 7.2.4. *If* $h \ll 1$*, then*

$$| \lambda_h - \lambda | \leq C(\|u_h - u\| + \inf_{v \in V_h} \|v - u\|_V^2), \tag{7.42}$$

$$\|u_h - u\|_V \leq C(\|u_h - u\| + \inf_{v \in V_h} \|v - u\|_V). \tag{7.43}$$

Consequently,

$$\lim_{h \to 0} \|u - u_h\|_V = 0. \tag{7.44}$$

It is shown by the above result that we may obtain the H^1-convergence of $u_h \to u$. The next result tells the error estimate.

Theorem 7.2.5. *If* $h \ll 1$, *then*

$$\begin{aligned}
\|u_h - u\| &\le C(h + \|u_h - u\|_{H^1(\Omega)})\|u_h - u\|_{H^1(\Omega)}, && (7.45)\\
|\ \lambda_h - \lambda\ | &\le C(h + \|u_h - u\|_{H^1(\Omega)}) \inf_{v \in V_h} \|v - u\|_V), && (7.46)\\
\|u_h - u\|_V &\le C \inf_{v \in V_h} \|v - u\|_V. && (7.47)
\end{aligned}$$

Proof. By a more sophisticated argument (see [69, 84] for more details), we have (7.45), which together with Theorem 7.2.4 produces (7.46) and (7.47). This completes the proof. □

We refer the readers to [81, 83, 84, 256] for the finite element analysis of other nonlinear eigenvalue problems resulting from electronic structure models.

7.3 Two-scale Discretization

In this section, we consider effective potential $\mathcal{V} = \mathcal{V}_{ext} + \mathcal{V}_0$ that satisfies $\mathcal{V}_0 \in L^\infty(\Omega)$ and

$$\mathcal{V}_{ext}(x) = -\sum_{j=1}^{N_{atom}} \frac{Z_j}{|x - r_j|} \tag{7.48}$$

with $r_i \in \Omega$, Z_j is some positive constant ($j = 1, 2, \cdots, N_{atom}$), and Ω is a bounded domain in $\mathbb{R}^3$.

The central computation in solving the Kohn-Sham equation is the repeated solution of (7.6) with some effective potential $\mathcal{V}$ that has a singular part as (7.48) when the exchange-correlation potential is approximated by X_α or local density approximation (LDA), for instance.

Define

$$a(w, v) = \int_\Omega \frac{1}{2}\nabla w \nabla v + \mathcal{V} w v \, \mathrm{d}x \quad \text{for all } w, v \in H_0^1(\Omega).$$

The weak form of (7.6) is: Find $(\lambda, u) \in \mathbb{R} \times H_0^1(\Omega)$ such that $\|u\| = 1$ and

$$a(u, v) = \lambda(u, v) \quad \text{for all } v \in H_0^1(\Omega). \tag{7.49}$$

The materials of this section are adapted from Gong, Shen, Zhang, and Zhou [134].

7.3.1 Regularity

To study the eigenpair of (7.49), we need the following result.

Lemma 7.3.1. *There is a constant $C > 0$ such that*

$$\|w\|_{H^1(\Omega)}^2 - C^{-1}\|w\|^2 \le 2a(w,w) \quad \textit{for all } w \in H_0^1(\Omega). \tag{7.50}$$

Proof. Using the uncertainty principle lemma (see page 169 of [217]):

$$\int_{\mathbb{R}^3} \frac{w^2(x)}{|x|^2}\,\mathrm{d}x \le 4\int_{\mathbb{R}^3} |\nabla w|^2\,\mathrm{d}x \quad \text{for all } w \in C_0^\infty(\mathbb{R}^3), \tag{7.51}$$

we obtain

$$\int_\Omega \frac{w(x)v(x)}{|x|}\,\mathrm{d}x \le 4\|\nabla w\|\|v\| \quad \text{for all } w,v \in H_0^1(\Omega), \tag{7.52}$$

which together with the Young's inequality produces

$$\begin{aligned}
&\sum_{j=1}^{N_{atom}} Z_j \int_\Omega \frac{w^2(x)}{|x-r_j|}\,\mathrm{d}x \\
\le\ & \frac{\|\nabla w\|^2}{2} + (8N_{atom}\sum_{j=1}^{N_{atom}} Z_j^2)\|w\|^2 \ \text{ for all } w \in H_0^1(\Omega).
\end{aligned}$$

Thus we obtain (7.50) from the definition of $a(\cdot,\cdot)$ and the assumption $\mathcal{V}_0 \in L^\infty(\Omega)$. This completes the proof. $\square$

It is seen from Lemma 7.3.1 that there is $\lambda > 0$ such that

$$C^{-1}\|w\|_{H^1(\Omega)}^2 \le a_\lambda(w,w) \quad \text{for all } w \in H_0^1(\Omega) \tag{7.53}$$

for some constant $C > 0$, where

$$a_\mu(w,v) = a(w,v) + \mu(w,v), \ \ w,v \in H_0^1(\Omega).$$

Note that (7.49) is equivalent to

$$a_\lambda(u,v) = E(u,v) \quad \text{for all } v \in H_0^1(\Omega) \tag{7.54}$$

with $E = \lambda + \mu$. Hence (7.49) has a countable sequence of real eigenvalues and the corresponding eigenfunctions in $H_0^1(\Omega)$.

Although the coefficient $\mathcal{V}$ of (7.6) is singular, we have the following result.

Theorem 7.3.2. *If $(\lambda, u) \in \mathbb{R} \times H_0^1(\Omega)$ is an eigenpair of (7.49), then $u \in H_0^1(\Omega) \cap W^{2,p}(\Omega)(2 \le p < q_0)$ for some $q_0 \in (2,3)$.*

Proof. Thanks to (7.51), we have that $\mathcal{V}u \in L^2(\Omega)$. Thus, we get from the regularity of Poisson's equation [128, 139] that

$$u = \left(-\frac{1}{2}\Delta\right)^{-1}(-\mathcal{V}u + \lambda u) \in H^2(\Omega),$$

which together with Sobolev embedding theorem leads to that $u \in C(\Omega)$.

Note that if R_0 is the diameter of Ω, then from

$$\begin{aligned}\int_\Omega \frac{u^p(x)}{|x-r_j|^p}\,\mathrm{d}x &\leq \|u\|^p_{L^\infty(\Omega)}\int_\Omega \frac{1}{|x-r_j|^p}\,\mathrm{d}x \\ &\leq C\|u\|^p_{L^\infty(\Omega)}\int_0^{r_j+R_0}\frac{1}{t^{p-2}}\,\mathrm{d}t,\end{aligned}$$

we obtain that $\mathcal{V}u \in L^p(\Omega)(2 \leq p < 3)$. Therefore there exists $q_0 \in (2,3)$ such that (see, e.g., [128, 139])

$$u = \left(-\frac{1}{2}\Delta\right)^{-1}(-\mathcal{V}u + \lambda u) \in W^{2,p}(\Omega) \quad \text{for all } p \in [2, q_0),$$

due to $-\mathcal{V}u + \lambda u \in L^p(\Omega)(2 \leq p < 3)$. This completes the proof. □

7.3.2 Scheme

Let $V_h \subset H^1_0(\Omega)$ be the piecewise linear finite element space associated with a shape-regular finite element mesh $\mathcal{T}_h$ over Ω. A standard finite element scheme for (7.49) may be viewed as a one-scale discretization: Find a pair of $(\lambda_h, u_h) \in \mathbb{R} \times V_h$ such that $\|u_h\| = 1$ and

$$a(u_h, v) = \lambda_h(u_h, v) \quad \text{for all } v \in V_h, \tag{7.55}$$

or equivalently

$$a_\mu(u_h, v) = E_h(u_h, v) \quad \text{for all } v \in V_h \tag{7.56}$$

with $E_h = \lambda_h + \mu$. One sees from (7.53) that (7.55) has a finite sequence of eigenvalues and the corresponding eigenfunctions in V_h.

Combining Theorem 7.3.2 and Babuška-Osborn theory (Theorems 1.4.7, 1.4.8, and 3.3.1; see also, e.g., [23]), we have the following error estimates for the one-scale discretization.

Theorem 7.3.3. *Let (λ, u) be a solution of (7.49). Then there is an associated solution (λ_h, u_h) of (7.55) satisfying*

$$\lambda_h - \lambda + \|u - u_h\| + h\|u - u_h\|_{H^1(\Omega)} \leq Ch^2. \tag{7.57}$$

To reduce the computational cost, we will now introduce a two-scale discretization scheme. The two-scale finite element discretization approach for eigenvalue problems may be dated back to [191] (see also a general formwork in [249] when $a(\cdot,\cdot)$ is a positive symmetric definite bilinear form). In this subsection, we will modify and generalize the standard two-scale finite element discretization approach in [249] to solve (7.49). With our two-scale scheme, the solution of an eigenvalue problem with singular coefficient on a fine grid is reduced to the solution of an eigenvalue problem with singular coefficient on a much coarser grid and a solution of linear algebraic system associated with Poisson's equation on the fine grid.

Let $H \gg h$ and assume that $V_H \subset V_h$. We consider the approximation of any eigenvalue λ of (7.49). Here and hereafter we let λ_H be a finite element eigenvalue of (7.55) corresponding to V_H and satisfy

$$|\lambda_H - \lambda| \leq CH^2. \tag{7.58}$$

Our two-scale discretization scheme for (7.49) is constructed as follows:

Step 1. Find $(\lambda_H, u_H) \in \mathbb{R} \times V_h$ such that $\|u_H\| = 1$ and

$$a(u_H, v) = \lambda_H(u_H, v) \quad \text{for all } v \in V_H.$$

Step 2. Find $u^h \in V_h$ satisfying

$$\frac{1}{2}(\nabla u^h, \nabla v) = \lambda_H(u_H, v) - (\mathcal{V}u_H, v) \quad \text{for all } v \in V_h.$$

Step 3. Compute the Rayleigh quotient:

$$\lambda^h = \frac{a(u^h, u^h)}{(u^h, u^h)}.$$

It is seen from Babuška-Osborn theory (Section 1.4.2) that, associated with the eigenfunction u_H obtained by Step 1 in the two-scale scheme, there exists an exact eigenfunction u of (7.49) satisfying $\|u\| = 1$ and

$$\|u - u_H\| + H\|u - u_H\|_{H^1(\Omega)} \leq CH^2. \tag{7.59}$$

For this two-scale scheme, the resulting approximation still maintains an optimal accuracy. Indeed, we have the following theorem.

Theorem 7.3.4. *Let (λ^h, u^h) be obtained from the two-scale discretization scheme. If $H = \mathcal{O}(h^{1/2})$, then there exists an eigenpair of* (7.49) *satisfying $\|u\| = 1$ and* (7.59) *such that*

$$|\lambda - \lambda^h| + h\|u - u^h\|_{H^1(\Omega)} \leq Ch^2. \tag{7.60}$$

Proof. Let $P_h : H_0^1(\Omega) \to V_h$ be defined by

$$a_\mu(w - P_h w, v) = 0 \;\text{ for all } v \in V_h,\; w \in H_0^1(\Omega),$$

then (see Cea's Lemma (Lemma 2.3.1) and the Aubin-Nitsche Lemma (Theorem 3.2.4))

$$\|w - P_h w\| + h\|\nabla(w - P_h w)\| \leq Ch^2. \tag{7.61}$$

We obtain from (7.52) that

$$|(\mathcal{V}(u_H - P_h u), v)| \leq C\|u_H - P_h u\|\|v\|_{H^1(\Omega)} \quad \text{for all } v \in H_0^1(\Omega). \tag{7.62}$$

From the construction of u^h, we immediately obtain

$$\begin{aligned} \frac{1}{2}(\nabla(u^h - P_h u), \nabla v) &= (\lambda_H - \lambda)(u, v) + \lambda_H(u_H - u, v) \\ &\quad +(\mathcal{V}(P_h u - u_H), v) \text{ for all } v \in V_h, \end{aligned}$$

which together with (7.62) leads to

$$\begin{aligned} &\|\nabla(u^h - P_h u)\| \\ \leq \ & C\left(|\lambda_H - \lambda| + \lambda_H\|u_H - u\| + \|u_H - P_h u\|\right). \end{aligned} \tag{7.63}$$

Using (7.58), (7.59), and the inverse inequality, we then get

$$\|\nabla(u^h - P_h u)\| \leq CH^2 + C\|u_H - P_h u\|.$$

Thus, combining (7.61) and (7.59), we arrive at

$$\|\nabla(u^h - P_h u)\| \leq CH^2,$$

and

$$\|\nabla(u^h - u)\| \leq CH^2, \tag{7.64}$$

which together with Lemma 7.2.2, completes the proof. □

We may obtain similar results for the following scheme (c.f. [249]):

Step 1. Find $(\lambda_H, u_H) \in \mathbb{R} \times V_H$ such that $\|u_H\| = 1$ and

$$a(u_H, v) = \lambda_H(u_H, v) \quad \text{for all } \ v \in V_H.$$

Step 2. Find $u^h \in V_h$ satisfying

$$a(u^h, v) = \lambda_H(u_h, v) \quad \text{for all } \ v \in V_h.$$

Step 3. Compute the Rayleigh quotient:

$$\lambda^h = \frac{a(u^h, u^h)}{(u^h, u^h)}.$$

Finally, we mention that there are other efficient schemes for solving (7.49). For instance, we may refer to [250] for local and parallel versions of the two-scale finite element schemes, [85, 106, 123, 193] for multilevel discretizations, and [104, 120, 194, 205, 224] for correction approaches.

Chapter 8

Adaptive Finite Element Approximations

8.1 Introduction

An adaptive mesh-refining algorithm usually consists of the following loop:

Solve $\rightarrow$ **Estimate** $\rightarrow$ **Mark** $\rightarrow$ **Refine**.

Solve. This step computes the piecewise polynomial finite element approximation with respect to a given mesh.

Estimate. Given a partition $\mathcal{T}_h$ and the corresponding output from the "Solve" step, "Estimate" computes some a posteriori error estimator.

Mark. We will replace the subscript h (or h_k) by an iteration counter k whenever convenient afterwards. Based on the a posteriori error indicators, "Mark" gives a strategy to choose a subset of elements $\mathcal{M}_k$ of $\mathcal{T}_k$ for refinement. One of the most widely used marking strategies to enforce error reduction is the so-called Dörfler strategy. A weaker strategy, which is called "Maximum Strategy," only requires that the set of marked elements $\mathcal{M}_k$ contains at least one element of $\mathcal{T}_k$ holding the largest value estimator. Note that the most commonly used marking strategies, e.g., Dörfler strategy and Equidistribution strategy, fulfill this condition.

Refine. Given the partition $\mathcal{T}_k$ and the set of marked elements $\mathcal{M}_k$, "Refine" produces a new partition $\mathcal{T}_{k+1}$ by refining all elements in $\mathcal{M}_k$ at least one time. Usually, people restrict themselves to a shape-regular bisection for the refinement. Defining

$$\mathcal{R}_{\mathcal{T}_k \to \mathcal{T}_{k+1}} = \mathcal{T}_k \backslash (\mathcal{T}_k \cap \mathcal{T}_{k+1}) \tag{8.1}$$

as the set of refined elements, we see that $\mathcal{M}_k \subset \mathcal{R}_{\mathcal{T}_k \to \mathcal{T}_{k+1}}$. Note that usually more than the marked elements in $\mathcal{M}_k$ are refined in order to keep the mesh conforming.

It is important in adaptive finite element computations to construct efficient and reliable error estimators. In this chapter, we mainly focus on construction and analysis of the residual-based a posteriori error estimators for finite element approximations to Laplace eigenvalue problems, from which we then present an adaptive finite element method. We start from the approximation to Poisson's equation and then the Laplace eigenvalue problem based on a so-called perturbation argument [105, 145].

8.2 A Posteriori Error Analysis for Poisson's Equation

Recall that Poisson's equation is

$$-\triangle u = f \quad \text{in } \Omega \tag{8.2}$$

with homogeneous Dirichlet boundary condition

$$u = 0 \quad \text{on } \partial\Omega, \tag{8.3}$$

where f is a given function and u is unknown. The associated weak formulation is: Find $u \in H_0^1(\Omega)$ such that

$$a(u, v) = (f, v) \quad \text{for all } v \in H_0^1(\Omega), \tag{8.4}$$

where

$$a(u, v) := (\nabla u, \nabla v).$$

Consider a shape-regular triangulation $\mathcal{T}_h$ of domain $\Omega \subset \mathbb{R}^n (n = 2, 3)$ with polygonal boundary $\partial\Omega$ with mesh size h. We define

$$V_h = \left\{ v \in H_0^1(\Omega) : v\,|_K \text{ is affine for all } K \in \mathcal{T}_h \right\},$$

a test and trial space. The Galerkin solution $u_h \in V_h$ satisfies

$$\int_\Omega \nabla u_h \nabla v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x \ \text{ for all } v \in V_h. \tag{8.5}$$

To state the a posteriori error estimates, we introduce some notation. Let $\partial\mathcal{T}_h$ be the set of all the interior edges or faces of the mesh $\mathcal{T}_h$ and

$$\partial\mathcal{T}_h(K) = \{F \in \partial\mathcal{T}_h : F \subset \overline{K}\}.$$

8.2.1 Residual Estimators

We need the following results (cf. Bernardi and Girault [30] and Clément [90]):

Lemma 8.2.1. *For any $K \in \mathcal{T}_h$, which is a shape-regular mesh, there exists a macro-element $\sigma_K \subset \Omega$ being a union of elements of $\mathcal{T}_h$ that contains K and satisfies $h_{\sigma_K} \leq Ch_K$. Moreover, there exists an operator $\Pi_h : L^2(\Omega) \longrightarrow V_h$ such that $\Pi_h H_0^1(\Omega) \subset V_h$ and*

$$\begin{aligned} \|w - \Pi_h w\|_{L^2(K)} &\leq Ch_K \|\nabla w\|_{L^2(\sigma_K)} \quad \textit{for all } w \in H^1(K), K \in \mathcal{T}_h, \\ \|\nabla \Pi_h w\|_{L^2(K)} &\leq C\|\nabla w\|_{L^2(\sigma_K)} \quad \textit{for all } w \in H^1(K), K \in \mathcal{T}_h. \end{aligned}$$

For $F \in \partial\mathcal{T}_h$, we set

$$\omega_F = \cup\{K' \in \mathcal{T}_h : F \in \partial T^h(K')\}.$$

Let $\mathbf{n}_F$ be a unit vector normal to F, and define for $v \in V_h$

$$\left[\frac{\partial v}{\partial \mathbf{n}_F}\right]_F = \lim_{s\to 0^+} \mathbf{n}_F^t((\nabla v)(x + s\mathbf{n}_F) - (\nabla v)(x - s\mathbf{n}_F)),$$

$$J_F(v) = \left|\left[\frac{\partial v}{\partial \mathbf{n}_F}\right]_F\right|,$$

that is, $J_F(v)$ is the jump across F in the normal component of ∇v. For $K \in \mathcal{T}_h$, we introduce $\eta_K(v), \eta^K(v)$ given by

$$\eta_K(v) = \|hR_K(v)\|_{L^2(K)} + \frac{1}{2}\left(\sum_{F\in\partial T^h(K)} \|h^{1/2}J_F(v)\|^2_{L^2(F)}\right)^{1/2} \tag{8.6}$$

and

$$\eta^K(v) = \|hR^K(v)\|_{L^2(K)} + \frac{1}{2}\left(\sum_{F\in\partial T^h(K)} \|h^{1/2}J_F(v)\|^2_{L^2(F)}\right)^{1/2}, \tag{8.7}$$

respectively. Here

$$R_K(v) = f_h + \triangle v \quad \text{and} \quad R^K(v) = f + \triangle v$$

with $f_h \in P_K^r$.

One sees that $\eta_K(u_h)$ and $\eta^K(u_h)$ are computable in terms of the finite element solution u_h.

We state the following basic results, which can be proved by the standard scaling arguments.

Lemma 8.2.2. *Let $K \in \mathcal{T}_h$, which is a shape-regular mesh, and $F \in \partial\mathcal{T}_h$.*

1. There exists a polynomial $\lambda_K \in H_0^1(K)$ such that for all $v \in P_K^r$, there hold

$$\|\lambda_K v\|_{L^2(K)}^2 \leq C\|v\|_{L^2(K)}^2 \leq C(v, \lambda_K v)_K, \tag{8.8}$$

$$\|\nabla(\lambda_K v)\|_{L^2(K)} \leq C\|h^{-1}v\|_{L^2(K)}. \tag{8.9}$$

2. There exists a polynomial $\mu_{_F} \in H_0^1(\omega_{_F})$ such that for all $v \in P_{_F}^r$, there hold

$$\|v\|_{L^2(F)}^2 \quad \leq \quad C(v, \mu_{_F} v)_{_F}, \tag{8.10}$$

$$\|\mu_{_F} v\|_{L^2(\omega_{_F})} \quad \leq \quad C\|h^{1/2}v\|_{L^2(F)}, \tag{8.11}$$

$$\|\nabla(\mu_{_F} v)\|_{L^2(\omega_{_F})} \quad \leq \quad C\|h^{-1/2}v\|_{L^2(F)}. \tag{8.12}$$

Proof. Let $x_0, x_1, x_d (d = 2, 3)$ be the nodes of K, and $\varphi_0, \varphi_1, \cdots, \varphi_d$ be the associated basis satisfying $\varphi_i(x_j) = \delta_{ij}(i, j = 0, 1, 2 \cdots, d)$. We see that

$$\lambda_K = \prod_{i=0}^{d} \varphi_i, \quad \mu_K = \prod_{i \neq j} \varphi_i$$

match the requirements, where F is not the face containing nodal x_j. $\square$

8.2.2 Upper Bound

First, we want to present an a posteriori error estimator for the upper bound.

