Finite Element Methods for Eigenvalue Problems

Book · July 2016

DOI: 10.1201/9781315372419

2 authors:

Jiguang. Sun
Michigan Technological University
112 PUBLICATIONS **2,335** CITATIONS

SEE PROFILE

Aihui Zhou
Chinese Academy of Sciences
184 PUBLICATIONS **4,794** CITATIONS

SEE PROFILE

Jiguang Sun and Aihui Zhou

Finite Element Methods for Eigenvalue Problems

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 [79, 217, 112, 13, 29, 43]. Due to the flexibility in treating complex structures and rigorous theoretical justification, finite element methods, including conforming finite elements, non-conforming 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 [238, 47, 123, 80, 243, 213, 138, 139, 140, 177, 72, 116, 117, 144, 201, 169, 22, 23, 81, 29, 181, 149, 46, 42, 176, 240, 128, 12, 58, 36, 76, 77, 162, 220, 75], in particular, the book chapter by Babuška and Osborn [24]. However, to the authors' opinion, so far, there does not exist a self-contained, systematic, and up-to-date treatment. This is the motivation of the book.

We start with functional analysis including the 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 detail account to 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, 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 text books. For example, we only choose the typical triangular mesh in two dimensions and tetrahedron mesh in three dimensions. There are many other meshes such as rectangular meshes or hexahedra 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 mentioned 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 [24]. 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 [37], discontinuous

Galerkin method [12]. 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 modal 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 [36]. 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 non-selfadjoint. 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 non-selfadjointness. 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, the transmission eigenvalue problem of electromagnetic is 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 then present and analyze a two-scale finite element discretization.

Adaptive finite element element methods have been an important topic, which is discussed in Chapter 8. The Laplace eigenvalue problem is used as a modal 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 Poisson equation and moves on to the Laplace eigenvalue problems 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 non-linear eigen-

value 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 is 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 prior 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. Laplace eigenvalue problem, 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 [48, 10, 15, 73], eigenvalue problem of elasticity [200], waveguide band structures [49, 129, 209], Stokes eigenvalue problems [214, 198, 93, 145, 196, 120, 155, 248], etc.

Of course, many interesting topics are not discussed or fully discussed. We list some of them here: discontinuous Galerkin methods [12, 59], the bounds on eigenvalues approximated by finite element methods [78, 152, 26, 155, 153], multi-level/multi-grid methods [250, 162, 164, 258, 248, 194], superconvergence [154, 193, 196], computation of a large number of eigenvalues or eigenvalue cluster [147, 41, 124], spectra pollution [38, 37, 111, 187], nonlinear eigenvalue problems [236, 239, 88, 71, 86], etc.

This book can be used as the graduate text book 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 last two years. We hope the presentation can draw some attention of researchers on these interesting research topics.

We would like to thank many people who helped us in the preparation of this book. The class of Topics of Computational Mathematics at Michigan Technological University (MTU) suggested many corrections and improvements. Graduate students at 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 supports. The Department of Mathematics at Michigan Technological University(MTU), MTU Research Excellence Fund, and The Institute of Computational Mathematics and Scientific/Engineering Computing at 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
$\mathcal{C}_0^k(\Omega)$	the set of k times continuously differentiable functions with compact support in Ω
$\mathcal{C}_0^k(\overline{\Omega})$	the set of k times continuously differentiable functions which have bounded and uniformly continuous derivatives up to order k with compact support in Ω
$\mathcal{C}_0^\infty(\Omega)$	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 imbedding
$x_n \rightharpoonup x$	$\{x_n\}$ 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 m-th 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 text books such as [180, 252, 81], 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 exists $a, b > 0$ 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 shall 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 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

$$Tx_n \to Tx \quad \text{in } Y \text{ as } n \to \infty.$$

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 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. Let $f \in X'$.

Definition 1.1.22. *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,

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

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 [180] 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 [199]).

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*

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

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*

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

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 [81]).

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, [81])

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

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 [81]) 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 [180].

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 $\{x_n\} \subset X$, $\{Tx_n\}$ 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 1.1.1. *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 1.1.8. *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 1.1.9. *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 1.1.10. *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 [17]).

Theorem 1.1.11. *(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 operate $z - T$ 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 clusions 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 [204]).

Theorem 1.1.12. *Let $T : X \to X$ be a compact, self-adjoint, linear operator on a Hilbert space X. Then there exists 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 sums has infinitely many terms,

(6) let $W =$ 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 [204]).

