

VARIATIONAL FORMULATION OF ELLIPTIC PARTIAL DIFFERENTIAL
EQUATIONS AND BASICS IN FINITE ELEMENT METHOD



Addis Ababa University

College of Natural and Computational Sciences
Department of Mathematics

A project submitted to the Department of Mathematics in partial
fulfillment of the requirement for the degree of Master of Science in
Mathematics

By : Erimyas Kefyalew
Advisor : Tsegaye Gedif(Ph.D)

July 12, 2016
Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY
DEPARTMENT OF MATHEMATICS

The undersigned hereby certify that they have read and recommend to the department of mathematics for acceptance of this project entitled “**Variational Formulation of Elliptic Partial Differential Equations and Basics in Finite Element Method**” by **Erimyas Kefyalew** in partial fulfillment of the requirements for the degree of Master of Science in Mathematics.

Dated: July, 2016

Advisor:

Dr. Tsegaye Gedif

Examining committee:

Dr. Tesfa Biset

Dr. Mengistu Goa

July, 2016

Table of contents

Acknowledgements	ii
Abstract	iii
Notations	iv
Introduction	v
1 PRELIMINARIES	1
1.1 Definitions	1
1.2 Function Spaces	2
1.3 Linear Operators on Normed Spaces	5
1.3.1 Operators	5
1.3.2 Continuous linear operators	6
1.3.3 Linear functionals	6
1.4 Sobolev Spaces	9
1.4.1 Weak Derivatives	9
1.4.2 Sobolev Spaces	10
2 Variational Formulations of Elliptic Boundary Value Problems	14
2.1 Review on Partial differential equation	14
2.2 A model boundary value problem	15
2.3 Some general results on existence and uniqueness	17
2.4 The Lax-Milgram Lemma	20
2.5 Weak formulations of linear elliptic boundary value problems	22
2.5.1 Problems with homogeneous Dirichlet boundary conditions	22
2.5.2 Problems with non-homogeneous Dirichlet boundary conditions	23
2.5.3 Problems with Neumann boundary conditions	24
3 Basics in Finite Element Method	26
3.1 One-dimensional examples	26
3.1.1 Linear elements for a second-order problem	26
3.1.2 High order elements and the condensation technique	29
3.1.3 Reference element technique	30
3.2 Basics in finite element method	32
3.2.1 Continuous linear elements	33

3.2.2	Affine-equivalent finite elements	38
3.2.3	Finite element spaces	40
3.3	Error estimates of finite element interpolations	41
3.3.1	Local interpolations	41
3.3.2	Local interpolation error estimates	41
3.3.3	Global interpolation error estimates	43
3.4	Convergence and error estimates	44
3.5	Application of Finite Element Method for two dimensional problem . .	46
4	Conculsion	54
	Appendix-A	55
	References	58

Acknowledgement

First and foremost I would like to give my thanks to God who gave me the chance to live and have all such experiences in learning up to this level of which I didn't think of.

Secondly, my heart felt gratitude goes to my advisor Dr. Tsegaye Gedife for his constructive comments and friendly approaches all through the work.

Finally, I would also like to express my deepest thanks to all my instructors for the long lasting support they gave me during my course work.

Abstract

Elliptic partial differential equations appear frequently in various fields of science and engineering. These involve equilibrium problems and steady state phenomena. The most common example of such equation is poisson's equation. Most of these physical problems are very hard to solve analytically, instead, they can be solved numerically using computational methods. The finite element method is the most popular numerical method for solving elliptic boundary value problems. In this project, we introduce the concept of weak formulation, the finite element method, the finite element interpolation theory and its application in error estimates of finite element solutions of linear elliptic boundary value problems. This project also include the numerical solution of a two dimensional poisson equation with dirichlet boundary conditions by finite element method.

Keywords: Weak formulation, Finite Element Method, Poisson Equation

Notations

FEM	Finite Element Method.
PDE	Partial Differential Equation.
$\mathbb{R}^n$	The Euclidean space of n -dimensional for $n > 1$.
$\mathcal{S}(\mathbb{R}^n)$	the Schwartz space of all rapidly decreasing infinitely differentiable functions in $\mathbb{R}^n$.
x^T	transpose of the vector $x \in \mathbb{R}^n, d \in \mathbb{N}$.
α	multi-index (a d -tuple): $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d)$
$ \alpha $	$:= \alpha _1$ order (or length) of the multi-index $\alpha \in \mathbb{N}^d, d \in \mathbb{N}$.
Ω	open set in $\mathbb{R}^n$.
$\partial\Omega = \Gamma$	the boundary of Ω .
$\bar{\Omega}$	closure of Ω .
Γ_D	part of the boundary on which Dirichlet conditions are prescribed.
Γ_N	part of the boundary on which Neumann conditions are prescribed.
$\frac{\partial}{\partial \nu}$	normal derivative.
D^α	partial derivative notation.
D_w^α	weak derivative notation.
$diam(K)$	diameter of K .
V_h	finite element spaces.
K	element (triangle, rectangle, etc.).
h_K	diameter of K (element).
$\hat{K}$	reference element.
K_h	triangulation (partition).

Introduction

Elliptic partial differential equations (Elliptic PDEs) are one of the most important tools of mathematical modeling available today. They are used to describe such diverse physical phenomena like heat transfer, diffusion, mechanics of elastic and plastic materials, fluid mechanics, electrostatics and -dynamics, and many more. However, only few partial differential equations in certain simple settings permit an analytic solution. It is therefore no surprise that the numerical solution, or to be more precise, the approximation of solutions to partial differential equations by numerical methods, has become one of the main areas of research in computational mathematics ever since the development of computers has made such computations feasible for problems on large scales. The finite element method, frequently abbreviated by FEM, was developed in the fifties in the aircraft industry, after the concept had been independently outlined by mathematicians at an earlier time. The finite element method was first introduced by Courant in 1943(*Courant*, 1943). From the 1950s to the 1970s, it was developed by engineers and mathematicians into a general method for the numerical solution of partial differential equations.

The finite element method is today one of the most established approaches for the numerical solution of PDEs and has both a broad foundation in mathematical theory, where it is still an active area of research, as well as an excellent track record in practical applications in science and industry. Instead of treating the partial differential equation in its classical formulation, the core idea behind the FEM in its modern formulation is to pass to a so-called weak formulation, which takes the form of a variational equation in Hilbert or Banach spaces. The functional spaces are then discretized by replacing them with finite-dimensional subspaces of functions which are usually piecewise polynomials, or maps of piecewise polynomials, with respect to the “elements” of a predetermined mesh which typically consists of simplices, quadrilaterals, or hexahedra. In this way, a discretized formulation of the partial differential equation is obtained. It can easily be viewed as a system of equations which, for linear partial differential equations, is again linear. Methods from numerical linear algebra are then used to solve these linear and sparse systems. For many classes of problems, rigorous error estimates for the resulting approximation to the exact solution are known. The techniques used to derive these error estimates stem from functional analysis.

The advantages of the finite element method are that general boundary conditions, complex geometry, and variable material properties can be relatively easily handled. Also, the clear structure and versatility of the finite element method makes it possible to develop general purpose software for applications. Furthermore, it has a solid theoretical foundation that gives added reliability, and in many situations it is possible to obtain concrete error estimates in finite element solutions.

Chapter 1

PRELIMINARIES

1.1 Definitions

Definition 1.1.1. A norm on a vector spaces V is a mapping $\|\cdot\| : V \rightarrow \mathbb{R}$ that satisfies

1. $\|u\| \geq 0$, with equality if and only if $u = 0$.
2. $\|\lambda u\| = |\lambda| \|u\|$
3. $\|u + v\| \leq \|u\| + \|v\|$, for all $u, v \in V$ and $\lambda \in \mathbb{R}$

The space V equipped with the norm $\|\cdot\|$, $(V, \|\cdot\|)$ is called a normed linear space or a normed space.

Definition 1.1.2. Given a linear space V , a semi-norm $|\cdot|$ is a function from V to $\mathbb{R}$ with the properties of a norm except that $|v| = 0$ does not necessarily imply $v = 0$.

Definition 1.1.3. Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on a normed space V are said to be equivalent if there are positive constants c_1 and c_2 such that

$$c_1 \|v\|_1 \leq \|v\|_2 \leq c_2 \|v\|_1 \quad \forall v \in V$$

Definition 1.1.4. Let V be a normed space. A sequence $\{v_n\} \subset V$ is called a Cauchy sequence if for any $\epsilon > 0$ there exists a number $N(\epsilon)$ such that

$$\|v_m - v_n\| < \epsilon \quad \forall m, n > N(\epsilon)$$

Definition 1.1.5. A normed space is said to be complete if every Cauchy sequence from the space converges to an element in the space. A complete normed space is called a Banach space.

Let A be a subset of a normed space V . Then,

- (i) The set A is said to be closed in V if and only if $v_n \in A$ and $v_n \rightarrow v$ imply that $v \in A$.
- (ii) The closure $\bar{A}$ of A is the smallest closed set in V containing A .
- (iii) The set A is dense in V if for every $v \in V$ there exists a sequence $\{v_n\}$ in A such that $v_n \rightarrow v$.

(iv) A is said to be bounded if for some constant M , $\|v\| \leq M$ for every $v \in A$.

Definition 1.1.6. Let V be a vector space. A mapping $(\cdot, \cdot) : V \times V \rightarrow \mathbb{K}$ is said to be an inner product on V if and only if for all vectors u, v and w and scalars α :

1. $(u + v, w) = (u, w) + (v, w)$
2. $(\alpha u, v) = \alpha(u, v)$
3. $(u, v) = \overline{(v, u)}$ (the bar denote complex conjugate)
4. $(u, u) \geq 0$ and $(u, u) = 0 \Leftrightarrow u = 0$

The space V together with the inner product $(\cdot, \cdot)$ is called an inner product space. We simply say V is an inner product space. When $\mathbb{K} = \mathbb{R}$, V is called a real inner product space, whereas if $\mathbb{K} = \mathbb{C}$, V is a complex inner product space. An inner product $(\cdot, \cdot)$ induces a norm through the formula

$$\|v\| = \sqrt{(v, v)} \quad v \in V$$

Definition 1.1.7. A complete inner product space is called a Hilbert space.

Theorem 1.1.1. (Cauchy-Schwarz Inequality)

Let V be an inner product space. Then

$$|(u, v)| \leq \|u\| \|v\| \quad \forall u, v \in V$$

and the equality holds if and only if u and v are linearly dependent.

Proof. See [2, page 137] □

1.2 Function Spaces

For $d \in \mathbb{N}$ we call a vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d)$, $\alpha_i \in \mathbb{N}$, multi index with the absolute value $|\alpha| = \alpha_1 + \dots + \alpha_d$ and with the factorial $\alpha! = \alpha_1! \dots \alpha_d!$. For $x \in \mathbb{R}^d$ we can therefore write

$$x^\alpha = x_1^{\alpha_1} \dots x_d^{\alpha_d}$$

If u is a sufficient smooth real valued function, then we can write partial derivatives as

$$D^\alpha u(x) := \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \dots \left(\frac{\partial}{\partial x_d}\right)^{\alpha_d} u(x_1, \dots, x_d)$$

We denote by $C(\Omega)$ the space of all real-valued functions that are continuous on Ω . Since Ω is open, a function from the space $C(\Omega)$ is not necessarily bounded; consider, for example, the continuous function $v(x) = \ln x$ on $(0, 1)$. We denote further by $C(\overline{\Omega})$ the space of functions that are bounded and uniformly continuous on Ω . The notation $C(\overline{\Omega})$ is consistent with the fact that a bounded and uniformly continuous function on Ω has a unique continuous extension to $\overline{\Omega}$. The space $C(\overline{\Omega})$ is a Banach space with the norm

$$\|v\|_{C(\overline{\Omega})} = \sup\{|v(x)| : x \in \Omega\} \equiv \max\{|v(x)| : x \in \overline{\Omega}\}$$

Let $\Omega \subseteq \mathbb{R}^d$ be some open subset and assume $k \in \mathbb{N}_0$. $C^k(\Omega)$ is the space of functions which are bounded and k times continuously differentiable in Ω . In particular, for $u \in C^k(\Omega)$ the norm

$$\|u\|_{C^k(\Omega)} := \sum_{|\alpha| \leq k} \sup_{x \in \Omega} |D^\alpha u(x)|$$

is finite. Correspondingly, $C^\infty(\Omega)$ is the space of functions which are bounded and infinitely often continuously differentiable.

$$C^\infty(\Omega) = \bigcap_{m=0}^{\infty} C^m(\Omega) \equiv \{v \in C(\Omega) \mid v \in C^k(\Omega) \quad \forall k \in \mathbb{Z}_+\}$$

$$C^\infty(\overline{\Omega}) = \bigcap_{m=0}^{\infty} C^m(\overline{\Omega}) \equiv \{v \in C(\overline{\Omega}) \mid v \in C^k(\overline{\Omega}) \quad \forall k \in \mathbb{Z}_+\}$$

For a function $u(x)$ defined for $x \in \Omega$ we denote

$$\text{supp } u := \overline{\{x \in \Omega : u(x) \neq 0\}}$$

to be the support of the function u . Then,

$$C_0^\infty(\Omega) := \{u \in C^\infty(\Omega) : \text{supp } u \subset \Omega\}$$

is the space of $C^\infty(\Omega)$ functions with compact support.

A function v defined on Ω is said to be Lipschitz continuous if for some constant c , there holds the inequality

$$|v(x) - v(y)| \leq c\|x - y\| \quad \forall x, y \in \Omega$$

In this formula, $\|x - y\|$ denotes the standard Euclidean distance between x and y . The smallest possible constant in the above inequality is called the Lipschitz constant of v , and is denoted by $\text{Lip}(v)$. The Lipschitz constant is characterized by the relation

$$\text{Lip}(v) = \sup\left\{\frac{|v(x) - v(y)|}{\|x - y\|} \mid x, y \in \Omega, x \neq y\right\}$$

More generally, a function v is said to be Hölder continuous with exponent $\beta \in (0, 1]$ if for some constant c ,

$$|v(x) - v(y)| \leq c\|x - y\|^\beta \quad \text{for } x, y \in \Omega$$

The Hölder space $C^{0,\beta}(\overline{\Omega})$ is defined to be the subspace of $C(\overline{\Omega})$ functions that are Hölder continuous with the exponent β . With the norm

$$\|v\|_{C^{0,\beta}(\overline{\Omega})} = \|v\|_{C(\overline{\Omega})} + \sup\left\{\frac{|v(x) - v(y)|}{\|x - y\|^\beta} \mid x, y \in \Omega, x \neq y\right\}$$

the space $C^{0,\beta}(\overline{\Omega})$ becomes a Banach space. When $\beta = 1$, the Hölder space $C^{0,1}(\Omega)$ consists of all the Lipschitz continuous functions.

For $k \in \mathbb{Z}_+$ and $\beta \in (0, 1]$, we similarly define the Hölder space

$$C^{k,\beta}(\overline{\Omega}) = \{v \in C^k(\overline{\Omega}) \mid D^\alpha v \in C^{0,\beta}(\overline{\Omega}), \text{ for all } \alpha \text{ with } |\alpha| = k\};$$

this is a Banach space with the norm

$$\|v\|_{C^{k,\beta}(\bar{\Omega})} = \|v\|_{C^k(\bar{\Omega})} + \sum_{|\alpha|=k} \sup\left\{\frac{|D^\alpha v(x) - D^\alpha v(y)|}{\|x - y\|^\beta} \mid x, y \in \Omega, x \neq y\right\}.$$

The boundary of an open set $\Omega \subset \mathbb{R}^d$ is defined as

$$\Gamma := \partial\Omega = \bar{\Omega} \cap (\mathbb{R}^d \setminus \Omega)$$

We require that for $d \geq 2$ the boundary $\Gamma = \partial\Omega$ can be represented locally as the graph of a Lipschitz function using different systems of Cartesian coordinates for different parts of Γ , as necessary. The simplest case occurs when there is a function $\gamma : \mathbb{R}^{d-1} \rightarrow \mathbb{R}$ such that

$$\Omega := \{x \in \mathbb{R}^d : x_d < \gamma < \tilde{x} \text{ for all } \tilde{x}(x_1, \dots, x_{d-1}) \in \mathbb{R}^{d-1}\}$$

If $\gamma(\cdot)$ is Lipschitz,

$$|\gamma(\tilde{x}) - \gamma(\tilde{y})| \leq L|\tilde{x} - \tilde{y}| \text{ for all } \tilde{x}, \tilde{y} \in \mathbb{R}^{d-1}$$

then Ω is said to be a Lipschitz hypograph with boundary

$$\Omega := \{x \in \mathbb{R}^d : x_d < \gamma(\tilde{x}) \text{ for all } \tilde{x}(x_1, \dots, x_{d-1}) \in \mathbb{R}^{d-1}\}$$

Definition 1.2.1. An open set $\Omega \subset \mathbb{R}^d$, $d \geq 2$, is called a *Lipschitz domain* if its boundary $\Gamma = \partial\Omega$ is compact and if there exist finite families $\{W_i\}$ and $\{\Omega_j\}$ having the following properties:

- The family W_j is a finite open cover of Γ , i.e. $W_j \subset \mathbb{R}^d$ is an open subset and $\Gamma \subseteq \bigcup_j W_j$.
- Each Ω_j can be transformed to a Lipschitz hypograph by a rigid motion, i.e. by rotations and translations.
- For all j the equality $W_j \cap \Omega = W_j \cap \Omega_j$ is satisfied.

The local representation of a Lipschitz boundary $\Gamma = \partial\Omega$, i.e. the choice of families W_j and Ω_j , is in general not unique. If the parametrizations satisfy $\gamma \in C^k(\mathbb{R}^{d-1})$ or $\gamma \in C^{k,\kappa}(\mathbb{R}^{d-1})$ we call the boundary k times differentiable or Hölder continuous, respectively. If this holds only locally, we call the boundary piecewise smooth.

By $L_p(\Omega)$ we denote the space of all equivalence classes of measurable functions on whose powers of order p are integrable. The associated norm

$$\|u\|_{L_p(\Omega)} := \left\{ \int_{\Omega} |u(x)|^p dx \right\}^{\frac{1}{p}} \quad 1 \leq p < \infty$$

Two elements $u, v \in L_p(\Omega)$ are identified with each other if they are different only on a set K of zero measure $\mu(K) = 0$. In what follows we always consider one represent $u \in L_p(\Omega)$. In addition, $L_\infty(\Omega)$ is the space of functions u which are measurable and bounded almost everywhere with the norm

$$\|u\|_{L_\infty(\Omega)} := \text{ess sup}\{|u(x)|\} := \inf_{k \subset \Omega, \mu(k)=0} \sup_{x \in \Omega \setminus k} |u(x)|$$

The spaces $L_p(\Omega)$ are Banach spaces with respect to the norm $\|\cdot\|_{L_p(\Omega)}$. There holds the Minkowski inequality

$$\|u + v\|_{L_p(\Omega)} \leq \|u\|_{L_p(\Omega)} + \|v\|_{L_p(\Omega)} \quad \text{for all } u, v \in L_p(\Omega)$$

For $u \in L_p(\Omega)$ and $v \in L_q(\Omega)$ with adjoint parameters p and q , i.e.

$$\frac{1}{p} + \frac{1}{q} = 1$$

we further have Hölders inequality

$$\int_{\Omega} |u(x)v(x)| dx \leq \|u\|_{L_p(\Omega)} \|v\|_{L_q(\Omega)} \quad (1.1)$$

Definition 1.2.2. Let S be a subset of a normed space V . We say S has an open covering by a collection of open sets $\{U_{\alpha} | \alpha \in \Lambda\}$, Λ an index set, if

$$S \subset \bigcup_{\alpha \in \Lambda} U_{\alpha}$$

We say S is compact if for every open covering $\{U_{\alpha}\}$ of S , there is a finite subcover $\{U_{\alpha_j} | j = 1, \dots, m\} \subset \{U_{\alpha} | \alpha \in \Lambda\}$ which also covers S .