Theorem 8.2.3. *There holds*

$$\|\nabla(u - u_h)\| \leq C\left(\sum_{K \in \mathcal{T}_h} (\eta^K(u_h))^2\right)^{1/2}. \tag{8.13}$$

Proof. We see that for any $\phi \in H_0^1(\Omega)$ and $v \in V_h$, there holds

$$\begin{aligned} a(u - u_h, \phi) &= a(u - u_h, \phi - v) \\ &= \sum_{K \in \mathcal{T}_h} \left(\int_K R_K(u_h)(\phi - v)\,\mathrm{d}x \right. \\ &\qquad \left. - \sum_{F \in \partial T^h(K)} \int_F \mathbf{n}_{_F}^{^T} \nabla u_h(\phi - v)\,\mathrm{d}s \right). \end{aligned} \tag{8.14}$$

So we need to estimate the two terms in (8.14). Note that for any $F \in \partial T^h(K)$, there exists $K' \in \mathcal{T}_h$ such that $F \in \partial T^h(K')$. Thus

$$
\begin{aligned}
& \inf_{v\in V_h} \left| \sum_{K\in\mathcal{T}_h} \sum_{F\in\partial T^h(K)} \int_F \mathbf{n}_{_F}^T \nabla u_h (\phi - v)\, \mathrm{d}s \right| \\
\leq \quad & \frac{1}{2} \inf_{v\in V_h} \sum_{F\in\partial\mathcal{T}_h} \int_F J_{_F}(u_h) \mid \phi - v \mid \mathrm{d}s \\
\leq \quad & C \inf_{v\in V_h} \sum_{F\in\partial\mathcal{T}_h} \|J_{_F}(u_h)\|_{L^2(F)} \|\phi - v\|_{L^2(F)}.
\end{aligned}
$$

And we then have

$$
\begin{aligned}
& \inf_{v\in V_h} \left| \sum_{K\in\mathcal{T}_h} \sum_{F\in\partial T^h(K)} \int_F \mathbf{n}_{_F}^T \nabla u_h (\phi - v)\, \mathrm{d}s \right| \\
\leq \quad & C \inf_{v\in V_h} \sum_{F\in\partial\mathcal{T}_h} h_K^{1/2} \|J_{_F}(u_h)\|_{L^2(F)} \left(\|h^{-1}(\phi - v)\|_{L^2(F)} + \|\phi - v\|_{H^1(F)} \right) \\
\leq \quad & C \sum_{F\in\partial\mathcal{T}_h} \|h^{1/2} J_{_F}(u_h)\|_{L^2(F)} \|\nabla\phi\|_{L^2(\sigma_{_F})} \\
\leq \quad & C \left(\sum_{F\in\partial\mathcal{T}_h} \|h^{1/2} J_{_F}(u_h)\|^2_{L^2(F)} \right)^{1/2} \|\nabla\phi\|.
\end{aligned}
$$

Namely,

$$
\begin{aligned}
& \inf_{v\in V_h} \left| \sum_{K\in\mathcal{T}_h} \sum_{F\in\partial T^h(K)} \int_F \mathbf{n}_{_F}^T \nabla u_h (\phi - v)\, \mathrm{d}s \right| \\
\leq \quad & C \left(\sum_{F\in\partial\mathcal{T}_h} \|h^{1/2} J_{_F}(u_h)\|^2_{L^2(F)} \right)^{1/2} \|\nabla\phi\|. \qquad (8.15)
\end{aligned}
$$

Since

$$
\begin{aligned}
& \inf_{v\in V_h} \sum_{K\in\mathcal{T}_h} \int_K R_K(u_h)(\phi - v) \\
\leq \quad & C \sum_{K\in\mathcal{T}_h} \|h R_K(u_h)\|_{L^2(K)} \|\nabla\phi\|_{L^2(\sigma_K)} \\
\leq \quad & C \left(\sum_{K\in\mathcal{T}_h} \|h R_K(u_h)\|^2_{L^2(K)} \right)^{1/2} \|\nabla\phi\|,
\end{aligned}
$$

we conclude from (8.14), (8.15), and the above estimation that

$$| a(u-u_h,\phi) | \leq C\left(\sum_{K\in\mathcal{T}_h}(\eta^K(u_h))^2\right)^{1/2}\|\nabla\phi\|.$$

This completes the proof. □

We mention that (8.13) can be localized in the sense of ignoring some higher order global term. We refer to [248] for more details.

8.2.3 Lower Bound

Then we turn to show the lower bound for the error, which is localized.

Theorem 8.2.4. *There holds*

$$\|\nabla(u-u_h)\| \leq C\left(\sum_{K\in\mathcal{T}_h}(\eta_K(u_h))^2+\|h(f-f_h)\|^2_{L^2(K)}\right)^{1/2}. \tag{8.16}$$

Moreover, for any $K\in\mathcal{T}_h$,

$$\|hR_K(u_h)\|_{L^2(K)} \leq C\left(\|\nabla(u-u_h)\|_{L^2(K)}+\|h(f-f_h)\|_{L^2(K)}\right) \tag{8.17}$$

and for any $F\in\partial\mathcal{T}_h$,

$$\begin{aligned}&\|h^{1/2}J_{_F}(u_h)\|_{L^2(F)}\\ &\leq C\left(\|\nabla(u-u_h)\|_{L^2(\omega_{_F})}+\left(\sum_{K'\subset\omega_{_F}}\|h(f-f_h)\|^2_{L^2(K')}\right)^{1/2}\right).\end{aligned} \tag{8.18}$$

Proof. For $K\in\mathcal{T}_h$, setting $\phi=\phi_K\equiv\lambda_K R_K(u_h)$ in

$$a(u-u_h,\phi)=\sum_{K\in\mathcal{T}_h}\left(\int_K R_K(u_h)\phi-\sum_{F\in\partial T^h(K)}\int_F \mathbf{n}_{_F}^{^T}\nabla u_h\phi\right), \tag{8.19}$$

we have

$$(R_K(u_h),\phi_K)_K = a(u-u_h,\phi_K)+(f_h-f,\phi_K)_K. \tag{8.20}$$

Lemma 8.2.2 implies that

$$\begin{aligned}a(u-u_h,\phi_K) &\leq C\|\nabla(u-u_h)\|_{L^2(K)}\|\nabla\phi_K\|_{L^2(K)}\\ &\leq C\|\nabla(u-u_h)\|_{L^2(K)}\|h^{-1}R_K(u_h)\|_{L^2(K)}\end{aligned}$$

and we get (8.17).

Next considering any $F \in \partial\mathcal{T}_h$ and setting $\phi = \phi_{_F} := \mu_F J_{_F}(u_h)$ in (8.19), we obtain

$$\begin{aligned} & -(J_{_F}(u_h), \phi_{_F})_F \\ = \quad & a(u - u_h, \phi_{_F}) - \sum_{K' \subset \omega_{_F}} (R_{K'}(u_h) + f - f_h, \phi_{_F})_K. \end{aligned}$$

Note that Lemma 8.2.2 also leads to

$$\begin{aligned} & a(u - u_h, \phi_{_F}) \le C\|\nabla(u - u_h)\|_{L^2(\omega_{_F})}\|\nabla\phi_{_F}\|_{L^2(\omega_{_F})} \\ \le \quad & C\|\nabla(u - u_h)\|_{L^2(\omega_{_F})}\|h^{-1/2}J_{_F}(u_h)\|_{L^2(F)} \end{aligned}$$

and

$$\begin{aligned} & \sum_{K' \subset \omega_{_F}} (R_{K'}(u_h) + f - f_h, \phi_{_F})_K \\ \le \quad & C\alpha\left(\sum_{K' \subset \omega_{_F}} \|hR_{K'}(u_h)\|^2_{L^2(K')} + \|h(f - f_h)\|^2_{L^2(K')}\right)^{1/2} \\ \le \quad & C\alpha\left(\|\nabla(u - u_h)\|^2_{L^2(\omega_{_F})} + \sum_{K' \subset \omega_{_F}} \|h(f - f_h)\|^2_{L^2(K')}\right)^{1/2}, \end{aligned}$$

where

$$\alpha = \|h^{-1}\phi_{_F}\|_{L^2(\omega_{_F})}.$$

Therefore, Lemma 8.2.2 and the above inequality produce (8.18). □

8.3 A Posteriori Error Analysis for the Laplace Eigenvalue Problem

The Dirichlet eigenvalue problem is: Find λ and u such that

$$-\triangle u = \lambda u \quad \text{in } \Omega, \tag{8.21}$$

where u satisfies the boundary condition (8.3). The weak formulation for the eigenvalue problem is to find $\lambda \in \mathbb{R}$ and non-trivial $u \in H^1_0(\Omega)$ such that

$$a(u, v) = \lambda(u, v) \quad \text{for all } v \in H^1_0(\Omega), \tag{8.22}$$

where $a(u, v) = (\nabla u, \nabla v)$.

Assume that $\Omega \subset \mathbb{R}^n (n = 2, 3)$ is covered by a regular triangular (tetrahedron) mesh $\mathcal{T}$. Let V_h be the finite element space using certain Lagrange elements with zero values for the nodes on $\partial\Omega$.

Recall that the discrete Dirichlet eigenvalue problem is: Find $(\lambda_h, u_h) \in \mathbb{R} \times V_h$ such that

$$a(u_h, v) = \lambda_h (u_h, v) \quad \text{for all } v \in V_h. \tag{8.23}$$

We see from Babuška and Osborn theory (Section 1.4.2) that for any eigenvector u_h of (8.23), there is an eigenpair (λ, u) of (8.22) satisfying

$$\|u - u_h\| + h\|\nabla(u - u_h)\| \leq Ch \min_{v \in V_h} \|\nabla(u - v)\|$$

and

$$\lambda \leq \lambda_h \leq \lambda + C \min_{v \in V_h} \|\nabla(u - v)\|.$$

Let $P_h : H_0^1(\Omega) \to V_h$ be the Galerkin projection defined by

$$a(P_h u - u, v) = 0 \quad \text{ for all } (u, v) \in H_0^1(\Omega) \times V_h. \tag{8.24}$$

We recall the Cea's Lemma (Lemma 2.3.1) and the Aubin-Nitsche Lemma (Theorem 3.2.4) and have that

$$\|u - P_h u\| + h\|\nabla(u - P_h u)\| \leq Ch \min_{v \in V_h} \|\nabla(u - v)\|. \tag{8.25}$$

There are some close relationships between the Ritz-Galerkin projection P_h of the eigenvector and the finite element approximation to the eigenvector.

Theorem 8.3.1. *There holds*

$$\|\nabla(P_h u - u_h)\| \leq C\left(\lambda - \lambda_h + \lambda\|u - u_h\|\right). \tag{8.26}$$

Proof. We get (8.26) from identity

$$a(P_h u - u_h, v) = (\lambda - \lambda_h)(u_h, v) + \lambda(u - u_h, v) \quad \text{ for all } v \in V_h. \tag{8.27}$$

□

Let $T : L^2(\Omega) \to H_0^1(\Omega)$ be defined as

$$a(Tw, v) = (w, v) \quad \text{for all } v \in H_0^1(\Omega).$$

Then (8.23) can be written as

$$u_u = \lambda_h T u_h.$$

We have for $w^h = \lambda_h u_h$ that

$$u_h = P_h w^h. \tag{8.28}$$

The following conclusion is simple but useful [105].

Theorem 8.3.2. *There exists $\kappa(h) \in (0,1)$ such that $\kappa(h) \to 0$ as $h \to 0$ and*

$$\begin{aligned} & \|\nabla(u-u_h)\| \\ = \quad & \|\nabla(w^h - P_h w^h)\| + \mathcal{O}(\kappa(h))\|\nabla(u-u_h)\|. \end{aligned} \tag{8.29}$$

Proof. By the definition of w^h and (8.28), we have

$$\begin{aligned} u - w^h &= \lambda T u - \lambda_h T u_h \\ &= (\lambda - \lambda_h) T u + \lambda_h T(u - u_h). \end{aligned}$$

Note that

$$\|u - u_h\| \leq Ch\|\nabla(u-u_h)\|,$$

and

$$\lambda - \lambda_h \leq C\|\nabla(u-u_h)\|^2.$$

We get (8.29) when we set

$$\kappa(h) = h + \|\nabla(u-u_h)\|.$$

This completes the proof. □

For $K \in \mathcal{T}_h$, we now introduce $\eta_K(v)$ by

$$\eta_K(v) = \|hR_K(v)\|_{L^2(K)} + \frac{1}{2}\left(\sum_{F\in\partial T^h(K)} \|h^{1/2}J_{F}(v)\|^2_{L^2(F)}\right)^{1/2}, \tag{8.30}$$

where

$$R_K(v) = \lambda_h u_h + \triangle v.$$

Given a subset $\mathcal{T}' \subset \mathcal{T}_h$, we define the error estimator $\eta_h(u_h, \mathcal{T}')$ by

$$\eta_h^2(u_h, \mathcal{T}') = \sum_{K\in\mathcal{T}'} \eta_K^2(u_h). \tag{8.31}$$

Theorem 8.3.3. *Let h_0 be small enough and $h \in (0, h_0]$. There exist constants C_1 and C_2, which only depend on the shape regularity constant γ^*, such that*

$$C_2\eta_h^2(u_h, \mathcal{T}_h) \leq \|\nabla(u-u_h)\|^2 \leq C_1\eta_h^2(u_h, \mathcal{T}_h). \tag{8.32}$$

Proof. Recall that $-\triangle w^h = \lambda_h u_h$. From (8.13) and (8.16) we have

$$\|\nabla(w^h - P_h w^h)\|^2 \leq \tilde{C}_1\eta_h^2(P_h w^h, \mathcal{T}_h) \tag{8.33}$$

and

$$\tilde{C}_2\eta_h^2(P_h w^h, \mathcal{T}_h) \leq \|\nabla(P_h w^h - w^h)\|^2 \tag{8.34}$$

when we set $f = f_h = \lambda_h u_h$. Thus we obtain (8.32) from (8.28), (8.29), (8.33), and (8.34). In particular, we may choose C_1, C_2, and C_3 satisfying

$$C_1 = \tilde{C}_1(1 + \tilde{C}\tilde{\kappa}(h_0))^2,\ C_2 = \tilde{C}_2(1 - \tilde{C}\tilde{\kappa}(h_0))^2. \tag{8.35}$$

This completes the proof. □

We should point out that for non-piecewise constant partial differential operators, some oscillation terms must be involved in the a posteriori error estimators for the finite element approximations of associated eigenvalue problems [105].

8.4 Adaptive Algorithm

For convenience in the following discussion, we replace the subscript h by an iteration counter called k.

With the a posteriori error estimators constructed in the above section, we can design the following adaptive finite element algorithm to solve the Laplace eigenvalue problem (8.21):

Choose a parameter $0 < \theta < 1$:

1. Pick an initial mesh $\mathcal{T}_0$ and let $k = 0$.
2. Solve (8.23) on $\mathcal{T}_k$ for the discrete solution u_k.
3. Compute the local indicators $\{\eta_k(u_k, K) : K \in \mathcal{T}_k\}$.
4. Construct $\mathcal{M}_k \subset \mathcal{T}_k$ by **Marking Strategy** and parameter θ.
5. Refine $\mathcal{T}_k$ to get a new conforming mesh $\mathcal{T}_{k+1}$ by **REFINE**.
6. Let $k = k + 1$ and go to Step 2.

We point out that **Marking Strategy** is crucial for our adaptive computations. The so-called Dörfler strategy is stated as follows:

Given a parameter $0 < \theta < 1$:

1. Construct a subset $\mathcal{M}_k$ of $\mathcal{T}_k$ by selecting some elements in $\mathcal{T}_k$ such that

$$\eta_k(u_k, \mathcal{M}_k) \geq \theta \eta_k(u_k, \mathcal{T}_k). \tag{8.36}$$

2. Mark all the elements in $\mathcal{M}_k$.

The Maximum strategy, however, is to construct the set of marked elements $\mathcal{M}_k$ containing at least one element $K_k^{\max} \in \mathcal{M}_k$ such that

$$\eta_k(u_k, K_k^{\max}) = \max_{K \in \mathcal{T}_k} \eta_k(u_k, K). \tag{8.37}$$

We see that the Dörfler strategy and Equidistribution strategy fulfill this condition.

The above adaptive algorithm and the marking strategy are set for the approximation to simple eigenvalues. For adaptive finite element approximations to multiple eigenvalues or eigenvalue clusters, we have to do some relevant modifications [103].

It requires several sophisticated techniques to carry out the analysis of the convergence and complexity of adaptive finite element approximations, which is out of the scope of this book. When the initial finite element mesh is sufficiently fine, people are able to prove the convergence and get the complexity of adaptive finite element approximations for eigenvalue problems of a class of elliptic partial differential operators. We refer to [103, 105, 145] for details, in which some relationship between the two level finite element approximations and the perturbation arguments are particularly used.

We will point out that the Maximum strategy is sufficient to get the convergence of the adaptive finite element approximations [124]. To obtain the convergence rate of the approximations, we require the Dörfler strategy; while to show the complexity, the marked $\mathcal{M}_k$ should satisfy (8.36) with minimal cardinality.

We will also mention that the similar conclusions hold true for nonlinear eigenvalue problems in modeling electronic structures [80, 82].

Chapter 9

Matrix Eigenvalue Problems

9.1 Introduction

Finite element methods for eigenvalue problems lead to matrix eigenvalue problems. They are usually generalized eigenvalue problems, which are large and sparse. There exist many algorithms in numerical linear algebra for matrices with different properties. It is always helpful to know these properties before we choose the algebraic eigenvalue solver. Excellent books on matrix computation include, for example, [132, 221, 235]. See also the survey paper [133] and references therein. The material in this chapter is classical and can be found in [132, 79, 221].

We start with some fundamentals of matrices. Let $\mathbb{C}^n$ be the complex n-dimensional space of column vectors. Let $\mathbf{x} \in \mathbb{C}^n$. We have that

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad \mathbf{x}^H = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n).$$

The scalar product on $\mathbb{C}^n \times \mathbb{C}^n$ is defined as

$$(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i \overline{y}_i = \mathbf{y}^H \mathbf{x}.$$

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if $(\mathbf{x}, \mathbf{y}) = 0$.

Let the set of vectors $\{\mathbf{x}_i\}_{i=1}^n$ be a basis for $\mathbb{C}^n$. We say the basis is orthonormal if

$$(\mathbf{x}_i, \mathbf{x}_j) = \delta_{ij}, \quad i, j = 1, 2, \ldots, n.$$

If the basis is not orthonormal, then there exists an adjoint basis $\{\mathbf{y}_i\}_{i=1}^n$ of $\{\mathbf{x}_i\}_{i=1}^n$ such that $\{\mathbf{y}_i\}_{i=1}^n$ is a basis of $\mathbb{C}^n$ and

$$(\mathbf{x}_i, \mathbf{y}_j) = \delta_{ij}, \quad i, j = 1, 2, \ldots, n.$$

For a vector $\mathbf{x} \in \mathbb{C}^n$, the Hölder norm for $p \geq 1$ is defined as

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}.$$

In particular, we have that

$$\begin{aligned}
\|\mathbf{x}\|_1 &= |x_1| + |x_2| + \ldots + |x_n|, \\
\|\mathbf{x}\|_2 &= \left(x_1^2 + x_2^2 + \ldots + x_n^2 \right)^{1/2}, \\
\|\mathbf{x}\|_\infty &= \max_{i=1,\ldots,n} |x_i|.
\end{aligned}$$

We denote by $A_{n \times n} = (a_{i,j}), i, j = 1, 2, \ldots, n$, or simply A, an $n \times n$ matrix. Let A^T and A^H be the transpose and the conjugate transpose of A, respectively. We denote by $I_{n \times n}$, or simply I, the $n \times n$ identity matrix.

Definition 9.1.1. *An $n \times n$ matrix A is said to be*

- *symmetric if $A^T = A$,*
- *Hermitian if $A^H = A$,*
- *skew-symmetric if $A^T = -A$,*
- *skew-Hermitian if $A^H = -A$,*
- *normal if $A^H A = AA^H$,*
- *unitary if $A^H A = I$.*

The inverse matrix of A, if exists, is denoted by A^{-1} such that

$$A^{-1} A = AA^{-1} = I.$$

Let $\det(A)$ denote the determinant of A. A matrix A is said to be non-singular if $\det(A) \neq 0$. Otherwise, A is singular. For determinants, we have that

$$\det(AB) = \det(BA)$$

and

$$\det(\overline{A}) = \overline{\det(A)}.$$

A matrix $A \in \mathbb{C}^{n\times n}$ defines a mapping from $\mathbb{C}^n$ to $\mathbb{C}^n$. The induced operator 2-norm of A is defined as

$$\|A\| = \max_{\mathbf{x}\in\mathbb{C}^n, \mathbf{x}\neq 0} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2}. \tag{9.1}$$

Similarly, $\|A\|_1$ and $\|A\|_\infty$ are defined as

$$\|A\|_1 = \max_{\mathbf{x}\in\mathbb{C}^n, \mathbf{x}\neq \mathbf{0}} \frac{\|A\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|$$

and

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{N} |a_{ij}| = \|A^H\|_1,$$

respectively.

The condition number of a regular matrix A is defined as

$$\text{cond}(A) = \|A\| \cdot \|A^{-1}\|.$$

Note that the norm can be 2-norm, 1-norm, or ∞-norm.

The null space (or kernel) of A is

$$\text{Ker}(A) = \{\mathbf{x} \in \mathbb{C}^n \mid A\mathbf{x} = \mathbf{0}\}$$

and the range of A is

$$R(A) = \{\mathbf{y} \in \mathbb{C}^n \mid \mathbf{y} = A\mathbf{x} \quad \text{for some } \mathbf{x} \in \mathbb{C}^n\}.$$

A matrix P is called a projection if $P^2 = P$. We have that

$$\mathbb{C}^n = R(P) + \text{Ker}(P).$$

Let M be a closed subspace of $\mathbb{C}^n$. The orthogonal complement of M is defined as

$$M^\perp = \{\mathbf{x} \in \mathbb{C}^n \mid (\mathbf{x}, \mathbf{y}) = 0 \text{ for all } \mathbf{x} \in M\}.$$

For an $n \times n$ matrix A, we have that

$$\text{Ker}(A^H) = (R(A))^\perp, \quad R(A^H) = (\text{Ker}(A))^\perp.$$

The characteristic polynomial of a matrix A is defined as

$$p(\lambda) = \det(A - \lambda I). \tag{9.2}$$

Definition 9.1.2. *A complex number λ is called an eigenvalue of A if there exists a nonzero vector $\mathbf{x}$ such that*

$$A\mathbf{x} = \lambda\mathbf{x}.$$

The vector $\mathbf{x}$ is called an eigenvector associated with λ.

It is well known that λ is a root of the characteristic polynomial $p(\lambda)$ of A. We denote the set of eigenvalues of A by $\sigma(A)$. The algebraic multiplicity of an eigenvalue λ is the multiplicity of λ as a root of $p(\lambda)$. If λ has algebraic multiplicity 1, it is said to be simple. The geometric multiplicity is the dimension of the eigenspace. The algebraic multiplicity of λ is always larger than or equal to its geometric multiplicity. An eigenvalue is said to be defective if its algebraic multiplicity is larger than its geometric multiplicity. A matrix A is said to be defective if it has at least one defective eigenvalue.

Let $\sigma(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. We have that

$$\det(A) = \lambda_1 \lambda_2 \ldots \lambda_n$$

and

$$\text{trace}(A) = \lambda_1 + \ldots + \lambda_n.$$

Definition 9.1.3. *Two matrices A and B are said to be similar if there is a nonsingular matrix X such that*

$$A = XBX^{-1}.$$

X is called a similarity transformation.

It is well known that similar matrices A and B have the same characteristic polynomial, eigenvalues, and algebraic and geometric multiplicities. An eigenvalue decomposition of a square matrix A is the factorization

$$A = X\Lambda X^{-1} \quad \text{or} \quad AX = X\Lambda,$$

where X is nonsingular and Λ is diagonal.

Definition 9.1.4. *The matrix A is said to be unitarily diagonalizable if there exists a unitary matrix Q such that*

$$A = Q\Lambda Q^H.$$

A matrix A is unitarily diagonalizable if and only if it is normal.

Definition 9.1.5. *A subspace $S \subset \mathbb{C}^n$ is said to be invariant for A if*

$$Ax \in S \quad \textit{for all } x \in S.$$

Lemma 9.1.1. *If $A \in \mathbb{C}^{n\times n}$ is partitioned as*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22} \end{bmatrix},$$

where $A_{11} \in \mathbb{C}^{p\times p}$, $A_{12} \in \mathbb{C}^{p\times q}$, $A_{22} \in \mathbb{C}^{q\times q}$, and $p + q = n$, then

$$\sigma(A) = \sigma(A_{11}) \cup \sigma(A_{22}).$$

Theorem 9.1.2. *Let A be a Hermitian matrix, i.e., $A^H = A$. We have that*

1. *The eigenvalues of A are real.*
2. *A is unitarily similar to a real diagonal matrix.*
3. *The eigenvalues of A satisfy*

$$\lambda_k = \min_{S, dim(S)=n-k+1} \max_{\mathbf{x}\in S, \mathbf{x}\neq\mathbf{0}} \frac{(A\mathbf{x},\mathbf{x})}{(\mathbf{x},\mathbf{x})}.$$

A Hermitian matrix A is said to be positive definite if

$$(A\mathbf{x},\mathbf{x}) > 0 \quad \text{for all } \mathbf{0} \neq \mathbf{x} \in \mathbb{C}^n.$$

A Hermitian matrix A is said to be semi-positive definite if

$$(A\mathbf{x},\mathbf{x}) \geq 0 \quad \text{for all } \mathbf{0} \neq \mathbf{x} \in \mathbb{C}^n.$$

For any matrix A, AA^H and A^HA are Hermitian semi-positive definite.