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 [199, 204] 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) $\mathcal{C}_0^k(\Omega)$: the set of k times continuously differentiable functions with compact support in Ω;

(3) $\mathcal{C}_0^k(\overline{\Omega})$: 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, with compact support in Ω;

(5) $L^p(\Omega), 1 \le p < \infty$: 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 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

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

Let $0 < \gamma < 1$. We call f a Lipschitz continuous function on Ω if

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

for some constant $C > 0$. A function f is said to be Hölder continuous with exponent γ if

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

for some constant $C > 0$. The γth Hölder simi-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

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

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

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

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

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

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

associated with the norm

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

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 imbedded in X. An embedding is compact if I is compact, written as $W^{s,p}(\Omega) \hookrightarrow X$.*

Compact embedding plays 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 [204]).

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 $m \geq 1$ and $j \geq 0$ and let $p \in \mathbb{R}$ with $1 \leq p < \infty$. Then the following embeddings are compact:*

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

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

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

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

(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 $s \ge 0$ and $1 \le p < \infty$. 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 $H^s(\Omega) = W^{s,2}(\Omega), s \ge 0$. Furthermore, the following embedding theorem holds.

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

1.2.2 Negative Norm

The negative Sobolev norm [182] is useful to study the regularity of the solutions of partial differential equations. We present some results following [252].

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

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

The negative norm of f is defined as

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

By Schwarz' inequality, we have

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

Thus we immediately have that

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

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 [252]) 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 [204] 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 [45]).

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*

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

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

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

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 [204]. 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

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

The following property of sesquilinear forms on $X \times X$ are 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*

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

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 [204]).

Theorem 1.3.1. *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

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

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*

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

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 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 [204]).

Theorem 1.3.2. *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*

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

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

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. 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 $dim Y = dim Z < \infty$, 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 [167])

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

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-Rappaz

We first present the abstract converge theory due to Descloux, Nassif, and Rappaz [116, 117]. 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 $u \in E(X)$
$$\lim_{h\to 0} d(u, E_h(X_h)) = 0, \tag{1.20}$$
in particular, if $\Gamma \cap \sigma(T) \neq \emptyset$,

(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 [116]:

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 1.4.1. *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 based on [116] since the techniques will be used later.

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

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

For h small enough, **P1** implies

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

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

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

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 [24]. It plays a key role in the convergence analysis for several finite element methods for the Laplace eigenvalue eigenvalue problem, the biharmonic eigenvalue, and the Maxwell's eigenvalue problem. The materials presented here are taken from Sections 7 and 8 of [24].

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 $u \in N((\lambda I - T)^j)$.

Remark 1.4.2. *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 1.1.12)).*

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 [24]. 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 [24] claims that $R(E)$ can be approximated by $R(E_h)$ correctly.

Theorem 1.4.2. *(Theorem 7.1 in [24]) 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 1.4.3. *(Theorem 7.2 in [24]) 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

$$\dim R(E) = \dim R(E_h) = m.$$

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

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

and

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

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

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

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

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

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 [24] (Theorem 7.3 therein).

Theorem 1.4.4. *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*

$$|\lambda - \lambda_{j,h}| \leq C \left\{ \sum_{j,k=1}^{m} |\langle (T - T_h)\phi_i, \phi_k \rangle| + \|(T - T_h)_{R(E)}\| \|(T' - T'_h)|_{R(E')} \right\}^{1/r}$$

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 $\|w'_h\| \leq \|E\|$. Noting

that $(T'-\lambda)^r w'_h = 0$, we obtain

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

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

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

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 1.4.5. *(Theorem 7.4 in [24]) 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

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

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 1.4.2 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 1.4.4 completes the proof. □

When X is a Hilbert space and T, T_h are selfadjoint, 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 materials in this section is based on Section 8 of [24]. 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 imbedded 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 $a(\cdot, \cdot)$ 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 $\lambda Tu = u$.

Remark 1.4.3. *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

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

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 1.4.6. *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 1.4.7. *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 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

$$0 < \lambda_1 \leq \lambda_2 \leq \dots$$

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

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

(2) Minimum-maximum principle:

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

(3) Maximum-minimum principle:

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

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

$$\lambda_k \leq \lambda_{k,h}, \quad k = 1, \ldots, N, N = \dim S_h$$

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$\{u_k, \ldots, u_{k+n-1}\}$ and $E_h = E_h(\lambda_k^{-1})$ be the orthogonal projection of H on to span$\{u_{k,h}, \ldots, u_{k+n-1,h}\}$. Then we have the following estimates:

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

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:

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

For more discussion of the results in this section, we refer the readers to [24] and references therein.