1.3 Linear Operators on Normed Spaces

1.3.1 Operators

Given two sets V and W , an operator T from V to W is a rule which assigns to each element in a subset of V a unique element in W . The domain $\mathcal{D}(T)$ of T is the subset of V where T is defined,

$$\mathcal{D}(T) = \{v \in V \mid T(v) \text{ is defined}\},$$

and the range $\mathcal{R}(T)$ of T is the set of the elements in W generated by T ,

$$\mathcal{R}(T) = \{w \in W \mid w = T(v) \text{ for some } v \in \mathcal{D}(T)\}.$$

It is also useful to define the null set, the set of the zeros of the operator,

$$\mathcal{N}(T) = \{v \in V \mid T(v) = 0\}.$$

Definition 1.3.1. Let V and W be two normed spaces. An operator $T : V \rightarrow W$ is continuous at $v \in \mathcal{D}(T)$ if

$$\{v_n\} \subset \mathcal{D}(T) \text{ and } v_n \rightarrow v \text{ in } V \Rightarrow T(v_n) \rightarrow T(v) \text{ in } W.$$

T is said to be continuous if it is continuous over its domain $\mathcal{D}(T)$.

1.3.2 Continuous linear operators

Definition 1.3.2. Let V and W be two linear spaces. An operator $L : V \rightarrow W$ is said to be linear if

$$L(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 L(v_1) + \alpha_2 L(v_2) \quad \forall v_1, v_2 \in V, \forall \alpha_1, \alpha_2 \in \mathbb{K}.$$

For a linear operator L , we usually write $L(v)$ as Lv .

Proposition 1.3.1. Let V and W be normed spaces, $L : V \rightarrow W$ a linear operator. Then L is bounded if and only if there exists a constant $\gamma \geq 0$ such that

$$\|Lv\|_W \leq \gamma \|v\|_V \quad \forall v \in V$$

Proof. see [1, page 56] □

Theorem 1.3.1. Let V and W be normed spaces, $L : V \rightarrow W$ a linear operator. Then L is continuous on V if and only if it is bounded on V .

Proof. see [1, page 56] □

We use the notation $\mathcal{L}(V, W)$ for the set of all the continuous linear operators from a normed space V to another normed space W . In the special case $W = V$, we use $\mathcal{L}(V)$ to replace $\mathcal{L}(V, V)$. From the above theorem we see that for a linear operator, boundedness is equivalent to continuity. Thus if $L \in \mathcal{L}(V, W)$, it is meaningful to define

$$\|L\|_{V, W} = \sup_{0 \neq v \in V} \frac{\|Lv\|_W}{\|v\|_V} \quad (1.2)$$

Theorem 1.3.2. The set $\mathcal{L}(V, W)$ is a linear space, and (1.2) defines a norm over the space.

The norm (1.2) is usually called the operator norm of L , which enjoys the following compatibility property

$$\|Lv\|_W \leq \|L\|_{V, W} \|v\|_V \quad \forall v \in V \quad (1.3)$$

Example 1.3.1. Let V be a linear space. Then the identity operator $I : V \rightarrow V$ belongs to $\mathcal{L}(V)$, and $\|I\| = 1$.

Theorem 1.3.3. Let V be a normed space, and W be a Banach space. Then $\mathcal{L}(V, W)$ is a Banach space.

Proof. see [1, page 59] □

1.3.3 Linear functionals

An important special case of linear operators is when they take on scalar values. Let V be a normed space, and $W = \mathbb{K}$, the set of scalars associated with V . The elements in $\mathcal{L}(V, \mathbb{K})$ are called linear functionals. Since $\mathbb{K}$ is complete, $\mathcal{L}(V, \mathbb{K})$ is a Banach space. This space is usually denoted as V' and it is called the dual space of V . Usually we use lower case letters, such as ℓ , to denote a linear function. In this project, since we use exclusively linear functionals which are bounded, we use the term linear functionals to refer to only bounded linear functionals.

Example 1.3.2. Let $\Omega \subset \mathbb{R}^d$ be a bounded open set. It is a well-known result that for $1 \leq p < \infty$, the dual space of $L_p(\Omega)$ can be identified with $L_q(\Omega)$. Here q is the conjugate exponent of p , defined by the relation

$$\frac{1}{p} + \frac{1}{q} = 1$$

By convention, $q = \infty$ when $p = 1$. In other words, given an $\ell \in (L_p(\Omega))'$, there is a function $u \in L_q(\Omega)$, uniquely determined a.e., such that

$$\ell(v) = \int_{\Omega} u(x)v(x) dx \quad \forall v \in L_p(\Omega) \quad (1.4)$$

Conversely, for any $u \in L_q(\Omega)$, the rule

$$u \mapsto \int_{\Omega} u(x)v(x) dx, \quad v \in L_p(\Omega)$$

defines a bounded linear functional on $L_p(\Omega)$. It is convenient to identify $\ell \in (L_p(\Omega))'$ and $u \in L_q(\Omega)$, related as in (1.4). Then we write

$$(L_p(\Omega))' = L_q(\Omega), \quad 1 \leq p < \infty$$

Moreover, $L_{\infty}(\Omega)$ is the dual space of $L_1(\Omega)$, but $L_1(\Omega)$ is not the dual space of $L_{\infty}(\Omega)$. For $p = 2$ we have $L_2(\Omega)$ to be the space of all square integrable functions, and Hölders inequality (1.1) turns out to be the CauchySchwarz inequality

$$\int_{\Omega} |u(x)v(x)| dx \leq \|u\|_{L_2(\Omega)} \|v\|_{L_2(\Omega)}$$

Moreover, for $u, v \in L_2(\Omega)$ we can define the inner product

$$(u, v)_{L_2(\Omega)} := \int_{\Omega} u(x)v(x) dx$$

and with

$$(u, u)_{L_2(\Omega)} = \|u\|_{L_2(\Omega)}^2 \quad \text{for all } u \in L_2(\Omega)$$

we conclude that $L_2(\Omega)$ is Hilbert space.

Definition 1.3.3. Let V be a Banach space. An operator $P \in L(V)$ with the property $P^2 = P$ is called a projection operator.

Theorem 1.3.4. (The Projection Theorem)

Let M be a closed subspace of the Hilbert space H . Then $M^{\perp}$ is also a closed subspace, and

$$H = M \oplus M^{\perp}$$

Further, in the decomposition $f = g + h$ where $g \in M$, $h \in M^{\perp}$, g is the element in M closest to f . Where $M^{\perp} := \{v \in H : (x, v) = 0 \quad \forall x \in M\}$.

Proof. see [3, page 32]

□

Definition 1.3.4. Let V be a real or complex linear space, $K \subset V$. The set K is said to be convex if

$$u, v \in K \Rightarrow \lambda u + (1 - \lambda)v \in K \quad \forall \lambda \in (0, 1).$$

Definition 1.3.5. Let K be a convex set in a linear space V . A function $f : K \rightarrow \mathbb{R}$ is said to be convex if

$$f(\lambda u + (1 - \lambda)v) \leq \lambda f(u) + (1 - \lambda)f(v) \quad \forall u, v \in K, \forall \lambda \in [0, 1].$$

The function f is strictly convex if the above inequality is strict for $u \neq v$ and $\lambda \in (0, 1)$.

Definition 1.3.6. Let V be a real normed space, and A and B non-empty sets in V . The sets A and B are said to be separated if there is a non-zero linear continuous functional ℓ on V and a number $\alpha \in \mathbb{R}$ such that

$$\ell(u) \leq \alpha \quad \forall u \in A,$$

$$\ell(v) \geq \alpha \quad \forall v \in B.$$

If the inequalities are strict, then we say the sets A and B are strictly separated.

Theorem 1.3.5. Let V be a real normed space, A and B be two non-empty disjoint convex subsets of V such that one of them is compact, and the other is closed. Then the sets A and B can be strictly separated.

Definition 1.3.7. If V and W are vector spaces, a bilinear form $a : V \times W \rightarrow \mathbb{R}$ is defined to be an operator with the properties,

$$a(\alpha u + \beta w, v) = \alpha a(u, v) + \beta a(w, v) \quad u, w \in V, v \in W$$

$$a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(u, w) \quad u \in V, v, w \in W$$

where α and β are real numbers.

Theorem 1.3.6. (Hahn-Banach theorem)

Let V_0 be a subspace of a normed space V , and $\ell : V_0 \rightarrow \mathbb{R}$ be linear and bounded. Then there exists an extension $\hat{\ell} \in V'$ of ℓ with $\hat{\ell}(v) = \ell(v) \quad \forall v \in V_0$, and $\|\hat{\ell}\| = \|\ell\|$.

Proof. see [2, page 221] □

Theorem 1.3.7. (Riesz representation theorem)

Let V be a real or complex Hilbert space, $\ell \in V'$. Then there is a unique $u \in V$ for which

$$\ell(v) = (v, u) \quad \forall v \in V \tag{1.5}$$

In addition,

$$\|\ell\| = \|u\| \tag{1.6}$$

Proof. see [1, page 82] □

1.4 Sobolev Spaces

1.4.1 Weak Derivatives

A sequence $\{\varphi_k\}$ in $C_0^\infty(\Omega)$ is said to converge to φ in $C_0^\infty(\Omega)$ if

- (a) there exists a compact set K in Ω such that φ_k vanishes outside K for any k , and
- (b) for each multi-index α , $D^\alpha \varphi_k \rightarrow D^\alpha \varphi$ uniformly in Ω .

The space $C_0^\infty(\Omega)$ endowed with this notion of convergence is called the space of test functions and is denoted by $D(\Omega)$.

Definition 1.4.1. A distribution on Ω is a continuous linear functional on $D(\Omega)$. That is, a linear functional ℓ on $D(\Omega)$ is a distribution if and only if

$$\varphi_k \rightarrow \varphi \text{ in } D(\Omega) \text{ implies } \langle \ell, \varphi_k \rangle \rightarrow \langle \ell, \varphi \rangle.$$

The space of distributions is denoted by $D'(\Omega)$.

Definition 1.4.2. Let $1 \leq p < \infty$. A function $v : \Omega \subset \mathbb{R}^d \rightarrow \mathbb{R}$ is said to be locally p -integrable, $v \in L_{loc}^p(\Omega)$, if for every $x \in \Omega$, there is an open neighborhood Ω' of x such that $\overline{\Omega'} \subset \Omega$ and $v \in L_p(\Omega')$.

Lemma 1.4.1. (Generalized Variational Lemma) Let $v \in L_{loc}^1(\Omega)$ with Ω a nonempty open set in $\mathbb{R}^d$. If

$$\int_{\Omega} v(x) \varphi(x) dx = 0 \quad \forall \varphi \in C_0^\infty(\Omega)$$

, then $v = 0$ a.e. on Ω .

Definition 1.4.3. Let $f \in L_{loc}^1(\Omega)$ be locally integrable function on the open set $\Omega \subseteq \mathbb{R}^n$. Given a multi-index $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$, if there exist a locally integrable function $g \in \mathcal{L}_{loc}^1(\Omega)$ such that

$$\int_{\Omega} f D^\alpha \varphi dx = (-1)^{|\alpha|} \int_{\Omega} D^\alpha f \varphi dx = (-1)^{|\alpha|} \int_{\Omega} g \varphi dx$$

for all test functions, then we say that g is the weak α -th derivatives of f , and write $g = D^\alpha f$.

Example 1.4.1. For $n = 1, \Omega = \mathbb{R}$, $|x|'_w = \text{sgn } x$

Proof. for all $\varphi \in D(\Omega)$

$$\begin{aligned} \int_{-\infty}^{\infty} |x| \varphi'(x) dx &= \int_{-\infty}^0 x \varphi'(x) dx + \int_0^{\infty} x \varphi'(x) dx \\ &= -x \varphi(x) \Big|_{-\infty}^0 + x \varphi(x) \Big|_0^{\infty} + \int_{-\infty}^0 \varphi(x) dx - \int_0^{\infty} \varphi(x) dx \\ &= \int_{-\infty}^0 \varphi(x) dx - \int_0^{\infty} \varphi(x) dx \end{aligned} \tag{1}$$

$$\begin{aligned} (-1) \int_{-\infty}^{\infty} \text{sgn } x \varphi(x) dx &= (-1) \left[- \int_{-\infty}^0 \varphi(x) dx + \int_0^{\infty} \varphi(x) dx \right] \\ &= \int_{-\infty}^0 \varphi(x) dx - \int_0^{\infty} \varphi(x) dx \end{aligned} \tag{2}$$

Hence from (1) and (2) we can see that $|x|'_w = \text{sgn } x$ □

1.4.2 Sobolev Spaces

Definition 1.4.4. Let $\Omega \subset \mathbb{R}^n$ be an open set, $1 \leq p \leq \infty$, and $k \geq 0$ be an integer. The Sobolev spaces $W^{k,p}(\Omega)$ of order k is defined by

$$W^{k,p}(\Omega) = \{f \in L_p(\Omega) : D^\alpha f \in L_p(\Omega), \text{ for all multi-indices } \alpha \text{ such that } |\alpha| \leq k\}$$

Definition 1.4.5. For $f \in W^{k,p}(\Omega)$, then $W^{k,p}(\Omega)$ norm is

$$\|f\|_{W^{k,p}(\Omega)} = \left(\sum_{|\alpha| \leq k} \|D^\alpha f\|_{L_p(\Omega)}^p \right)^{\frac{1}{p}} \quad \text{if } 1 \leq p < \infty \quad (1.7)$$

and

$$\|f\|_{W^{k,p}(\Omega)} = \max_{|\alpha| \leq k} \|D^\alpha f\|_{L_\infty(\Omega)} \quad \text{if } p = \infty \quad (1.8)$$

Definition 1.4.6. The standard seminorm over the space $W^{k,p}(\Omega)$

$$|f|_{W^{k,p}(\Omega)} = \begin{cases} \left(\sum_{|\alpha|=k} \|D^\alpha f\|_{L_p(\Omega)}^p \right)^{\frac{1}{p}} & 1 \leq p < \infty \\ \max_{|\alpha|=k} \|D^\alpha f\|_{L_\infty(\Omega)} & p = \infty. \end{cases}$$

Remark 1.4.1. We can replace $\|v\|_{W^{k,p}(\Omega)}$ by the simpler notation $\|v\|_{k,p,\Omega}$ or even $\|v\|_{k,p}$ when no confusion results.

Theorem 1.4.1. The space $W^{k,p}(\Omega)$ is a Banach space.

Proof. Let $f_1, f_2 \in W^{k,p}(\Omega)$. For $|\alpha| \leq k$, call $D^\alpha f_1, D^\alpha f_2$ their weak derivatives. Then, for any $\lambda, \mu \in \mathbb{R}$, the linear combination $\lambda f_1 + \mu f_2$ is locally integrable function. Its weak derivatives are

$$D^\alpha(\lambda f_1 + \mu f_2) = \lambda D^\alpha f_1 + \mu D^\alpha f_2$$

Therefore, $D^\alpha(\lambda f_1 + \mu f_2) \in L_p(\Omega)$ for every $|\alpha| \leq k$. This proves that $W^{k,p}(\Omega)$ is a vector spaces.

Next, we show that (1.7) and (1.8) are norm.

Indeed, for $\lambda \in \mathbb{R}$ and $f \in W^{k,p}(\Omega)$ one has

$$\|\lambda f\|_{W^{k,p}(\Omega)} = |\lambda| \|f\|_{W^{k,p}(\Omega)}$$

$$\|f\|_{W^{k,p}(\Omega)} \geq \|f\|_{L(p)} \geq 0$$

with equality holding if and only if $f = 0$.

Moreover, if $f_1, f_2 \in W^{k,p}(\Omega)$, then $1 \leq p < \infty$ Minkowski's inequality yields

$$\begin{aligned} \|f_1 + f_2\|_{W^{k,p}(\Omega)} &= \left(\sum_{|\alpha| \leq k} \|D^\alpha f_1 + D^\alpha f_2\|_{L_p(\Omega)}^p \right)^{\frac{1}{p}} \\ &\leq \left(\sum_{|\alpha| \leq k} (\|D^\alpha f_1\|_{L_p(\Omega)} + \|D^\alpha f_2\|_{L_p(\Omega)})^p \right)^{\frac{1}{p}} \\ &\leq \left(\sum_{|\alpha| \leq k} \|D^\alpha f_1\|_{L_p(\Omega)}^p \right)^{\frac{1}{p}} + \left(\sum_{|\alpha| \leq k} \|D^\alpha f_2\|_{L_p(\Omega)}^p \right)^{\frac{1}{p}} \\ &= \|f_1\|_{W^{k,p}(\Omega)} + \|f_2\|_{W^{k,p}(\Omega)} \end{aligned}$$

In the case $p = \infty$, the above computation is repalced by

$$\begin{aligned}\|f_1 + f_2\|_{W^{k,p}(\Omega)} &= \sum_{|\alpha| \leq k} (\|D^\alpha f_1 + D^\alpha f_2\|_{L_\infty(\Omega)}) \\ &\leq \sum_{|\alpha| \leq k} \|D^\alpha f_1\|_{L_\infty(\Omega)} + \sum_{|\alpha| \leq k} \|D^\alpha f_2\|_{L_\infty(\Omega)} \\ &= \|f_1\|_{W^{k,p}(\Omega)} + \|f_2\|_{W^{k,p}(\Omega)}\end{aligned}$$

To conclude the proof, we need to show that the spaces $W^{k,p}(\Omega)$ is complete, hence is a Banach spaces.

Let $\{f_j\}_{j=1}^\infty$ be cauchy sequence in $W^{k,p}(\mathbb{R}^n)$.

Then $\{D^\alpha f_j\}_{j=1}^\infty$ are cauchy sequence in $L_p(\mathbb{R}^n)$ for $|\alpha| \leq k$. Hence there are $f^\alpha \in L_p(\mathbb{R}^n)$ with for $|\alpha| \leq k$. Hence there are $f^\alpha \in L_p(\mathbb{R}^n)$ with

$$D^\alpha f_j \rightarrow f^\alpha \text{ in } L_p(\mathbb{R}^n), |\alpha| \leq k, \text{ and } f^0 = f.$$

It follows from

$$\int_{\mathbb{R}^n} D^\alpha f_j(x) \varphi(x) dx = (-1)^{|\alpha|} \int_{\mathbb{R}^n} f_j(x) D^\alpha \varphi(x) dx, \quad \varphi \in \mathcal{S}(\mathbb{R}^n)$$

And Hölder's inequality applied to $D^\alpha f_j \rightarrow f^\alpha$, $f_j \rightarrow f \in L_p(\mathbb{R}^n)$ and $\varphi, D^\alpha \varphi \in \mathcal{S}(\mathbb{R}^n)$ that

$$\int_{\mathbb{R}^n} D^\alpha f^\alpha(x) \varphi(x) dx = (-1)^{|\alpha|} \int_{\mathbb{R}^n} f(x) D^\alpha \varphi(x) dx, \quad \varphi \in \mathcal{S}(\mathbb{R}^n)$$

Then $f^\alpha = D^\alpha f$, $|\alpha| \leq k$, and $f \in W^{k,p}(\mathbb{R}^n)$ with $f_j \rightarrow f$ in $W^{k,p}(\mathbb{R}^n)$ for $j \rightarrow \infty$. Consequently, $W^{k,p}(\mathbb{R}^n)$ is Banach space. $\square$

Definition 1.4.7. We denote the k -th order Sobolev spaces in $L_2(\Omega)$ by

$$H^k(\Omega) \equiv W^{k,2}(\Omega).$$

And the Sobolev spaces $W^{k,2}(\Omega)$ equipped with the scalar product

$$(f, g)_{W^{k,2}(\Omega)} = \int_{\Omega} \sum_{|\alpha| \leq k} \partial^\alpha f(x) \overline{\partial^\alpha g(x)} dx$$

becomes Hilbert spaces.