The square roots of eigenvalues of A^HA for a general rectangular matrix A are called the singular values of A. A simple calculation

$$\|A\|_2^2 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{(A\mathbf{x},A\mathbf{x})}{(\mathbf{x},\mathbf{x})} = \max_{\mathbf{x}\neq\mathbf{0}} \frac{(A^HA\mathbf{x},\mathbf{x})}{(\mathbf{x},\mathbf{x})}$$

shows that the 2-norm of A is equal to the largest singular value of A.

Definition 9.1.6. *An upper Hessenberg matrix has zero entries below the first subdiagonal. A lower Hessenberg matrix has zero entries above the first superdiagonal.*

9.2 Iterative Methods for Real Symmetric Matrices

We have seen several self-adjoint eigenvalue problems including the Laplacian eigenvalue problem and the biharmonic eigenvalue problem. Certain finite element methods for such problems, e.g., conforming finite element discretizations, lead to positive definite symmetric matrix eigenvalue problems. We introduce several iterative methods in this section for symmetric matrices.

9.2.1 Power Iteration

The power iteration is a very simple method to compute the eigenvalue with largest norm and the associated eigenvector. For simplicity, we assume that A is symmetric and the largest eigenvalue of A is simple, i.e.,

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n| \geq 0.$$

The power iteration can be simply stated as follows.

Power Iteration:

1. choose a random vector $\mathbf{x}_0$ such that $\|\mathbf{x}_0\| = 1$.

2. for $k = 1, 2, \ldots$

$$\mathbf{y} = A\mathbf{x}_{k-1};$$
$$\mathbf{x}_k = \mathbf{y}/\|\mathbf{y}\|;$$
$$\lambda_k = \mathbf{x}_k^T A\mathbf{x}_k.$$

It is easy to show that the eigenvalue has the following convergence rate (Theorem 27.1 of [235])

$$|\lambda_k - \lambda_1| = O\left(\frac{|\lambda_2|}{|\lambda_1|}\right)^{2k}.$$

9.2.2 Inverse Power Iteration

The power can only find the largest eigenvalue. A modification of the power iteration can be used to find interior eigenvalues. Suppose we want to find an eigenvalue λ closest to a regular value z of A. Letting $\mathbf{x}$ be the eigenvector associated with λ, from $A\mathbf{x} = \lambda\mathbf{x}$, we have that

$$(A - zI)\mathbf{x} = (\lambda - z)\mathbf{x},$$

which implies

$$(A - zI)^{-1}\mathbf{x} = (\lambda - z)^{-1}\mathbf{x}.$$

Employing the power iteration to compute the largest eigenvalue of $(A - zI)^{-1}$, we obtain the so-called inverse iteration.

Inverse Iteration:

1. choose a random vector $\mathbf{x}_0$ such that $\|\mathbf{x}_0\| = 1$.

2. for $k = 1, 2, \ldots$

solve

$$(A - zI)\mathbf{y} = \mathbf{x}_{k-1};$$

$$\mathbf{x}_k = \mathbf{y}/\|\mathbf{y}\|;$$
$$\lambda_k = \mathbf{x}_k^T A\mathbf{x}_k.$$

9.2.3 Rayleigh Quotient Iteration

Another iteration method is Rayleigh quotient iteration. For a vector $\mathbf{x} \in \mathbb{R}^n$, we define the Rayleigh quotient

$$r(\mathbf{x}) := \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}}. \tag{9.3}$$

In fact, $r(\mathbf{x})$ minimize $\|(A - \lambda I)\mathbf{x}\|_2$ (see Section 8.2.3 of [132] or Lecture 27 of [235]). If $\mathbf{x}$ is an eigenvector of A associated with the eigenvalue λ, we have that $r(\mathbf{x}) = \lambda$.

Straightforward calculation shows that

$$\nabla r(\mathbf{x}) = \frac{2}{\mathbf{x}^T \mathbf{x}}(A\mathbf{x} - r(\mathbf{x})\mathbf{x}).$$

Hence $\nabla r(\mathbf{x}) = \mathbf{0}$ if $\mathbf{x}$ is an eigenvector of A. On the other hand, if $\mathbf{x} \neq 0$ and $\nabla r(\mathbf{x}) = \mathbf{0}$, we have that

$$A\mathbf{x} = r(\mathbf{x})\mathbf{x}.$$

Hence $(r(\mathbf{x}), \mathbf{x})$ is an eigenpair of A.

Rayleigh Quotient Iteration:

1. choose a random vector $\mathbf{x}_0$ such that $\|\mathbf{x}_0\| = 1$.
2. for $k = 0, 1, 2, \ldots$

    $\lambda^{(k)} = r(\mathbf{x}_k)$;

    solve

    $$(A - \lambda^{(k)} I)\mathbf{y}_{k+1} = \mathbf{x}_k;$$

    $\mathbf{x}_{k+1} = \frac{\mathbf{y}_{k+1}}{\|\mathbf{y}_{k+1}\|}$.

9.3 The Arnoldi Method

Finite element discretization of eigenvalue problems leads to large, spare generalized matrix eigenvalue problems, sometime non-Hermitian. In this section, we present a popular method, called the Arnoldi method, following [184].

9.3.1 The QR Method

We start with the QR method.

Theorem 9.3.1. *(Schur Decomposition [132]) Let $A \in \mathbb{C}^{n\times n}$. Then there is a unitary matrix Q and an upper triangular matrix R such that*

$$AQ = QR. \tag{9.4}$$

The diagonal elements of R are the eigenvalues of A.

The partial Shur decomposition is defined as follows. Let Q_k denote the leading k columns of Q. One has that

$$AQ_k = Q_k R_k.$$

This leads to the popular shifted QR iteration.

Shifted QR Iteration:

1. given A and $AV = VH, V^H V = I$, H is upper Hessenberg
2. for $j = 1, 2, \ldots$

   choose a shift $\mu = \mu_j$;
   factorize $[Q, R] = qr(H - \mu I)$;
   $H = Q^H H Q$;
   $V = VQ$.

9.3.2 Krylov Subspaces and Projection Methods

The Krylov subspace is given by

$$\mathcal{K}_k(A, \mathbf{v}_1) = \text{Span}\{\mathbf{v}_1, A\mathbf{v}_1, A^2\mathbf{v}_1, \ldots, A^{k-1}\mathbf{v}_1\}.$$

One attempts to formulate the best possible approximations to eigenvectors from this subspace. This can be done by imposing a Galerkin condition.

A vector $\mathbf{x} \in \mathcal{K}_k(A, \mathbf{v}_1)$ is called a Ritz vector with corresponding Ritz value θ if the following Galerkin condition holds:

$$(\mathbf{w}, A\mathbf{x} - \mathbf{x}\theta) = 0 \quad \text{for all } \mathbf{w} \in \mathcal{K}_k(A, \mathbf{v}_1).$$

Let W be a matrix whose columns form an orthonormal basis for $\mathcal{K}_k$. Let $\mathcal{P} = WW^H$ denote the related orthogonal projector onto $\mathcal{K}_k$ and define

$$\hat{A} = \mathcal{P}A\mathcal{P} = WGW^H,$$

where $G = W^H AW$.

Lemma 9.3.2. *For the quantities defined above, we have the following properties (Lemma 4.2.1 of [184]):*

1. $(\mathbf{x}, \theta)$ *is a Ritz pair if and only if* $\mathbf{x} = W\mathbf{s}$ *with* $G\mathbf{s} = \mathbf{s}\theta$*,*
2. *for all* $M \in \mathbb{C}^{n\times n}$ *such that* $M\mathcal{K}_k \subset \mathcal{K}_k$*,*

$$\|(I - \mathcal{P})AW\| = \|(A - \hat{A})W\| \le \|(A - M)W\|,$$

3. *the Ritz-pairs* $(\mathbf{x}, \theta)$ *and the minimum value* $\|(I - \mathcal{P})AW\|$ *are independent of the choice of orthonormal basis* W*.*

These facts are actually valid for any k dimensional subspace S. One also notes that $\mathbf{w} \in \mathcal{K}_k$ can be written as $\mathbf{w} = \phi(A)\mathbf{v}_1$, where $\phi(\cdot)$ is a polynomial of degree less than k.

In addition, let $\mathbf{v}_1$ be a linear combination of vectors spanning an invariant subspace of A. Then $\mathcal{K}_k$ is an invariant subspace for A.

9.3.3 The Arnoldi Factorization

Definition 9.3.1. *Let* $A \in \mathbb{C}^{n\times n}$*. The following form is called the* k*-step Arnoldi factorization of* A*:*

$$AV_k = V_k H_k + \mathbf{f}_k \mathbf{e}_k^T, \tag{9.5}$$

where $V_k \in \mathbb{C}^{n\times k}$ *has orthonormal columns,* $V_k^H \mathbf{f}_k = 0$*, and* $H_k \in \mathbb{C}^{k\times k}$ *is upper Hessenberg with non-negative subdiagonal elements. The columns of* V_k *are called the Arnoldi vectors. If* A *is Hermitian then* H_k *is real, symmetric, and tridiagonal. The relation* (9.5) *is then referred to as a* k*-step Lanczos factorization of* A*. The columns of* V_k *are called Lanczos vectors accordingly.*

An alternative for the Arnoldi factorization is as follows:

$$AV_k = (V_k, \mathbf{v}_{k+1}) \begin{pmatrix} H_k \\ \beta_k \mathbf{e}_k^T \end{pmatrix},$$

where

$$\beta_k = \|\mathbf{f}_k\| \quad \text{and} \quad \mathbf{v}_{k+1} = \frac{1}{\beta_k}\mathbf{f}_k.$$

If $H_k\mathbf{s} = \mathbf{s}\theta$ then the vector $\mathbf{x} = V_k\mathbf{s}$ is such that

$$\|A\mathbf{x} - \mathbf{x}\theta\| = \|(AV_k - V_k H_k)\mathbf{s}\| = |\beta_k \mathbf{e}_k^T \mathbf{s}|.$$

Then $|\beta_k \mathbf{e}_k^T \mathbf{s}|$ is called the Ritz estimate for the Ritz pair $(\mathbf{x}, \theta)$ as an approximation of an eigenpair for A. If $(\mathbf{x}, \theta)$ is a Ritz pair, we have that

$$\theta = \mathbf{s}^H H_k \mathbf{s} = (V_k\mathbf{s})^H A (V_k\mathbf{s}) = \mathbf{x}^H A\mathbf{x}.$$

The Rayleigh quotient residual is defined as

$$\mathbf{r}(\mathbf{x}) \equiv A\mathbf{x} - \mathbf{x}\theta.$$

It is easy to see that

$$\|\mathbf{r}(\mathbf{x})\| = |\beta_k \mathbf{e}_k^T \mathbf{s}|.$$

The following algorithm is the k-step Arnoldi factorization.

k-step Arnoldi Factorization:

1. choose $\mathbf{v}_1$;
2. compute $\mathbf{v}_1 = \mathbf{v}/\|\mathbf{v}_1\|$; $\mathbf{w} = A\mathbf{v}_1$; $\alpha_1 = \mathbf{v}_1^H\mathbf{w}$;
3. compute $\mathbf{f}_1 = \mathbf{w} - \mathbf{v}_1\alpha_1$; $V_1 = (\mathbf{v}_1)$; $H_1 = (\alpha_1)$;
4. for $j = 1, 2, \ldots, k-1$,

$$\beta_j = \|\mathbf{f}_j\|; \mathbf{v}_{j+1} = \mathbf{f}_j/\beta_j;$$
$$V_{j+1} = (V_j, \mathbf{v}_{j+1});$$
$$\hat{H}_h = \begin{pmatrix} H_j \\ \beta_j \mathbf{e}_j^T \end{pmatrix};$$
$$\mathbf{w} = A\mathbf{v}_{j+1};$$
$$\mathbf{h} = V_{j+1}^H \mathbf{w};$$
$$\mathbf{f}_{j+1} = \mathbf{w} - V_{j+1}\mathbf{h};$$
$$H_{j+1} = (\hat{H}_j, \mathbf{h}).$$

There is no way to know in advance how many steps are needed when a satisfactory approximation of eigenvalues by the Ritz values will be obtained. Furthermore, as k gets larger, the computation cost of the Arnoldi method is prohibitive. One approach to resolve this problem is to combine the implicit shifted QR scheme with a k-step Arnoldi factorization. The method is referred to as the implicitly restarted Arnoldi method (Section 4.4 of [184]).

Implicitly Restarted Arnoldi Method:

1. given an m-step Arnoldi factorization:

$$AV_m = V_m H_m + f_m e_m^T,$$

2. for $l = 1, 2, \ldots$

compute $\sigma(H_m)$ and select p shifts $\mu_1, \ldots, \mu_p$;

$\mathbf{q}^H = \mathbf{e}_m^T$;

for j=1,2, …, p,

- factor $[Q, R] = qr(H_m - \mu_j I)$;
- $H_m = Q^H H_m Q$, $V_m = V_m Q$;
- $\mathbf{q} = \mathbf{q}^H Q$;

$\mathbf{f}_k = \mathbf{v}_{k+1}\hat{\beta}_k + \mathbf{f}_m\sigma_k$;

$\mathbf{f}_k = \mathbf{v}_{k=1}\hat{\beta}_k = \mathbf{f}_m\sigma_k$;

$H_k = H_m(1:k, 1:k)$;

begin with the k-step Arnoldi factorization

- $AV_k = V_k H_k + \mathbf{f}_k \mathbf{e}_k^T$;
- apply p additional steps of the Arnoldi process to obtain

$$AV_m = V_m H_m + \mathbf{f}_m \mathbf{e}_m^T.$$

Many details need to be taken care of to implement an eigenvalue solver, for example, the storage, the stopping criterion, etc. There also exist many techniques to improve the performance of an eigenvalue solver, e.g., inflation/deflation, preconditioning, post-processing, etc. We refer the readers to [184] and the references therein for some discussion on these aspects.

Chapter 10

Integral Based Eigensolvers

10.1 Introduction

There are needs for computing eigenvalues of a nonlinear and/or non-Hermitian eigenvalue problem that lie in a given region in the complex plane. For example, a conforming finite element discretization of the fourth order reformulation of the transmission eigenvalue problem (6.19) leads to a quadratic matrix eigenvalue problem. In general, only a small fraction of interior eigenvalues are of interest. Other than some crude qualitative estimates, no spectral information is available. In addition, the distribution of the eigenvalues is very complicated in general (see Fig. 10.1). Most existing eigenvalue solvers are not suitable for these problems. Traditional methods such as shift and invert Arnoldi are handicapped by the lack of a priori eigenvalue estimate.

Recently, integral based methods [222, 214, 15, 31, 233, 141] have become popular (see also [131]). These methods are based on eigenprojections using contour integrals of the resolvent [19].

In this chapter, we first introduce two integral based methods: one is due to Sakurai and Sugiura [222], the other due to Polizzi [214]. These methods can be viewed as classical subspace iteration methods accelerated by approximate spectral projection [233].

In Section 10.2, we present a new recursive integral based method (RIM) proposed by Huang et al. [155]. The method uses the spectral projection to compute an indicator of a region to decide whether or not the region contains eigenvalues. The procedure continues recursively until the region is small enough and focuses on the desired eigenvalue(s). In Section 10.3, a similar idea is employed to solve a nonlinear integral eigenvalue problem. These methods are novel in the sense that they do not actually compute eigenvalues.

The demand for effective and efficient eigensolvers for nonlinear and/or nonself-adjoint eigenvalue problems is increasing. Methods like RIM have the potential to treat these problems, which, we believe, is a promising research area.

10.1.1 Sukurai–Sugiura Method

Sukurai and Sugiura consider the generalized eigenvalue problem in [222]

$$A\mathbf{x} = \lambda B\mathbf{x}, \tag{10.1}$$

where $A, B \in \mathbb{C}^{n\times n}$. Let $\lambda_1, \dots, \lambda_d$ be finite generalized eigenvalues for (10.1). For $z \in \mathbb{C}$, define an analytic function

$$f(z) := \mathbf{u}^H (zB - A)^{-1}\mathbf{v} \quad \text{for } \mathbf{u}, \mathbf{v} \in \mathbb{C}^n.$$

Let $J_d = \text{diag}(\lambda_1, \dots, \lambda_d)$. Considering the pencil $A - zB$, one has the following classical result regarding the Weierstrass' canonical form.

Theorem 10.1.1. *(Theorem 1 of [222]) Let $A - zB$ be a regular pencil of order n. Then there exist nonsingular matrices $P, Q \in \mathbb{C}^{n\times n}$ such that*

$$P(zB - A)Q = \begin{pmatrix} zI_d - J_d & 0 \\ 0 & zJ_{n-d} - I_{n-d} \end{pmatrix},$$

where J_d and J_{n-d} are in Jordan canonical form, J_{n-d} is nilpotent, and I_d denotes the identity matrix of order d.

Consequently, one can define two matrices P and Q such that

$$P^H = (\mathbf{p}_1, \dots, \mathbf{p}_n), \quad \mathbf{p}_1, \dots, \mathbf{p}_n \in \mathbb{C}^n$$

and

$$Q = (\mathbf{q}_1, \dots, \mathbf{q}_n), \quad \mathbf{q}_1, \dots, \mathbf{q}_n \in \mathbb{C}^n.$$

The following theorem is proved in [222]

Theorem 10.1.2. *Define*

$$v_j := \mathbf{u}^H \mathbf{q}_j \mathbf{p}_j^H \mathbf{v} \quad 1 \le j \le d.$$

Let K be the maximum size of Jordan blocks of J_{n-d}. If $A - \lambda B$ is regular and A is diagonalizable, then

$$f(z) = \sum_{j=1}^{d} \frac{v_j}{z - \lambda_j} + g(z), \tag{10.2}$$

where $g(z)$ is a polynomial of degree $K - 1$.

Let Γ be a simple closed curve enclosing m eigenvalues $\lambda_1, \ldots \lambda_m$. It is easy to see that $\lambda_1, \ldots \lambda_m$ are poles of $f(z)$. Define

$$\mu_k = \frac{1}{2\pi i}\int_\Gamma (z-\gamma)^k f(z)\,\mathrm{d}z, \quad k = 0, 1, \ldots, \tag{10.3}$$

where γ is inside Γ. Let H_m and $H_m^<$ be given by

$$H_m := [\mu_{i+j-2}]_{i,j=1}^m = \begin{pmatrix} \mu_0 & \mu_1 & \cdots & \mu_{m-1} \\ \mu_1 & \mu_2 & \cdots & \mu_m \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{m-1} & \mu_m & \cdots & \mu_{2m-2} \end{pmatrix}$$

and

$$H_m^< := [\mu_{i+j-1}]_{i,j=1}^m = \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_m \\ \mu_2 & \mu_3 & \cdots & \mu_{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_m & \mu_{m+1} & \cdots & \mu_{2m-1} \end{pmatrix},$$

respectively. Then the eigenvalues of the pencil $H_m^< - \lambda H_m$ and the eigenvalues of (10.1) have the following relation.

Theorem 10.1.3. *If $v_j \neq 0$ for $1 \leq j \leq m$, then the eigenvalues of the pencil $H_m^< - \lambda H_m$ are given by $\lambda_1 - \gamma, \ldots, \lambda_m - \gamma$.*

The above theorem implies that the eigenvalues can be obtained by solving the (much smaller) generalized eigenvalue problem

$$H_m^< \mathbf{x} = \lambda H_m \mathbf{x}.$$

Note that m coincides with the number of eigenvalues, counting multiplicity, inside Γ. To evaluate the eigenvector, define

$$\mathbf{s}_k := \frac{1}{2\pi i}\int_\Gamma (z-\gamma)^k (zB - A)^{-1}\mathbf{v}\,\mathrm{d}z, \quad k = 0, 1, \ldots.$$

Let

$$\sigma_j := \mathbf{p}_j^H \mathbf{v}, \quad j = 1, 2, \ldots, m.$$

Then it can be shown that

$$[\mathbf{s}_0, \ldots, \mathbf{s}_{m-1}] = [\sigma_1 \mathbf{q}_1, \ldots, \sigma_m \mathbf{q}_m] V_m^T,$$

where V_m is the Vandermonde matrix given by

$$V_m := \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 - \gamma & \lambda_2 - \gamma & \cdots & \lambda_m - \gamma \\ \vdots & \vdots & \ddots & \vdots \\ (\lambda_1 - \gamma)^{m-1} & (\lambda_2 - \gamma)^{m-1} & \cdots & (\lambda_m - \gamma)^{m-1} \end{pmatrix}.$$

If Γ is a circle, trapezoidal rule can be used to obtain accurate approximations of the integrals. Let Γ be a circle centered at $x_0 \in \mathbb{C}$ with radius ρ. The approximation of μ_k using the trapezoidal rule is simply

$$\mu_k \approx \hat{\mu}_k := \frac{1}{N} \sum_{j=0}^{N-1} (\omega_j - x_0)^{k+1}, \quad k = 0, 1, \dots, \tag{10.4}$$

where

$$\omega_j = x_0 + \rho e^{(2\pi i/N)j}, \quad j = 0, 1, \dots, N-1,$$

and N is the number of equally distributed points on Γ.

Let $\hat{H}_m$ and $\hat{H}_m^<$ be defined as

$$\hat{H}_m := [\hat{\mu}_{i+j-2}]_{i,j=1}^m$$

and

$$\hat{H}_m^< := [\hat{\mu}_{i+j-1}]_{i,h=1}^m,$$

respectively. Let $\psi_1, \dots, \psi_m$ be the eigenvalues of the pencil $\hat{H}_m^< - \lambda \hat{H}_m$. Then approximations for eigenvalues $\lambda_1, \dots, \lambda_m$ are

$$\hat{\lambda}_j = x_0 + \psi_j, \quad 1 \leq j \leq m.$$

Define

$$\mathbf{y}_j := (\omega_j B - A)^{-1} \mathbf{v}, \quad j = 0, 1, \dots, N-1,$$

and

$$\hat{\mathbf{s}}_k := \frac{1}{N} \sum_{j=0}^{N-1} (\omega_j - x_0)^{k+1} \mathbf{y}_j, \quad k = 0, 1, \dots.$$

One has that

$$f(\omega_j) = \mathbf{u}^H (w_j B - A)^{-1} \mathbf{v} = \mathbf{u}^H \mathbf{y}_j.$$

Let $\hat{V}_m$ be the Vandermode matrix given by

$$\hat{V}_m = \begin{pmatrix} 1 & 2 & \dots & m \\ \psi_1 & \psi_2 & \dots & \psi_m \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1^{m-1} & \psi_2^{m-1} & \dots & \psi_m^{m-1} \end{pmatrix}.$$

The approximations for the eigenvectors are given by

$$[\hat{\mathbf{q}}_1, \dots, \hat{\mathbf{q}}_m] = [\hat{\mathbf{s}}_0, \dots, \hat{\mathbf{s}}_{m-1}] \hat{V}_m^{-T}. \tag{10.5}$$

Note that the elements of $\hat{V}_m^{-T}$ can be obtained using the coefficients of the Lagrange polynomials

$$\phi_j(z) := \Pi_{l=1, l \neq j}^m \frac{z - \psi_l}{\psi_j - \psi_l}, \quad j = 1, 2, \dots, m.$$

The Sakurai-Sugiura method is as follows.