Chapter 2

Finite Elements

2.1 Introduction

In this chapter, we introduce fundamental concepts of finite elements. There are ample excellent literatures, for example, [90, 55, 45]. 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 [90].

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) $\mathcal{N} = \{N_1, \ldots, N_s\}$ 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 $\{\phi_1, \phi_2, \ldots, \phi_s\}$ 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 shall 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 [90, 55, 45, 204].

We start with some definitions of meshes following [45].

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 are constants $\tau > 0$ and $h_0 > 0$ independent of h such that

$$\tau h \leq h_f \quad \text{for all } f \in \mathcal{T}_h \text{ and } h_0 \geq h > 0.$$

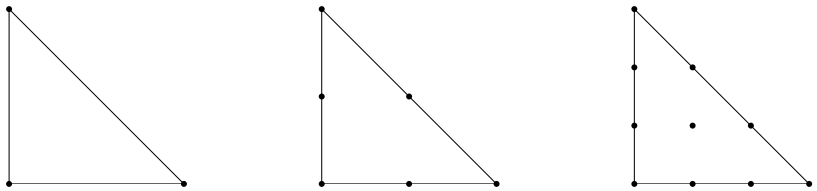

Figure 2.1: Left: Linear Lagrange element. Middle: Quadratic Lagrange element. Right: Cubic Lagrange element.

Let $\hat{K}$ be the reference element, i.e., a triangle whose vertices are $(0,0)$, $(1,0)$, and $(0,1)$ in $\mathbb{R}^2$, or a 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 $k > 2$,

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

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 function for which all $N_i \in \mathcal{N}$ are defined, the local interplant 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 [45]).

Theorem 2.1.1. *Let $t \geq 0$ 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 t 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 $t - 1$. Without loss of generality, we assume that one edge of K lies on the x-axis and it contains the points $z_1, \ldots z_{t+1}$. Then there is a univariate polynomial $p_0(x)$ with

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

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

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

The proof is complete. □

To define the conforming elements, we need the follow theorem from [45].

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:

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

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 $k \le 3$, 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 $1 \le i < j \le 3$. 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 [118], 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 [118].

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 [168]. 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 the 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., [90, 45, 55, 17, 204]. The presentation closely follows Section 2.3 of [204].

2.3.1 Cea's Lemma

Theorem 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	weight 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 ponts, $k = 3$: 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	weight 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 ponts, $k = 2$: 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 [204]) 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 1.3.1).

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. □

We have the well-posedness of both the continuous and discrete problems (Theorems 1.3.2 and 2.3.2). We shall move on to prove the error estimates.

Lemma 2.3.1. *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 1.3.1) since $b(\cdot,\cdot)$ satisfies the discrete Babuška-Brezzi condition (2.12) and for $v_h \in Z_h^\perp$ we have that

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

□

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

Theorem 2.3.3. *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 2.3.4. *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. 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 2.3.5. *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*

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

for some constant C *independent of* h*.*

Proof. For Theorems 2.3.3 and 2.3.4, we obtain

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

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 2.3.1, 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 $\text{diam}K$ 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 taken from [55].

Theorem 2.3.6. *Let $\rho h \leq \text{diam}K \leq h$, 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*

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

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

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

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

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

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). □

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 [241] without proofs.

Theorem 2.3.7. *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 2.3.8. *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 $u \in P_k(K)$ we have that*

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

Theorem 2.3.9. *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 [55]). 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 [45].

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 $\mathcal{T}_h = \{K_1, K_2, \ldots, K_M\}$ 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 interplant $I_h v$ 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 [45]).

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 $z_1, z_2, \ldots, z_s$ 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 $z_i, i = 1, \ldots, s$, 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

$$|v(z_i)| \leq C\|v\|_{H^t(\Omega)} \quad i = 1, 2, \ldots, s.$$

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

$$\|v_n\|_{L^2(\Omega)} = 1, \quad |||v_n||| \leq 1/n, \quad n = 1, 2, \ldots.$$

Due to the compact imbedding of $H^t(\Omega)$ into $H^{t-1}(\Omega)$ (see Theorem 1.2.2), there exists a subsequence of $\{v_n\}$, still denoted by $\{v_n\}$, 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 [45]).

Lemma 2.4.1. *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

$$V^k := V^k(\mathcal{T}) = \left\{ v \in L^2(\Omega) \,|\, v|_K \in \mathcal{P}_k \text{ for every } K \in \mathcal{T} \right\}.$$

By Theorem 2.1.2, there exists a unique interpolation operator

$$I_h : H^t(\Omega) \to V^k, \quad t \geq 2.$$

The following estimate holds for I_h (Theorem 6.4 of [45]).