Definition 1.4.8. The closure of $C_0^\infty(\Omega)$ in the norm of $W^{k,p}(\Omega)$ is denoted by $W_0^{k,p}(\Omega)$. So, $W_0^{k,p}(\Omega)$ is subspace in the spaces $W^{k,p}(\Omega)$. When $p = 2$, we denote $H_0^k(\Omega) \equiv W_0^{k,2}(\Omega)$

We interpret $W_0^{k,p}(\Omega)$ to be the space of all the functions v in $W^{k,p}(\Omega)$ with the property that

$$D^\alpha v(x) = 0 \quad \text{on } \partial\Omega, \quad \forall \alpha \text{ with } |\alpha| \leq k-1.$$

Definition 1.4.9. Let $s \geq 0$. Then we define $W_0^{s,p}(\Omega)$ to be the closure of the space $C_0^\infty(\Omega)$ in $W^{s,p}(\Omega)$. When $p = 2$, we have a Hilbert space $H_0^s(\Omega) \equiv W_0^{s,2}(\Omega)$. With the spaces $W_0^{s,p}(\Omega)$, we can then define Sobolev spaces with negative order.

Definition 1.4.10. Let $s \geq 0$, either an integer or a non-integer. Let $p \in [1, \infty)$ and denote its conjugate exponent p' defined by the relation $\frac{1}{p} + \frac{1}{p'} = 1$. Then we define $W^{-s,p'}(\Omega)$ to be the dual space of $W_0^{s,p}(\Omega)$. In particular, $H^{-s}(\Omega) \equiv W^{-s,2}(\Omega)$.

On several occasions later, we need to use in particular the Sobolev space $H^{-1}(\Omega)$, defined as the dual of $H_0^1(\Omega)$. Thus, any $\ell \in H^{-1}(\Omega)$ is a bounded linear functional on $H_0^1(\Omega)$:

$$|\ell(v)| \leq M\|v\|, \quad \forall v \in H_0^1(\Omega)$$

The norm of ℓ is

$$\|\ell\|_{H^{-1}(\Omega)} = \sup_{v \in H_0^1(\Omega)} \frac{\ell(v)}{\|v\|_{H_0^1(\Omega)}}$$

Any function $f \in L_2(\Omega)$ naturally induces a bounded linear functional $f \in H^{-1}(\Omega)$ by the relation

$$\langle f, v \rangle = \int_{\Omega} f v \, dx \quad \forall v \in H_0^1(\Omega)$$

Sometimes even when $f \in H^{-1}(\Omega) \setminus L_2(\Omega)$, we write $\int_{\Omega} f v \, dx$ for the duality pairing $\langle f, v \rangle$ between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$, although integration in this situation does not make sense.

Theorem 1.4.2. Assume Ω is a Lipschitz domain in $\mathbb{R}^d$, $1 \leq p < \infty$. Then there exists a continuous linear operator $\gamma : W^{1,p}(\Omega) \rightarrow L_p(\partial\Omega)$ with the following properties.

- (a) $\gamma v = v|_{\partial\Omega}$ if $v \in W^{1,p}(\Omega) \cap C(\bar{\Omega})$
- (b) For some constant $c > 0$, $\|\gamma v\|_{L_p(\partial\Omega)} \leq \|v\|_{W^{1,p}(\Omega)}$, $\forall v \in W^{1,p}(\Omega)$
- (c) The mapping $\gamma : W^{1,p}(\Omega) \rightarrow L_p(\partial\Omega)$ is compact; i.e., for any bounded sequence $\{v_n\}$ in $W^{1,p}(\Omega)$, there is a subsequence $\{v_{n'}\} \subset \{v_n\}$ such that $\{\gamma v_{n'}\}$ is convergent in $L_p(\partial\Omega)$.

The operator γ is called the trace operator, and γv can be called the generalized boundary value of v . The trace operator is neither an injection nor a surjection from $W^{1,p}(\Omega)$ to $L_p(\partial\Omega)$. The range $\gamma(W^{1,p}(\Omega))$ is a space smaller than $L_p(\partial\Omega)$, namely $W^{1-\frac{1}{p},p}(\partial\Omega)$, a positive order Sobolev space over the boundary.

When we discuss weak formulations of boundary value problems later in this project, we need to use traces of the $H^1(\Omega)$ functions, that form the space $H^{\frac{1}{2}}(\partial\Omega)$; in other words,

$$H^{\frac{1}{2}}(\partial\Omega) = \gamma(H^1(\Omega))$$

Correspondingly, we can use the following as the norm for $H^{\frac{1}{2}}(\partial\Omega)$:

$$\|g\|_{H^{\frac{1}{2}}(\partial\Omega)} = \inf_{\substack{v \in H^1(\Omega) \\ \gamma v = g}} \|v\|_{H^1(\Omega)}$$

Proposition 1.4.1. (Poincaré inequality) There exists a constant C_{Ω} such that

$$\|u\|_{L_2(\Omega)} \leq C_{\Omega} \|\nabla u\|_{L_2(\Omega)} \quad \forall u \in H_0^1(\Omega),$$

where the constant C_{Ω} is a constant depending on the diameter of Ω .

Proof. See [6] □

Theorem 1.4.3. (*Divergence theorem*) Let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$ and $Q : \Omega \rightarrow \mathbb{R}^n$ be a vector field whose components are in $H^1(\Omega)$. The following equality holds

$$\int_{\Omega} \nabla \cdot Q \, dx = \int_{\partial\Omega} Q \cdot \nu \, ds$$

where ν is the unit outward pointing normal to $\partial\Omega$.

Proposition 1.4.2. For $Q \in H^1(\overline{\Omega}; \mathbb{R}^n)$, $g \in H^1(\Omega)$,

$$\int_{\Omega} Q \cdot \nabla g \, dx = \int_{\partial\Omega} g Q \cdot \nu \, ds - \int_{\Omega} g \nabla \cdot Q \, dx$$

Proof. By the divergence theorem we see that

$$\int_{\Omega} \nabla \cdot (Qg) \, dx = \int_{\partial\Omega} Q \cdot \nu g \, ds$$

Alternatively,

$$\nabla \cdot (Qg) = g \cdot \nabla Q + Q \cdot \nabla g$$

By combining these equality we get the desired formula. □

For example, if $Q = \nabla u$ and $g = v$, we have by integration by parts formula that

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\partial\Omega} v \nabla u \cdot \nu \, ds - \int_{\Omega} v \Delta u \, dx$$

where Δ is the Laplacian and noting that

$$\nabla u \cdot \nu = \frac{\partial u}{\partial \nu}$$

we obtain Green's formula

$$-\int_{\Omega} v \Delta u \, dx = -\int_{\partial\Omega} v \frac{\partial u}{\partial \nu} \, ds + \int_{\Omega} \nabla u \cdot \nabla v \, dx$$

Chapter 2

Variational Formulations of Elliptic Boundary Value Problems

2.1 Review on Partial differential equation

Definition 2.1.1. Let $u = u(x_1, \dots, x_n)$ be a function of n independent variables $x_1, \dots, x_n$. A Partial Differential Equation (PDE) is an equation that contains the independent variables $x_1, \dots, x_n$, the dependent variable or the unknown function $u : \Omega \rightarrow \mathbb{R}$, where Ω is an open subset of $\mathbb{R}^n$, $n \geq 2$ and its partial derivatives up to some order. It has the form

$$F(x_1, \dots, x_n, u, u_{x_1}, \dots, u_{x_n}, u_{x_1x_1}, \dots, u_{x_ix_j}, \dots) = 0$$

where F is a given function and $u_{x_j} = \partial u / \partial x_j$, $u_{x_ix_j} = \partial^2 u / \partial x_i \partial x_j$, $i, j = 1, \dots, n$ are the partial derivatives of u .

Let $\Omega \subseteq \mathbb{R}^n$ be an open set and $s^{n \times n}$ be the set of $n \times n$ real symmetric matrices. A second order partial differential equation on Ω in an unknown $u = u(x_1, \dots, x_n)$ is an equation of the form

$$F(x, u, Du, D^2u) = 0 \quad (2.1)$$

where $F : \Omega \times \mathbb{R} \times \mathbb{R}^n \times s^{n \times n} \rightarrow \mathbb{R}$. A typical point ϖ of $\Gamma = \Omega \times \mathbb{R} \times \mathbb{R}^n \times s^{n \times n}$ is given by $\varpi = (x, z, \xi, \eta)$ where $x \in \Omega$, $z \in \mathbb{R}$, $\xi \in \mathbb{R}^n$, $\eta \in s^{n \times n}$.

Definition 2.1.2. The second order partial differential equation (2.1) is called linear if it is of the form

$$\sum_{i,j=1}^n a_{ij}(x) D_{ij}u + \sum_{i=1}^n b_i(x) D_iu + c(x)u + d(x) = 0$$

A partial differential equation (2.1) and assume that F is differentiable in the η . We extend F to the whole space of $n \times n$ by say $F = (x, z, \xi, \eta) = F(x, z, \xi, \frac{1}{2}(A + A^T))$, where $(x, z, \xi, A) \in \Omega \times \mathbb{R} \times \mathbb{R}^n \times s^{n \times n} \rightarrow \mathbb{R}$ the $n \times n$ matrix $[F_{ij}(\varpi)]_{n \times n}$ is symmetric where $F_{ij} := \frac{\partial F}{\partial \eta_{ij}}$

Definition 2.1.3. The equation (2.1) is said to be elliptic at a point $\varpi = (x, z, \xi, \eta) \in \Gamma = \Omega \times \mathbb{R} \times \mathbb{R}^n \times s^{n \times n}$ if and only if the matrix $[F_{ij}(\varpi)]_{n \times n}$ is positive definite, that is $\sum_{i,j=1}^n \frac{\partial F}{\partial \eta_{ij}} \zeta_i \zeta_j > 0$ for all $\zeta \in \mathbb{R}^n \setminus \{0\}$.

Equivalently, the partial differential equation (2.1) is elliptic at ϖ if and only if the all Eigen values (they depend on ϖ) of $[F_{ij}(\varpi)]_{n \times n}$ are positive.

Example 2.1.1. *The equation*

$$-\Delta u = f \quad (\text{where } \Delta = \sum_{i=1}^3 \frac{\partial^2}{\partial x_i^2} \quad \text{in } \mathbb{R}^3)$$

is elliptic everywhere in $\mathbb{R}^3$.

A partial differential equation subject to certain conditions in the form of initial or boundary conditions is known as an initial value problem (IVP) or boundary value problem (BVP). The initial conditions, also known as Cauchy conditions, are the values of the unknown function u and of an appropriate number of its derivatives at the initial point, while the boundary conditions are the values on the boundary $\partial\Omega$ of the domain Ω under consideration. The three most important kinds of boundary conditions are:

- (i) Dirichlet conditions or boundary conditions of the first kind are the values of u prescribed at each point of the boundary $\partial\Omega$.
- (ii) Neumann conditions or boundary conditions of the second kind are the values of the normal derivative of u prescribed at each point of the boundary $\partial\Omega$.
- (iii) Robin conditions or mixed boundary conditions or boundary conditions of the third kind are the values of a linear combination of u and its normal derivative prescribed at each point of the boundary $\partial\Omega$.

Remark 2.1.1. *Dirichlet boundary conditions are sometimes called essential since they essentially influence the weak formulation: They determine the function space in which the solution is sought. On the other hand, Neumann boundary conditions do not influence the function space and can be naturally incorporated into the boundary integrals. Therefore they are called natural.*

Definition 2.1.4. *(Hadamard's well-posedness) A problem is said to be well-posed If*

1. *it has a unique solution,*
2. *the solution depends continuously on the given data.*

otherwise the problem is ill-posed.

Further reading see [4]

2.2 A model boundary value problem

To begin, Let Ω be a bounded domain in $\mathbb{R}^d$ with a Lipschitz continuous boundary Γ . The unit outward normal vector $n = (n_1, \dots, n_d)^T$ exists a.e. on Γ , and we will

use $\partial u/\partial n$ to denote the normal derivative of u on Γ . We use the following model boundary value problem as an illustrative example:

$$\begin{cases} -\Delta u = f, & \text{in } \Omega \\ u = 0, & \text{on } \Gamma. \end{cases} \quad (2.2)$$

Here Δ denotes the Laplacian operator, defined by

$$\Delta u = \frac{\partial^2 u}{\partial x_i \partial x_i}$$

The differential equation in (2.2) is called the Poisson equation. The Poisson equation can be used to describe many physical processes, e.g., steady state heat conduction, electrostatics, deformation of a thin elastic membrane.

A classical solution of the problem (2.2) is a smooth function $u \in C^2(\Omega) \cap C(\bar{\Omega})$ which satisfies the differential equation (2.2)₁ and the boundary condition (2.2)₂ pointwise. Necessarily we have to assume $f \in C(\Omega)$, but this condition, or even the stronger condition $f \in C(\bar{\Omega})$ does not guarantee the existence of a classical solution of the problem. One purpose of the introduction of the weak formulation is to remove the high smoothness requirement on the solution and as a result it is easier to have the existence of a (weak) solution.

To derive the weak formulation corresponding to (2.2), we temporarily assume it has a classical solution $u \in C^2(\Omega) \cap C(\bar{\Omega})$. We multiply the differential equation (2.2)₁ by an arbitrary function $v \in C_0^\infty(\Omega)$ (so-called smooth test functions), and integrate the relation on Ω ,

$$-\int_{\Omega} \Delta u \, v \, dx = \int_{\Omega} f v \, dx.$$

An integration by parts for the integral on the left side yields (recall that $v = 0$ on Γ)

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx \quad (2.3)$$

This relation was proved under the assumptions $u \in C^2(\Omega) \cap C(\bar{\Omega})$ and $v \in C_0^\infty(\Omega)$. However, for all the terms in the relation (2.3) to make sense, we only need to require the following regularities of u and v that is $u, v \in H^1(\Omega)$, assuming $f \in L_2(\Omega)$. Recalling the homogeneous Dirichlet boundary condition (2.2)₂, we thus seek a solution $u \in H_0^1(\Omega)$ satisfying the relation (2.3) for any $v \in C_0^\infty(\Omega)$. Since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, the relation (2.3) is then valid for any $v \in H_0^1(\Omega)$. Therefore, the weak formulation of the boundary value problem (2.2) is

$$u \in H_0^1(\Omega), \quad \int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx \quad \forall v \in H_0^1(\Omega) \quad (2.4)$$

Actually, we can even weaken the assumption, $f \in L_2(\Omega)$. It is enough for us to assume $f \in H^{-1}(\Omega) = (H_0^1(\Omega))'$, as long as we interpret the integral $\int_{\Omega} f v \, dx$ as the duality pairing $\langle f, v \rangle$ between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$. We adopt the convention of using $\int_{\Omega} f v \, dx$ for $\langle f, v \rangle$ when $f \in H^{-1}(\Omega)$ and $v \in H_0^1(\Omega)$.

We have shown that if u is a classical solution of (2.2), then it is also a solution of the weak formulation (2.4). Conversely, suppose u is a weak solution with the additional

regularity $u \in C^2(\Omega) \cap C(\overline{\Omega})$ and $f \in C(\Omega)$. Then for any $v \in C_0^\infty(\Omega) \subset H_0^1(\Omega)$ from (2.4) we obtain

$$\int_{\Omega} (-\Delta u - f)v \, dx = 0$$

Then we must have $-\Delta u = f$ in Ω , i.e., the differential equation $(2.2)_1$ is satisfied. Also u satisfies the homogeneous Dirichlet boundary condition pointwisely.

Thus we have shown that the boundary value problem (2.2) and the variational problem (2.3) are formally equivalent. In case the weak solution u does not have the regularity $u \in C^2(\Omega) \cap C(\overline{\Omega})$. We will say u formally solves the boundary value problem (2.2).

We let $V = H_0^1(\Omega)$, $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ the bilinear form defined by

$$a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, dx \quad \text{for } u, v \in V,$$

and $\ell : V \rightarrow \mathbb{R}$ the linear functional defined by

$$\ell(v) = \int_{\Omega} f v \, dx \quad \text{for } v \in V.$$

Then the weak formulation of the problem (2.2) is to find $u \in V$ such that

$$a(u, v) = \ell(v) \quad \forall v \in V \tag{2.5}$$

We define a differential operator A associated with the boundary value problem (2.2) by

$$A : H_0^1(\Omega) \rightarrow H^{-1}(\Omega), \quad \langle u, v \rangle = a(u, v) \quad \forall u, v \in H_0^1(\Omega).$$

Here, $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$. Then the problem (2.5) can be viewed as a linear operator equation

$$Au = \ell \quad \text{in } H^{-1}(\Omega)$$

A formulation of the type (2.2) in the form of a partial differential equation and a set of boundary conditions is referred to as a classical formulation of a boundary value problem, whereas a formulation of the type (2.5) is known as a weak formulation. One advantage of weak formulations over classical formulations is that questions related to existence and uniqueness of solutions can be answered more satisfactorily.

2.3 Some general results on existence and uniqueness

We first present some general ideas and results on existence and uniqueness for a linear operator equation of the form

$$u \in V, \quad Lu = f, \tag{2.6}$$

where $L : \mathcal{D}(L) \subset V \rightarrow W$, V and W are Hilbert spaces, and $f \in W$. Notice that the solvability of the equation is equivalent to the condition $\mathcal{R}(L) = W$, whereas the uniqueness of a solution is equivalent to the condition $\mathcal{N}(L) = \{0\}$.

A very basic existence result is the following theorem.

Theorem 2.3.1. *Let V and W be Hilbert spaces, $L : \mathcal{D}(L) \subset V \rightarrow W$ a linear operator. Then $\mathcal{R}(L) = W$ if and only if $\mathcal{R}(L)$ is closed and $\mathcal{R}(L)^\perp = \{0\}$.*

Proof. If $\mathcal{R}(L) = W$, then obviously $\mathcal{R}(L)$ is closed and $\mathcal{R}(L)^\perp = \{0\}$. Now assume $\mathcal{R}(L)$ is closed and $\mathcal{R}(L)^\perp = \{0\}$, but $\mathcal{R}(L) \neq W$. Then $\mathcal{R}(L)$ is a closed subspace of W . Let $w \in W \setminus \mathcal{R}(L)$. By Theorem(1.3.5), the compact set $\{w\}$ and the closed convex set $\mathcal{R}(L)$ can be strictly separated by a closed hyperplane, i.e., there exists a $w^* \in W'$ such that $\langle w^*, w \rangle > 0$ and $\langle w^*, Lv \rangle \leq 0$ for all $v \in \mathcal{D}(L)$. Since L is a linear operator, $\mathcal{D}(L)$ is a subspace of V . Hence, $\langle w^*, Lv \rangle = 0$ for all $v \in \mathcal{D}(L)$. Therefore, $0 \neq w^* \in \mathcal{R}(L)^\perp$. This is a contradiction. $\square$

Let us see under what conditions $\mathcal{R}(L)$ is closed. We first introduce an important generalization of the notion of continuity.