Sakurai-Sugiura Algorithm:

1. given $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$, N, m, x_0, ρ
2. compute $\omega_j = x_0 + \rho e^{2\pi j i/N}, j = 0, \ldots, N-1$
3. compute $\mathbf{y}_j = (\omega_j B - A)^{-1}\mathbf{v}, j = 0, \ldots, N-1$
4. compute $f_h = \mathbf{u}^H \mathbf{y}_j, j = 0, \ldots, N-1$
5. compute $\hat{\mu}_k$ using (10.4)
6. compute the eigenvalues $\psi_1, \ldots, \psi_m$ of the pencil $\hat{H}_m^{<} - \lambda \hat{H}_m$
7. compute $\hat{\mathbf{q}}_1, \ldots, \hat{\mathbf{q}}_m$ using (10.5)
8. compute $\hat{\lambda}_j = x_0 + \psi_j, j = 1, \ldots, m$

10.1.2 Polizzi's Method

The problem considered by Polizzi [214] is the following generalized eigenvalue problem arising from electronic structure calculations

$$A\mathbf{x} = \lambda B\mathbf{x}, \tag{10.6}$$

where A is $n \times n$ real symmetric or Hermitian and B is an $n \times n$ symmetric positive definite matrix. The goal is to compute all eigenvalues in a given interval $(\lambda_{min}, \lambda_{max})$.

The Green's function, i.e., the resolvent for the generalized eigenvalue problem, is defined as

$$G(\sigma) = (\sigma B - A)^{-1}.$$

Let Γ be a circle centered at $(\lambda_{min} + \lambda_{max})/2$ with radius $r = (\lambda_{max} - \lambda_{min})/2$. The spectrum projection,

$$P = -\frac{1}{2\pi i}\int_\Gamma (\sigma B - A)^{-1}\mathrm{d}z,$$

is referred to as the reduced density matrix. Assume that there are m eigenvalues inside Γ, counting multiplicity. Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be the set of associated eigenvectors. In addition, let $\mathbf{Y}_{n\times m} = [\mathbf{y}_1, \ldots, \mathbf{y}_m]$, where $\mathbf{y}_1, \ldots, \mathbf{y}_m$ are linearly independent random vectors. One obtains a new set of independent vectors

$$Q_{n\times m} = [\mathbf{q}_1, \ldots, \mathbf{q}_1] = P\mathbf{Y}.$$

Then the eigenvalue problem of finding eigenvalues of (10.6) inside Γ becomes a reduced generalized eigenvalue problem

$$A_Q\Phi = \lambda B_Q\Phi \tag{10.7}$$

where

$$A_Q = Q^T A Q \quad \text{and} \quad B_Q = Q^T B Q.$$

The matrix Q can be obtained using suitable numerical integration such as Gauss-Legendre quadrature. If A is Hermitian, the spectrum projection is given by

$$P = -\frac{1}{2\pi i}\int_{\Gamma^+} G(z) - G(\overline{z})\mathrm{d}z,$$

where Γ^+ is the upper half circle. When A is real symmetric,

$$P = -\frac{1}{\pi}\int_{\Gamma^+} \mathcal{I}(G(z))\mathrm{d}z,$$

where $\mathcal{I}$ denotes the imaginary part.

To be specific, when A is Hermitian, using N-point Gauss-Legendre quadrature for Γ^+ with x_j being the jth Gauss node and ω_j being the weight, one obtains

$$Q = -\sum_{j=1}^{N} \frac{1}{4}\omega_j r\left[e^{i\theta_j}G(x_j) + e^{-i\theta_j}G(\overline{x}_j)\right]Y,$$

where

$$\theta_j = -\frac{\pi}{2}(j-1) \quad \text{and} \quad x_j = \frac{\lambda_{min}+\lambda_{max}}{2} + re^{i\theta_j}.$$

When A is real symmetric, one has that

$$Q = -\sum_{j=1}^{N} \frac{1}{2}\omega_j \mathcal{R}\left[re^{i\theta_j}G(x_j)\right],$$

where $\mathcal{R}$ denotes the real part.

Polizzi's algorithm, called FEAST, can be described as follows (Fig. 2 of [214]).

Algorithm FEAST:

1. generate $m_0 > m$ random vectors $Y_{n\times m_0}$
2. set $Q_{n\times m_0} = \mathbf{0}$, and
$$r = \frac{\lambda_{max}-\lambda_{min}}{2}$$
3. for $j = 1, \ldots, N_j$

   3.1 compute $\theta_j = -\pi/2(j-1)$

   3.2 compute
   $$z_j = \frac{\lambda_{min}+\lambda_{max}}{2} + re^{i\theta_j}$$

   3.3 solve $(z_jB - A)Q_j = Y$ for $Q_j \in \mathbb{C}^{n\times m_0}$

3.4 compute
$$Q = Q - (\omega_j/2)\mathcal{R}\left[re^{i\theta_j}G(x_j)\right]$$

4. construct $A_Q = Q^T AQ$ and $B_Q = Q^T BQ$
5. solve $A_Q\Phi = \lambda B_Q\Phi$ to obtain m_0 eigenvalues and eigenvectors $\Phi_{m_0\times m_0}$
6. compute $X_{n\times x_0} = Q_{n\times m_0}\Phi_{m_0\times m_0}$
7. check convergence for the trace of the eigenvalues. If refinement is needed, compute $Y = BX$ and go to step 2

10.2 The Recursive Integral Method

The methods introduced in the previous two sections depend on some estimation on the locations, number of eigenvalues, and dimensions of eigenspaces. At a certain stage of the methods, one needs to solve eigenvalue problems of a smaller size. However, there are many eigenvalue problems with no a priori information available. For example, in Fig. 10.1, we show transmission eigenvalues with an extremely complicated spectrum.

In this section, we introduce a recursive integral method (RIM) to compute eigenvalues, which can be viewed as an eigensolver without actually computing the eigenvalues [155]. In particular, it can be viewed as a general eigensolver for problems with the following features:

1) the problem is non-Hermitian,
2) spectrum is complicated,
3) no a priori information, such as number of eigenvalues, is available,
4) interior eigenvalues (real and/or complex) are needed.

RIM recursively searches a region in the complex plane for eigenvalues using approximate eigenprojection. It is well suited to the transmission eigenvalue problem which typically seeks certain interior eigenvalues.

We start by recalling some classical results of operator theory (see, e.g., [165]). Let $T : X \to X$ be an operator on a complex Hilbert space X.

Let Γ be a simple closed curve on the complex plane $\mathbb{C}$ lying in $\rho(T)$ which contains m eigenvalues, counting multiplicity, of T: $\lambda_j, j = 1,\ldots,m$. The spectrum projection
$$P = \frac{1}{2\pi i}\int_\Gamma R_z(T)\mathrm{d}z$$
projects an element $x \in X$ onto the space of generalized eigenfunctions $\mathbf{u}_j, j =$

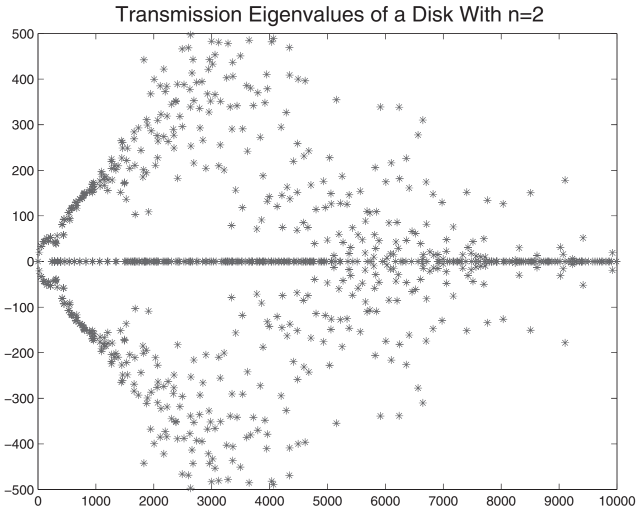


Figure 10.1: A sample eigenvalue problem with complicated spectrum distribution: transmission eigenvalues of a disc with radius $1/2$ and index of refraction $n = 2$.

$1, \ldots, m$ associated with $\lambda_j, j = 1, \ldots, m$. We know that the projection P depends only on $\lambda_j, j = 1, \ldots, m$, inside Γ and the associated eigenfunctions [165].

Let $\mathbf{f} \in X$ be randomly chosen. If there are no eigenvalues inside Γ, we have that $P\mathbf{f} = 0$. Otherwise, if there are m eigenvalues $\lambda_i, i = 1, \ldots, m$, $P\mathbf{f} \neq 0$ provided that $\mathbf{f}$ has components in $\mathbf{u}_i, i = 1, \ldots, m$. Thus $P\mathbf{f}$ can be used to decide if a region contains eigenvalues of T or not.

We are now ready to introduce RIM. Our goal is to find all the eigenvalues of T inside Γ to a certain precision such that the interior of Γ, denoted by S.

Let $\{z_j, \omega_j\}, j = 1, \ldots, W$, be a suitable quadrature rule for Γ. We approximate $P\mathbf{f}$ by

$$P\mathbf{f} \approx \frac{1}{2\pi i} \sum_{j=1}^{W} \omega_j R_{z_j}(T)\mathbf{f}. \tag{10.8}$$

Other than computing $R_{z_j}\mathbf{f} = (z_j - T)^{-1}\mathbf{f}$, we define $\mathbf{x}_j, j = 1, \ldots, W$, and solve

$$(z_j - T)\mathbf{x}_j = \mathbf{f}, \quad j = 1, \ldots, W.$$

Then we have

$$P\mathbf{f} = \sum_{j=1}^{W} \mathbf{x}_j.$$

The key observation for RIM is that $\|P\mathbf{f}\|$ can be used as an indicator if there are eigenvalues in S, i.e.,

(i) if $\|P\mathbf{f}\| = O(1)$, there exists at least one eigenvalue in S;

(ii) if $\|P\mathbf{f}\| = o(1)$, there is no eigenvalue in S.

In practice, we do need a threshold σ to distinguish between $\|P\mathbf{f}\| = O(1)$ and $\|P\mathbf{f}\| = o(1)$. We postpone the discussion on the appropriate value for σ at this moment.

If S contains eigenvalue(s), we partition S into subregions and recursively repeat this procedure. The process terminates when each eigenvalue is isolated within a sufficiently small subregion, i.e., the size of the region $h(S)$ is smaller than the required tolerance ϵ.

A general algorithm of RIM can be described as follows.

RIM$(S, \epsilon, \mathbf{f})$

Input:

a region S,

tolerance ϵ,

a randomly chosen $\mathbf{f}$

Output:

λ, eigenvalue(s) of T in S

1. Approximate $P\mathbf{f}$ by (10.8);
2. Decide if S contains eigenvalue(s) using $\|P\mathbf{f}\|$:
   - No. exit
   - Yes. compute the size $h(S)$ of S
     - if $h(S) > \epsilon$, partition S into subregions $S_i, i = 1, \ldots N$

       for $i = 1$ to N

       RIM$(S_i, \epsilon, \mathbf{f})$

       end
     - if $h(S) \leq \epsilon$, output the eigenvalue λ and exit

Since the finite element methods of eigenvalue problems lead to generalized matrix eigenvalue problems, we specialize RIM to the case of matrix pencils. Consider the generalized eigenvalue problem

$$A\mathbf{x} = \lambda B\mathbf{x}, \tag{10.9}$$

where $A, B \in \mathbb{C}^{n\times n}$, $\lambda \in \mathbb{C}$ is a scalar, and $\mathbf{x} \in \mathbb{C}^n$. The resolvent is

$$R_z(A, B) = (zB - A)^{-1} \tag{10.10}$$

for z in the resolvent set of the matrix pencil (A, B). The projection onto the generalized eigenspace associated to eigenvalues enclosed by Γ is given by

$$P(A,B) = \frac{1}{2\pi i}\int_\Gamma (zB - A)^{-1}\,\mathrm{d}z. \tag{10.11}$$

The projection of a vector $\mathbf{f} \in \mathbb{C}^n$ onto the generalized eigenspace is approximated by

$$\begin{aligned} P\mathbf{f} &= \frac{1}{2\pi i}\int_\Gamma R_z(A,B)\mathbf{f}\,\mathrm{d}z \\ &\approx \frac{1}{2\pi i}\sum_{j=1}^{W}\omega_j R_{z_j}(A,B)\mathbf{f} \\ &= \frac{1}{2\pi i}\sum_{j=1}^{W}\omega_j \mathbf{x}_j, \end{aligned} \tag{10.12}$$

where $\mathbf{x}_j$'s are the solutions of the following linear systems

$$(z_j B - A)\mathbf{x}_j = \mathbf{f}, \quad j = 1, \ldots, W. \tag{10.13}$$

Similar to the continuous case, if there are no eigenvalues inside Γ, then $P = 0$ and thus $P\mathbf{f} = \mathbf{0}$ for all $\mathbf{f} \in \mathbb{C}^n$.

10.2.1 Implementation

We discuss some details of the implementation of RIM for the matrix eigenvalue problems. We choose the search region S to be a rectangle in the complex plane and Γ is its boundary. In particular, we assume that the width and length are comparable. Otherwise, one can pre-divide S into smaller rectangles to satisfy the above assumption. We call S *admissible* if the indicator function $\|P\mathbf{f}\| > \sigma$ where σ is the threshold value we will specify later. We recursively divide an admissible rectangle S into non-overlapping sub-rectangles and compute the indicator function until certain precision is reached.

There are several keys in the implementation of RIM:

1) a suitable quadrature rule for the contour integral;

2) a mechanism to solve (10.13);

3) an effective rule to decide if a region S contains eigenvalues or not.

Since the region S is a rectangle, we use the midpoint of each edge as the quadrature point and four points in total. Note that other contour integral methods use many more points. For example, twenty-five quadrature points are used in [31].

To solve the linear systems (10.13), we use the Matlab "\" command, which is considered to be exact. However, for mid-size problems of a few ten thousands, we find that some iterative solvers, such as "lsqr" with tolerance 10^{-4}, also work for the numerical examples.

Now we discuss the rule to decide if S is admissible or not. We denote by $\mathcal{R}(P)$ the range of P, which coincides with the finite dimensional generalized eigenspace associated with the eigenvalues inside Γ. Let $\boldsymbol{\phi}_j, j = 1, \ldots, m$, be an orthonormal basis of $\mathcal{R}(P)$. Let $\mathbf{f}$ be a randomly chosen vector such that

$$\mathbf{f}|_{\mathcal{R}(P)} = \sum_{j=1}^{m} a_j \boldsymbol{\phi}_j. \tag{10.14}$$

Then we have that

$$P\mathbf{f} = \sum_{j=1}^{m} a_i \mathbf{x}_j.$$

We mentioned above using $\|P\mathbf{f}\|$ to decide if a region contains eigenvalues. However, there are two concerns we need to address for the robustness of the algorithm.

(i) $\|P\mathbf{f}\|$ can be relatively small when there is an eigenvalue(s) inside Γ.

(ii) $\|P\mathbf{f}\|$ can be relatively large when there is no eigenvalue inside Γ.

Case (i) can happen if $\|\mathbf{f}|_{\mathcal{R}(P)}\|$ is small, i.e., $\sum_{j=1}^{M} a_i^2$ is small. Our solution is to project $P\mathbf{f}$ once again and set the indicator as

$$\sigma_S = \left\| P\left(\frac{P\mathbf{f}}{\|P\mathbf{f}\|}\right) \right\|. \tag{10.15}$$

Case (ii) happens if there exists eigenvalue(s) that lies outside Γ but close to Γ. In fact, this must happen when RIM zooms into the neighborhood of an eigenvalue. Fortunately, RIM has an interesting *self-correction* property that fixes such problems on subsequent iterations. This property is observed in the numerical experiments.

Here are some details in the actual implementation.

1. The search region S is a rectangle;
2. We use Matlab "\" to solve the linear systems;
3. We use one point quadrature for each edge of S;
4. We use one randomly chosen vector $\mathbf{f}$;
5. We project $\mathbf{f}$ twice and compute the indicator using (10.15);
6. We use $\sigma = 1/10$ as the threshold value, i.e., if $\sigma_S > 1/10$, S is admissible.

RIM for generalized matrix eigenvalue problems $A\mathbf{x} = \lambda B\mathbf{x}$ can be described as follows.

M-RIM$(A, B, S, \epsilon, \mathbf{f}, \sigma)$

Input:

- matrix pencil (A, B)
- search region S
- precision ϵ
- random vector $\mathbf{f}$
- threshold value σ

Output:

- generalized eigenvalue(s) λ inside S

1. Compute σ_S
2. Decide if S contains eigenvalue(s)
   - If $\sigma_S < \sigma$, exit
   - Otherwise, compute the diameter $h(S)$ of S
     - if $h(S) > \epsilon$, partition S into subregions $S_i, i = 1, \ldots I$
       for $i = 1$ to I
       M-RIM$(A, B, S_i, \epsilon, \mathbf{f})$
       end
     - if $h(S) \leq \epsilon$
       set λ to be the center of S
       output λ
       exit

10.2.2 Numerical Examples

We use the transmission eigenvalue problem discussed in Chapter 6 as an example. In particular, the mixed finite element in Section 6.5 leads to a generalized non-Hermitian eigenvalue problem

$$A\mathbf{x} = \lambda B\mathbf{x}. \tag{10.16}$$

In general, there exist complex eigenvalues. Classical methods are not effective for (10.16). In fact, this problem is our motivation to develop RIM.

We assume that the initial search region S is a rectangle. We present examples to show several properties of RIM.

1. **Effectiveness:**

   Example 1: We consider a disc Ω with radius $1/2$ and index of refraction $n(x) = 16$ as in Section 6.5. A regular mesh with $h \approx 0.05$ is used to generate two 1018×1018 matrices A and B for (10.16). We choose a search region given by

$$S = [3, 9] \times [-3, 3].$$

The tolerance is set to be $\epsilon = 1.0e - 3 \times (\pm 1 \pm i)$. The exact generalized eigenvalues of (10.16) in S are

$$\lambda_1 = 3.99453902, \quad \lambda_2 = 6.93505399, \quad \lambda_3 = 6.93971914.$$

We set $\epsilon = 1.0e - 3$. RIM successfully computes 3 eigenvalues

$$\lambda_1 = 3.99462891 \pm 10^{-3} \pm 10^{-3}i,$$
$$\lambda_2 = 6.93505859 \pm 10^{-3} \pm 10^{-3}i,$$
$$\lambda_3 = 6.93994140 \pm 10^{-3} \pm 10^{-3}i.$$

As another search region, we choose

$$S = [22, 25] \times [-8, 8].$$

There exist two eigenvalues in S.

$$\lambda_1 = 24.15856715 + 5.69011376i, \quad \lambda_2 = 24.15856715 - 5.69011376i.$$

RIM outputs the following

$$\lambda_1 = 24.15881348 - 5.69030762i \pm 10^{-3} \pm 10^{-3}i,$$
$$\lambda_2 = 24.15881348 + 5.69006348i \pm 10^{-3} \pm 10^{-3}i.$$

The search regions explored by RIM are shown in Fig. 10.2. The algorithm refines near the eigenvalues until the precision is reached.

Example 2: Let Ω be the unit square and $n(x) = 16$ with $h \approx 0.05$. The matrices A and B are 1298×1298. The first search region is given by

$$S = [6, 9] \times [-1, 1].$$

The exact eigenvalues are given by

$$\lambda_1 = 6.04952764, \quad \lambda_2 = 6.05117989, \quad \lambda_3 = 8.36856820.$$

RIM correctly computes the following eigenvalues

$$\lambda_1 = 6.04931641 \pm 10^{-3} \pm 10^{-3}i,$$
$$\lambda_2 = 6.05126953 \pm 10^{-3} \pm 10^{-3}i,$$
$$\lambda_3 = 8.36865234 \pm 10^{-3} \pm 10^{-3}i.$$

The second search region is given by

$$S = [20, 21] \times [-6, 6].$$

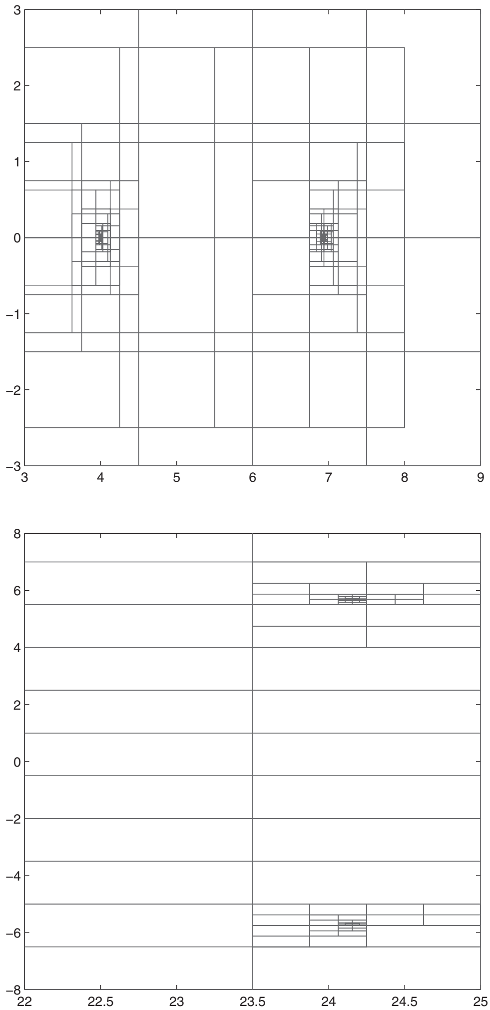


Figure 10.2: The regions explored by RIM for the disc with radius $1/2$, $n(x) = 16$, and $\epsilon = 1.0e - 3 \times (\pm 1 \pm i)$. Top: the search region is given by $S = [3, 9] \times [-3, 3]$. Bottom: the search region is given by $S = [22, 25] \times [-8, 8]$.

The exact eigenvalues are

$$\lambda_1 = 20.57378570 + 5.12722497i,$$
$$\lambda_2 = 20.57378570 - 5.12722497i.$$

The eigenvalues computed by RIM are

$$\lambda_1 = 20.57373047 - 5.12744141i \pm 10^{-3} \pm 10^{-3}i,$$
$$\lambda_2 = 20.57373047 + 5.12646484i \pm 10^{-3} \pm 10^{-3}i.$$

We plot the search regions in Fig. 10.3. The top picture is for $S = [6, 9] \times [-1, 1]$. The bottom picture is for $S = [20, 21] \times [-6, 6]$.

2. **Robustness:**

   We demonstrate the robustness of RIM related to the use of only one random vector and the choice of the threshold value. We first check the use of one random vector in the algorithm. Let

$$S_1 = [3.9, 4.1] \times [-0.1, 0.1],$$
$$S_2 = [24.1, 24.2] \times [5.6, 5.7],$$

   for **Example 1**, and

$$S_3 = [6.04, 6.06] \times [-0.01, 0.01],$$
$$S_4 = [20.5, 20.6] \times [5.1, 5.2],$$

   for **Example 2**. Each region has an eigenvalue inside.