Theorem 2.4.2. *Let $t \geq 2$, and suppose $\mathcal{T}_h$ is a shape-regular triangulation of Ω. Then there exists a constant C such that*

$$\|u - I_h u\|_{m,h} \leq C h^{t-m} |u|_{H^t(\Omega)} \quad \textit{for } u \in H^t(\Omega), \quad 0 \leq m \leq t,$$

where I_h is the interpolation by a piecewise polynomial of degree up to $t-1$.

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 $v \in H^m(\Omega)$, 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

$$\|B\| \leq r(K)/\rho(\hat{K}),$$

which implies

$$\|B\| \leq h/\rho(\hat{K}) \leq 4h.$$

Thus the following holds

$$|u - I_h u|_{H^l(K)} \leq Ch^{t-l}|u|_{H^t(K)}.$$

Summing over l from 0 to m, we obtain that

$$\|u - I_h u\|_{H^m(K)} \leq Ch^{t-m}|u|_{H^t(K)} \quad \text{for all } u \in H^t(K), K \in \mathcal{T}_h.$$

The theorem follows immediately. □

Finally we present a classical result of polynomial interpolation, which can be found in many classical finite element books, for example, [90, 55].

Theorem 2.4.3. *Let $v \in H^{k+1}(K)$. There exists a constant C such that*

$$\inf_{p \in \mathcal{P}_k} \|v + p\|_{H^{k+1}(K)} \leq C|v|_{H^{k+1}(K)}.$$

2.5 Appendix: Implementing Finite Element Methods

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 lights 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.

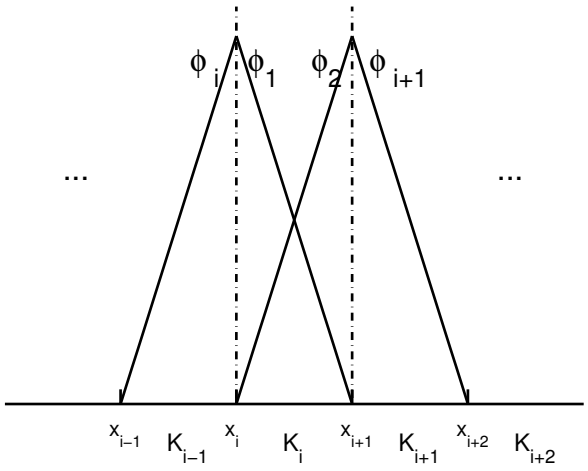

Figure 2.2: Linear Lagrange basis functions in one dimension.

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:

$$(\phi_j', \phi_j'), \quad (\phi_j', \phi_i'), \quad (\phi_i', \phi_j'), \quad (\phi_i', \phi_i'),$$

which contribute to S_{jj}, S_{ij}, S_{ji} and S_{ii}, 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.

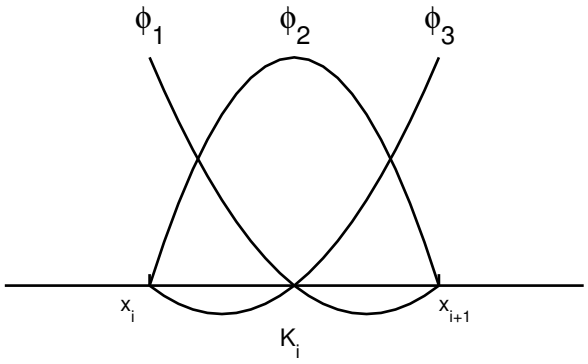

Figure 2.3: Quadratic Lagrange basis functions in one dimension.

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 has 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 the Matlab command '*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 complicate in higher dimensions.

3. The local matrices are computed exactly. However, for higher dimensional problems, techniques such as affine mapping are needed.

4. There is 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 tetrahedra 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. [137].

The problem has been studied by many researchers, see e.g., [104]. The theory and numerical methods are well-developed. Due to the simplicity of both theory and implementation, it serves well as the first modal problem to study finite element methods for eigenvalue problems. There are many finite element methods proposed for the Laplace eigenvalue problem in literature [37, 36, 12, 197]. In the 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 complicate 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., the Poisson's equation. Given a function f, find u such that

$$-\triangle u = f \quad \text{in } \Omega \tag{3.1}$$

with 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 [45]).