Definition 2.3.1. *Let V and W be Banach spaces. An operator $T : \mathcal{D}(T) \subset V \rightarrow W$ is said to be a closed operator if for any sequence $\{v_n\} \subset \mathcal{D}(T)$, $v_n \rightarrow v$ and $T(v_n) \rightarrow w$ imply $v \in \mathcal{D}(T)$ and $w = T(v)$.*

Theorem 2.3.2. *Let V and W be Hilbert spaces, $L : \mathcal{D}(L) \subset V \rightarrow W$ a linear closed operator. Assume for some constant $c > 0$, the following a priori estimate holds:*

$$\|Lv\|_W \geq c\|v\|_V \quad \forall v \in \mathcal{D}(L) \quad (2.7)$$

which is usually called a stability estimate. Also assume $\mathcal{R}(L)^\perp = \{0\}$. Then for each $f \in W$, the equation (2.6) has a unique solution.

Proof. Let us verify that $\mathcal{R}(L)$ is closed. Let $\{f_n\}$ be a sequence in $\mathcal{R}(L)$, converging to f . Then there is a sequence $\{v_n\} \subset \mathcal{D}(L)$ with $f_n = Lv_n$. By(2.7),

$$c\|v_n - v_m\|_V \leq \|f_n - f_m\|_W$$

Thus $\{v_n\}$ is a Cauchy sequence in V . Since V is a Hilbert space, the sequence $\{v_n\}$ converges : $v_n \rightarrow v \in V$. Now L is assumed to be closed, we conclude that $v \in \mathcal{D}(L)$ and $f = Lv \in \mathcal{R}(L)$. So we can invoke Theorem(2.3.1) to obtain the existence of a solution. Let $v \in \mathcal{N}(L)$ implies $Lv = 0$, so that $\|Lv\| = 0$ by stability estimate(2.7) we obtain $c\|v\|_V \leq \|Lv\|_W = 0$. Hence, $v = 0$. Thus $\mathcal{N}(L) = \{0\}$. Therefore there is a unique solution. $\square$

Example 2.3.1. *Let V be a Hilbert space, $L \in \mathcal{L}(V, V')$ be strongly monotone, i.e., for some constant $c > 0$,*

$$\langle Lv, v \rangle \geq c\|v\|_V^2 \quad \forall v \in V.$$

Then (2.7) holds because from the monotonicity,

$$\|Lv\|_{V'}\|v\|_V \geq c\|v\|_V^2$$

which implies

$$\|Lv\|_{V'} \geq c\|v\|_V.$$

Also $\mathcal{R}(L)^\perp = \{0\}$, since from $v \perp \mathcal{R}(L)$ we have

$$c\|v\|_V^2 \leq \langle Lv, v \rangle = 0,$$

and hence $v = 0$. Therefore from Theorem (2.3.2), under the stated assumptions, for any $f \in V'$, there is a unique solution $u \in V$ to the equation $Lu = f$ in V' .

Example 2.3.2. As a concrete example, we consider the weak formulation of the model elliptic boundary value problem (2.6). Here, $\Omega \subset \mathbb{R}^d$ is an open bounded set with a Lipschitz boundary $\partial\Omega$, $V = H_0^1(\Omega)$ with the norm $\|v\|_V = |v|_{H^1(\Omega)}$, and $V' = H^{-1}(\Omega)$. Given $f \in H^{-1}(\Omega)$, consider the problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (2.8)$$

We define the operator $L : V \rightarrow V'$ by

$$\langle Lu, v \rangle = \int_{\Omega} \nabla u \cdot \nabla v \, dx \quad u, v \in V.$$

Then L is linear, continuous, and strongly monotone; indeed, we have

$$\langle Lv, v \rangle = \|v\|_V^2 \quad \forall v \in V.$$

Proof.

$$\begin{aligned} \langle L(\alpha u + \beta w), v \rangle &= \int_{\Omega} \nabla(\alpha u + \beta w) \cdot \nabla v \, dx \\ &= \int_{\Omega} (\alpha \nabla u + \beta \nabla w) \cdot \nabla v \, dx \\ &= \int_{\Omega} (\alpha \nabla u \cdot \nabla v + \beta \nabla w \cdot \nabla v) \, dx \\ &= \alpha \int_{\Omega} \nabla u \cdot \nabla v \, dx + \beta \int_{\Omega} \nabla w \cdot \nabla v \, dx \\ &= \alpha \langle Lu, v \rangle + \beta \langle Lw, v \rangle \\ &= \langle \alpha Lu, v \rangle + \langle \beta Lw, v \rangle \\ &= \langle \alpha Lu + \beta Lw, v \rangle \end{aligned}$$

Thus, $L(\alpha u + \beta w) = \alpha Lu + \beta Lw \quad \forall u, v \text{ and } w \in V \text{ and } \forall \alpha, \beta \in \mathbb{R}$. Hence L is linear. Next, let us show continuity

$$\begin{aligned} \langle Lu, v \rangle &= \int_{\Omega} \nabla u \cdot \nabla v \, dx \\ &\leq \int_{\Omega} |\nabla u \cdot \nabla v| \, dx \\ &\leq \int_{\Omega} \|\nabla u\| \|\nabla v\| \, dx \quad (\text{by Cauchy Schwarz inequality for } \mathbb{R}^2) \\ &\leq \|\nabla u\|_{L_2(\Omega)} \|\nabla v\|_{L_2(\Omega)} \quad (\text{by Cauchy Schwarz inequality for } L_2(\Omega)) \end{aligned}$$

Since,

$$\begin{aligned} \|\nabla u\|_{L_2(\Omega)}^2 &= \int_{\Omega} \nabla u \cdot \nabla u \, dx \\ &= \int_{\Omega} (\nabla u \cdot \nabla u + u^2) \, dx \\ &= \|u\|_V^2 \end{aligned}$$

and similarly for v , it follows that

$$\langle Lu, v \rangle \leq \|u\|_V \|v\|_V \quad \text{for all } u, v \in V$$

implies,

$$\sup_{v \neq 0} \frac{\langle Lu, v \rangle}{\|v\|_V} \leq \|u\|_V.$$

Thus, $\|Lu\| \leq \|u\|_V$. Hence L is bounded. Therefore L is continuous.

Now let us show strong monotonicity of L . From Poincaré's inequality we get

$$\langle Lu, u \rangle \geq c^2 \|u\|_{H^1(\Omega)}^2 \quad \text{for all } u \in H_0^1(\Omega)$$

Therefore, L is strongly monotone. Thus from Example (2.3.1), for any $f \in H^{-1}(\Omega)$, there is a unique $u \in H_0^1(\Omega)$ such that

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \langle f, v \rangle \quad \forall v \in V.$$

i.e., the boundary value problem (2.8) has a unique weak solution. $\square$

2.4 The Lax-Milgram Lemma

The Lax-Milgram Lemma is employed frequently in the study of linear elliptic boundary value problems of the form (2.5). For a real Banach space V , let us first explore the relation between a linear operator $A : V \rightarrow V'$ and a bilinear form $a : V \times V \rightarrow \mathbb{R}$ related by

$$\langle Au, v \rangle = a(u, v) \quad \forall u, v \in V. \quad (2.9)$$

The bilinear form $a(\cdot, \cdot)$ is continuous if and only if there exists $M > 0$ such that

$$|a(u, v)| \leq M \|u\| \|v\| \quad \forall u, v \in V.$$

Theorem 2.4.1. *There exists a one-to-one correspondence between linear continuous operators $A : V \rightarrow V'$ and continuous bilinear forms $a : V \times V \rightarrow \mathbb{R}$, given by the formula (2.9).*

Proof. If $A \in \mathcal{L}(V, V')$, then $a : V \times V \rightarrow \mathbb{R}$ defined in (2.9) is bilinear and bounded:

$$|a(u, v)| \leq \|Au\| \|v\| \leq \|A\| \|u\| \|v\| \quad \forall u, v \in V.$$

Conversely, let $a(\cdot, \cdot)$ be given as a continuous bilinear form on V . For any fixed $u \in V$, the map $v \mapsto a(u, v)$ defines a linear continuous operator on V . Thus, there is an element $Au \in V'$ such that (2.9) holds. From the bilinearity of $a(\cdot, \cdot)$, we obtain the linearity of A . From the boundedness of $a(\cdot, \cdot)$, we obtain the boundedness of A . $\square$

With a linear operator A and a bilinear form a related through (2.9), many properties of the linear operator A can be defined through those of the bilinear form a , or vice versa. Some examples are (assuming V is a real Hilbert space):

- a is bounded ($a(u, v) \leq M \|u\| \|v\| \quad \forall u, v \in V$) if and only if A is bounded ($\|Av\| \leq M \|v\| \quad \forall v \in V$).

- a is positive ($a(v, v) \geq 0 \quad \forall v \in V$) if and only if A is positive ($\langle Av, v \rangle \geq 0 \quad \forall v \in V$).
- a is strictly positive ($a(v, v) > 0 \quad \forall 0 \neq v \in V$) if and only if A is strictly positive ($\langle Av, v \rangle > 0 \quad 0 \neq v \in V$).
- a is strongly positive or V -elliptic ($a(v, v) \geq \alpha \|v\|^2 \quad \forall v \in V$) if and only if A is strongly positive ($\langle Av, v \rangle \geq \alpha \|v\|^2 \quad \forall v \in V$).
- a is symmetric ($a(u, v) = a(v, u) \quad \forall u, v \in V$) if and only if A is symmetric ($\langle Au, v \rangle = \langle Av, u \rangle \quad \forall u, v \in V$).

Theorem 2.4.2. (*Lax-Milgram Lemma*)

Assume V is a Hilbert space, $a(\cdot, \cdot)$ is a bounded, V -elliptic bilinear form on V , $\ell \in V'$. Then there is a unique solution of the problem

$$u \in V, \quad a(u, v) = \ell(v) \quad \forall v \in V. \quad (2.10)$$

Proof. (Existence)

Let $A : V \rightarrow V'$ be the linear operator associated with the bilinear form $a(\cdot, \cdot)$; see (2.9). Then A is bounded and strongly positive: $\forall v \in V$,

$$\|Av\| \leq M\|v\|,$$

$$\langle Av, v \rangle \geq \alpha \|v\|^2.$$

Denote $\mathcal{J} : V' \rightarrow V$ the isometric dual mapping from the Riesz representation theorem. Then

$$a(u, v) = \langle Au, v \rangle = (\mathcal{J}Au, v) \quad \forall u, v \in V,$$

and

$$\|\mathcal{J}Au\| = \|Au\| \quad \forall u \in V.$$

And, Let $L = \mathcal{J}A : V \rightarrow V$. We recall that $\mathcal{R}(L) = V$ if and only if $\mathcal{R}(L)$ is closed and $\mathcal{R}(L)^\perp = \{0\}$.

To show $\mathcal{R}(L)$ is closed, we let $\{u_n\} \subset \mathcal{R}(L)$ be a sequence converging to u . Then $u_n = \mathcal{J}Aw_n$ for some $w_n \in V$. We have

$$\|u_n - u_m\| = \|\mathcal{J}A(w_n - w_m)\| = \|A(w_n - w_m)\| \geq \alpha \|w_n - w_m\|.$$

Hence $\{w_n\}$ is a Cauchy sequence and so has a limit $w \in V$. Then

$$\|u_n - \mathcal{J}Aw\| = \|\mathcal{J}A(w_n - w)\| = \|A(w_n - w)\| \leq M\|w_n - w\| \rightarrow 0.$$

Hence, $u = \mathcal{J}Aw \in \mathcal{R}(L)$ and $\mathcal{R}(L)$ is closed.

Now suppose $u \in \mathcal{R}(L)^\perp$. Then for any $v \in V$,

$$0 = (\mathcal{J}Av, u) = a(v, u).$$

Taking $v = u$ above, we have $a(u, u) = 0$. By the V -ellipticity of $a(\cdot, \cdot)$, we conclude $u = 0$. So we can invoke Theorem (2.3.1) to obtain the existence of a solution.

(Uniqueness)

Suppose u_1, u_2 are two solutions of (2.10), i.e;

$$a(u_1, v) = \ell(v),$$

$$a(u_2, v) = \ell(v) \quad \forall v \in V.$$

Subtraction and linearity give

$$a(u_2, v) - a(u_1, v) = a(u_2 - u_1, v) = 0 \quad \forall v \in V.$$

In particular, choose $v = u_2 - u_1$, then we see

$$a(u_2 - u_1, u_2 - u_1) = 0.$$

By the V -ellipticity of $a(\cdot, \cdot)$,

$$\begin{aligned} 0 &\geq \alpha \|u_2 - u_1\|_V^2, \\ \Rightarrow 0 &\geq \|u_2 - u_1\|_V, \\ \Rightarrow u_2 &= u_1 \end{aligned}$$

□

Example 2.4.1. *Applying the Lax-Milgram Lemma, we conclude that the boundary value problem (2.8) has a unique weak solution $u \in H_0^1(\Omega)$.*

2.5 Weak formulations of linear elliptic boundary value problems

In this section, we formulate and analyze weak formulations of some linear elliptic boundary value problems. To present the ideas clearly, we will frequently use boundary value problems associated with the Poisson equation

$$-\Delta u = f$$

and the Helmholtz equation

$$-\Delta u + u = f$$

as examples.

2.5.1 Problems with homogeneous Dirichlet boundary conditions

So far, we have studied the model elliptic boundary value problem corresponding to the Poisson equation with the homogeneous Dirichlet boundary condition

$$-\Delta u = f \quad \text{in } \Omega, \tag{2.11}$$

$$u = 0 \quad \text{in } \Gamma, \tag{2.12}$$

where $f \in L_2(\Omega)$. The weak formulation of the problem is

$$u \in V, \quad a(u, v) = \ell(v) \quad \forall v \in V. \tag{2.13}$$

Here

$$\begin{aligned} V &= H_0^1(\Omega), \\ a(u, v) &= \int_{\Omega} \nabla u \cdot \nabla v \, dx \quad \text{for } u, v \in V \\ \ell(v) &= \int_{\Omega} f v \, dx \quad \text{for } v \in V \end{aligned}$$

The problem (2.13) has a unique solution $u \in V$ by the Lax-Milgram Lemma.

2.5.2 Problems with non-homogeneous Dirichlet boundary conditions

Suppose that instead of (2.12) the boundary condition is

$$u = g \quad \text{on } \Gamma. \quad (2.14)$$

To derive a weak formulation, we first assume the boundary value problem (2.11)-(2.14) has a classical solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$. Multiplying the equation (2.11) by a test function v with certain smoothness which validates the following calculations, and integrating over Ω , we have

$$\int_{\Omega} -\Delta u v \, dx = \int_{\Omega} f v \, dx.$$

Integrate by parts,

$$-\int_{\Gamma} \frac{\partial u}{\partial \nu} v \, ds + \int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx$$

We now assume $v = 0$ on Γ so that the boundary integral term vanishes; the boundary integral term would otherwise be difficult to deal with under the expected regularity condition $u \in H^1(\Omega)$ on the weak solution. Thus we arrive at the relation

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx$$

if v is smooth and $v = 0$ on Γ . For each term in the above relation to make sense, we assume $f \in L_2(\Omega)$, and let $u \in H^1(\Omega)$ and $v \in H_0^1(\Omega)$. Recall that the solution u should satisfy the boundary condition $u = g$ on Γ . We observe that it is necessary to assume $g \in H^{1/2}(\Omega)$. Finally, we obtain the weak formulation for the boundary value problem (2.11)-(2.14):

$$u \in H^1(\Omega), \quad u = g \quad \text{on } \Gamma, \quad \int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx \quad \forall v \in H_0^1(\Omega). \quad (2.15)$$

For the weak formulation (2.15), though, we cannot apply Lax-Milgram Lemma directly since the trial function u and the test function v do not lie in the same space. There is a standard way to get rid of this problem. Since $g \in H^{1/2}(\Omega)$ and $\gamma(H^1(\Omega)) = H^{1/2}(\Omega)$ (recall that γ is the trace operator), we have the existence of a function $G \in H^1(\Omega)$ such that $\gamma G = g$. We remark that finding the function G in practice may be nontrivial. Thus, setting

$$u = w + G,$$

the problem may be transformed into one of seeking w such that

$$w \in H_0^1(\Omega), \quad \int_{\Omega} \nabla w \cdot \nabla v \, dx = \int_{\Omega} (f v - \nabla G \cdot \nabla v) \, dx \quad \forall v \in H_0^1(\Omega) \quad (2.16)$$

The classical form of the boundary value problem for w is

$$\begin{aligned} -\Delta w &= f + \Delta G \quad \text{in } \Omega, \\ w &= 0 \quad \text{on } \Gamma. \end{aligned}$$

Applying the Lax-Milgram Lemma, we have a unique solution $w \in H_0^1(\Omega)$ of the problem (2.16). Then we set $u = w + G$ to get a solution u of the problem (2.15). Notice that the choice of the function G is not unique, so the uniqueness of the solution u of the problem (2.15) does not follow from the above argument. Nevertheless we can show the uniqueness of u by a standard approach. Assume both u_1 and u_2 are solution of the problem (2.15). Then the difference $u_1 - u_2$ satisfies

$$u_1 - u_2 \in H_0^1(\Omega), \quad \int_{\Omega} \nabla(u_1 - u_2) \cdot \nabla v \, dx = 0 \quad \forall v \in H_0^1(\Omega)$$

Taking $v = u_1 - u_2$, we obtain

$$\int_{\Omega} |\nabla(u_1 - u_2)|^2 \, dx = 0$$

Thus, $\nabla(u_1 - u_2) = 0$ a.e. in Ω , and hence $u_1 - u_2 = c$ a.e. in Ω . Using the boundary condition $u_1 - u_2 = 0$ a.e. Γ , we see that $u_1 = u_2$ a.e. in Ω .

2.5.3 Problems with Neumann boundary conditions

Consider next the Neumann problem of determining u which satisfies

$$\begin{cases} -\Delta u + u = f & \text{in } \Omega, \\ \partial u / \partial \nu = g & \text{on } \Gamma. \end{cases} \quad (2.17)$$

Here f and g are given functions in Ω and on Γ , respectively, and $\partial / \partial \nu$ denotes the normal derivative on Γ . Again we first derive a weak formulation.