   To see how the choice of different random vectors affect the indicators, we use 100 random vectors. The results are shown in Table 10.1. The second, third, fourth, and fifth columns are the average, minimum, maximum, and the standard deviation of the indicator, respectively. We can see that all the random vectors give similar indicators. The standard deviation is very small. We also show the indicators in Fig. 10.4. All the values are $O(1)$ indicating the existence of an eigenvalue inside the region. The results demonstrate that one random test vector is enough to obtain the indicator of the region correctly.

S	average	min	max	std
S_1	0.63662546	0.63662432	0.63662669	2.42494379e-07
S_2	0.82076270	0.82076270	0.82076270	3.48933530e-11
S_3	0.63667811	0.63662296	0.63674302	4.23597573e-05
S_4	0.53606809	0.53606809	0.53606809	5.68226051e-11

   Table 10.1: The indicators for different regions with eigenvalues inside.

   Next we show the behavior of the indicator when the regions contain no eigenvalues. Let

$$S_5 = [3.7, 3.9] \times [-0.1, 0.1],$$
$$S_6 = [24.0, 24.1] \times [5.6, 5.7],$$
$$S_7 = [6.02, 6.04] \times [-0.01, 0.01],$$
$$S_8 = [20.4, 20.5] \times [5.1, 5.2].$$

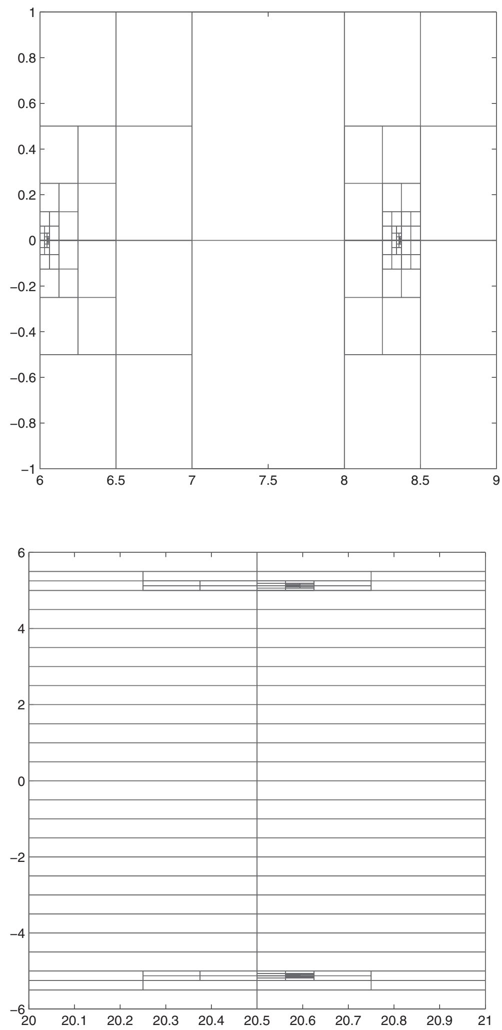


Figure 10.3: The regions explored by RIM for the unit square with $n(x) = 16$ and $\epsilon = 1.0e - 3 \times (\pm 1 \pm i)$. Top: the search region is given by $S = [6, 9] \times [-1, 1]$. Bottom: the search region is given by $[20, 21] \times [-6, 6]$.

In Table 10.2, it can be seen that the indicators are very small for all four examples and various random test vectors.

Although the above regions contain no eigenvalues, they are rather close to

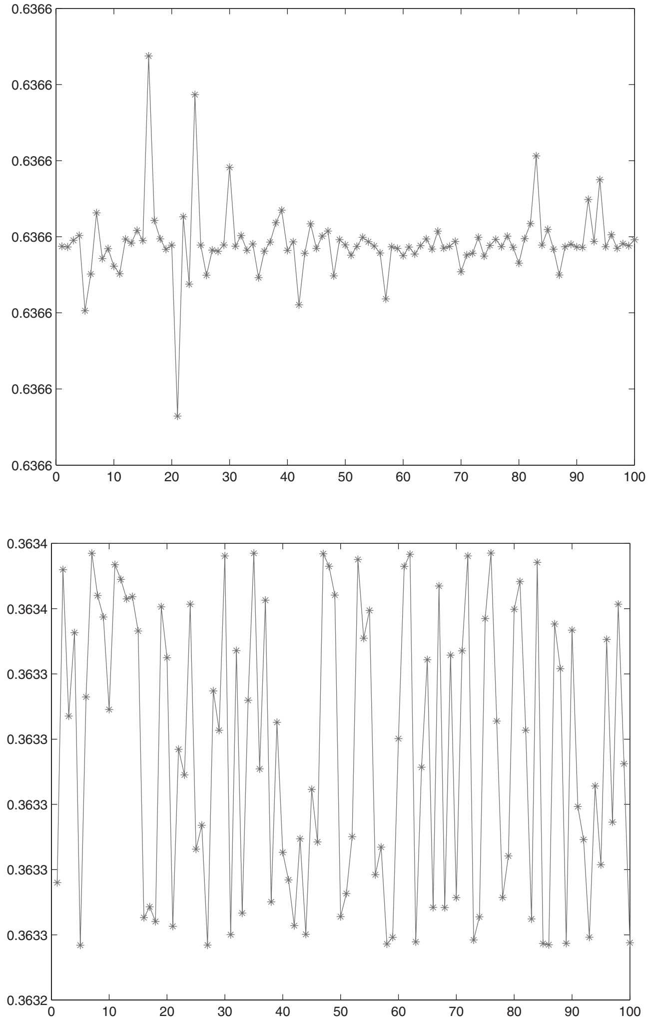


Figure 10.4: The indicators for different regions with eigenvalues inside using 100 random vectors. The indicators are almost the same for different random vectors. Top: $[3.9, 4.1] \times [-0.1, 0.1]$. Bottom: $[6.04, 6.06] \times [-0.01, 0.01]$.

them. We also choose regions that have some distance from the eigenvalue.

S	average	min	max	std
S_5	0.04778437	0.04778398	0.04778539	1.48221826e-07
S_6	0.02227906	0.02227906	0.02227906	6.92810350e-12
S_7	0.04143107	0.03354195	0.04701297	4.44534110e-03
S_8	0.01615291	0.01615294	0.01615294	4.94631162e-11

Table 10.2: The indicators for different regions without eigenvalues inside.

Let

$$
\begin{aligned}
S_8 &= [20.4, 20.5] \times [25.1, 25.2],\\
S_9 &= [-0.1, 0.1] \times [0.9, 1.1],\\
S_{10} &= [24.0, 24.1] \times [20.6, 20.7],\\
S_{11} &= [6.02, 6.04] \times [5.99, 6.01].
\end{aligned}
$$

In Table 10.3, we can see that the indicators are much smaller ($\approx o(1)$). Again, the minimum, maximum, and the standard deviation show that the indicators are stable with respect to the choice of different random test vectors.

S	average	min	max	std
S_9	1.98729099e-04	2.01897954e-05	2.49329490e-04	5.38612705e-05
S_{10}	8.01335639e-11	7.71374586e-11	8.35343994e-11	1.28959584e-12
S_{11}	1.11538517e-12	4.71967623e-13	2.58764561e-12	4.06962505e-13
S_{12}	2.47668028e-11	2.33329418e-11	2.62893133e-11	5.92424005e-13

Table 10.3: The indicators for different regions without eigenvalues inside.

The choice of the threshold value σ is important to the success of RIM. It is easy to see that $\sigma_S \in [0, 1]$. Ideally, we should have $\sigma = 1$ if there are eigenvalues in the search region and $\sigma = 0$ otherwise. However, approximation of the contour integral, including the quadrature and linear solver, introduces errors. Furthermore, it is very likely that eigenvalues are close to Γ or even on Γ. In fact, this is the case whenever the search region is close to the eigenvalues.

In the algorithm the threshold value is $1/10$. The above examples show that $1/10$ can easily distinguish between cases when the eigenvalue are inside and outside the region. However, an eigenvalue might be on the edge or even is a corner of the region. We show in the following that RIM is robust to treat such cases.

We first check the case when the eigenvalue is on the edge of the search region. Let two rectangles be given by

$$S_{13} = [3.99, 4.00] \times [-0.01, 0.00]$$

and

$$S_{14} = [3.99, 4.00] \times [0.00, 0.01]$$

sharing the edge for **Example 1**. Since the eigenvalue

$$\lambda_1 = 3.994628906250000$$

is real, it locates on the real axis. In Table 10.4, we show the indicators. We can see that both regions are admissible. Next, we choose S_{15} and S_{16} such that the sharing edge goes through a complex eigenvalue. It can be seen that the indicator for S_{16} is smaller than $1/10$. Fortunately, this is fine since S_{15} is caught and we will not miss the eigenvalue.

S	indicator
$S_{13} = [3.99, 4.00] \times [-0.01, 0.00]$	0.52275012
$S_{14} = [3.99, 4.00] \times [0.00, 0.01]$	0.52275012
$S_{15} = [24.15881348, 24.17] \times [5.68, 5.70]$	0.48810370
$S_{16} = [24.15, 24.15881348] \times [5.68, 5.70]$	0.08569820

Table 10.4: The indicators when the eigenvalue is on the edge of the search region.

Next we consider the extreme case when an eigenvalue is a corner of the search region. We know from above that search regions S_{17}, S_{18}, S_{19}, and S_{20} (see Table 10.5) share a corner, which is an eigenvalue. Similarly, S_{17}, S_{18}, S_{19}, and S_{20} share a corner as an eigenvalue. For both cases, we see that $1/10$ is a good choice as the threshold value.

S	indicator
$S_{17} = [3.99453902, 4.01] \times [-0.01, 0.0]$	0.70164096
$S_{18} = [3.98, 3.99453902] \times [-0.01, 0.0]$	0.91502267
$S_{19} = [3.98, 3.99453902] \times [0.00, 0.01]$	0.25047340
$S_{20} = [3.99453902, 4.01] \times [0.00, 0.01]$	0.25047335
$S_{21} = [24.152, 24.15856715] \times [5.688, 5.69011376]$	0.43892705
$S_{22} = [24.152, 24.15856715] \times [5.69011376, 5.700]$	0.12732395
$S_{23} = [24.15856715, 24.161] \times [5.69011376, 5.700]$	0.12732395
$S_{24} = [24.15856715, 24.161] \times [5.688, 5.69011376]$	0.19531957

Table 10.5: The indicators when the eigenvalue is a corner of the search region.

3. **Self-correction Property:**

   The choice of threshold value is related to a nice property of RIM, which we call *self-correction property*. We illustrate this as follows. If a quadrature point is close to an eigenvalue λ, the linear system is ill-conditioned. In particular, when an eigenvalue is right outside the search region S, the indicator function χ_S could be large because either the linear solver or quadrature rule is not sufficiently accurate. RIM will take such regions as admissible at the beginning. Fortunately, after a few subdivisions, RIM discards these regions. We demonstrate this interesting *self-correction property* using two examples.

We use matrices A and B from **Example 1** and focus on the eigenvalue located around 3.9945. We choose the initial search region

$$S^0 = [4.0, 4.2] \times [0, 0.2].$$

Note that there is no eigenvalue in S^0. However, RIM computes

$$\chi_{S^0} = 0.11666587, \tag{10.17}$$

indicating that S^0 is admissible and RIM continues to recursively explore S^0 by dividing it into the four rectangles

$$\begin{aligned} S_1^1 &= [4.0, 4.1] \times [0, 0.1], \\ S_2^1 &= [4.0, 4.1] \times [0.1, 0.2], \\ S_3^1 &= [4.1, 4.2] \times [0, 0.2], \\ S_4^1 &= [4.1, 4.2] \times [0.1, 0.2], \end{aligned}$$

with indicators

$$\begin{aligned} \chi_{S_1^1} &= 0.10687367, \\ \chi_{S_2^1} &= 0.00609138, \\ \chi_{S_3^1} &= 0.00561028, \\ \chi_{S_4^1} &= 0.00182170. \end{aligned}$$

RIM discards S_2^1, S_3^1, and S_4^1 and retains S_1^1 as admissible.

The four rectangles by dividing S_1^1 are

$$\begin{aligned} S_1^2 &= [4.0, 4.05] \times [0.0, 0.05], \\ S_2^2 &= [4.0, 4.05] \times [0.05, 0.10], \\ S_3^2 &= [4.05, 4.10] \times [0.0, 0.05], \\ S_4^2 &= [4.05, 4.10] \times [0.05, 0.10], \end{aligned}$$

with indicator values

$$\begin{aligned} \chi_{S_1^2} &= 0.08957099, \quad & \chi_{S_2^2} &= 0.00579253, \\ \chi_{S_3^2} &= 0.00494816, \quad & \chi_{S_4^2} &= 0.00169434. \end{aligned}$$

At this stage, RIM discards all the regions. Let us see one more level. Suppose

$\chi_{S_1^2}$ is subdivided into

$$\begin{aligned}S_1^3 &= [4.0, 4.025] \times [0, 0.025],\\ S_2^3 &= [4.0, 4.025] \times [0.025, 0.05],\\ S_3^3 &= [4.025, 4.05] \times [0, 0.025],\\ S_4^3 &= [4.025, 4.05] \times [0.025, 0.05],\end{aligned}$$

with indicator values

$$\begin{aligned}\chi_{S_1^3} &= 0.06258907,\\ \chi_{S_2^3} &= 0.00519080,\\ \chi_{S_3^3} &= 0.00388825,\\ \chi_{S_4^3} &= 0.00146650.\end{aligned}$$

Hence even if we use an smaller threshold value, RIM eventually discards S.

Example 3: The same experiment is conducted for a search region around the complex eigenvalue $\lambda = 24.1586 + 5.690i$ with initial search region

$$S = [24.16, 24.96] \times [5.30, 6.10],$$

which does not contain any eigenvalues, but close to the eigenvalue. Indicator values are in Table 10.6. Note that RIM does eventually conclude that there are no eigenvalues in the region.

4. **Close eigenvalues:**

Since RIM uses a tolerance, it separates nearby eigenvalues provided the tolerance is less than the distance between them. We need more decimal digits for the following examples.

For **Example 1**, there are two close eigenvalues

$$\lambda_1 = 6.935053985844653, \quad \lambda_2 = 6.939719143809611.$$

With $\epsilon = 3.0e - 2 \times (\pm 1 \pm i)$, RIM fails to separate the eigenvalues and we obtain only one eigenvalue

$$\lambda_1 = 6.942500000000000 + 0.002500000000000i \pm 3 \times 10^{-2} \pm 3 \times 10^{-2}i.$$

However, with $\epsilon = 1.0e - 4 \times (\pm 1 \pm i)$, RIM separates the eigenvalues and we obtain

$$\begin{aligned}\lambda_1 &= 6.939716796875000 + 0.000009765625000i \pm 10^{-4} \pm 10^{-4}i,\\ \lambda_2 &= 6.935126953124999 + 0.000009765625000i \pm 10^{-4} \pm 10^{-4}i.\end{aligned}$$

$S_1^1 = [24.16, 24.56] \times [5.30, 5.70]$	0.825
$S_2^1 = [24.16, 24.56] \times [5.70, 6.10]$	0.195
$S_3^1 = [24.56, 24.96] \times [5.30, 5.70]$	5.418e-11
$S_4^1 = [24.56, 24.96] \times [5.70, 6.10]$	4.119e-11
$S_1^2 = [24.16, 24.36] \times [5.30, 5.50]$	9.216e-11
$S_2^2 = [24.16, 24.36] \times [5.50, 5.70]$	3.682
$S_3^2 = [24.36, 24.56] \times [5.30, 5.50]$	8.712e-14
$S_4^2 = [24.36, 24.56] \times [5.50, 5.70]$	5.870e-11
$S_1^3 = [24.16, 24.26] \times [5.50, 5.60]$	1.742e-11
$S_2^3 = [24.16, 24.26] \times [5.60, 5.70]$	7.806
$S_3^3 = [24.26, 24.36] \times [5.50, 5.60]$	1.476e-13
$S_4^3 = [24.26, 24.36] \times [5.60, 5.70]$	6.755e-11
$S_1^4 = [24.16, 24.21] \times [5.60, 5.65]$	6.558e-10
$S_2^4 = [24.16, 24.21] \times [5.65, 5.70]$	2.799
$S_3^4 = [24.21, 24.26] \times [5.60, 5.65]$	1.378e-13
$S_4^4 = [24.21, 24.26] \times [5.65, 5.70]$	8.229e-11
$S_1^5 = [24.16, 24.185] \times [5.65, 5.675]$	1.159e-8
$S_2^5 = [24.16, 24.185] \times [5.675, 5.70]$	1.556
$S_3^5 = [24.185, 24.21] \times [5.65, 5.675]$	4.000e-13
$S_4^5 = [24.185, 24.21] \times [5.675, 5.70]$	8.648e-11
$S_1^6 = [24.16, 24.185] \times [5.65, 5.675]$	5.574e-06
$S_2^6 = [24.16, 24.1725] \times [5.6875, 5.70]$	0.095
$S_3^6 = [24.185, 24.21] \times [5.65, 5.675]$	4.304e-12
$S_4^6 = [24.185, 24.21] \times [5.675, 5.70]$	2.628e-11

Table 10.6: The indicators on different search regions.

Example 6: This example comes from a finite element discretization of the Neumann eigenvalue problem:

$$-\triangle u = \lambda u, \qquad \text{in } \Omega, \tag{10.18a}$$

$$\frac{\partial u}{\partial \nu} = 0, \qquad \text{on } \partial\Omega, \tag{10.18b}$$

where Ω is the unit square. It has an eigenvalue π^2 of multiplicity 2. We use linear Lagrange elements on a triangular mesh with $h \approx 0.025$ to discretize and obtain a generalized eigenvalue problem

$$A\mathbf{x} = \lambda B\mathbf{x}, \tag{10.19}$$

where the stiffness matrix A and mass matrix B are 2075×2075. The discretization has broken the symmetry. In (10.19), the eigenvalue of multiplicity 2 has been approximated by a very close pair of eigenvalues given by

$$\lambda_1 = 9.872899741642826 \quad \text{and} \quad \lambda_2 = 9.872783160389966.$$

With $\epsilon = 1.0e-3 \times (\pm 1 \pm i)$ RIM fails to separate the eigenvalues and we obtain only one eigenvalue

$$\lambda_1 \approx 9.872680664062500.$$

However, with $\epsilon = 1.0e-9 \times (\pm 1 \pm i)$, RIM separates the eigenvalues and we obtain

$$\lambda_1 \approx 9.872899741516449 \quad \text{and} \quad \lambda_2 \approx 9.872783160419203.$$

The search regions explored by RIM with different tolerances are shown in Fig. 10.5.

Example 7: As the last example we compute the eigenvalues of the 40×40 Wilkinson matrix

$$A = \begin{pmatrix} 19 & -1 & & & & & & & \\ -1 & 18 & -1 & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & -1 & 0 & -1 & & & & \\ & & & -1 & 1 & -1 & & & \\ & & & & \ddots & \ddots & \ddots & & \\ & & & & & -1 & 19 & -1 \\ & & & & & & -1 & 20 \end{pmatrix},$$

which is known to have very close eigenvalues. With $\epsilon = 1.0e-14 \times (\pm 1 \pm i)$ and the search region $S = [-2, 10] \times [-2, 10]$, RIM accurately distinguishes the close eigenvalues. The results are shown in Table 10.7 and Fig. 10.6.

1	-1.125441522046458	11	5.000236265619321
2	0.253805817279499	12	5.999991841327017
3	0.947534367500339	13	6.000008352188331
4	1.789321352320258	14	6.999999794929806
5	2.130209219467361	15	7.000000207904748
6	2.961058880959172	16	7.999999996191775
7	3.043099288071971	17	8.000000003841876
8	3.996047997334983	18	8.999999999945373
9	4.004353817323874	19	9.000000000054399
10	4.999774319815003	20	9.999999999999261

Table 10.7: The first twenty computed Wilkinson eigenvalues by RIM.

We have seen that RIM can effectively find all eigenvalues in a region when neither the location or number of eigenvalues is known. The key difference between RIM and other counter integral based methods in the literature is that RIM only tests

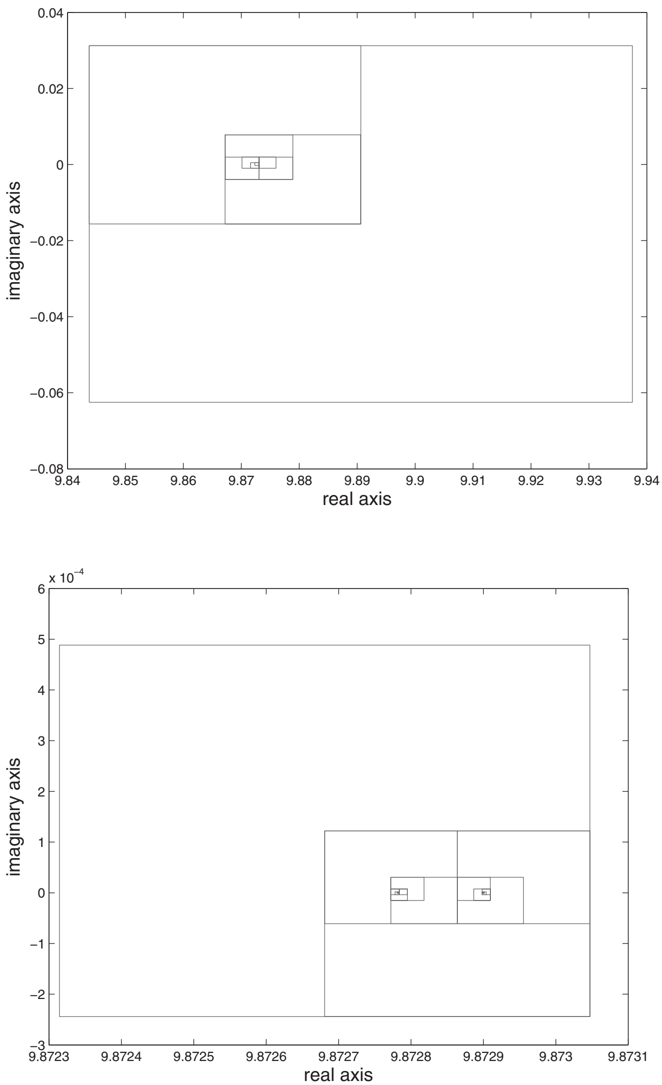


Figure 10.5: The regions explored by RIM. The search region is given by $S = [1, 10] \times [-1, 1]$. Top: $\epsilon = 1.0e - 3 \times (\pm 1 \pm i)$. Bottom: $\epsilon = 1.0e - 9 \times (\pm 1 \pm i)$.

if a region contains eigenvalues. Consequently, accuracy requirements on quadratures, linear solvers, and the number of test vectors may be significantly reduced.