Theorem 3.1.1. *Suppose Ω is contained in an n-dimensional cube with side length s. Then*

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

Consequently, the coercivity of $a(\cdot,\cdot)$ holds:

$$a(u,u) = \|\nabla u\|^2 \geq \alpha \|u\|_{H^1(\Omega)}^2 \quad \text{for all } u \in H_0^1, \tag{3.4}$$

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 [69] (see also Chapter 8 of [141]).

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, $\alpha_0 = 1$ 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 $u \in H_0^1(\Omega)$ such that

$$-\triangle u = \lambda u \quad \text{in } \Omega. \tag{3.5}$$

The variational formulation is to find $\lambda \in \mathbb{R}$ and anon-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 known 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)

$$\inf_{v_h \in V_h} \|u - v_h\| \leq C h^{\min\{k+1, r\}} \|u\|_{H^r(\Omega)}, \tag{3.7}$$

$$\inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \leq C h^{\min\{k, r-1\}} \|u\|_{H^r(\Omega)}. \tag{3.8}$$

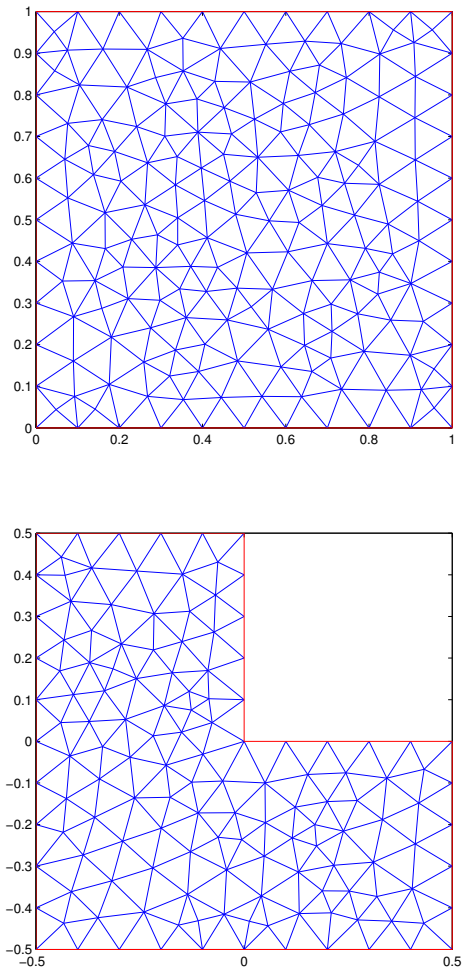

Figure 3.1: Two polygonal domains with triangular meshes. Top: unit square (convex). Bottom: L-shaped domain (non-convex).

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:

$$\|u - P_h u\|_{H^1(\Omega)} \leq Ch^{\alpha} \|u\|_{H^{1+\alpha}(\Omega)} \leq Ch^{\alpha} \|f\|_{H^{-1+\alpha}(\Omega)}. \tag{3.9}$$

The discrete problem for the 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 [45].

Theorem 3.2.2. *Suppose $\mathcal{T}_h$ is a family of shape-regular triangulations of Ω. Let u be the solution of the 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*

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch^\tau \|f\|. \tag{3.12}$$

Proof. From Céa's Lemma 2.3.1,

$$\|u - u_h\|_{H^1(\Omega)} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}.$$

Then (3.8) implies that

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} &\leq Ch^\tau \|u\|_{H^{\tau+1}(\Omega)} \\ &\leq Ch^\tau \|f\|_{H^{-1+\tau}(\Omega)}, \end{aligned}$$

where we have used (3.9). By the result on negative norm (Section 1.2.2), we have that

$$\|f\|_{H^{-1+\tau}(\Omega)} \leq \|f\|,$$

and thus

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch^\tau \|f\|.$$

□

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 [19, 211]. The following theorem is taken from [45] (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. Then the finite element solution in $V_h \subset V$ satisfies*

$$\|u - u_h\|_H \leq C\|u - u_h\|_V \sup_{g\in H, g\neq 0} \left\{ \frac{1}{|g|} \inf_{v\in V_h} \|\phi_g - v\|_V \right\},$$

where, for every $g \in H$, $\phi_g \in V$ denotes the corresponding unique solution of the equation

$$a(w, \phi_g) = (g, w) \quad \textit{for all } w \in V. \tag{3.14}$$

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

$$\begin{aligned} \|u - u_h\|_H &= \sup_{g\in H, g\neq 0} \frac{(g, u - u_h)}{\|g\|_H} \\ &\leq C\|u - u_h\|_V \sup_{g\in H, g\neq 0} \left\{ \inf_{v_h\in V_h} \frac{\|\phi_g - v_h\|_V}{\|g\|_H} \right\}. \end{aligned}$$