Assume $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$ is a classical solution of the problem (2.17). Multiplying (2.17)1 by an arbitrary test function v with certain smoothness for the following calculations to make sense, integrating over Ω and performing an integration by parts, we obtain

$$\int_{\Omega} (\nabla u \cdot \nabla v + u v) \, dx = \int_{\Omega} f v \, dx + \int_{\Gamma} \frac{\partial u}{\partial \nu} v \, ds$$

Then, substitution of the Neumann boundary condition (2.17)2 in the boundary term leads to the relation

$$\int_{\Omega} (\nabla u \cdot \nabla v + u v) \, dx = \int_{\Omega} f v \, dx + \int_{\Gamma} g v \, ds$$

Assume $f \in L_2(\Omega)$, $g \in L_2(\Gamma)$. For each term in the above relation to make sense, it is natural to choose the space $H^1(\Omega)$ for both the trial function u and the test function v . Thus, the weak formulation of the boundary value problem (2.17) is

$$u \in H^1(\Omega) \quad \int_{\Omega} (\nabla u \cdot \nabla v + u v) \, dx = \int_{\Omega} f v \, dx + \int_{\Gamma} g v \, ds \quad \forall v \in H^1(\Omega) \quad (2.18)$$

This problem has the form (2.13), where $V = H^1(\Omega)$, $a(\cdot, \cdot)$ and $\ell(\cdot)$ are defined by

$$\begin{aligned} a(u, v) &= \int_{\Omega} (\nabla u \cdot \nabla v + u v) \, dx, \\ \ell(v) &= \int_{\Omega} f v \, dx + \int_{\Gamma} g v \, ds, \end{aligned}$$

respectively. Applying the Lax-Milgram Lemma, we can show that the weak formulation (2.18) has a unique solution $u \in H^1(\Omega)$.

It is more delicate to study the pure Neumann problem for the Poisson equation

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ \partial u / \partial \nu = g & \text{on } \Gamma. \end{cases} \quad (2.19)$$

where $f \in L_2(\Omega)$ and $g \in L_2(\Gamma)$ are given. In general, the problem (2.19) does not have a solution, and when the problem has a solution u , any function of the form $u + c, c \in \mathbb{R}$, is a solution. This suggests that in formulating the weak version of this problem we should restrict ourselves to the subspace

$$V = \{v \in H^1(\Omega) \mid \int_{\Omega} v \, dx = 0\}$$

Formally, the corresponding weak formulation is

$$u \in H^1(\Omega) \quad \int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx + \int_{\Gamma} g v \, ds \quad \forall v \in H^1(\Omega) \quad (2.20)$$

An application of Equivalent Norm Theorem shows that over the space V , $|\cdot|_1$ is a norm equivalent to the norm $\|\cdot\|_1$. The bilinear form $a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, dx$ is both continuous and V -elliptic. So there is a unique solution to the problem

$$u \in V \quad \int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx + \int_{\Gamma} g v \, ds \quad \forall v \in V.$$

Chapter 3

Basics in Finite Element Method

In this chapter, we introduce the concept of the finite element method, the finite element interpolation theory and its application in error estimates of finite element solutions of linear elliptic boundary value problems.

3.1 One-dimensional examples

To have some idea of the finite element method, in this section we examine some examples on solving one-dimensional boundary value problems. These examples exhibit various aspects of the finite element method in the simple context of one-dimensional problems.

3.1.1 Linear elements for a second-order problem

Let us consider a finite element method to solve the boundary value problem

$$\begin{cases} u'' + u = f & \text{in } \Omega = (0, 1), \\ u(0) = 0, \quad u'(1) = b, \end{cases} \quad (3.1)$$

where $f \in L_2(0, 1)$ and $b \in \mathbb{R}$ are given. Let

$$V = H_0(0, 1) = \{v \in H^1(0, 1) \mid v(0) = 0\},$$

a subspace of $H^1(0, 1)$. The weak formulation of the problem is

$$u \in V, \quad \int_0^1 (u'v' + uv) dx = \int_0^1 f v dx + bv(1) \quad \forall v \in V. \quad (3.2)$$

Applying the Lax-Milgram Lemma, we see that the problem (3.2) has a unique solution. Let us develop a finite element method for the problem. For a natural number N , we partition the set $\overline{\Omega}=[0,1]$ into N parts:

$$\overline{\Omega} = \bigcup_{i=1}^N K_i,$$

where $K_i = [x_{i-1}, x_i]$, $1 \leq i \leq N$, are called the elements, and the x_i , $0 \leq i \leq N$, are called the nodes, $0 = x_0 < x_1 < \cdots < x_N = 1$. In this example, we have a Dirichlet

condition at the node x_0 . Denote $h_i = x_i - x_{i-1}$, and $h = \max_{1 \leq i \leq N} h_i$. The value h is called the mesh size or mesh parameter. We use continuous piecewise linear functions for the approximation, i.e., we choose

$$V_h = \{v_h \in V \mid v_h|_{K_i} \in \mathbb{P}_1(K_i), 1 \leq i \leq N\}.$$

We know that for a piecewisely smooth function v_h , $v_h \in H^1(\Omega)$ if and only if $v_h \in C(\overline{\Omega})$. Thus a more transparent yet equivalent definition of the finite element space is

$$V_h = \{v_h \in C(\overline{\Omega}) \mid v_h|_{K_i} \in \mathbb{P}_1(K_i), 1 \leq i \leq N, v_h(0) = 0\}.$$

For the basis functions of the space v_h , we introduce hat functions associated with the nodes $x_1, \dots, x_N$. For $i = 1, \dots, N-1$, let

$$\phi_i(x) = \begin{cases} (x - x_{i-1})/h_i, & x_{i-1} \leq x \leq x_i, \\ (x_{i+1} - x)/h_{i+1}, & x_i \leq x \leq x_{i+1}, \\ 0, & \text{otherwise,} \end{cases} \quad (3.3)$$

and for $i = N$,

$$\phi_N(x) = \begin{cases} (x - x_{N-1})/h_N, & x_{N-1} \leq x \leq x_N, \\ 0, & \text{otherwise.} \end{cases} \quad (3.4)$$

These functions are continuous and piecewise linear (see Figure 3.1). It is easy to see they are linearly independent. The first order weak derivatives of the basis functions are piecewise constants. Indeed, for $i = 1, \dots, N-1$

$$\phi'_i(x) = \begin{cases} 1/h_i, & x_{i-1} \leq x \leq x_i, \\ -1/h_{i+1}, & x_i \leq x \leq x_{i+1}, \\ 0, & x < x_{i-1} \text{ or } x > x_{i+1}, \end{cases}$$

and for $i = N$,

$$\phi'_N(x) = \begin{cases} 1/h_N, & x_{N-1} \leq x \leq x_N, \\ 0, & x < x_{N-1}. \end{cases}$$

Then the finite element space can be expressed as

$$V_h = \text{span}\{\phi_i \mid 1 \leq i \leq N\},$$

i.e., any function in V_h is a linear combination of the hat functions $\{\phi_i\}_{i=1}^N$. The corresponding finite element method is

$$u_h \in V_h, \quad \int_0^1 (u'_h v'_h + u_h v_h) dx = \int_0^1 f v_h dx + b v_h(1) \quad \forall v_h \in V_h, \quad (3.5)$$

which admits a unique solution by another application of the Lax-Milgram Lemma. Write

$$u_h = \sum_{j=1}^N u_j \phi_j.$$

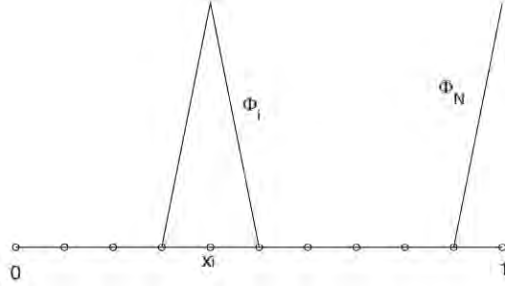


Figure 3.1: Piecewise linear basis functions

Note that $u_j = u_h(x_j)$, $1 \leq j \leq N$. We see that the finite element method (3.5) is equivalent to the following linear system for the unknowns $u_1, \dots, u_N$:

$$\sum_{j=1}^N u_j \int_0^1 (\phi'_i \phi'_j + \phi_i \phi_j) dx = \int_0^1 f \phi_i dx + b \phi_i(1), \quad 1 \leq i \leq N. \quad (3.6)$$

Let us find the coefficient matrix of the system (3.6) in the case of a uniform partition, i.e., $h_1 = \dots = h_N = h$. The following formulas are useful for this purpose.

$$\begin{aligned} \int_0^1 \phi'_i \phi'_{i-1} dx &= \frac{-1}{h}, \quad 2 \leq i \leq N, \\ \int_0^1 (\phi'_i)^2 dx &= \frac{2}{h}, \quad 1 \leq i \leq N-1, \\ \int_0^1 \phi_i \phi_{i-1} dx &= \frac{h}{6}, \quad 2 \leq i \leq N, \\ \int_0^1 (\phi_i)^2 dx &= \frac{2h}{3}, \quad 1 \leq i \leq N-1, \\ \int_0^1 (\phi'_N)^2 dx &= \frac{1}{h}, \\ \int_0^1 (\phi_N)^2 dx &= \frac{h}{3}. \end{aligned}$$

We see that in matrix/vector notation, in the case of a uniform partition, the finite element system (3.6) can be written as

$$A u = b,$$

where

$$u = (u_1, \dots, u_N)^T$$

is the unknown vector,

$$A = \begin{pmatrix} \frac{2}{3}h + \frac{2}{h} & \frac{h}{6} - \frac{1}{h} & & \\ \frac{h}{6} - \frac{1}{h} & \frac{2}{3}h + \frac{2}{h} & \frac{h}{6} - \frac{1}{h} & \\ \ddots & \ddots & \ddots & \\ & \frac{h}{6} - \frac{1}{h} & \frac{2}{3}h + \frac{2}{h} & \frac{h}{6} - \frac{1}{h} \\ & & \frac{h}{6} - \frac{1}{h} & \frac{2}{3}h + \frac{2}{h} \end{pmatrix}_{N \times N} \quad (3.7)$$

is the stiffness matrix, and

$$b = \left(\int_0^1 f \phi_1 dx, \dots, \int_0^1 f \phi_{N-1} dx, \int_0^1 f \phi_N dx + b \right)^T$$

is the load vector. The matrix A is sparse, thanks to the small supports of the basis functions. One distinguished feature of the finite element method is that the basis functions are constructed in such a way that their supports are as small as possible, so that the corresponding stiffness matrix is as sparse as possible.

Remark 3.1.1. *The matrix A is said to be sparse, if most of its entries are zero; otherwise the matrix is said to be dense. Sparseness of the matrix can be utilized for two purposes. First, the stiffness matrix is less costly to form (observing that the computation of each entry of the matrix involves a domain integration and sometimes a boundary integration as well). Second, if the coefficient matrix is sparse, then the linear system can usually be solved more efficiently.*

3.1.2 High order elements and the condensation technique

We still consider the finite element method for solving the boundary value problem (3.1). This time we use piecewise quadratic functions. So the finite element space is

$$V_h = \{v_h \in V \mid v_h|_{K_i} \text{ is quadratic}\}$$

Equivalently,

$$V_h = \{v_h \in C(\overline{\Omega}) \mid v_h|_{K_i} \text{ is quadratic, } v_h(0) = 0\}.$$

Let us introduce a basis for the space v_h . We denote the mid-points of the subintervals by $x_{i-1/2} = (x_{i-1} + x_i)/2$, $1 \leq i \leq N$. Associated with each node x_i , $1 \leq i \leq N-1$, we define

$$\phi_i(x) = \begin{cases} 2(x - x_{i-1})(x - x_{i-1/2})/h_i^2, & x \in [x_{i-1}, x_i], \\ 2(x_{i+1} - x)(x_{i+1/2} - x)/h_{i+1}^2, & x \in [x_i, x_{i+1}], \\ 0, & \text{otherwise.} \end{cases} \quad (3.8)$$

Associated with x_N , we define

$$\phi_N(x) = \begin{cases} 2(x - x_{N-1})(x - x_{N-1/2})/h_N^2, & x \in [x_{N-1}, x_N], \\ 0, & \text{otherwise.} \end{cases} \quad (3.9)$$

We also need basis functions associated with the mid-points $x_{i-1/2}$, $1 \leq i \leq N$,

$$\psi_{i-1/2} = \begin{cases} 4(x_i - x)(x - x_{i-1})/h_i^2, & x \in [x_{i-1}, x_i], \\ 0, & \text{otherwise.} \end{cases} \quad (3.10)$$

We notice that a mid-point basis function is non-zero only in one element. Now the finite element space can be represented as

$$V_h = \text{span}\{\phi_i, \psi_{i-1/2} \mid 1 \leq i \leq N\},$$

and we write

$$u_h = \sum_{j=1}^N u_j \phi_j + \sum_{j=1}^N u_{j-1/2} \psi_{j-1/2}.$$

The finite element system

$$\begin{cases} a(u_h, \phi_i) = \ell(\phi_i) & 1 \leq i \leq N, \\ a(u_h, \psi_{i-1/2}) = \ell(\psi_{i-1/2}) & 1 \leq i \leq N. \end{cases}$$

can be written as, in the matrix/vector notation,

$$M_{11}u + M_{12}\tilde{u} = b_1, \quad (3.11)$$

$$M_{21}u + M_{22}\tilde{u} = b_2. \quad (3.12)$$

Here, $u = (u_1, \dots, u_N)^T$, $\tilde{u} = (u_{1/2}, \dots, u_{N-1/2})^T$, $M_{11} = (a(\phi_j, \phi_i))_{N \times N}$ is a tridiagonal matrix, $M_{12} = (a(\psi_{j-1/2}, \phi_i))_{N \times N}$ is a matrix with two diagonals, $M_{21} = M_{12}^T$, and $D_{22} = (a(\psi_{j-1/2}, \psi_{i-1/2}))_{N \times N}$ is a diagonal matrix with positive diagonal elements. We can eliminate $\tilde{u}$ from the system (3.11)-(3.12) easily. From (3.12), we have

$$\tilde{u} = D_{22}^{-1}(b_2 - M_{21}u).$$

This relation is substituted into (3.11),

$$Mu = b, \quad (3.13)$$

where $M = M_{11} - M_{12}D_{22}^{-1}M_{21}$ is a tridiagonal matrix, $b = b_1 - M_{12}D_{22}^{-1}b_2$. It can be shown that M is positive definite. As a result we see that for the finite element solution with quadratic elements, we only need to solve a tridiagonal system of order N , just like in the case of using linear elements in Subsection 3.1. The procedure of eliminating $\tilde{u}$ from (3.11)-(3.12) to form a smaller size system (3.13) is called condensation.

This condensation technique is especially useful in using high order elements to solve higher dimensional problems.

3.1.3 Reference element technique

Here we introduce the reference element technique which plays an important role for higher dimensional problems.

Consider a clamped beam, initially occupying the region $[0, 1]$, which is subject to the action of a transversal force of density f . Denote u the deflection of the beam. Then the boundary value problem is

$$\begin{cases} u^{(4)} = f & \text{in } \Omega = (0, 1), \\ u(0) = u'(0) = u(1) = u'(1) = 0. \end{cases} \quad (3.14)$$

The weak formulation of the problem is

$$u \in V, \quad \int_0^1 u''v'' dx = \int_0^1 fv dx \quad \forall v \in V, \quad (3.15)$$

where $V = H_0^2(0, 1)$. If we use the conforming finite element method, i.e., choose the finite element space v_h to be a subspace of V , then any function in V_h must be C^1 continuous. Suppose V_h consists of piecewise polynomials of degree less than or equal to p . The requirement that a finite element function be C^1 is equivalent

to the C^1 continuity of the function across the interior nodal points $\{x_i\}_{i=1}^{N-1}$, which places $2(N-1)$ constraints. Additionally, the Dirichlet boundary conditions impose 4 constraints. Hence,

$$\dim(V_h) = (p+1)N - 2(N-1) - 4 = (p-1)N - 2.$$

Now it is evident that the polynomial degree p must be at least 2. However, with $p = 2$, we cannot construct basis functions with small supports. Thus we should choose p to be at least 3. For $p = 3$, our finite element space is taken to be

$$V_h = \{v_h \in C^1(\bar{\Omega}) \mid v_h|_{K_i} \in \mathbb{P}_3(K_i), 1 \leq i \leq N, v_h(x) = v'_h(x) = 0 \text{ at } x = 0, 1\}.$$

It is then possible to construct basis functions with small supports using interpolation conditions of the function and its first derivative at the interior nodes $\{x_i\}_{i=1}^{N-1}$. More precisely, associated with each interior node x_i , there are two basis functions ϕ_i and ψ_i satisfying the interpolation conditions

$$\phi_i(x_j) = \delta_{ij}, \quad \phi'_i(x_j) = 0$$

$$\psi_i(x_j) = 0, \quad \psi'_i(x_j) = \delta_{ij}$$

A more plausible approach to constructing the basis functions is to use the reference element technique. To this end, let us choose $\hat{K} = [0, 1]$ as the reference element. Then the mapping

$$F_i : \hat{K} \rightarrow K_i, \quad F_i(\hat{x}) = x_{i-1} + h_i \hat{x}$$

is a bijection between $\hat{K}$ and K_i . Over the reference element $\hat{K}$, we construct cubic functions Φ_0, Φ_1, Ψ_0 and Ψ_1 satisfying the interpolation conditions

$$\begin{aligned} \Phi_0(0) &= 1, & \Phi_0(1) &= 0, & \Phi'_0(0) &= 0, & \Phi'_0(1) &= 0, \\ \Phi_1(0) &= 0, & \Phi_1(1) &= 1, & \Phi'_1(0) &= 0, & \Phi'_1(1) &= 0, \\ \Psi_0(0) &= 0, & \Psi_0(1) &= 0, & \Psi'_0(0) &= 1, & \Psi'_0(1) &= 0, \\ \Psi_1(0) &= 0, & \Psi_1(1) &= 0, & \Psi'_1(0) &= 0, & \Psi'_1(1) &= 1. \end{aligned}$$

It is not difficult to find these functions,

$$\begin{aligned} \Phi_0(\hat{x}) &= (1 + 2\hat{x})(1 - \hat{x})^2, \\ \Phi_1(\hat{x}) &= (3 - 2\hat{x})\hat{x}^2, \\ \Psi_0(\hat{x}) &= \hat{x}(1 - \hat{x})^2, \\ \Psi_1(\hat{x}) &= -(1 - \hat{x})\hat{x}^2. \end{aligned}$$

These functions, defined on the reference element, are called shape functions. With the shape functions, it is an easy matter to construct the basis functions with the aid of the mapping functions $\{F_i\}_{i=1}^{N-1}$. We have

$$\phi_i(x) = \begin{cases} \Phi_1(F_i^{-1}(x)), & x \in K_i, \\ \Phi_0(F_{i+1}^{-1}(x)), & x \in K_{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$

and

$$\psi_i(x) = \begin{cases} h_i \Psi_1(F_i^{-1}(x)), & x \in K_i, \\ h_{i+1} \Psi_0(F_{i+1}^{-1}(x)), & x \in K_{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$

Once the basis functions are available, it is a routine work to form the finite element system. We emphasize that the computations of the stiffness matrix and the load are done on the reference element. For example, by definition,

$$a_{i-1,i} = \int_0^1 (\phi_{i-1})''(\phi_i)'' dx = \int_{K_i} (\phi_{i-1})''(\phi_i)'' dx;$$

using the mapping function F_i and the definition of the basis functions, we have

$$\begin{aligned} a_{i-1,i} &= \int_{\hat{K}} (\Phi_0)'' h_i^{-2} (\Phi_1)'' h_i^{-2} h_i d\hat{x} \\ &= \frac{1}{h_i^3} \int_{\hat{K}} 6(2\hat{x} - 1)6(1 - 2\hat{x}) d\hat{x} \\ &= -\frac{12}{h_i^3} \end{aligned}$$

Remark 3.1.2. *For higher dimensional problems, the use of the reference element technique is essential for both theoretical error analysis and practical implementation of the finite element method. The computations of the stiffness matrix and the load vector involve a large number of integrals which usually cannot be computed analytically. With the reference element technique, all the integrals are computed on a single region the reference element, and therefore numerical quadratures are needed on the reference element only.*

3.2 Basics in finite element method

In this section, we discuss basic aspects of finite elements. We restrict our discussion to two-dimensional problems; most of the discussion can be extended to higher-dimensional problems straightforwardly. We will assume the domain Ω is a polygon so that it can be partitioned into straight-sided triangles and quadrilaterals. When Ω is a general domain with a curved boundary, it cannot be partitioned into straight-sided triangles and quadrilaterals, and usually curved-sided elements need to be used. We will emphasize the use of the reference element technique to estimate the error; for this reason, we need a particular structure on the finite elements, namely, we will consider only affine families of finite elements. In general, bilinear functions are needed for a bijective mapping between a four-sided reference element and a general quadrilateral. So a further restriction is that we will mostly use triangular elements, for any triangle is affine equivalent to a fixed triangle the reference triangle. First, we need a triangulation of the domain Ω into subsets. Here we consider triangular subsets. We say $\mathcal{T}_h = \{K\}$ is a triangulation, a mesh, or a partition of the domain Ω into triangular elements if the following properties hold:

1. $\bar{\Omega} = \bigcup_{K \in \mathcal{T}_h} K$
2. Each K is a triangle.
3. For distinct $K_1, K_2 \in \mathcal{T}_h$, $\overset{\circ}{K}_1 \cap \overset{\circ}{K}_2 = \emptyset$
4. For distinct $K_1, K_2 \in \mathcal{T}_h$, $K_1 \cap K_2$ is empty, or a common vertex, or a common side of K_1 and K_2 .