RIM is a non-classical eigenvalue solver which is well suited to problems that only require eigenvalues. In particular, the method not only works for matrix eigen-

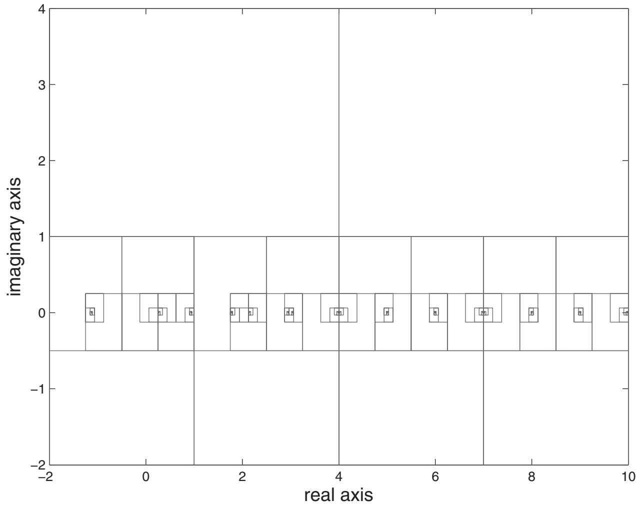


Figure 10.6: The regions explored by RIM for the Wilkinson matrix ($\epsilon = 1.0e-14$).

value problems resulting from suitable numerical approximations of PDE-based eigenvalue problems, but also eigenvalue problems which can not be easily cast as a matrix eigenvalue problem (e.g., see [31, 173]).

10.3 An Integral Eigenvalue Problem

In this section, we solve a nonlinear integral eigenvalue problem using spectrum projection. The problem is a reformulation of the transmission eigenvalue problems. Using a boundary element method (BEM), the integral equations are discretized and a generalized eigenvalue problem of dense matrices is obtained. The matrices are significantly smaller than those from finite element methods. It is shown that, if zero is a generalized eigenvalue of the new system, the corresponding wavenumber k is a transmission eigenvalue.

We present a probing method based on the spectrum projection using contour integrals based on [252]. The contour is chosen to be a small circle centered at 0. Then a numerical quadrature is used to compute the spectrum projection of a random vector. The norm of the projected vector is used as an indicator to decide whether zero is an eigenvalue or not. The idea is similar to RIM. However, it has a very distinct feature in the sense that it only tests if a small disc centered at 0 contains an eigenvalue or not.

10.3.1 Boundary Integral Formulation

In contrast to the assumption on Ω in other sections of the book, we assume that $\Omega \subset \mathbb{R}^2$ is an open bounded domain with C^2 boundary $\partial\Omega$. In addition, we assume the index of refraction n is a constant greater than 1. We recall that the transmission eigenvalue problem is to find $k \in \mathbb{C}$ such that there exist non-trivial solutions w and v satisfying

$$\Delta w + k^2 n w = 0, \qquad \text{in } \Omega, \tag{10.20a}$$

$$\Delta v + k^2 v = 0, \qquad \text{in } \Omega, \tag{10.20b}$$

$$w - v = 0, \qquad \text{on } \partial\Omega, \tag{10.20c}$$

$$\frac{\partial w}{\partial \nu} - \frac{\partial v}{\partial \nu} = 0, \qquad \text{on } \partial\Omega, \tag{10.20d}$$

where ν is the unit outward normal to $\partial\Omega$.

In the following, we describe an integral formulation of the transmission eigenvalue problem following [98] (see also [173]). Let Φ_k be the Green's function given by

$$\Phi_k(x, y) = \frac{i}{4} H_0^{(1)}(k|x - y|),$$

where $H_0^{(1)}$ is the Hankel function of the first kind of order 0. The single and double layer potentials are defined as

$$(S_k\phi)(x) = \int_{\partial\Omega} \Phi_k(x, y)\phi(x)\, \mathrm{d}s(y),$$
$$(K_k\phi)(x) = \int_{\partial\Omega} \frac{\partial \Phi_k}{\partial \nu(y)}(x, y)\phi(x)\, \mathrm{d}s(y),$$

where ϕ is the density function.

Let $(v, w) \in H^1(\Omega) \times H^1(\Omega)$ be a solution to (10.20). Denote by $k_1 = \sqrt{n}k$ and set

$$\alpha := \frac{\partial v}{\partial \nu}\Big|_{\partial\Omega} = \frac{\partial w}{\partial \nu}\Big|_{\partial\Omega} \in H^{-1/2}(\partial\Omega),$$
$$\beta := v|_{\partial\Omega} = w|_{\partial\Omega} \in H^{1/2}(\partial\Omega).$$

Then v and w have the following integral representation

$$v = S_k\alpha - K_k\beta, \qquad \text{in } \Omega, \tag{10.21a}$$

$$w = K_{k_1}\alpha - K_{k_1}\beta, \qquad \text{in } \Omega. \tag{10.21b}$$

Let $u := w - v$. Then $u|_{\partial\Omega} = 0$ and $\frac{\partial u}{\partial \nu}|_{\partial\Omega} = 0$. The boundary conditions of (10.20) imply that the transmission eigenvalues are k's such that

$$Z(k)\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = 0, \tag{10.22}$$

where

$$Z(k) = \begin{pmatrix} S_{k_1} - S_k & -K_{k_1} + K_k \\ -K'_{k_1} + K'_k & T_{k_1} - T_k \end{pmatrix}.$$

The potentials K'_k, T_k are given by

$$(K'_k\phi)(x) = \int_{\partial\Omega} \frac{\partial \Phi_k}{\partial \nu(x)}(x,y)\phi(y)\,\mathrm{d}s(y), \tag{10.23a}$$

$$(T_k\psi)(x) = \frac{\partial}{\partial \nu(x)} \int_{\partial\Omega} \frac{\partial \Phi_k}{\partial \nu(y)}(x,y)\phi(y)\,\mathrm{d}s(y). \tag{10.23b}$$

It is shown in [98] that

$$Z(k) := H^{-3/2}(\partial\Omega) \times H^{-1/2}(\partial\Omega) \to H^{3/2}(\partial\Omega) \times H^{1/2}(\partial\Omega)$$

is of Fredholm type with index zero and analytic on $\mathbb{C} \setminus \mathbb{R}^-$.

From (10.22), k is a transmission eigenvalue if 0 is an eigenvalue of $Z(k)$. Unfortunately, $Z(k)$ is compact. The eigenvalues of $Z(k)$ accumulate at zero, which makes it impossible to distinguish between zero and other eigenvalues numerically. The work-around proposed in [97] is to consider a generalized eigenvalue problem

$$Z(k)\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \lambda B(k) \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \tag{10.24}$$

where $B(k) = Z(ik)$. Since there does not exist purely imaginary transmission eigenvalues [95], the accumulation point is shifted to -1. Then 0 becomes isolated.

Now we describe the boundary element discretization of the potentials and refer the readers to [210, 223] for more details. The boundary $\partial\Omega$ is partitioned into element segments. Suppose that $\partial\Omega$ is discretized into N segments

$$\partial\Omega_1, \partial\Omega_2, ..., \partial\Omega_N$$

by nodes $x_1, x_2, ..., x_N$ and $\partial\Omega = \cup_{i=1}^N \partial\Omega_i$. Let $\{\psi_j\}, j = 1,2,...,N$, be piecewise constant basis functions and $\{\varphi_j\}, j = 1,2,...,N$, be piecewise linear basis functions. We seek an approximate solution α_h and β_h in the form

$$\alpha_h = \sum_{j=1}^N \alpha_j \psi_j$$

and

$$\beta_h = \sum_{j=1}^N \beta_j \varphi_j.$$

The discrete problem becomes

$$\begin{aligned} (V_{k,h} - V_{k_1,h})\vec{\alpha} + (-K_{k,h} + K_{k_1,h})\vec{\beta} &= 0, \\ (K'_{k,h} - K'_{k_1,h})\vec{\alpha} + (W_{k,h} - W_{k_1,h})\vec{\beta} &= 0, \end{aligned}$$

where $\vec{\alpha} = (\alpha_1, ..., \alpha_N)^T$, $\vec{\beta} = (\beta_1, ..., \beta_N)^T$, and $V_{k,h}, K_{k,h}, K'_{k,h}, W_{k,h}$ are matrices with entries

$$\begin{aligned} V_{k,h}(i,j) &= \int_{\partial\tilde{\Omega}} (S_k\psi_j)\psi_i \, \mathrm{d}s, \\ K_{k,h}(i,j) &= \int_{\partial\tilde{\Omega}} (K_k\varphi_j)\psi_i \, \mathrm{d}s, \\ K'_{k,h}(i,j) &= \int_{\partial\tilde{\Omega}} (K'_k\psi_j)\varphi_i \, \mathrm{d}s, \\ W_{k,h}(i,j) &= \int_{\partial\tilde{\Omega}} (T_k\varphi_j)\varphi_i \, \mathrm{d}s. \end{aligned}$$

In the above matrices, we can use series expansions of the first kind Hankel function as

$$\begin{aligned} H_0^{(1)}(x) &= \sum_{m=0}^{\infty} \frac{(-1)^m}{(m!)^2}\left(\frac{x}{2}\right)^{2m} + \frac{2i}{\pi}\sum_{m=0}^{\infty} \frac{(-1)^m}{(m!)^2}\left(\frac{x}{2}\right)^{2m}\left(\ln\frac{x}{2} + c_e\right) \\ &\quad -\frac{2i}{\pi}\sum_{m=0}^{\infty} \frac{(-1)^m}{(m!)^2}\left(\frac{x}{2}\right)^{2m}\left(1 + \frac{1}{2} + \frac{1}{m}\right), \end{aligned}$$

where c_e is the Euler constant. Thus,

$$\begin{aligned} H_0^{(1)}(k|x-y|) &= \sum_{m=0}^{\infty}\left(C_5(m) + C_6(m)\ln\frac{k}{2}\right)k^{2m}|x-y|^{2m} \\ &\quad + C_6(m)\ln|x-y|k^{2m}|x-y|^{2m}, \end{aligned}$$

where

$$\begin{aligned} C_5(m) &= \frac{(-1)^m}{2^{2m}(m!)^2}\left[1 + \frac{2c_e i}{\pi} - \frac{2i}{\pi}\left(1 + \frac{1}{2} + \frac{1}{m}\right)\right], \\ C_6(m) &= \frac{(-1)^m i}{2^{2m-1}(m!)^2\pi}. \end{aligned}$$

We also need the following integrals which can be computed exactly.

$$\begin{aligned} Int_7(m) &= \int_{-1}^{1}\int_{-1}^{1}(\xi_1 - \xi_2)^{2m}\mathrm{d}\xi_2\mathrm{d}\xi_1 \\ &= \frac{2^{2m+2}}{(2m+1)(m+1)}, \end{aligned}$$

$$\begin{aligned} Int_8(m) &= \int_{-1}^{1}\int_{-1}^{1}(\xi_1 - \xi_2)^{2m}\ln|\xi_1 - \xi_2|\mathrm{d}\xi_2\mathrm{d}\xi_1 \\ &= \frac{2^{2m+2}\ln 2}{(2m+1)(m+1)} - \frac{(4m+3)2^{2m+3}}{(2m+1)^2(2m+2)^2}, \end{aligned}$$

$$\begin{aligned} Int_9(m) &= \int_{-1}^{1}\int_{-1}^{1}(\xi_1-\xi_2)^{2m}\xi_1\xi_2\mathrm{d}\xi_2\mathrm{d}\xi_1 \\ &= \sum_{l=0}^{2m}\frac{(-1)^l C_{2m}^l}{(l+2)(2m+2-l)}[1-(-1)^l]^2, \end{aligned}$$

and

$$\begin{aligned} Int_{10}(m) &= \int_{-1}^{1}\int_{-1}^{1}(\xi_1-\xi_2)^{2m}\xi_1\xi_2\ln|\xi_1-\xi_2|\mathrm{d}\xi_2\mathrm{d}\xi_1 \\ &= \frac{-m2^{2m+2}\ln 2}{(2m+1)(m+1)(m+2)} \\ &\qquad +\frac{1}{(2m+1)(m+1)}\left[\frac{2^{2m+3}}{2m+3}-\frac{2^{2m+2}}{(m+2)^2}-\frac{2^{2m+1}}{m+1}\right] \\ &\qquad\qquad +\frac{1}{2(m+1)^2(2m+1)^2}\cdot Q, \end{aligned}$$

where

$$Q=\sum_{l=0}^{2m+1}C_{2m+1}^l\left[\frac{(2m+1)^2}{l+2}(1-(-1)^l)-\frac{4m+3}{l+3}(1-(-1)^{l+1})\right].$$

Now we consider

$$\begin{aligned} V_{k,h}(i,j) &= \int_{\partial\tilde{\Omega}}(V_k\psi_j)\psi_i\mathrm{d}s \\ &= \int_{\partial\tilde{\Omega}}\int_{\partial\tilde{\Omega}}\Phi_k(x,y)\psi_j(y)\psi_i(x)\mathrm{d}s_y\mathrm{d}s_x \\ &= \int_{\partial\Omega_i}\int_{\partial\Omega_j}\Phi_k(x,y)\psi_j(y)\psi_i(x)\mathrm{d}s_y\mathrm{d}s_x. \end{aligned}$$

The integral over $\partial\Omega_i\times\partial\Omega_j$ can be calculated as

$$\begin{aligned} &\int_{\partial\Omega_i}\int_{\partial\Omega_j}\Phi_k(x,y)\psi_j(y)\psi_i(x)\mathrm{d}s_y\mathrm{d}s_x \\ =\;& \frac{i}{4}\int_{\partial\Omega_i}\int_{\partial\Omega_j}H_0^{(1)}(k|x-y|)\psi_j(y)\psi_i(x)\mathrm{d}s_y\mathrm{d}s_x \\ =\;& \frac{iL_iL_j}{16}\int_{-1}^{1}\int_{-1}^{1}H_0^{(1)}(k|x(\xi_1)-y(\xi_2)|)\mathrm{d}\xi_2\mathrm{d}\xi_1, \end{aligned}$$

where

$$\begin{aligned} x(\xi_1) &= x_i+\frac{1+\xi_1}{2}(x_{i+1}-x_i), \\ y(\xi_2) &= x_j+\frac{1+\xi_2}{2}(x_{j+1}-x_j). \end{aligned}$$

When $i \neq j$, it can be calculated by Gaussian quadrature rule. When $i = j$, we have

$$
\begin{aligned}
&\frac{iL_i^2}{16}\int_{-1}^{1}\int_{-1}^{1} H_0^{(1)}(k|x(\xi_1)-y(\xi_2)|)\mathrm{d}\xi_2\mathrm{d}\xi_1\\
&= \frac{iL_i^2}{16}\sum_{m=0}^{\infty}\frac{k^{2m}L_i^{2m}}{2^{2m}}\left(C_5(m)+C_6(m)\ln\frac{kL^i}{4}\right)\int_{-1}^{1}\int_{-1}^{1}(\xi_1-\xi_2)^{2m}\mathrm{d}\xi_2\mathrm{d}\xi_1\\
&\qquad + \frac{iL_i^2}{16}\sum_{m=0}^{\infty}\frac{k^{2m}L_i^{2m}}{2^{2m}}C_6(m)\int_{-1}^{1}\int_{-1}^{1}(\xi_1-\xi_2)^{2m}\ln|\xi_1-\xi_2|\mathrm{d}\xi_2\mathrm{d}\xi_1\\
&= \sum_{m=0}^{\infty}\frac{ik^{2m}L_i^{2m+2}}{2^{2m+4}}\left[\left(C_5(m)+C_6(m)\ln\frac{kL^i}{4}\right)Int_7(m)+C_6(m)Int_8(m)\right].
\end{aligned}
$$

The following regularization formulation is needed to discretize the hypersingular boundary integral operator

$$
W_k\beta(x) = -\frac{d}{ds_x}V_k(\frac{d\beta}{ds})(x) - k^2\nu_x\cdot V_k(\beta\nu)(x). \tag{10.25}
$$

We refer the readers to [150] for details of the discretization.

The above boundary element method leads to the following generalized eigenvalue problem

$$
A\mathbf{x} = \lambda B\mathbf{x}, \tag{10.26}
$$

where $A, B \in \mathbb{C}^{n\times n}$, $\lambda \in \mathbb{C}$ is a scalar, and $\mathbf{x} \in \mathbb{C}^n$.

To compute transmission eigenvalues, the following method is proposed in [97]. A searching interval for wave numbers is discretized. For each k, the boundary integral operators $Z(k)$ and $Z(ik)$ are discretized to obtain (10.26). Then all eigenvalues of (10.26) are computed and arranged such that

$$
0 \leq |\lambda_1(k)| \leq |\lambda_2(k)| \leq \dots .
$$

If k is a transmission eigenvalue, $|\lambda_1|$ is very close to 0. If one plots the inverse of $|\lambda_1(k)|$ against k, the transmission eigenvalues are located at spikes (see Fig. 10.9).

10.3.2 A Probing Method

The method in [97] only uses the smallest eigenvalue. Hence it is not necessary to compute all eigenvalues of (10.24). In fact, there is no need to know the exact value of λ_1. The only thing we need is that, if k is a transmission eigenvalue, the generalized eigenvalue problem (10.24) has an isolated eigenvalue close to 0. This motivates us to propose a probing method to test if 0 is a generalized eigenvalue of (10.24). The method does not compute the actual eigenvalue and only solves a couple of linear systems. The workload is reduced significantly in two dimensions. Much more savings are expected in three dimensions.

We recall the the spectrum projection of the generalized eigenvalue problem

$$
P_k(A,B) = \frac{1}{2\pi i}\int_\Gamma (zB-A)^{-1}\mathrm{d}z. \tag{10.27}
$$

We write P_k to emphasize that P depends on the wavenumber k. Let $\mathbf{f} \in \mathbb{C}^n$ be randomly chosen. As we discussed before, if there are no eigenvalues inside Γ, we have that $P\mathbf{f} = 0$. Therefore, $P_k\mathbf{f}$ can be used to decide if Γ encloses eigenvalues or not.

The approximation of $P_k\mathbf{f}$ is computed by a quadrature rule

$$\begin{aligned} P_k\mathbf{f} &= \frac{1}{2\pi i}\int_\Gamma R_z(A,B)\mathbf{f}\mathrm{d}z \\ &\approx \frac{1}{2\pi i}\sum_{j=1}^{W}\omega_j R_{z_j}(A,B)\mathbf{f} \\ &= \frac{1}{2\pi i}\sum_{j=1}^{W}\omega_j \mathbf{x}_j, \end{aligned} \tag{10.28}$$

where w_j are weights and z_j are quadrature points. Here $\mathbf{x}_j$'s are the solutions of the following linear systems

$$(z_jB - A)\mathbf{x}_j = \mathbf{f}, \quad j = 1, \ldots, W. \tag{10.29}$$

Similar to the previous section, we project the random vector twice, i.e., we compute $P_k^2\mathbf{f}$.

For a fixed wavenumber k, the algorithm of the probing method is as follows.

Input:

a small circle Γ center at the origin with radius $r \ll 1$,

a random $\mathbf{f}$,

Output:

0 - k is not a transmission eigenvalue,

1 - k is a transmission eigenvalue,

1. compute $P_k^2\mathbf{f}$ by (10.28);
2. decide if Γ contains an eigenvalue:

   No. output 0.

   Yes. output 1.

10.3.3 Numerical Examples

We start with an interval (a,b) of wavenumbers and uniformly divide it into K subintervals. At each wavenumber

$$k_j = a + jh, \quad j = 0, 1, \ldots, K, \; h = \frac{b-a}{K},$$

$m = 0$	1.9880	3.7594	6.5810
$m = 1$	2.6129	4.2954	5.9875
$m = 2$	3.2240	4.9462	6.6083

Table 10.8: TEs of a disc with radius $r = 1/2$ and index of refraction $n = 16$.

we employ the boundary element method to discretize the potentials. We choose $N = 32$ and end up with a generalized eigenvalue problem (10.26) with 64×64 matrices A and B.

To test whether 0 is a generalized eigenvalue of (10.26), we choose Γ to be a circle of radius $1/100$. Then we use 16 uniformly distributed quadrature points on Γ and evaluate the eigenprojection (10.28). If at a wavenumber k_j, the projection is of $O(1)$, then k_j is a transmission eigenvalue. For the actual computation, we use a threshold value $\sigma = 1/2$ to decide if k_j is a transmission eigenvalue or not, i.e., k_j is a transmission eigenvalue if

$$\|P_{k_j}^2 \mathbf{f}\| / \|P_{k_j} \mathbf{f}\| \geq \sigma$$

and not otherwise.

Let Ω be a disc with radius $1/2$. The index of refraction is $n = 16$. In this case, the exact transmission eigenvalues are known [95]. For convenience, we list the eigenvalues again in Table (10.8).

We choose the interval to be $(1.5, 3.5)$ and uniformly divide it into 2000 subintervals. At each k_j we compute the projection (10.28) twice. The probing method finds three eigenvalues in $(1.5, 3.5)$

$$k_1 = 1.988, \quad k_2 = 2.614, \quad k_3 = 3.228,$$

which approximate the exact eigenvalues (the first column of Table (10.8)) accurately. We also plot the log of $|P^2 \mathbf{f}|$ against the wavenumber k in Fig. 10.7. The method is robust since the eigenvalues can be easily identified.

We repeat the experiment by choosing $n = 9$ and $(a, b) = (3, 5)$. The other parameters keep the same. The following eigenvalues are obtained

$$k_1 = 3.554, \quad k_2 = 4.360.$$

The log of $|P^2 \mathbf{f}|$ against the wavenumber k is shown in Fig. 10.8.

Finally, we compare the above method with the method in [97]. We implement the algorithm in [97]. We take $n = 16$ and compute for 2000 wavenumbers. The CPU time in seconds is shown in Table 10.9. Note that all the computation is done using Matlab R2014a on a MacBook Pro with a 3 GHz Intel Core i7 and 16 GB memory. We can see that the proposed method saves more time if the size of the generalized eigenvalue problem is larger. We expect that it has a greater advantage for three-dimensional problems since the size of the matrices is much larger than in two-dimensional cases.

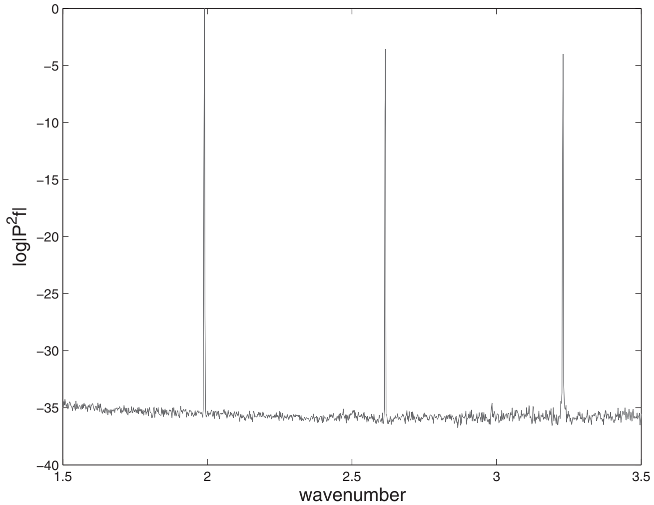


Figure 10.7: The plot of $\log|P^2\mathbf{f}|$ against the wavenumber k for $n = 16$.