□

Application of the above theorem to the 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 $u \in H^s(\Omega), 1 \leq s \leq 2$ and $\tau = \min\{k, s-1\}$. Then*

$$\|u - u_h\| \leq Ch^\tau \|u - u_h\|_{H^1(\Omega)}.$$

Furthermore, if $f \in H^{-1+\tau}(\Omega)$ so that $u \in H^{1+\tau}(\Omega)$,

$$\|u - u_h\| \leq Ch^{2\tau} \|f\|_{H^{-1+\tau}(\Omega)} \leq Ch^{2\tau}\|f\|.$$

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

$$\sup_{g \in H, g \neq 0} \left\{ \inf_{v_h \in V_h} \frac{\|\phi_g - v_h\|_{H^1(\Omega)}}{\|g\|} \right\} \leq Ch^\tau.$$

Applying the Aubin-Nitsche Lemma (Theorem 3.2.4) and (3.7), we obtain that

$$\|u - u_h\| \leq Ch^\tau \|u - u_h\|_{H^1(\Omega)}.$$

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,

$$\|Tf - T_h f\| \leq Ch^{2\tau}\|f\|,$$

which implies

$$\|T - T_h\| \leq Ch^{2\tau}.$$

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 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 $\lambda_h^1, \ldots, \lambda_h^m$ be 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 1.4.3, 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 [36]). 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

$$a(Tf, v) = (f, v) \quad \text{for all } v \in H_0^1(\Omega).$$

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 $T_h : H_0^1(\Omega) \to H_0^1(\Omega)$ as

$$a(T_h f, v) = (f, v) \quad \text{for all } v_h \in V_h \subset H_0^1(\Omega).$$

The selfadjointness of T and T_h can be derived in the same way as above. From Corollary 3.2.3, we see that T_h converges to T uniformly. In addition, one has that

$$\|T - T_h\| \le Ch^\tau.$$

The following argument is an alternative proof for Theorem 3.3.2. Since T and T_h are self-adjoint, again from Theorem 1.4.3, we only need to approximate

$$\sum_{j,k=1}^{m} \left| ((T - T_h)\phi_j, \phi_k)_{H_0^1(\Omega)} \right|,$$

where $\{\phi_1, \ldots, \phi_m\}$ is a basis for the eigenspace $R(E)$. Let $u, v \in R(E)$ corresponding to the eigenvalue λ. Since $v = \lambda Tv$, one has that

$$\|v\|_{H^{1+\tau}(\Omega)} \le C\|v\|_{H^1(\Omega)}.$$

Thus we have that

$$\begin{aligned}
\left| ((T - T_h)u, v))_{H_0^1(\Omega)} \right| &= C\,|a((T - T_h)u, v)| \\
&= C \inf_{v_h \in V_h} |a((T - T_h)u, v - v_h)| \\
&\le C\|(T - T_h)u\|_{H^1(\Omega)} \inf_{v_h \in V_h} \|v - v_h\|_{H^1(\Omega)} \\
&\le Ch^\tau \|u\|_{H^1(\Omega)} h^\tau \|v\|_{H^{1+\tau}(\Omega)} \\
&\le Ch^{2\tau} \|u\|_{H^1(\Omega)} \|v\|_{H^{1+\tau}(\Omega)}.
\end{aligned}$$

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 validate 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.