The second property is introduced just for convenience in the following discussion; rectangular and general quadrilateral elements are also widely used. In the third property, $\overset{\circ}{K}$ denotes the interior of a set K . The fourth property is called a regularity condition. Each $K \in \mathcal{T}_h$ is called an element. For a triangulation of a three-dimensional domain into tetrahedral, hexahedral or pentahedral elements, the regularity condition requires that the intersection of two distinct elements is empty, or a common vertex, or a common side, or a common face of the two elements. For an arbitrary element K , we denote

$$h_K = \text{diam}(K) = \max\{\|x - y\| \mid x, y \in K\}$$

and

$$\rho_K = \text{diameter of the largest sphere inscribed in } K.$$

Since K is a triangular element, h_K is the length of the longest side. The quantity h_K describes the size of K , while the ratio h_K/ρ_K is an indication whether the element is flat. We denote $h = \max_{K \in \mathcal{T}_h} h_K$.

3.2.1 Continuous linear elements

Suppose the boundary value problem over Ω under consideration is of second order. Then the Sobolev space $H^1(\Omega)$ or its subsets are used in the weak formulation. As an example, we consider solving the following Neumann boundary value problem:

$$-\Delta u + u = f \quad \text{in } \Omega, \quad (3.16)$$

$$\frac{\partial u}{\partial \nu} = g \quad \text{on } \Gamma, \quad (3.17)$$

where $f \in L_2(\Omega)$ and $g \in L_2(\Gamma)$ are given. The weak formulation of the boundary value problem is

$$u \in V, \quad \int_{\Omega} (\nabla u \cdot \nabla v + uv) dx = \int_{\Omega} f v dx + \int_{\Gamma} g v ds, \quad \forall v \in V. \quad (3.18)$$

where the space $V = H^1(\Omega)$. We construct a linear element space V_h of continuous piecewise linear functions for V :

$$V_h = \{v_h \in V \mid v_h|_K \in \mathbb{P}_1 \quad \forall K \in \mathcal{T}_h\} \quad (3.19)$$

Since the restriction of v_h on each element K is smooth, a necessary and sufficient condition for $v_h \in H^1(\Omega)$ is $v_h \in C(\overline{\Omega})$. So equivalently,

$$V_h = \{v_h \in C(\overline{\Omega}) \mid v_h|_K \in \mathbb{P}_1 \quad \forall K \in \mathcal{T}_h\} \quad (3.20)$$

Let v_h be a piecewise linear function. Then $v_h \in C(\overline{\Omega})$ if and only if

$$v_h|_{K_1} = v_h|_{K_2} \quad \text{on } K_1 \cap K_2 \quad \forall K_1, K_2 \in \mathcal{T}_h \text{ with } K_1 \cap K_2 \neq \emptyset \quad (3.21)$$

For (3.21) to hold, it is natural to define $v_h|_K$ by its values at the three vertices of K , for any $K \in \mathcal{T}_h$. Thus, let us determine any function $v_h \in V_h$ by its values at all the vertices. Let $\{x_i\}_{i=1}^{N_h} \subset \overline{\Omega}$ be the set of all the vertices of the elements $K \in \mathcal{T}_h$. Then

a basis of the space V_h is $\{\phi_i\}_{1 \leq i \leq N_h}$ where the basis function ϕ_i is associated with the vertex x_i , i.e., it satisfies the following conditions:

$$\phi_i \in V_h, \quad \phi_i(x_j) = \delta_{ij} \quad 1 \leq i, j \leq N_h. \quad (3.22)$$

We can then write

$$V_h = \text{span}\{\phi_i\}_{1 \leq i \leq N_h} \quad (3.23)$$

For any element $K \in \mathcal{T}_h$ containing x_i as a vertex, ϕ_i is a linear function on K , whereas if x_i is not a vertex of K , then $\phi_i(x) = 0$ for $x \in K$. If we write

$$v_h = \sum_{i=1}^N v_i \phi_i, \quad (3.24)$$

then using the property (3.22) we obtain

$$v_i = v_h(x_i) \quad 1 \leq i \leq N_h \quad (3.25)$$

We say the vertices $\{x_i\}_{1 \leq i \leq N_h}$ are the nodes of the linear element space V_h defined in (3.19) or (3.20). The finite element method for (3.18) is

$$u_h \in V_h, \quad \int_{\Omega} (\nabla u_h \cdot \nabla v_h + u_h v_h) dx = \int_{\Omega} f v_h dx + \int_{\Gamma} g v_h ds \quad \forall v_h \in V_h \quad (3.26)$$

Write the finite element solution as

$$u_h = \sum_{j=1}^{N_h} u_j \phi_j$$

Then from (3.26), we have the following discrete system for the coefficients $\{u_j\}_{1 \leq j \leq N_h}$:

$$\sum_{j=1}^{N_h} u_j \int_{\Omega} (\nabla \phi_j \cdot \nabla \phi_i + \phi_j \phi_i) dx = \int_{\Omega} f \phi_i dx + \int_{\Gamma} g \phi_i ds \quad 1 \leq i \leq N_h \quad (3.27)$$

We now discuss construction of the basis functions $\{\phi_i\}_{1 \leq i \leq N_h}$. For this purpose, it is convenient to use barycentric coordinates. Let K be a triangle and denote its three vertices by $a_j = (a_{1j}, a_{2j})^T, 1 \leq j \leq 3$, numbered counter-clockwise. We define the barycentric coordinates $\{\lambda_j(x)\}_{1 \leq j \leq 3}$, associated with the vertices $\{a_j\}_{1 \leq j \leq 3}$, to be the affine functions satisfying

$$\lambda_j(a_i) = \delta_{ij}, \quad 1 \leq i, j \leq 3 \quad (3.28)$$

Then by the uniqueness of linear interpolation, we have

$$\sum_{i=1}^3 \lambda_i(x) a_i = x \quad (3.29)$$

$$\sum_{i=1}^3 \lambda_i(x) = 1 \quad (3.30)$$

The equations (3.29) and (3.30) constitute a linear system for the three unknowns $\{\lambda_i\}_{i=1}^3$,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad (3.31)$$

Remark 3.2.1. The determinant of the coefficient matrix A of this system is twice the area of the triangle K . Since the triangle K is non-degenerate (i.e., the interior of K is nonempty), the matrix A is non-singular and so the system (3.31) uniquely determines the barycentric coordinates $\{\lambda_i\}_{i=1}^3$.

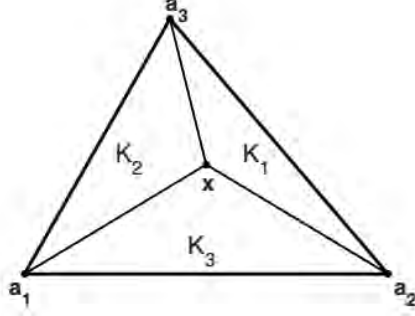


Figure 3.2: Interpretation of barycentric coordinates

By Cramers rule, we have the following formulas:

$$\begin{aligned}\lambda_1(x) &= \frac{1}{\det A} \begin{vmatrix} x_1 & a_{12} & a_{13} \\ x_2 & a_{22} & a_{23} \\ 1 & 1 & 1 \end{vmatrix}, \\ \lambda_2(x) &= \frac{1}{\det A} \begin{vmatrix} a_{11} & x_1 & a_{13} \\ a_{21} & x_2 & a_{23} \\ 1 & 1 & 1 \end{vmatrix}, \\ \lambda_3(x) &= \frac{1}{\det A} \begin{vmatrix} a_{11} & a_{12} & x_1 \\ a_{21} & a_{22} & x_2 \\ 1 & 1 & 1 \end{vmatrix}.\end{aligned}$$

For $x \in K$, let K_1 be the triangle with the vertices x, a_2, a_3 , and similarly define two other triangles K_2 and K_3 as shown in Figure (3.2). Then from the above formulas, we get the following geometric interpretation of the barycentric coordinates:

$$\lambda_i(x) = \frac{\text{area}(K_i)}{\text{area}(K)}, \quad 1 \leq i \leq 3, \quad x \in K.$$

Since $\{\lambda_i\}_{i=1}^3$ are affine functions and $\lambda_i(a_j) = \delta_{ij}$, the following relation holds:

$$\lambda_i(t_1 a_1 + t_2 a_2 + (1 - t_1 - t_2) a_3) = t_1 \delta_{i1} + t_2 \delta_{i2} + (1 - t_1 - t_2) \delta_{i3}, \quad 1 \leq i \leq 3.$$

This formula is useful in computing the barycentric coordinates of points on the plane. In particular, λ_1 is a constant along any line parallel to the side $\overline{a_2 a_3}$, and on such a parallel line, the value of λ_1 is proportional to the signed scaled distance of the line containing $\overline{a_2 a_3}$ (the sign is positive for points on the same side as a_1 of the line containing $\overline{a_2 a_3}$, and negative on the other side). See Figure 3.3, where $a_{12} = (a_1 + a_2)/2$ and $a_{31} = (a_3 + a_1)/2$ are side mid-points.

A basis function ϕ_i is non-zero on an element $K \in \mathcal{T}_h$ if and only if the associated node x_i is a vertex of K . Without loss of generality, let a_1 be the node associated with ϕ_i . Then on K , $\phi_i(x) = \lambda_1(x)$. Thus, with the barycentric coordinates, it is easy

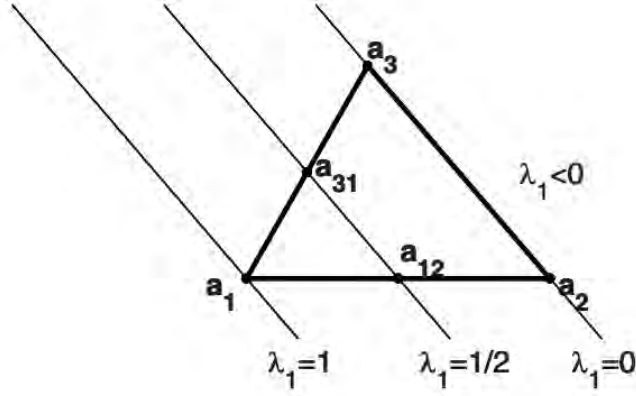


Figure 3.3: Value of barycentric coordinate λ_1

to compute the basis functions $\{\phi_i\}_{1 \leq i \leq N_h}$ defined in (3.22). Also, note that over the element K , the function v_h of (3.24) is given by the simple formula

$$v_h = \sum_{i=1}^3 v_h(a_i) \lambda_i. \quad (3.32)$$

Back to the finite element system (3.27), we notice that for general functions f and g , the right hand sides can be computed only through numerical integrations. As an example, consider the calculation of the integral $\int_{\Omega} f \phi_i dx$, which is the summation of elemental level integrals of the form $\int_K f \phi_i dx$ where ϕ_i is non-zero on K . Rather than doing numerical integration on each element, we introduce a change of variables so that all the elemental level integrals are calculated over a fixed triangle $\hat{K}$, called a reference element. In this way, we only need numerical quadratures over $\hat{K}$ in computing integrals of the form $\int_{\Omega} f \phi_i dx$. The reference element for triangular elements is usually taken to be either an equilateral or right isosceles triangle, as shown in Figure 3.4.

For the equilateral triangular reference element with $\hat{a}_1 = (-1, 0)$, $\hat{a}_2 = (1, 0)$, and $\hat{a}_3 = (0, \sqrt{3})$, the barycentric coordinates are

$$\begin{aligned} \hat{\lambda}_1(\hat{x}) &= \frac{1}{2} \left(1 - \hat{x}_1 - \frac{\hat{x}_2}{\sqrt{3}} \right), \\ \hat{\lambda}_2(\hat{x}) &= \frac{1}{2} \left(1 + \hat{x}_1 - \frac{\hat{x}_2}{\sqrt{3}} \right), \\ \hat{\lambda}_3(\hat{x}) &= \frac{\hat{x}_2}{\sqrt{3}}. \end{aligned}$$

For the right isosceles triangular reference element with $\hat{a}_1 = (0, 0)$, $\hat{a}_2 = (1, 0)$, and $\hat{a}_3 = (0, 1)$, the barycentric coordinates are

$$\hat{\lambda}_1(\hat{x}) = 1 - \hat{x}_1 - \hat{x}_2, \quad \hat{\lambda}_2(\hat{x}) = \hat{x}_1, \quad \hat{\lambda}_3(\hat{x}) = \hat{x}_2$$

For definiteness, let us choose the right isosceles triangle reference element in the following discussion. We construct an affine mapping F_K from $\hat{K}$ to K such that $a_i = F_K(\hat{a}_i)$, $1 \leq i \leq 3$. It is easy to verify that

$$x = F_K(\hat{x}) = a_1 + B_K \hat{x}, \quad (3.33)$$

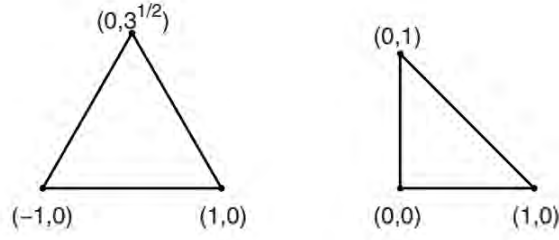


Figure 3.4: Reference triangular elements in $\mathbb{R}^2$

where the matrix

$$B_K = \begin{pmatrix} a_{12} - a_{11} & a_{13} - a_{11} \\ a_{22} - a_{21} & a_{23} - a_{21} \end{pmatrix} \quad (3.34)$$

From now on, we will relate $x \in K$ and $\hat{x} \in \hat{K}$ by the relation (3.33). Moreover, we will relate a function v defined on K with a function $\hat{v}$ defined on $\hat{K}$ by the relation

$$v(x) = \hat{v}(\hat{x}) \quad (3.35)$$

An integral over an element K can be transformed to one over the reference element $\hat{K}$ as follows:

$$\int_K v(x) dx = \det(B_K) \int_{\hat{K}} v(F_K(\hat{x})) d\hat{x}, \quad (3.36)$$

which is then approximated by applying a numerical quadrature over the fixed region $\hat{K}$. The boundary integral term in (3.27) can be handled similarly, and the calculations are done on the sides of the reference element $\hat{K}$. Actually, the finite element space (3.23) can be constructed from a single function space over $\hat{K}$ together with the mappings (3.33). We use the symbol $\hat{X}$ for the set of functions $\hat{v} \in \mathbb{P}_1(\hat{K})$ that are determined by their values at the three vertices $\hat{a}_1, \hat{a}_2$, and $\hat{a}_3$. Introduce the basis functions $\hat{\phi}_i(\hat{x}) = \hat{\lambda}_i(\hat{x}), 1 \leq i \leq 3$, associated with the vertices. Any $\hat{v} \in \hat{X}$ can be expressed as

$$\hat{v}(\hat{x}) = \sum_{i=1}^3 \hat{v}(\hat{a}_i) \hat{\phi}_i(\hat{x})$$

Now consider a general element K with the three vertices a_1, a_2 , and a_3 . Then for any $v_h \in V_h$, $v_h|_K \in \mathbb{P}_1(K)$ can be written as

$$v_h(x) = \sum_{i=1}^3 v_h(a_i) \lambda_i(x) = \sum_{i=1}^3 v_h(a_i) \hat{\phi}_i(\hat{x}), \quad x \in K.$$

In this way, the restriction of any function $v_h \in V_h$ on any element K can be obtained from a corresponding function in $\hat{X}$. Since numerical integrations in constructing finite element systems are done on the reference element, the finite element method is completely determined by the space $\hat{X}$ over the reference element $\hat{K}$ and the mapping functions $\{F_K\}_{K \in \mathcal{T}_h}$. Consequently, it is actually not necessary to construct the basis functions $\{\phi_i\}_{1 \leq i \leq N_h}$ of the finite element space V_h .

3.2.2 Affine-equivalent finite elements

Given finite element partitions of the polygon $\bar{\Omega}$, we now consider construction of finite element spaces based on affine-equivalent finite elements. We introduce a function space $\hat{X}$ over the reference element $\hat{K}$, including a description on how a function in $\hat{X}$ is determined, and then construct a corresponding function space on a general element K by using the mapping function (3.33). Although it is possible to choose any finite dimensional function space as $\hat{X}$, the overwhelming choice for $\hat{X}$ for practical use is a polynomial space. It is convenient to use the barycentric coordinates to represent polynomials. For example, if $\hat{X}_1$ consists of functions from $\mathbb{P}_1(\hat{K})$ that are determined by their values at the vertices $\{\hat{a}_i\}_{i=1}^3$. Then for any $\hat{v} \in \hat{X}_1$, we have the representation

$$\hat{v}(\hat{x}) = \sum_{i=1}^3 \hat{v}(\hat{a}_i) \hat{\lambda}_i(\hat{x}).$$

In this case, the vertices $\{\hat{a}_i\}_{i=1}^3$ are called the nodes, the function values $\{\hat{v}(\hat{a}_i)\}_{i=1}^3$ are called the parameters (used to determine the linear function). A quadratic function has six coefficients and we need six interpolation conditions to determine it. For this, we introduce the side mid-points,

$$\hat{a}_{ij} = \frac{1}{2}(\hat{a}_i + \hat{a}_j), \quad 1 \leq i < j \leq 3$$

Then we introduce the space $\hat{X}_2$ of $\mathbb{P}_2(\hat{K})$ functions that are determined by their values at the vertices $\{\hat{a}_i\}_{i=1}^3$ and the side mid-points $\{\hat{a}_{ij}\}_{1 \leq i < j \leq 3}$. For any $\hat{v} \in \hat{X}_2$, we have the representation formula

$$\hat{v}(\hat{x}) = \sum_{i=1}^3 \hat{v}(\hat{a}_i) \hat{\lambda}_i(\hat{x})(2\hat{\lambda}_i(\hat{x}) - 1) + \sum_{1 \leq i < j \leq 3} 4\hat{v}(\hat{a}_{ij}) \hat{\lambda}_i(\hat{x}) \hat{\lambda}_j(\hat{x}) \quad (3.37)$$

This formula is derived from the observations that

- (1) for each i , $1 \leq i \leq 3$, $\hat{\lambda}_i(\hat{x})(2\hat{\lambda}_i(\hat{x}) - 1)$ is a quadratic function that takes on the value 1 at $\hat{a}_i$, and the value 0 at the other vertices and the side mid-points;
- (2) for $1 \leq i < j \leq 3$, $4\hat{\lambda}_i(\hat{x})\hat{\lambda}_j(\hat{x})$ is a quadratic function that takes on the value 1 at $\hat{a}_{ij}$, and the value 0 at the other side mid-points and the vertices. In this case, the vertices and the side mid-points are called the nodes, the function values at the nodes are called the parameters (used to determine the quadratic function).