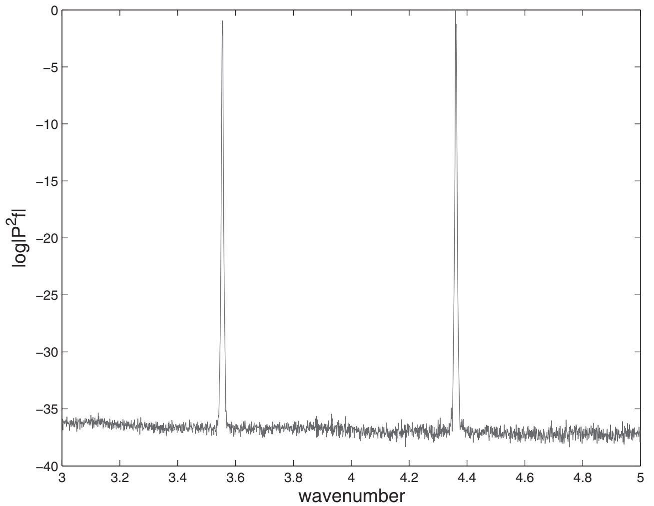


Figure 10.8: The plot of $\log|P^2\mathbf{f}|$ against the wavenumber k for $n = 9$.

size	probing method	method in [97]	ratio
64×64	1.741340	5.742839	3.30
128×128	5.653961	31.152448	5.51
256×256	25.524530	224.435704	8.79
512×512	130.099433	1822.545973	14.01

Table 10.9: Comparison of the probing method and the method in [97]. The first column is the size of the matrix problem. The second column is the time used by the proposed method in seconds. The second column is the time used by the method given in [97]. The fourth column is the ratio.

We also show the log plot of $1/|\lambda_1|$ by the method of [97] in Figure 10.9. Comparing with Figures 10.7 and 10.8, it is clear that the probing method has a much narrower span indicating that the probing method is more effective.

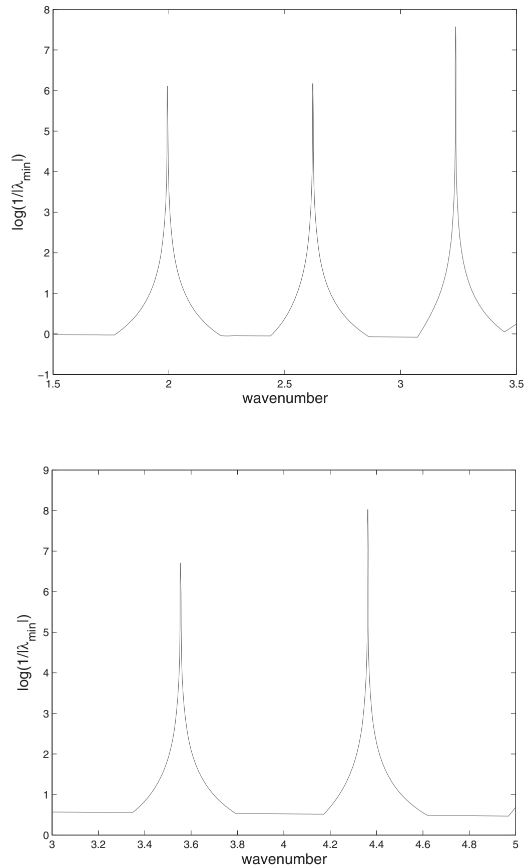


Figure 10.9: Log plot of $1/|\lambda_{min}|$. Top: $n = 16$. Bottom: $n = 9$.

Bibliography

[1] M. Abramowitz and I.A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1965.

[2] A.Buffa, P. Houston, and I. Perugia. Discontinuous Galerkin computation of the Maxwell eigenvalues on simplicial meshes. *J. Comput. Appl. Math.*, 204(2):317–333, 2007.

[3] R.A. Adams. *Sobolel Spaces, Volume 65 of Pure and Applied Mathematics*. Academic Press, New York, 1975.

[4] A. Adini and R.W. Clough. Analysis of plate bending by the finite element method. Technical Report G. 7337, NSF, 1961.

[5] T. Aktosun, D. Gintides, and V. Papanicolaou. The uniqueness in the inverse problem for transmission eigenvalues for the spherically symmetric variable-speed wave equation. *Inverse Problems*, 27(11):115004, 2011.

[6] J. An and J. Shen. A Fourier-spectral-element method for transmission eigenvalue problems. *J. Sci. Comput.*, 57:670–688, 2013.

[7] J. An and J. Shen. Efficient spectral methods for transmission eigenvalues and estimation of the index of refraction. *J. Math. Study*, 47(1):1–20, 2014.

[8] J. An and J. Shen. Spectral approximation to a transmission eigenvalue problem and its applications to an inverse problem. *Comput. Math. Appl.*, 69(10):1132–1143, 2015.

[9] A.B. Andreev and T.D. Todorov. Isoparametric finite-element approximation of a Steklov eigenvalue problem. *IMA J. Numer. Anal.*, 24:309–322, 2004.

[10] A.L. Andrew, K.W. Eric Chu, and P. Lancaster. Derivatives of eigenvalues and eigenvectors of matrix functions. *SIAM J. Matrix Anal. Appl.*, 14(4):903–926, 1993.

[11] P. F. Antonietti, A. Buffa, and I. Perugia. Discontinuous Galerkin approximation of the Laplace eigenproblem. *Comput. Methods Appl. Mech. Engrg.*, 195:3483–3503, 2006.

[12] P. Arbenz and R. Geus. A comparison of solvers for large eigenvalue problems occurring in the design of resonant cavities. *Numer. Linear Algebra Appl.*, 6:3–16, 1999.

[13] J.H. Argyris, I. Fried, and D.W. Scharpf. The TUBA family of plate elements for the matrix displacement method. *Aero. J. Roy. Aero. Soc.*, 72:701–709, 1968.

[14] M.G. Armentano and C. Padra. A posteriori error estimates for the Steklov eigenvalue problem. *Appl. Numer. Math.*, 58:593–601, 2008.

[15] J. Asakura, T. Sakurai, H. Tadano, T. Ikegami, and K. Kimura. A numerical method for nonlinear eigenvalue problems using contour integrals. *JSIAM Lett.*, 1:52–55, 2009.

[16] K. Atkinson and W. Han. *Theoretical Numerical Analysis: A Functional Analysis Framework*. Springer, New York, 3rd edition, 2009.

[17] K.E. Atkinson. *An Introduction to Numerical Analysis*. John Wiley & Sons, Inc., 2nd edition, 1989.

[18] J.P. Aubin. Behaviour of the error of the approximate solution of boundary value problems for linear elliptic operators by Galerkin's and finite difference methods. *Ann. Scuola Norm. Sup. Pisa*, 21:599–637, 1967.

[19] A.P. Austin, P. Kravanja, and L.N. Trefethen. Numerical algorithms based on analytic function values at roots of unity. *SIAM J. Numer. Anal.*, 52(4):1795–1821, 2014.

[20] I. Babuška, J. Osborn, and J. Pitkäranta. Analysis of mixed methods using mesh dependent norms. *Math. Comp.*, 35:1039–1062, 1980.

[21] I. Babuška and J. Osborn. Estimates for the errors in eigenvalue and eigenvector approximation by Galerkin methods, with particular attention to the case of multiple eigenvalues. *SIAM J. Numer. Anal.*, 24(6):1249–1276, 1987.

[22] I. Babuška and J. Osborn. Finite element-Galerkin approximation of the eigenvalues and eigenvectors of self-adjoint problems. *Math. Comp.*, 52(186):275–297, 1989.

[23] I. Babuška and J. Osborn. *Eigenvalue Problems, Handbook of Numerical Analysis, Vol. II, Finite Element Methods (Part 1), Edited by P.G. Ciarlet and J.L. Lions*. Elseveier Science Publishers B.V. (North-Holland), 1991.

[24] C. Bacuta, J. Bramble, and J. Pasciak. Shift theorems for the biharmonic Dirichlet problem. In *Recent Progress in Computational and Applied PDEs*, Proceedings for the CAPDE conference held in Zhangjiajie, China. Kluwer Academic/Plenum Publishers, 2001.

[25] G.R. Barrenechea, L. Boulton, and N. Boussaïd. Finite element eigenvalue enclosures for the Maxwell operator. *SIAM J. Sci. Comput.*, 36(6):A2887–A2906, 2014.

[26] G.P. Bazeley, Y.K. Cheung, B.M. Irons, and O.C. Zienkiewicz. Triangular elements in bending - conforming and nonconforming solutions. In *Proceedings of the Conference on Matrix Methods in Structural Mechanics*, pages 547–576, Wright Patterson A.F.B., Ohio, 1965.

[27] M. Bellalij, Y. Saad, and H. Sadok. Further analysis of the Arnoldi process for eigenvalue problems. *SIAM J. Numer. Anal.*, 48(2):393–407, 2010.

[28] A. Bermúdez, R. Durán, M. A. Muschietti, R. Rodríguez, and J. Solomin. Finite element vibration analysis of fluid-solid systems without spurious modes. *SIAM J. Numer. Anal.*, 32:1280–1295, 1995.

[29] A. Bermúdez and D.C. Pedreira. Mathematical analysis of a finite element method without spurious solutions for computation of dielectric waveguides. *Numer. Math.*, 61(1):39–57, 1992.

[30] C. Bernardi and V. Girault. A local regularization operator for triangular and quadrilateral finite elements. *SIAM J. Numer. Anal.*, 35:1893–1916, 1998.

[31] W.J. Beyn. An integral method for solving nonlinear eigenvalue problems. *Linear Algebra Appl.*, 436(10):3839–3863, 2012.

[32] P.E. Bjørstad and B.P. Tjøstheim. High precision solution of two fourth order eigenvalue problems. *Computing*, 63:97–107, 1997.

[33] H. Blum and R. Rannacher. On the boundary value problem of the biharmonic operator on domains with angular corners. *Math. Methods Appl. Sci.*, 2:556–581, 1980.

[34] D. Boffi. Fortin operator and discrete compactness for edge elements. *Numer. Math.*, 87(2):229–246, 2000.

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

[36] D. Boffi, F. Brezzi, and L. Gastaldi. On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form. *Math. Comp.*, 69:121–140, 2000.

[37] D. Boffi, R.G. Duran, and L. Gastaldi. A remark on spurious eigenvalues in a square. *Appl. Math. Lett.*, 12(3):107–114, 1999.

[38] D. Boffi, P. Fernandes, L. Gastaldi, and I. Perugia. Computational models of electromagnetic resonators: analysis of edge element approximation. *SIAM J. Numer. Anal.*, 36(4):1264–1290, 2004.

[39] D. Boffi and L. Gastaldi. Edge finite elements for the approximation of Maxwell resolvent operator. *M2AN Math. Model. Numer. Anal.*, 36(2):293–305, 2002.

[40] D. Boffi and L. Gastaldi. Some remarks on finite element approximation of multiple eigenvalues. *Appl. Numer. Math.*, 79:18–28, 2014.

[41] D. Boffi, F. Kikuchi, and J. Schöberl. Edge element computation of Maxwell's eigenvalues on general quadrilateral meshes. *Math. Models Methods Appl. Sci.*, 16:265–273, 2006.

[42] A. Bossavit. Solving Maxwell's equations in a closed cavity and the question of spurious modes. *IEEE Trans. Magnetics*, 26:702–705, 1990.

[43] A. Bossavit. *Computational Electromagnetism. Variational Formulations, Complementarity, Edge Elements*. Academic Press, San Diego, CA, 1998.

[44] D. Braess. *Finite Elements, Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, Cambridge, 2001.

[45] J.H. Brambel, T.V. Kolev, and J.E. Pasciak. The approximation of the Maxwell eigenvalue problem using a least-square method. *Math. Comp.*, 74(252):1575–1598, 2005.

[46] J. H. Bramble and J. E. Osborn. Rate of convergence estimates for nonself-adjoint eigenvalue approximations. *Math. Comp.*, 27:525–549, 1973.

[47] J.H. Bramble and J.E. Osborn. Approximation of Steklov eigenvalues of nonself-adjoint second order elliptic operators. In A.K. Aziz, editor, *Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, pages 387–408. Academic Press, New York, 1972.

[48] H. Brandsmeier, K. Schmidt, and C. Schwab. A multiscale hp-FEM for 2D photonic crystal bands. *J. Comput. Phys.*, 230(2):349–374, 2011.

[49] S.C. Brenner. C^0 interior penalty methods. In *Frontiers in Numerical Analysis - Durham 2010*, volume 85 of *Lect. Notes Comput. Sci. Eng.*, pages 79–147. Springer-Verlag, 2012.

[50] S.C. Brenner, S. Gu, T. Gudi, and L.-Y. Sung. A quadratic C^0 interior penalty method for linear fourth order boundary value problems with boundary conditions of the Cahn-Hilliard type. *SIAM J. Numer. Anal.*, 50:2088–2110, 2012.

[51] S.C. Brenner, F. Li, and L. Sung. Nonconforming Maxwell eigensolvers. *J. Sci. Comput.*, 40(1-3):51–85, 2009.

[52] S.C. Brenner, P. Monk, and J. Sun. *C^0 interior penalty Galerkin method for biharmonic eigenvalue problems*, volume 106 of *Lect. Notes Comput. Sci. Eng.*, pages 3–15. Springer, Switzerland, 2015.

[53] S.C. Brenner and M. Neilan. A C^0 interior penalty method for a fourth order elliptic singular perturbation problem. *SIAM J. Numer. Anal.*, 49(2):869–892, 2011.

[54] S.C. Brenner and L.R. Scott. *The Mathematical Theory of Finite Element Methods*. Number 15 in Texts in Applied Mathematics. Springer, New York, 3rd edition, 2008.

[55] S.C. Brenner and L. Sung. C^0 interior penalty methods for fourth order elliptic boundary value problems on polygonal domains. *J. Sci. Comput.*, 22/23:463–478, 2005.

[56] S.C. Brenner, K. Wang, and J. Zhao. Poincaré-Friedrichs inequalities for piecewise H^2 functions. *Numer. Funct. Anal. Optim.*, 25(5-6):463–478, 2004.

[57] A. Buffa, P. Houston, and I. Perugia. Discontinuous Galerkin computation of the Maxwell eigenvalues on simplicial meshes. *J. Comput. Appl. Math.*, 204:317–333, 2007.

[58] A. Buffa and I. Perugia. Discontinuous Galerkin approximation of the Maxwell eigenproblem. *SIAM J. Numer. Anal.*, 44:2198–2226, 2006.

[59] A. Buffa, I. Perugia, and T. Warburton. The mortar-discontinuous Galerkin method for the 2D Maxwell eigenproblem. *J. Sci. Comput.*, 40(1-3):86–114, 2009.

[60] F. Cakoni, D. Colton, and H. Haddar. On the determination of Dirichlet and transmission eigenvalues from far field data. *C. R. Math. Acad. Sci. Paris Ser. I*, 348:379–383, 2010.

[61] F. Cakoni, D. Colton, and P. Monk. *The Linear Sampling Method in Inverse Electromagnetic Scattering*. CBMS-NSF Regional Conference Series in Applied Mathematics 80. SIAM, Philadelphia, 2011.

[62] F. Cakoni, D. Colton, P. Monk, and J. Sun. The inverse electromagnetic scattering problem for anisotropic media. *Inverse Problems*, 26:074004, 2010.

[63] F. Cakoni and D. Gintides. New results on transmission eigenvalues. *Inverse Probl. Imaging*, 4(1):39–48, 2010.

[64] F. Cakoni, D. Gintides, and H. Haddar. The existence of an infinite discrete set of transmission eigenvalues. *SIAM J. Math. Analysis*, 42(1):237–255, 2010.

[65] F. Cakoni and H. Haddar. On the existence of transmission eigenvalues in an inhomogeneous medium. *Appl. Anal.*, 88(4):475–493, 2009.

[66] F. Cakoni and H. Haddar. Transmission eigenvalues in inverse scattering theory. *Inside Out II, G. Uhlmann editor*, 60:527–578, 2012.

[67] F. Cakoni and A. Kirsch. On the interior transmission eigenvalue problem. *Int. Jour. Comp. Sci. Math.*, 3(1-2):142–167, 2010.

[68] F. Cakoni, P. Monk, and J. Sun. Error analysis of the finite element approximation of transmission eigenvalues. *Comput. Methods Appl. Math.*, 14(4):419–427, 2014.

[69] E. Cancès, R. Chakir, and Y. Maday. Numerical analysis of nonlinear eigenvalue problems. *J. Sci. Comput.*, 45(1-3):90–117, 2010.

[70] C. Canuto. Eigenvalue approximations by mixed methods. *RAIRO Anal. Numér.*, 12:27–50, 1978.

[71] L. Cao, L. Zhang, W. Allegretto, and Y. Lin. Multiscale asymptotic method for Steklov eigenvalue equations in composite media. *SIAM J. Numer. Anal.*, 51(1):273–296, 2013.

[72] S. Caorsi, P. Fernandes, and M. Raffetto. On the convergence of Galerkin finite element approximations of electromagnetic eigenproblems. *SIAM J. Numer. Anal.*, 38:580–607, 2000.

[73] C. Carstensen, D. Gallistl, and M. Schedensack. Adaptive nonconforming Crouzeix-Raviart FEM for eigenvalue problems. *Math. Comp.*, 84(293):1061–1087, 2015.

[74] C. Carstensen and J. Gedicke. An oscillation-free adaptive FEM for symmetric eigenvalue problems. *Numer. Math.*, 118(3):401–427, 2011.

[75] C. Carstensen and J. Gedicke. An adaptive finite element eigenvalue solver of asymptotic quasi-optimal computational complexity. *SIAM J. Numer. Anal.*, 50(3):1029–1057, 2012.

[76] C. Carstensen and J. Gedicke. Guaranteed lower bounds for eigenvalues. *Math. Comp.*, 83(290):2605–2629, 2014.

[77] Z. Cendes and P. Silvester. Numerical solution of dielectric loaded waveguides: I-finite element analysis. *IEEE Trans. Microw. Theory Techn.*, 18:1124–1131, 1970.

[78] F. Chatelin. Convergence of approximation methods to compute eigenelements of linear operations. *SIAM J. Numer. Anal.*, 10:939–948, 1973.

[79] F. Chatelin. *Spectral approximation of linear operators*. Classics in Applied Mathematics 65. SIAM, Philadelphia, 2011.

[80] H. Chen, X. Dai, X. Gong, L. He, and A. Zhou. Adaptive finite element approximations for Kohn-Sham models. *Multiscale Model. Simul.*, 12:1828–1869, 2014.

[81] H. Chen, X. Gong, L. He, Z. Yang, and A. Zhou. Numerical analysis of finite dimensional approximations of Kohn-Sham models. *Adv. Comput. Math.*, 38:225–256, 2013.

[82] H. Chen, X. Gong, L. He, and A. Zhou. Adaptive finite element approximations for a class of nonlinear eigenvalue problems in quantum physics. *Adv. Appl. Math. Mech.*, 3(4):493–518, 2011.

[83] H. Chen, X. Gong, and A. Zhou. Numerical approximations of a nonlinear eigenvalue problem and applications to a density functional model. *Math. Methods Appl. Sci.*, 33:1723–1742, 2010.

[84] H. Chen, L. He, and A. Zhou. Finite element approximations of nonlinear eigenvalue problems in quantum physics. *Comput. Methods Appl. Mech. Engrg.*, 200:1846–1865, 2011.

[85] H. Chen, F. Liu, and A. Zhou. A two-scale higher order finite element discretization for Schrödinger equations. *J. Comput. Math.*, 27:315–337, 2009.

[86] L. Chen, M. Holst, and J. Xu. The finite element approximation of the nonlinear Poisson-Boltzmann equation. *SIAM J. Numer. Anal.*, 45(6):2298–2320, 2007.

[87] W. Chen. Eigenvalue approximation of the biharmonic eigenvalue problem by Ciarlet-Raviart scheme. *Numer. Methods Partial Differential Equations*, 21(3):512–520, 2005.

[88] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems, Classics in Applied Mathematics, Volume 40.* SIAM, Philadelphia, 2002.

[89] P.G. Ciarlet and P.A. Raviart. A mixed finite element method for the biharmonic equation. In *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 125–145, Academic Press, New York, 1974.

[90] P. Clément. Approximation by finite element functions using local regularization. *RAIRO Anal. Numer.*, 9:77–84, 1975.

[91] K.A. Cliffe, E. Hall, and P. Houston. Adaptive discontinuous Galerkin methods for eigenvalue problems arising in incompressible fluid flows. *SIAM J. Sci. Comput.*, 31:4607–4632, 2010.

[92] D. Colton and R. Kress. The construction of solutions to acoustic scattering problems in a spherically stratified medium II. *Quart. Jour. Mech. Appl. Math.*, 32:53–63, 1979.

[93] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory, Applied Mathematical Sciences, Volume 93.* Springer, New York, 2013.

[94] D. Colton and P. Monk. The inverse scattering problem for time-harmonic acoustic waves in an inhomogeneous medium. *Quart. Jour. Mech. Applied Math.*, 41:97–125, 1988.

[95] D. Colton, P. Monk, and J. Sun. Analytical and computational methods for transmission eigenvalues. *Inverse Problems*, 26(4):045011, 2010.

[96] D. Colton, L. Päivärinta, and J. Sylvester. The interior transmission problem. *Inverse Probl. Imaging*, 1(1):13–28, 2007.

[97] A. Cossonnière. *Valeurs propres de transmission et leur utilisation dans l'identification d'inclusions à partir de mesures électromagnétiques*. PhD thesis, Université de Toulouse, 2011.

[98] A. Cossonnière and H. Haddar. Surface integral formulation of the interior transmission problem. *J. Integral Equations Appl.*, 25(3):341–376, 2013.

[99] M. Costabel and M. Dauge. Singularities of electromagnetic fields in polyhedral domains. *Arch. Ration. Mech. Anal.*, 151(3):221–276, 2000.

[100] M. Costabel and M. Dauge. Weighted regularization of Maxwell equations in polyhedral domains. A rehabilitation of nodal finite elements. *Numer. Math.*, 2:239–277, 2002.

[101] M. Costabel and M. Dauge. Computation of resonance frequencies for Maxwell equations in non-smooth domains. In *Topics in Computational Wave Propagation*, number 31 in Lect. Notes Comput. Sci. Eng., pages 125–161. Springer, Berlin, 2003.

[102] R. Courant and D. Hilbert. *Methods of Mathematical Physics, Volume I*. Wiley-Interscience, 1962.

[103] X. Dai, L. He, and A. Zhou. Convergence and quasi-optimal complexity of adaptive finite element computations for multiple eigenvalues. *IMA J. Numer. Anal.*, 35:1934–1977, 2015.

[104] X. Dai, L. Shen, and A. Zhou. A local computational scheme for higher order finite element eigenvalue approximations. *Int. J. Numer. Anal. Model.*, 5:570–589, 2008.

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

[106] X. Dai and A. Zhou. Three-scale finite element discretizations for quantum eigenvalue problems. *SIAM J. Numer. Anal.*, 46:295–324, 2008.

[107] R.L. Dailey. Eigenvector derivatives with repeated eigenvalues. *AIAA J.*, 27:486–491, 1989.

[108] M. Dauge. *Elliptic Boundary Value Problems on Corner Domain*. Lecture Notes in Mathematics 1341. Springer-Verlag, Berlin-Heidelberg, 1988.

[109] E.B. Davies and M. Plum. Spectral pollution. *IMA J. Numer. Anal.*, 24(3):417–438, 2004.

[110] J. Davies, F. Fernandez, and G. Philippou. Finite element analysis of all modes in cavities with circular symmetry. *IEEE Trans. Microw. Theory Techn.*, 30:1975–1980, 1982.

[111] C.B. Davis. A partition of unity method with penalty for fourth order problems. *J. Sci. Comput.*, 60:228–248, 2014.

[112] L. Demkowicz and P. Monk. Discrete compactness and the approximation of Maxwell's equation in $\mathbb{R}^3$. *Math. Comp.*, 70:507–523, 2001.