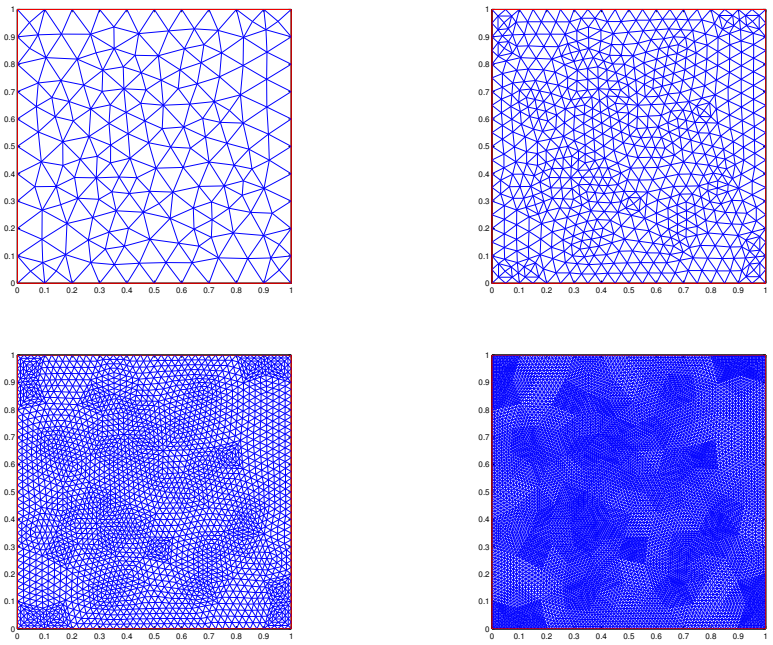

Figure 3.2: Sample uniformly refined unstructured meshes for the unit square.

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 element, we have $\tau = 1$ (see Corollary 3.2.5) and the second order convergence is observed (see Theorem 3.3.2).

h	λ_h	$\lvert\lambda_h - \lambda\rvert$	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 Dirichlet 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 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).

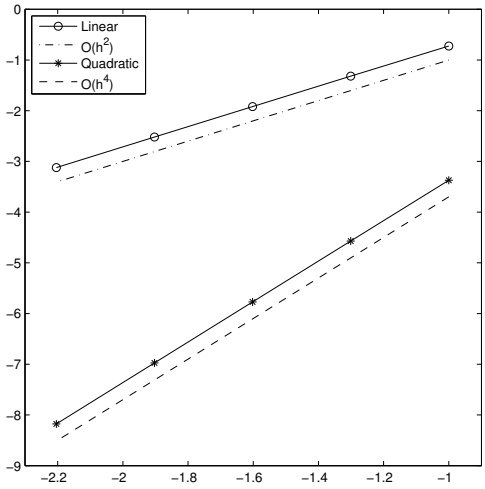

Figure 3.3: The log-log plot of the error of linear and quadratic Lagrange elements for the first eigenvalue of the unit square.

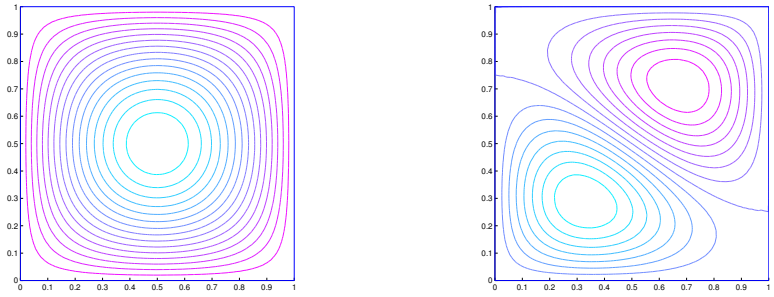

Figure 3.4: Eigenfunctions of the unit square. Left: the first eigenfunction. Right: the second eigenfunction.

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 validate 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).

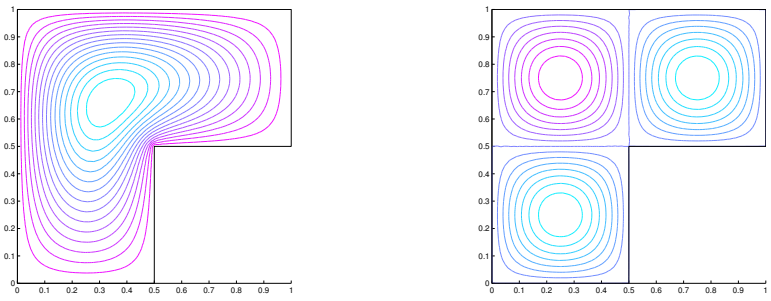

Figure 3.5: Dirichlet eigenfunctions of the L-shaped domain. Left: The first eigenfunction. Right: The third eigenfunction.