For a positive integer k , let $\mathbb{P}_k(\hat{K})$ be the space of polynomials of degree less than or equal to k on $\hat{K}$. Note that the dimension of the space is $\dim \mathbb{P}_k(\hat{K}) = (k+2)(k+1)/2$. To uniquely determine a $\hat{v} \in \mathbb{P}_k(\hat{K})$, we can use the values of v at the following nodal points:

$$\hat{N}_k = \left\{ \sum_{i=1}^3 t_i \hat{a}_i \mid \sum_{i=1}^3 t_i = 1, t_i \in \left\{ \frac{j}{k} \right\}_{0 \leq j \leq k}, 1 \leq i \leq 3 \right\} \quad (3.38)$$

In general, let $\hat{X}$ be a finite dimensional space over $\hat{K}$, with a dimension $\dim \hat{X} = I$, such that any function $\hat{v} \in \hat{X}$ is uniquely determined by its values at the I nodes

$\hat{x}_1, \dots, \hat{x}_I \in \hat{K}$. Then we have the following formula

$$\hat{v}(\hat{x}) = \sum_{i=1}^I \hat{v}(\hat{x}_i) \hat{\phi}_i(\hat{x})$$

The functions $\{\hat{\phi}_i\}_{i=1}^I$ form a basis for the space $\hat{X}$ with the property

$$\hat{\phi}_i(\hat{x}_j) = \delta_{ij}$$

We will then define function spaces over a general element. For the affine equivalent families of finite elements, there exists one or several reference elements, $\hat{K}$, such that each element K is the image of $\hat{K}$ under an invertible affine mapping $F_K : \hat{K} \rightarrow K$ of the form

$$F_K(\hat{x}) = T_K(\hat{x}) + b_K \quad (3.39)$$

The mapping F_K is a bijection between $\hat{K}$ and K , T_K is an invertible 2×2 matrix and b_K is a translation vector. For each element K , we establish a correspondence between functions defined on K and $\hat{K}$ through the use of the affine mapping F_K . For any function v defined on K , $\hat{v}$ denotes the corresponding function defined on $\hat{K}$ through $\hat{v} = v \circ F_K$. Conversely, for any function $\hat{v}$ on $\hat{K}$, we let v be the function on K defined by $v = \hat{v} \circ F_K^{-1}$. Thus we have the relation

$$v(x) = \hat{v}(\hat{x}) \quad \forall x \in K, \hat{x} \in \hat{K}, \text{ with } x = F_K(\hat{x})$$

We then define a finite dimensional function space X_K formally by the formula

$$X_K = \hat{X} \circ F_K^{-1} \quad (3.40)$$

where any function $v \in X_K$ corresponds to a function $\hat{v} \in \hat{X}$ with $v = \hat{v} \circ F_K^{-1}$. Implicit in the definition (3.40) is that functions in X_K are determined in the same way as functions in $\hat{X}$; e.g., if $\hat{X}$ consists of $\mathbb{P}_1(\hat{K})$ functions determined by their values at the three vertices of $\hat{K}$, then X_K consists of $\mathbb{P}_1(K)$ functions determined by their values at the three vertices of K . An immediate consequence of this definition is that if $\hat{X}$ contains polynomials of certain degree, then X_K contains polynomials of the same degree. Using the nodal points $\hat{x}_i, 1 \leq i \leq I$, of $\hat{K}$, we introduce the nodal points $x_i^K, 1 \leq i \leq I$, of K defined by

$$x_i^K = F_K(\hat{x}_i), \quad i = 1, \dots, I \quad (3.41)$$

Recall that $\{\hat{\phi}_i\}_{i=1}^I$ are the basis functions of the space $\hat{X}$ associated with the nodal points $\{\hat{x}_i\}_{i=1}^I$ with the property that

$$\hat{\phi}_i(\hat{x}_j) = \delta_{ij}$$

We define

$$\phi_i^K = \hat{\phi}_i \circ F_K^{-1}, \quad i = 1, \dots, I.$$

Then the functions $\{\phi_i^K\}_{i=1}^I$ have the property that

$$\phi_i^K(x_j^K) = \delta_{ij}$$

Hence, $\{\phi_i^K\}_{i=1}^I$ form a set of local basis functions on K .

Lemma 3.2.1. *For the affine map $F_K : \hat{K} \rightarrow K$ defined by (3.39), we have the bounds*

$$\|T_K\| \leq \frac{h_K}{\hat{\rho}} \quad \text{and} \quad \|T_K^{-1}\| \leq \frac{\hat{h}}{\rho_K}$$

Proof. By definition of the matrix norm,

$$\|T_K\| = \sup \left\{ \frac{\|T_K \hat{x}\|}{\|\hat{x}\|} \mid \hat{x} \neq 0 \right\}$$

Let us rewrite it in the equivalent form

$$\|T_K\| = \hat{\rho}^{-1} \sup \{ \|T_K \hat{z}\| \mid \|\hat{z}\| = \hat{\rho} \}$$

by taking $\hat{z} = \hat{\rho} \hat{x} / \|\hat{x}\|$. Now for any $\hat{z}$ with $\|\hat{z}\| = \hat{\rho}$, pick up any two vectors $\hat{x}$ and $\hat{y}$ that lie on the largest sphere $\hat{S}$ of diameter $\hat{\rho}$, which is inscribed in $\hat{K}$, such that $\hat{z} = \hat{x} - \hat{y}$. Then

$$\begin{aligned} \|T_K\| &= \hat{\rho}^{-1} \sup \left\{ \|T_K(\hat{x} - \hat{y})\| \mid \hat{x}, \hat{y} \in \hat{S} \right\} \\ &= \hat{\rho}^{-1} \sup \left\{ \|(T_K \hat{x} + b_K) - (T_K \hat{y} + b_K)\| \mid \hat{x}, \hat{y} \in \hat{S} \right\} \\ &\leq \hat{\rho}^{-1} \sup \{ \|x - y\| \mid x, y \in K \} \\ &\leq h_K / \hat{\rho} \end{aligned}$$

□

3.2.3 Finite element spaces

A global finite element function v_h is defined piecewise by the formula

$$v_h|_K \in X_K \quad \forall K \in \mathcal{T}_h$$

We then define a finite element space corresponding to the triangulation $\mathcal{T}_h$:

$$X_h = \{v_h \mid v_h|_K \in X_K, \forall K \in \mathcal{T}_h\}$$

Remark 3.2.2. *if $v_h \in H^1(\Omega)$ holds for $v_h \in X_h$. Since the restriction of v_h on each element K is a smooth function, a necessary and sufficient condition for $v_h \in H^1(\Omega)$ is $v_h \in C(\overline{\Omega})$. Then $X_h \subset H^1(\Omega)$ if and only if $X_h \subset C(\overline{\Omega})$. We remark that the condition $v_h \in C(\overline{\Omega})$ is guaranteed if v_h is continuous across any interelement boundary, i.e., if the condition (3.21) is satisfied.*

Assume $X_h \subset C(\overline{\Omega})$ is valid. Then for a second-order boundary value problem with Neumann boundary condition, $V = H^1(\Omega)$ and we use $V_h = X_h$ as the finite element space; for a second-order boundary value problem with the homogeneous Dirichlet condition, $V = H_0^1(\Omega)$ and we choose the finite element space to be

$$V_h = \{v_h \in X_h \mid v_h = 0 \text{ on } \Gamma\}$$

3.3 Error estimates of finite element interpolations

3.3.1 Local interpolations

We first introduce an interpolation operator $\hat{\Pi}$ for continuous functions on $\hat{K}$. Recall that $\{\hat{x}_i\}_{i=1}^I$ are the nodal points whereas $\{\hat{\phi}_i\}_{i=1}^I$ are the associated basis functions of the polynomial space $\hat{X}$ in the sense that $\hat{\phi}_i(\hat{x}_j) = \delta_{ij}$ is valid. We define

$$\hat{\Pi} : C(\hat{K}) \rightarrow \hat{X}, \quad \hat{\Pi}v = \sum_{i=1}^I v(\hat{x}_i) \hat{\phi}_i. \quad (3.42)$$

Evidently, $\hat{\Pi}v \in \hat{X}$ is uniquely determined by the interpolation conditions

$$\hat{\Pi}v(\hat{x}_i) = v(\hat{x}_i), \quad i = 1, \dots, I$$

On any element K , we define similarly the interpolation operator Π_K by

$$\Pi_K : C(K) \rightarrow X_K, \quad \Pi_K v = \sum_{i=1}^I v(x_i^K) \phi_i^K. \quad (3.43)$$

We see that $\Pi_K v \in X_K$ is uniquely determined by the interpolation conditions

$$\Pi_K v(x_i^K) = v(x_i^K), \quad i = 1, \dots, I$$

Theorem 3.3.1. *For the two interpolation operators $\hat{\Pi}$ and Π_K introduced above, we have $\hat{\Pi}(\hat{v}) = \Pi_K v \circ F_K$, i.e., $\hat{\Pi}\hat{v} = \widehat{\Pi_K v}$.*

Proof. From the definition (3.43), we have

$$\Pi_K v = \sum_{i=1}^I v(x_i^K) \phi_i^K = \sum_{i=1}^I \hat{v}(\hat{x}_i) \hat{\phi}_i^K$$

Since $\phi_i^K \circ F_K = \hat{\phi}_i$, we obtain

$$(\Pi_K v) \circ F_K = \sum_{i=1}^I \hat{v}(\hat{x}_i) \hat{\phi}_i = \hat{\Pi}\hat{v}.$$

□

3.3.2 Local interpolation error estimates

We now consider the finite element interpolation error over each element K .

Theorem 3.3.2. *Assume $x = T_K \hat{x} + b_K$ is a bijection from $\hat{K}$ to K . Then $v \in H^m(K)$ if and only if $\hat{v} \in H^m(\hat{K})$. Furthermore, for some constant c independent of K and $\hat{K}$, the estimates*

$$|\hat{v}|_{m, \hat{K}} \leq c \|T_K\|^m |\det T_K|^{-1/2} |v|_{m, K} \quad (3.44)$$

and

$$|v|_{m, K} \leq c \|T_K^{-1}\|^m |\det T_K|^{1/2} |\hat{v}|_{m, \hat{K}} \quad (3.45)$$

hold.

Proof. We only need to prove the inequality (3.44); the inequality (3.45) follows from (3.44) by interchanging the roles played by x and $\hat{x}$. Recall the multi-index notation: for $\alpha = (\alpha_1, \alpha_2)$,

$$\partial_{\hat{x}}^\alpha = \frac{\partial^{|\alpha|}}{\partial \hat{x}_1^{\alpha_1} \partial \hat{x}_2^{\alpha_2}}$$

,

$$\partial_x^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}}$$

By a change of variables, we have

$$\begin{aligned} |\hat{v}|_{m, \hat{K}}^2 &= \sum_{|\alpha|=m} \int_{\hat{K}} [\partial_{\hat{x}}^\alpha \hat{v}(\hat{x})]^2 d\hat{x} \\ &= \sum_{|\alpha|=m} \int_K [\partial_{\hat{x}}^\alpha \hat{v}(F_K^{-1})]^2 |\det T_K|^{-1} dx \end{aligned}$$

Since the mapping function is affine, for any multi-index α with $|\alpha| = m$, we have

$$\partial_{\hat{x}}^\alpha \hat{v} = \sum_{|\beta|=m} c_{\alpha, \beta}(T_K) \partial_x^\beta v$$

where each $c_{\alpha, \beta}(T_K)$ is a product of m entries of the matrix T_K . Thus

$$\sum_{|\alpha|=m} |\partial_{\hat{x}}^\alpha \hat{v}(F_K^{-1}(x))|^2 \leq c \|T_K\|^{2m} \sum_{|\alpha|=m} |\partial_x^\alpha v(x)|^2,$$

and so

$$\begin{aligned} |\hat{v}|_{m, \hat{K}}^2 &\leq c \sum_{|\alpha|=m} \int_K [\partial_x^\alpha v(x)]^2 \|T_K\|^{2m} (\det T_K)^{-1} dx \\ &= c \|T_K\|^{2m} (\det T_K)^{-1} |v|_{m, K}^2 \end{aligned}$$

from which the inequality (3.44) follows. $\square$

Theorem 3.3.3. *Let k and m be nonnegative integers with $k > 0$, $k + 1 \geq m$, and $\mathbb{P}_k(\hat{K}) \subset \hat{X}$. Let Π_K be the operators defined in (3.43). Then there is a constant c depending only on $\hat{K}$ and $\hat{\Pi}$ such that*

$$|v - \Pi_K v|_{m, K} \leq c \frac{h_K^{k+1}}{\rho_K^m} |v|_{k+1, K}, \quad \forall v \in H^{k+1}(K). \quad (3.46)$$

Proof. From Theorem 3.3.1 we have $\hat{v} - \hat{\Pi} \hat{v} = (v - \Pi_K v) \circ F_K$. Consequently, using (3.45) we obtain

$$|v - \Pi_K v|_{m, K} \leq c \|T_K^{-1}\|^m |\det T_K|^{1/2} |\hat{v} - \hat{\Pi} \hat{v}|_{m, \hat{K}}.$$

Using the estimate (??), we have

$$|v - \Pi_K v|_{m, K} \leq c \|T_K^{-1}\|^m |\det T_K|^{1/2} |\hat{v}|_{k+1, \hat{K}} \quad (3.47)$$

The inequality (3.44) with $m = k + 1$ is

$$|\hat{v}|_{k+1, \hat{K}} \leq c \|T_K\|^{k+1} |\det T_K|^{-1/2} |v|_{k+1, K}$$

So from (3.47), we obtain

$$|v - \Pi_K v|_{m, K} \leq c \|T_K^{-1}\|^m \|T_K\|^{k+1} |v|_{k+1, K}$$

The estimate (3.46) now follows from an application of Lemma(3.2.1). $\square$

The error estimate (3.46) is proved through the use of the reference element $\hat{K}$. The proof method can be termed the reference element technique. The error bound in (3.46) depends on two parameters h_K and ρ_K . It will be convenient to use the parameter h_K only in an interpolation error bound. For this purpose we introduce the notion of a regular family of finite elements. For a triangulation $\mathcal{T}_h$, we denote

$$h = \max_{K \in \mathcal{T}_h} h_K \quad (3.48)$$

often called the mesh parameter. The quantity h is a measure of how refined the mesh is. The smaller h is, the finer the mesh.

Definition 3.3.1. A family $\{\mathcal{T}_h\}_h$ of finite element partitions is said to be regular if

- (a) there exists a constant σ such that $h_K/\rho_K \leq \sigma$ for all elements $K \in \mathcal{T}_h$ and for any h ;
- (b) the mesh parameter h approaches zero.

Corollary 3.3.1. We keep the assumptions stated in Theorem 3.3.3. Furthermore, assume $\{\mathcal{T}_h\}_h$ is a regular family of finite elements. Then there is a constant c such that for any $\{\mathcal{T}_h\}$ in the family,

$$\|v - \Pi_K v\|_{m, K} \leq c h_K |v|_{k+1, K} \quad \forall v \in H^{k+1}(K) \quad \forall K \in \mathcal{T}_h$$

3.3.3 Global interpolation error estimates

We now estimate the finite element interpolation error of a continuous function over the entire domain Ω . For a function $v \in C(\overline{\Omega})$, we construct its global interpolant $\Pi_h v$ in the finite element space X_h by the formula

$$\Pi_h v|_K = \Pi_K v \quad \forall K \in \mathcal{T}_h$$

Let $\{x_i\}_{i=1}^{N_h} \subset \overline{\Omega}$ be the set of the nodes collected from the nodes of all the elements $K \in \mathcal{T}_h$. We have the representation formula

$$\Pi_h v = \sum_{i=1}^{N_h} v(x_i) \phi_i \quad (3.49)$$

for the global finite element interpolant. Here $\phi_i, i = 1, \dots, N_h$, are the global basis functions that span X_h . The basis function ϕ_i is associated with the node x_i , i.e., ϕ_i is a piecewise polynomial of degree less than or equal to k , and $\phi_i(x_j) = \delta_{ij}$. If the node

x_i is a vertex x_l^K of the element K , then $\phi_i|_K = \phi_l^K$. If x_i is not a node of K , then $\phi_i|_K = 0$. Thus the functions ϕ_i are constructed from local basis functions ϕ_i^K . In the context of the finite element approximation of a linear second order elliptic boundary value problem, there holds the Céa's inequality

$$\|u - u_h\|_{1,\Omega} \leq c \inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega}$$

Then

$$\|u - u_h\|_{1,\Omega} \leq c \|u - \Pi_h u\|_{1,\Omega}$$

and we need to find an estimate of the interpolation error $\|u - \Pi_h u\|_{1,\Omega}$

Theorem 3.3.4. *Assume that all the conditions of Corollary 3.3.1 hold. Then there exists a constant c independent of h such that*

$$\|v - \Pi_h v\|_{m,\Omega} \leq ch^{k+1-m} |v|_{k+1,m} \quad \forall v \in H^{k+1}(\Omega), m = 0, 1 \quad (3.50)$$

Proof. Since the finite element interpolant $\Pi_h u$ is defined piecewisely by $\Pi_h u|_K = \Pi_K u$, we can apply Corollary 3.3.1 with $m = 0$ and 1 to find

$$\begin{aligned} \|u - \Pi_h u\|_{m,\Omega}^2 &= \sum_{K \in \mathcal{T}_h} \|u - \Pi_K u\|_{m,K}^2 \\ &\leq \sum_{K \in \mathcal{T}_h} ch_K^{2(k+1-m)} |u|_{k+1,K}^2 \\ &\leq ch^{2(k+1-m)} |u|_{k+1,\Omega}^2 \end{aligned}$$

Taking the square root of the above relation, we obtain the error estimates(3.50) $\square$

3.4 Convergence and error estimates

As an example, we consider the convergence and error estimates for finite element approximations of a linear second-order elliptic problem over a polygonal domain. The function space V is a subspace of $H^1(\Omega)$; e.g., $V = H_0^1(\Omega)$ if the homogeneous Dirichlet condition is specified over the whole boundary, whereas $V = H^1(\Omega)$ if a Neumann condition is specified over the boundary. Let the weak formulation of the problem be

$$u \in V, \quad a(u, v) = \ell(v) \quad \forall v \in V \quad (3.51)$$

We assume all the assumptions required by the Lax-Milgram Lemma; then the problem (3.50) has a unique solution u . Let $V_h \subset V$ be a finite element space. Then the discrete problem

$$u_h \in V_h, \quad a(u_h, v_h) = \ell(v_h) \quad \forall v_h \in V_h \quad (3.52)$$

also has a unique solution $u_h \in V_h$ and there holds the Céas inequality

$$\|u - u_h\|_V \leq \inf_{v_h \in V_h} \|u - v_h\|_V \quad (3.53)$$

This inequality is a basis for convergence and error analysis.