[113] L. Demkowicz, P. Monk, C. Schwab, and L. Vardapetyan. Maxwell eigenvalues and discrete compactness in two dimensions. *Comput. Math. Appl.*, 40:589–605, 2000.

[114] J. Descloux, N. Nassif, and J. Rappaz. On spectral approximation. I. the problem of convergence. *RAIRO Anal. Numér.*, 12(2):97–112, 1978.

[115] J. Descloux, N. Nassif, and J. Rappaz. On spectral approximation. II. error estimates for the Galerkin method. *RAIRO Anal. Numér.*, 12(2):113–119, 1978.

[116] D.A. Dunavant. High degree efficient symmetrical Gaussian quadrature rules for the triangle. *Internat. J. Numer. Methods Engrg.*, 21(6):1129–1148, 1985.

[117] G. Engel, K. Garikipati, T.J.R. Hughes, M.G. Larson, L. Mazzei, and R.L. Taylor. Continuous/discontinuous finite element approximations of fourth order elliptic problems in structural and continuum mechanics with applications to thin beams and plates, and strain gradient elasticity. *Comput. Methods Appl. Mech. Engrg.*, 191:3669–3750, 2002.

[118] J. Evans and T. Hughes. Discrete spectrum analyses for various mixed discretizations of the Stokes eigenproblem. *Comput. Mech.*, 50(6):667–674, 2012.

[119] R.S. Falk and J.E. Osborn. Error estimates for mixed methods. *RAIRO Anal. Numér.*, 14:269–277, 1980.

[120] J. Fang, X. Gao, and A. Zhou. A finite element recovery approach to eigenvalue approximations with applications to electronic structure calculations. *J. Sci. Comput.*, 55:432–454, 2013.

[121] G.J. Fix. Eigenvalue approximation by the finite element method. *Advances in Math.*, 10:300–316, 1973.

[122] D. Gallistl. An optimal adaptive FEM for eigenvalue clusters. *Numer. Math.*, 130(3):467–496, 2015.

[123] X. Gao, F. Liu, and A. Zhou. Three-scale finite element eigenvalue discretizations. *BIT*, 48(3):533–562, 2008.

[124] E.M. Garau, P. Morin, and C. Zuppa. Convergence of adaptive finite element methods for eigenvalue problems. *Math. Models Methods Appl. Sci.*, 19(5):721–747, 2009.

[125] H. Geng, X. Ji, J. Sun, and L. Xu. C^0 IP methods for the transmission eigenvalue problem. *J. Sci. Comput.*, online, 2016.

[126] S. Giani and I.G. Graham. A convergent adaptive method for elliptic eigenvalue problems. *SIAM J. Numer. Anal.*, 47:1067–1091, 2009.

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

[128] D. Gilbarg and N.S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer, Berlin, Heidelberg, New York, 2001.

[129] D. Gintides and N. Pallikarakis. A computational method for the inverse transmission eigenvalue problem. *Inverse Problems*, 29(10):104010, 2013.

[130] G. Giorgi and H. Haddar. Computing estimates of material properties from transmission eigenvalues. *Inverse Problems*, 28(5):055009, 2012.

[131] S. Goedecker. Linear scaling electronic structure methods. *Rev. Modern Phys.*, 71:1085–1123, 1999.

[132] G.H. Golub and C.F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 3rd edition, 1996.

[133] G.H. Golub and H.A. van der Vorst. Eigenvalue computation in the 20th century. *J. of Comp. and Applied Math.*, 123:35–65, 2000.

[134] X. Gong, L. Shen, D. Zhang, and A. Zhou. Finite element approximations for Schrödinger equations with applications to electronic structure computations. *J. Comput. Math.*, 26(3):310–323, 2008.

[135] D.S. Grebenkov and B.-T. Nguyen. Geometrical structure of Laplacian eigenfunctions. *SIAM Rev.*, 55(4):601–667, 2013.

[136] R. D. Grigorieff. Diskrete approximation von eigenwertproblemen i: Qualitative konvergenz. *Numer. Math.*, 24:355–374, 1975.

[137] R.D. Grigorieff. Diskrete approximation von eigenwertproblemen ii: Konvergenzordnung. *Numer. Math.*, 24:415–433, 1975.

[138] R.D. Grigorieff. Diskrete approximation von eigenwertproblemen iii: Asymptotische entwicklungen. *Numer. Math.*, 25:79–97, 1975.

[139] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman, London, 1985.

[140] P. Grisvard. *Singularities in Boundary Value Problems*. Masson, Paris, 1992.

[141] S. Güttel, E. Polizzi, P. Tang, and G. Viaud. Zolotarev quadrature rules and load balancing for the FEAST eigensolver. *SIAM J. Sci. Comput.*, 37(4):A2100–A2122, 2015.

[142] W. Hackbusch. On the computation of approximate eigenvalues and eigenfunctions of elliptic operators by means of a multi-grid method. *SIAM J. Numer. Anal.*, 16:201–215, 1979.

[143] J. Han, Z. Zhang, and Y. Yang. A new adaptive mixed finite element method based on residual type a posterior error estimates for the Stokes eigenvalue problem. *Numer. Methods Partial Differential Equations*, 31:31–53, 2015.

[144] I. Harris, F. Cakoni, and J. Sun. Transmission eigenvalues and non-destructive testing of anisotropic magnetic materials with voids. *Inverse Problems*, 30(3):035016, 2014.

[145] L. He and A. Zhou. Convergence and complexity of adaptive finite element methods for elliptic partial differential equations. *Int. J. Numer. Anal. Model.*, 8(4):615–640, 2011.

[146] V. Heuveline. On the computation of a very large number of eigenvalues for selfadjoint elliptic operators by means of multigrid methods. *J. Comput. Phys.*, 184(1):321–337, 2003.

[147] R. Hiptmair. Finite elements in computational electromagnetism. *Acta Numer.*, 11:237–339, 2002.

[148] R. Hiptmair and P. D. Ledger. Computation of resonant modes for axisymmetric Maxwell cavities using hp-version edge finite elements. *Internat. J. Numer. Methods Engrg.*, 62:1652–1676, 2005.

[149] G. Hsiao, F. Liu, J. Sun, and X. Li. A coupled BEM and FEM for the interior transmission problem in acoustics. *J. Comput. Appl. Math.*, 235(17):5213–5221, 2011.

[150] G.C. Hsiao and L. Xu. A system of boundary integral equations for the transmission problem in acoustics. *Appl. Num. Math.*, 61:1017–1029, 2011.

[151] J. Hu, Y. Huang, and Q. Shen. The lower/upper bound property of approximate eigenvalues by nonconforming finite element methods for elliptic operators. *J. Sci. Comput.*, 58:574–591, 2014.

[152] J. Hu, Y. Huang, and Q. Shen. Constructing both lower and upper bounds for the eigenvalues of elliptic operators by nonconforming finite element methods. *Numer. Math.*, 131(2):273–302, 2015.

[153] H. Huang, S. Chang, C. Chien, and Z. Li. Superconvergence of high order FEMs for eigenvalue problems with periodic boundary conditions. *Comput. Methods Appl. Mech. Engrg.*, 198(30-32):2246–2259, 2009.

[154] P. Huang. Lower and upper bounds of Stokes eigenvalue problem based on stabilized finite element methods. *Calcolo*, 52(1):109–121, 2015.

[155] R. Huang, A. Struthers, J. Sun, and R. Zhang. Recursive integral method for transmission eigenvalues. arXiv:1503.04741.

[156] T.M. Huang, W.Q. Huang, and W.W. Lin. A robust numerical algorithm for computing Maxwell's transmission eigenvalue problems. *SIAM J. Sci. Comput. 37*, 37(5):A2403–A2423, 2015.

[157] K. Ishihara. A mixed finite element method for the biharmonic eigenvalue problems of plate bending. *Publ. Res. Inst. Math. Sci., Kyoto Univ.*, 14:399–414, 1978.

[158] K. Ishihara. On the mixed finite element approximation for the buckling of plates. *Numer. Math.*, 33:195–210, 1979.

[159] X. Ji, H. Geng, J. Sun, and L. Xu. C^0IPG for a fourth order eigenvalue problem. *Commun. Comput. Phys.*, 19(2):393–410, 2016.

[160] X. Ji and J. Sun. A multilevel method for transmission eigenvalues of anisotropic media. *J. Comput. Phys.*, 255:422–435, 2013.

[161] X. Ji, J. Sun, and T. Turner. A mixed finite element method for Helmholtz transmission eigenvalues. *ACM Trans. Math. Software*, 38(4):Algorithm 922, 2012.

[162] X. Ji, J. Sun, and H. Xie. A multigrid method for Helmholtz transmission eigenvalue problems. *J. Sci. Comput.*, 60(3):276–294, 2014.

[163] X. Ji, J. Sun, and Y. Yang. Optimal penalty parameter for C^0 IPDG. *Appl. Math. Lett.*, 37:112–117, 2014.

[164] J.-M. Jin. *The Finite Element Method in Electromagnetics*. Wiley, New York, 1993.

[165] T. Kato. *Perturbation Theory of Linear Operators*. Springer-Verlag, 1989.

[166] P. Keast. Moderate tetrahedral quadrature formulas. *Comput. Methods Appl. Mech. Engrg.*, 55(3):339–348, 1986.

[167] F. Kikuchi. Mixed and penalty formulations for finite element analysis of an eigenvalue problem in electromagnetism. *Comput. Methods Appl. Mech. Engrg.*, 64:509–521, 1987.

[168] F. Kikuchi. Mixed formulations for finite element analysis of magnetostatic and electrostatic problems. *Japan J. Appl. Math.*, 6(2):209–221, 1989.

[169] F. Kikuchi. On a discrete compactness property for the Nédélec finite elements. *J. Fac. Sci. Univ. Tokyo, Sect. 1A Math.*, 36:479–490, 1989.

[170] A. Kirsch. The denseness of the far field patterns for the transmission problem. *IMA J. Appl. Math.*, 37:213–226, 1986.

[171] A. Kirsch. On the existence of transmission eigenvalues. *Inverse Probl. Imaging*, 3(2):155–172, 2009.

[172] A. Kirsch and A. Lechleiter. The inside-outside duality for scattering problems by inhomogeneous media. *Inverse Problems*, 29:No. 104011, 2013.

[173] A. Kleefeld. A numerical method to compute interior transmission eigenvalues. *Inverse Problems*, 29:104012, 2013.

[174] A. V. Knyazev and J. E. Osborn. New a priori FEM error estimates for eigenvalues. *SIAM J. Numer. Anal.*, 43:2647–2667, 2006.

[175] W.G. Kolata. Approximation in variationally posed eigenvalue problems. *Numer. Math.*, 29:159–171, 1978.

[176] A. Konrad. On the reduction of the number of spurious modes in the vectorial finite-element solution of the three dimensional cavities and waveguides. *IEEE Trans. Microwave Theory Techn.*, 34(2):224–227, 1986.

[177] M. Koshiba and M. Suzuki. Vectorial finite-element method without any spurious solutions for dielectric waveguiding problems using transverse magnetic-field component. *IEEE Trans. Microwave Theory Techn.*, 34(11):1120–1124, 1986.

[178] K. Kreyszig. *Introductory Functional Analysis with Applications*. John Wiley & Sons, 1989.

[179] M.G. Larson. A posteriori and a priori error analysis for finite element approximations of self-adjoint elliptic eigenvalue problems. *SIAM J. Numer. Anal.*, 38:608–625, 2000.

[180] P. Lax. On Cauchy's problem for hyperbolic equations and the differentiability of solutions of elliptic equations. *Comm. Pure Appl. Math.*, 8:615–633, 1955.

[181] A. Lechleiter and S. Peters. Determining transmission eigenvalues of anisotropic inhomogeneous media from far field data. *Commun. Math. Sci.*, 13(7):1803–1827, 2015.

[182] A. Lechleiter and S. Peters. The inside-outside duality for inverse scattering problems with near field data. *Inverse Problems*, 31(8):085004, 2015.

[183] A. Lechleiter and M. Rennoch. Inside-outside duality and the determination of electromagnetic interior transmission eigenvalues. *SIAM J. Math. Anal.*, 47(1):684–705, 2015.

[184] R.B. Lehoucq, D.C. Sorensen, and C. Yang. *ARPACK users' guide. Solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*. Software, Environments, and Tools, 6. SIAM, Philadelphia, 1998.

[185] Y.-J. Leung and D. Colton. Complex transmission eigenvalues for spherically stratified media. *Inverse Problems*, 28(7):075005, 2012.

[186] M. Levitin and E. Shargorodsky. Spectral pollution and second-order relative spectra for self-adjoint operators. *IMA J. Numer. Anal.*, 24(3):393–416, 2004.

[187] T. Li, W. Huang, W.W. Lin, and J. Liu. On spectral analysis and a novel algorithm for transmission eigenvalue problems. *J. Sci. Comput.*, 64(1):83–108, 2015.

[188] E.H. Lieb. Density functionals for Coulomb systems. *Int. J. Quantum Chem.*, 24:243–277, 1983.

[189] E.H. Lieb and M. Loss. *Analysis*. American Mathematical Society, 2nd edition, 2001.

[190] E.H. Lieb, R. Seiringer, and J. Yangvason. Bosons in a trap: a rigorous derivation of the Gross-Pitaevskii energy functional. *Phys. Rev. A*, 61(4):043602–043614, 1999.

[191] Q. Lin and G. Xie. Accelerating the finite element method in eigenvalue problems (in Chinese). *Kexue Tongbao*, 26:449–452, 1981.

[192] Q. Lin and H. Xie. A superconvergence result for mixed finite element approximations of the eigenvalue problem. *ESAIM Math. Model. Numer. Anal.*, 46(4):797–812, 2012.

[193] Q. Lin and H. Xie. A multi-level correction scheme for eigenvalue problems. *Math. Comp.*, 84(291):71–88, 2015.

[194] F. Liu, M. Stynes, and A. Zhou. Postprocessed two-scale finite element discretizations, Part I. *SIAM J. Numer. Anal.*, 49:1947–1971, 2011.

[195] H. Liu, W. Gong, S. Wang, and N. Yan. Superconvergence and a posteriori error estimates for the Stokes eigenvalue problems. *BIT*, 53(3):665–687, 2013.

[196] X. Liu and S. Oishi. Verified eigenvalue evaluation for the Laplacian over polygonal domains of arbitrary shape. *SIAM J. Numer. Anal.*, 51(3):1634–1654, 2013.

[197] C. Lovadina, M. Lyly, and R. Stenberg. A posteriori estimates for the Stokes eigenvalue problem. *Numer. Methods Partial Differential Equations*, 25:244–257, 2009.

[198] W. McLean. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, Cambridge, 2000.

[199] S. Meddahi, D. Mora, and R. Rodríguez. Finite element spectral analysis for the mixed formulation of the elasticity equations. *SIAM J. Numer. Anal.*, 51(2):1041–1063, 2013.

[200] B. Mercier, J. Osborn, J. Rappaz, and P.-A. Raviart. Eigenvalue approximation by mixed and hybrid methods. *Math. Comp.*, 36:427–453, 1981.

[201] P. Monk. A mixed finite element method for the biharmonic equation. *SIAM J. Numer. Anal.*, 24:737–749, 1987.

[202] P. Monk. *Finite Element Methods for Maxwell's Equations*. Clarendon Press, Oxford, 2003.

[203] P. Monk and J. Sun. Finite element methods of Maxwell transmission eigenvalues. *SIAM J. Sci. Comput.*, 34:B247–B264, 2012.

[204] L. Morley. The triangular equilibrium problem in the solution of plate bending problems. *Aero. Quart.*, 19:149–169, 1968.

[205] A. Naga, Z. Zhang, and A. Zhou. Enhancing eigenvalue approximation by gradient recovery. *SIAM J. Sci. Comput.*, 28(4):1289–1300, 2006.

[206] S.A. Nazarov and B.A. Plamenevsky. *Elliptic Problems in Domains with Piecewise Smooth Boundaries*. Walter de Gruyter, Berlin, 1994.

[207] S.A. Nazarov, K. Ruotsalainen, and P. Uusitalo. Bound states of waveguides with two right-angled bends. *J. Math. Phys.*, 56(2):021505, 2015.

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

[209] J. A. Nitsche. Ein Kriterium für die Quasioptimalität des Witzchen Verfahrens. *Numer. Math.*, 11:346–348, 1968.

[210] F. Olver, D. Lozier, R. Boisvert, and C. Clark, editors. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[211] J. Osborn. Spectral approximation for compact operators. *Math. Comp.*, 29:712–725, 1975.

[212] J. Osborn. Approximation of the eigenvalue of a nonself-adjoint operator arising in the study of the stability of stationary solutions of the Navier-Stokes equations. *SIAM J. Numer. Anal.*, 13:185–197, 1976.

[213] L. Päivärinta and J. Sylvester. Transmission eigenvalues. *SIAM J. Math. Anal.*, 40:738–753, 2008.

[214] E. Polizzi. Density-matrix-based algorithms for solving eigenvalue problems. *Phys. Rev. B*, 79:115112, 2009.

[215] R. Rannacher. Nonconforming finite element methods for eigenvalue problems in linear plate theory. *Numer. Math.*, 13:23–42, 1979.

[216] R. Rannacher. On nonconforming and mixed finite element method for plate bending problems. The linear case. *RAIRO Anal. Numér.*, 13:369–387, 1979.

[217] M. Reed and B. Simon. *Methods of Modern Mathematical Physics, II: Fourier Analysis, Self-adjointness*. Academic Press, San Diego, 1975.

[218] R. Rodríguez and P. Venegas. Numerical approximation of the spectrum of the curl operator. *Math. Comp.*, 83(286):553–577, 2014.

[219] B. Rynne and B. Sleeman. The interior transmission problem and inverse scattering from inhomogeneous media. *SIAM J. Math. Anal.*, 22:1755–1762, 1991.

[220] Y. Saad. Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices. *Linear Algebra Appl.*, 34:269–295, 1980.

[221] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. SIAM, Philadelphia, 2nd edition, 2011.

[222] T. Sakurai and H. Sugiura. A projection method for generalized eigenvalue problems using numerical integration. *J. Comput. Appl. Math.*, 159(1):119–128, 2003.

[223] S. Sauter and C. Schwab. *Boundary Element Methods*. Springer Series in Computational Mathematics. Springer, 2011.

[224] L. Shen and A. Zhou. A defect correction scheme for finite element eigenvalues with applications to quantum chemistry. *SIAM J. Sci. Comput.*, 28(1):321–338, 2006.

[225] Z.C. Shi. Error estimates of Morley element. *Chinese J. Numer. Math. & Appl.*, 12:9–15, 1990.

[226] R.P. Silvester and R.L. Ferrari. *Finite Element Methods for Electrical Engineers*. Cambridge University Press, Cambridge, 3rd edition, 1996.

[227] J. Sun. Estimation of transmission eigenvalues and the index of refraction from Cauchy data. *Inverse Problems*, 27:015009, 2011.

[228] J. Sun. Iterative methods for transmission eigenvalues. *SIAM J. Numer. Anal.*, 49(5):1860–1874, 2011.

[229] J. Sun. An eigenvalue method using multiple frequency data for inverse scattering problems. *Inverse Problems*, 28:025012, 2012.

[230] J. Sun. A new family of high regularity elements. *Numer. Methods Partial Differential Equations*, 28:1–16, 2012.

[231] J. Sun. A mixed FEM for the quad-curl eigenvalue problem. *Numer. Math.*, 132(1):185–200, 2016.

[232] J. Sun and L. Xu. Computation of the Maxwell's transmission eigenvalues and its application in inverse medium problems. *Inverse Problems*, 29:104013, 2013.

[233] P. Tang and E. Polizzi. FEAST as a subspace iteration eigensolver accelerated by approximate spectral projection. *SIAM J. Matrix Anal. Appl.*, 35(2):354–390, 2014.

[234] F. Tisseur and K. Meerbergen. The quadratic eigenvalue problem. *SIAM Rev.*, 43(2):235–286, 2001.

[235] L.N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, Philadelphia, 3rd edition, 1997.

[236] G.M. Vaĭnikko. On the rate of convergence of certain approximation methods of Galerkin type in eigenvalue problems. *Amer. Math. Soc. Transl.*, 36:249–259, 1970.

[237] A. Veeser. Convergent adaptive finite elements for the nonlinear Laplacian. *Numer. Math.*, 92(4):743–770, 2002.

[238] M. Wang and J. Xu. The Morley element for fourth order elliptic equations in any dimensions. *Numer. Math.*, 103(1):155–169, 2006.

[239] T. Warburton and M. Embree. The role of the penalty in the local discontinuous Galerkin method for Maxwell's eigenvalue problem. *Comput. Methods Appl. Mech. Engrg.*, 195:3205–3223, 2006.

[240] T. Warburton and J.S. Hesthaven. On the constants in hp-finite element trace inverse inequality. *Comput. Methods Appl. Mech. Engrg.*, 192(25):2765–2773, 2003.

[241] G. Wei and H.K. Xu. Inverse spectral analysis for the transmission eigenvalue problem. *Inverse Problems*, 29(11):115012, 2013.

[242] H.F. Weinberger. *Variational methods for eigenvalue approximation*. Number 15 in Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, 1974.

[243] G.N. Wells and N.T. Dung. A C^0 discontinuous Galerkin formulation for Kirchhoff plates. *Comput. Methods Appl. Mech. Engrg.*, 196:3370–3380, 2007.

[244] H. Whitney. *Geometric Integration Theory*. Princeton University Press, Princeton, 1957.

[245] C. Wieners. Bounds for the N lowest eigenvalues of fourth-order boundary value problems. *Computing*, 192:29–41, 1997.

[246] X. Wu and W. Chen. Error estimates of the finite element method for interior transmission problems. *J. Sci. Comput.*, 57(2):331–348, 2013.

[247] H. Xie. A type of multilevel method for the Steklov eigenvalue problem. *IMA J. Numer. Anal.*, 34(2):592–608, 2014.

[248] J. Xu and A. Zhou. Local and parallel finite element algorithms based on two-grid discretizations. *Math. Comp.*, 69:881–909, 2000.

[249] J. Xu and A. Zhou. A two-grid discretization scheme for eigenvalue problems. *Math. Comp.*, 70(233):17–25, 2001.

[250] J. Xu and A. Zhou. Local and parallel finite element algorithms for eigenvalue problems. *Acta Math. Appl. Sin. Engl. Ser.*, 18(2):185–200, 2002.

[251] K. Yosida. *Functional Analysis*. Springer-Verlag, Berlin, Heidelberg, New York, 1980.

[252] F. Zeng, J. Sun, and L. Xu. A probing method for transmission eigenvalue problem. submitted.

[253] F. Zeng, T. Turner, and J. Sun. Some results on electromagnetic transmission eigenvalues. *Math. Methods Appl. Sci.*, 38(1):155–163, 2015.

[254] B. Zheng, Q. Hu, and J. Xu. A nonconforming finite element method for fourth order curl equation in $\mathbb{R}^3$. *Math. Comp.*, 80(276):1871–1886, 2011.

[255] A. Zhou. An analysis of finite-dimensional approximations for the ground state solution of Bose-Einstein condensates. *Nonlinearity*, 17(2):541–550, 2004.

[256] A. Zhou. Finite dimensional approximations for the electronic ground state solution of a molecular system. *Math. Methods Appl. Sci.*, 30(4):429–447, 2007.

[257] J. Zhou, X. Hu, Y. Huang, and L. Chen. Two-grid methods for Maxwell eigenvalue problems. *SIAM J. Numer. Anal.*, 52(4):2027–2047, 2014.

Index