3.5.1 Generating 2D 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.

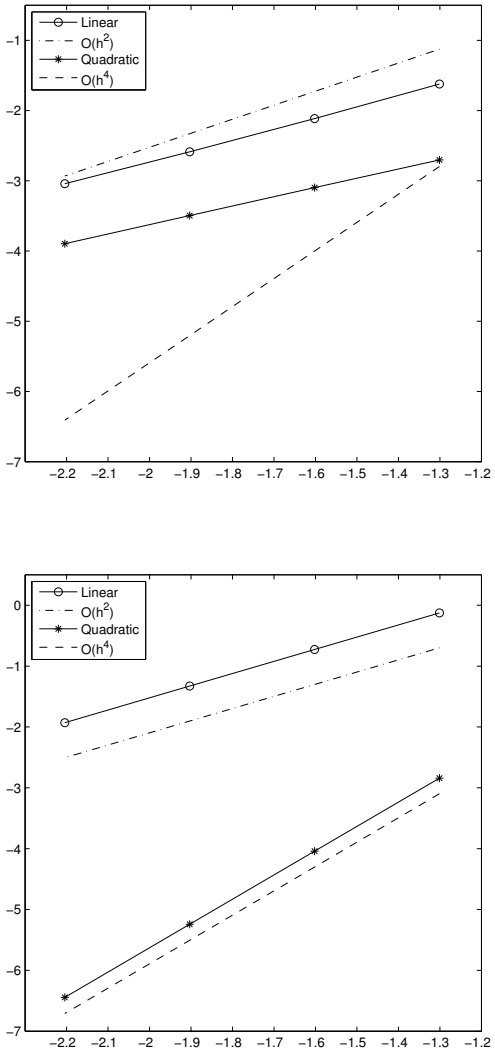

Figure 3.6: The log-log plot for the error for the L-shaped domain. Top: the first eigenvalue. Bottom: the third eigenvalue.

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 including

```
pdecirc(XC,YC,RADIUS,LABEL)
```

The command draws a circle with center in (XC,YC), RADIUS radius, and label 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 in (XC,YC), x- and y-axis radius (RADIUSX,RADIUSY), rotated counter-clockwise 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 (name) 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 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 counter clockwise order. The fourth row contains the subdomain number.

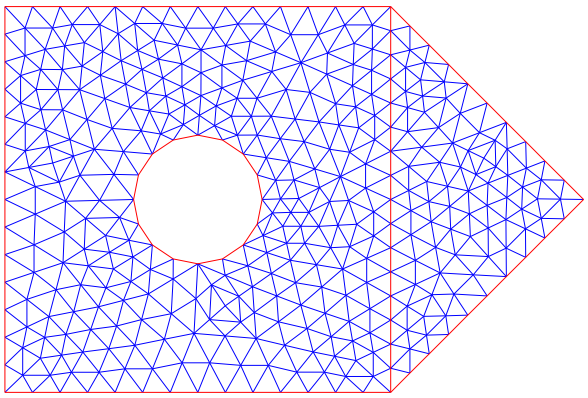

Figure 3.7: A domain and its triangular mesh obtained by the combination of simple geometries using Matlab PDEtool.

(3) the edge matrix 'e' is $7 \times p$ 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.

Note that 'e' only contains edges which coincide 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 generation more complicate 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.

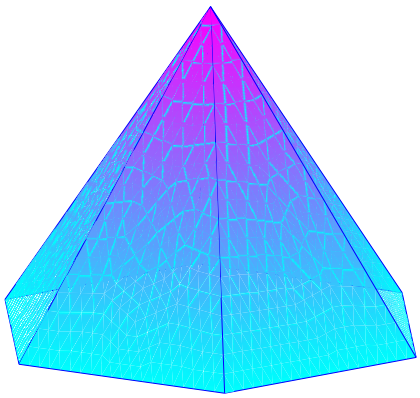

Figure 3.8: Linear Lagrange basis function.

3.5.2 Matrices Assembly

We consider the implementation of (3.16) using 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 degree 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.

Let $\{\phi_1, \phi_2, \ldots, \phi_N\}$ 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

$$a\left(\sum_{i=1}^{N} u_i \phi_i, \phi_j\right) = \lambda_h \left(\sum_{i=1}^{N} u_i \phi_i, \phi_j\right), \quad j = 1, \ldots, N.$$

Using the definition of $a(\cdot,\cdot)$, we have that

$$\sum_{i=1}^{N} (\nabla \phi_i, \nabla \phi_j) u_i = \lambda_h \sum_{i=1}^{N} (\phi_i, \phi_j) u_i, \quad j = 1, \ldots, N.$$

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 we call them the global index. In other words, the global basis functions ϕ_I, ϕ_J, ϕ_L has 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 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 basis function on K. For linear Lagrange element, it is enough to use 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 correspond 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 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 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 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 is 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 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 set 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 the 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'.

f. Line 22 computes the affine transformation.

g. Line 23 computes the area of triangle 'it'.

h. Lines 24-26 compute the values of gradients of the basis functions of triangle 'it'.

i. Lines 27-31 compute the local stiffness matrix.

j. Line 32 distributes the local stiffness matrix to the global stiffness matrix.

k. Lines 33-37 compute the local mass matrix.

l. 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 on \hat{T} (2nd order exactly)
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];
```