Theorem 3.4.1. *We keep the assumptions mentioned above. Let $k > 0$ be an integer, and let $\{V_h\} \subset V$ be affine-equivalent finite element spaces of piecewise polynomials of degree less than or equal to k , corresponding to a regular family of triangulations of $\bar{\Omega}$. Then the finite element method converges:*

$$\|u - u_h\|_V \rightarrow 0 \quad \text{as } h \rightarrow 0$$

Assume $u \in H^{k+1}(\Omega)$. Then there exists a constant c such that the following error estimate holds:

$$\|u - u_h\|_{1,\Omega} \leq ch^k |u|_{k+1,\Omega} \quad (3.54)$$

Proof. We take $v_h = \Pi_h u$ in Céas inequality (3.53),

$$\|u - u_h\|_{1,\Omega} \leq c \|u - \Pi_h u\|_{1,\Omega}$$

Using the estimate (3.50) with $m = 1$, we obtain the error estimate (3.54). The convergence of the finite element solution under the basic solution regularity $u \in V$ follows from the facts that smooth functions are dense in the space V and for a smooth function, its finite element interpolants converge (with a convergence order k). $\square$

Example 3.4.1. *Consider the problem*

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \Gamma \end{aligned}$$

The corresponding variational formulation is: Find $u \in H_0^1(\Omega)$ such that

$$\int_{\Omega} \nabla u \cdot \nabla v = \int_{\Omega} f v \, dx \quad \forall v \in H_0^1(\Omega)$$

This problem has a unique solution. Similarly, the discrete problem of finding $u_h \in V_h$ such that

$$\int_{\Omega} \nabla u_h \cdot \nabla v_h = \int_{\Omega} f v_h \, dx \quad \forall v_h \in V_h$$

has a unique solution. Here V_h is a finite element subspace of $H_0^1(\Omega)$, consisting of piecewise polynomials of degree less than or equal to k , corresponding to a regular triangulation of $\bar{\Omega}$. If $u \in H^{k+1}(\Omega)$, then the error is estimated by

$$\|u - u_h\|_{1,\Omega} \leq ch^k \|u\|_{k+1,\Omega}.$$

3.5 Application of Finite Element Method for two dimensional problem

[7]The Poisson equation is given by

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y) \text{ or } \Delta u = f(x, y) \quad (3.55)$$

Here the Laplace operator denotes $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$. Equation (3.55) is elliptic second order linear partial differential equation. In (3.55), the dependent variable $u(x, y)$ is a function of its arguments and depends on the independent variables x and y . The function $f(x, y)$ is known as a source function. For the purpose of numerical solution of the Poisson equation (3.55) we consider a plane region defined by $R = \{(x, y) : 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1\}$. We will also impose Dirichlet boundary conditions for the dependent variable $u(x, y)$ on the region R defined as

$$u(0, y) = 0, u(x, 0) = 0, u(x, 1) = 0, u(1, y) = 0 \quad (3.56)$$

The region R together with the Dirichlet boundary conditions (3.56) is illustrated in Figure 3.5. The x -variable increases along the positive horizontal axis while the y -variable increases along the positive vertical axis.

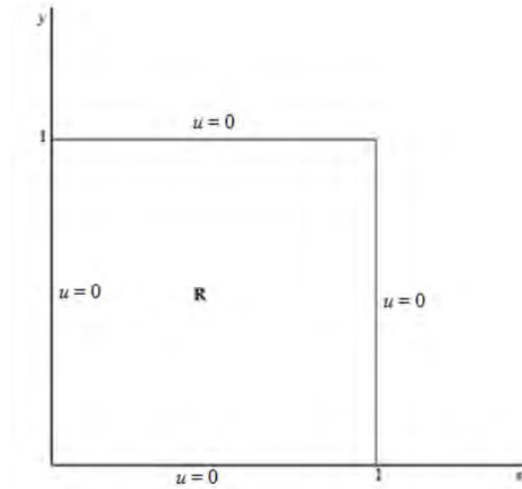


Figure 3.5: Rectangular region R with boundary conditions

We now split the region R in Figure 3.5 into a finite number of rectangular elements. The selection of step lengths $h = (1/4)$ in x -direction and $k = (1/4)$ in y -direction will split the plane region R into 16 rectangular elements. The nodes and the sides in and on the region R can be classified into two groups, viz. interior and exterior. The interior nodes and sides lie inside the region R while the exterior nodes and sides lie on the boundary of the region R . Further all the interior nodes and sides are common to adjacent rectangular elements. Of course this fact is exempted for the exterior nodes and sides. The nodes of the region R are numbered and shown in Figure 3.6.

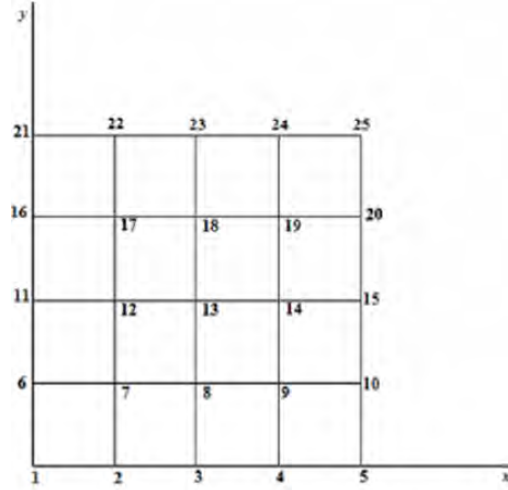


Figure 3.6: Rectangular elements with node numbers.

Here our problem is to find numerical values of the function $u(x, y)$ at the interior node points of the region R provided that the Poisson equation (3.55) and the Dirichlet conditions given in (3.56) are satisfied. Finite element methods are essentially methods for finding approximate solutions of partial differential equations defined in a finite region or domain. Finite element method involves the following steps in solving a partial differential equation together with boundary conditions:

- i. Divide the domain in to a finite number of elements.
- ii. Drive the weak formulation corresponding to the given problem.
- iii. Calculate the stiffness matrix and load vector for each element in the domain.
- iv. Calculate global stiffness matrix and assemble. Calculate global load vector and assemble
- v. Solve the resulting system of algebraic linear equations subject to the satisfaction of the boundary conditions.

The given domain R is divided into 32 congruent right angled and isosceles triangles. The nodes of the triangles are represented using unenclosed numbers. These unenclosed numbers are called global node numbers. Similarly, the triangles are represented by en-rectangle numbers. These enclosed numbers are called element numbers. The en-rectangle number is a number that is enclosed by a rectangle. These details are shown in Figure 3.7.

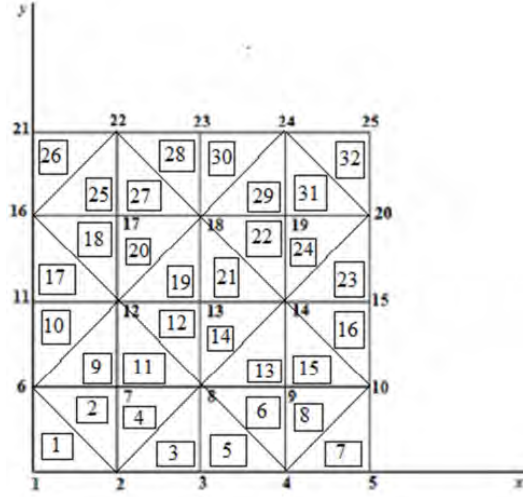


Figure 3.7: Division of domain R into isosceles and right angled triangles.

The weak formulation of the problem defined in (3.55) and (3.56) requires that the function $u(x, y)$:

- (i) is a member of Sobolev space $H_0^1(R)$ i.e., $u \in H_0^1(R)$ and
- (ii) has to satisfy the condition $[-\int_R \nabla u \cdot \nabla v \, dR] = [\int_R f v \, dR], \forall v \in H_0^1(R)$

Here, Sobolev space $H_0^1(R)$ is defined as a set of functions w satisfying the conditions

- (i) $\int_R w^2 \, dR < \infty$
- (ii) $\int_R (w_x)^2 \, dR < \infty$
- (iii) $\int_R (w_y)^2 \, dR < \infty$
- (iv) $w = 0$ on the boundary of the region R

Now, the weak formulation of equation (3.55) is

Multiply both sides of equation (3.55) by arbitrary test function $v = v(x, y) \in H_0^1(R)$ and integrate over the region R , we get

$$\int_R \Delta u \, v \, dR = \int_R f v \, dR \quad (3.57)$$

To reduce the order of derivatives in (3.57) we use Greens formula. The formula states as follows

$$\int_R \Delta u \, v \, dR = - \int_R \nabla u \cdot \nabla v \, dR + \int_{\partial R} \frac{\partial u}{\partial n} v \, d\gamma \quad (3.58)$$

Here, ∂R , $d\gamma$, ∇u and n denotes respectively boundary of region R , line segment on the boundary of R , gradient of u and outward unit normal on the boundary of R .

Substituting equation (3.58) in (3.57) we have

$$- \int_R \nabla u \cdot \nabla v \, dR + \int_{\partial R} \frac{\partial u}{\partial n} v \, d\gamma = \int_R f v \, dR \quad (3.59)$$

Since $v \in H_0^1(R)$ the condition : $\int_{\partial R} \frac{\partial u}{\partial n} v d\gamma = 0$ is valid. Thus equation (3.59) becomes $-\int_R \nabla u \cdot \nabla v dR = \int_R f v dR$

The finite element problem for the approximation of Poisson equation (3.55) together with boundary condition (3.56) requires finding $u_d \in V_d$ satisfying the following condition:

$$-\int_R \nabla u_d \cdot \nabla v_d dR = \int_R f v_d dR \quad \forall v_d \in V_d \quad (3.60)$$

Here in (3.60) the space V_d approximates $H_0^1(R)$ and defined as

$$V_d = \{v_d \in C(\bar{R}) : v_d|_{K_m} \in \mathbb{P}_r, v_d|_{\partial R} = 0, \forall m \in \{1, 2, \dots, 32\}\}$$

Where $C(\bar{R})$ = the space of continous function on the closure of R , that is. $\bar{R}$.

∂R = boundary of the region R .

$\mathbb{P}_r$ = the space of plynomials of degree less than or equal to r .

K_m =the m^{th} -triangular elements of the region R .

d = $\max_{m \in \{1, 2, \dots, 32\}} \text{diameter}(K_m) = \max_{x, y \in K_m} |x - y|$

Furthermore, each function $v_d \in V_d$ is characterized, uniquely, by the values it takes at the nodes $N_j, \forall j = 1, 2, 3$ of the triangular elements. A basis in the space V_d can be the set of the characterstic Lagrangian functions $\varphi_i \in V_d, i = 1, 2, 3$ suchthat

$$\varphi_i(N_j) = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$$

where, $i = 1, 2, 3$ and N_j =node j of K_m . The function φ_i is called shape function. The function $v_d \in V_d$ can be expressed through a linear combination of the basis functions of V_d in the following way

$$v_d(x, y) = \sum_{j=1}^3 v_j \varphi_j(x, y), \quad v_j = v_h(N_j) \quad (3.61)$$

By expressing the discrete solution u_d in terms of the basis $\{\varphi_i\}$, we have

$$u_d(x, y) = \sum_{i=1}^3 u_i \varphi_i(x, y), \quad u_i = u_h(N_i) \quad (3.62)$$

We now assign numbers 1, 2, and 3 to the nodes of each triangle element. Starting with right angled vertex, number the nodes in anti clockwise direction using 1, 2, and 3 respectively. This numbering procedure will help us in having all elements the same stiffness matrix since these elements have the same geometry. The detailed derivation of equation of stiffness matrix and load vector is given in the following way: consider the m^{th} triangular element K_m . Equation (3.60) can be defined on triangular element K_m as

$$-\int_{K_m} [\nabla u_d \cdot \nabla v_d + f v_d] dx dy = 0, m = 1, 2, \dots, 32 \quad (3.63)$$

Substituting (3.61) and (3.62) in equation (3.63), we get

$$-\int_{K_m} \left[\nabla \left(\sum_{i=1}^3 u_i \varphi_i(x, y) \right) \cdot \nabla \left(\sum_{j=1}^3 v_j \varphi_j(x, y) \right) + f \left(\sum_{j=1}^3 v_j \varphi_j(x, y) \right) dx dy \right] = 0$$

Which is equivalent to

$$\sum_{j=1}^3 v_j \left[\sum_{i=1}^3 u_i - \int_{K_m} \nabla \varphi_i \nabla \varphi_j dx dy - \int_{K_m} f \varphi_j dx dy \right] = 0$$

Since this equation holds true for all $v_d \in V_d$, we have

$$\sum_{i=1}^3 u_i - \int_{K_m} \nabla \varphi_i \nabla \varphi_j dx dy = \int_{K_m} f \varphi_j dx dy, \quad j = 1, 2, 3.$$

Here for each $m \in \{1, 2, \dots, 32\}$ we have the stiffness matrix and the load vector respectively are $A_{i,j}^{K_m} = - \int_{K_m} \nabla \varphi_i \nabla \varphi_j dx dy$ and $F_j^{K_m} = \int_{K_m} f \varphi_j dx dy$, $i, j = 1, 2, 3$. Thus, the stiffness matrix for each triangular element is calculated using the relation

$$A_{i,j}^{K_m} = - \int_{K_m} \nabla \varphi_i \nabla \varphi_j dx dy, \quad i, j = 1, 2, 3.$$

K_m = The m^{th} triangular element in the domain R . Here $m = 1, \dots, 32$.

A^{K_m} = Stiffness matrix of the m^{th} triangular element.

$A_{i,j}^{K_m}$ = the (i, j) entry of the stiffness matrix A^{K_m} .

φ_l = Shape functions at the nodes of the element K_m . Here $l = 1, 2, 3$.

Thus the stiffness matrix for each element is

Table 1. Stiffness matrix representing every triangular element of the region R .

Local node numbers			
Local node numbers	1	2	3
1	-1	0.5	0.5
2	0.5	-0.5	0
3	0.5	0	-0.5

The (p, q) entry of the global matrix of element K_m takes the (i, j) entry of the stiffness matrix of element K_m where $p, q = 1, 2, \dots, 25$ are global node numbers, and $i, j = 1, 2, 3$ are corresponding local node numbers for the global node numbers and $m = 1, 2, \dots, 32$. The sum of the 32 global matrices gives the assembled global matrix and the assembled global matrix is given in Table 2.

Table 2. Assembled global stiffness matrix.

Global node numbers		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Global node numbers	1	-1	0.5	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0.5	-2	0.5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0.5	-2	0.5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	0.5	-2	0.5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0.5	-1	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0.5	0	0	0	0	-2	1	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	7	0	1	0	0	0	1	-4	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	8	0	0	1	0	0	0	1	-4	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	9	0	0	0	1	0	0	0	1	-4	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	10	0	0	0	0	0.5	0	0	0	1	-2	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0
	11	0	0	0	0	0	0.5	0	0	0	0	-2	1	0	0	0	0.5	0	0	0	0	0	0	0	0	0
	12	0	0	0	0	0	0	1	0	0	0	1	-4	1	0	0	0	1	0	0	0	0	0	0	0	0
	13	0	0	0	0	0	0	0	1	0	0	0	1	-4	1	0	0	0	1	0	0	0	0	0	0	0
	14	0	0	0	0	0	0	0	0	1	0	0	0	1	-4	1	0	0	0	1	0	0	0	0	0	0
	15	0	0	0	0	0	0	0	0	0	0.5	0	0	0	1	-2	0	0	0	0	0.5	0	0	0	0	0
	16	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	-2	1	0	0	0	0.5	0	0	0	0
	17	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	-4	1	0	0	0	1	0	0	0
	18	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	-4	1	0	0	0	1	0	0
	19	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	-4	1	0	0	0	1	0
	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	1	-2	0	0	0	0	0.5
	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	-1	0.5	0	0	0
	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.5	-2	0.5	0	0
	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.5	-2	0.5	0
	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.5	-2	0.5
	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0.5	-1

For the constant $f(x, y) = 1$, the load vector for each triangular element K_m is denoted by F^{K_m} and calculated from the relation

$$F_j^{K_m} = \int_{K_m} \varphi_j dx dy, \quad \forall j = 1, 2, 3.$$

Here $F_j^{K_m}$ denotes the j -entry of the load vector $F_j^{K_m}$. Also, $\varphi_l (l = 1, 2, 3)$ represents a shape function at the three nodes of any triangular element K_m where $m = 1, 2, \dots, 32$. Using two dimensional four point rule Gauss quadrature formula, each element has the same load vector

Table 3. Load vector corresponds to each triangular element

Local node numbers	Load vector
1	0.0104
2	0.0104
3	0.0104

The p entry of the global load vector of element K_m takes the i entry of the load vector of element K_m where $p = 1, 2, \dots, 25$ are global node numbers, $i = 1, 2, 3$ are corresponding local node numbers for the global node numbers and $m = 1, 2, \dots, 32$. The sum of the 32 global load vectors gives the assembled global load vector and the assembled global load vector is

Table 4. Assembled global load vectors

Global node numbers	1	0.0104	Global node numbers	13	0.0417
	2	0.0417		14	0.0833
	3	0.0208		15	0.0208
	4	0.0417		16	0.0417
	5	0.0104		17	0.0417
	6	0.0417		18	0.0833
	7	0.0417		19	0.0417
	8	0.0833		20	0.0417
	9	0.0417		21	0.0104
	10	0.0417		22	0.0417
	11	0.0208		23	0.0208
	12	0.0833		24	0.0417
Rows through 1 - 12			25		0.0104
Rows through 1 - 12			Rows through 13 - 25		

Applying the boundary conditions (3.56) the assembled global matrix and the load vector becomes respectively

Table 5. Reduced assembled global matrix and load vector.

	Global node numbers												
		7	8	9	12	13	14	17	18	19			
Global node numbers	7	-4	1	0	1	0	0	0	0	0	Global node numbers	7	0.0417
	8	1	-4	1	0	1	0	0	0	0		8	0.0833
	9	0	1	-4	0	0	1	0	0	0		9	0.0417
	12	1	0	0	-4	1	0	1	0	0		12	0.0833
	13	0	1	0	1	-4	1	0	1	0		13	0.0417
	14	0	0	1	0	1	-4	0	0	1		14	0.0833
	17	0	0	0	1	0	0	-4	1	0		17	0.0417
	18	0	0	0	0	1	0	1	-4	1		18	0.0833
	19	0	0	0	0	0	1	0	1	-4		19	0.0417

Using Gauss Seidel method successive approximate solution for $u(x, y)$ at the free nodes of the region R is obtained. The results are given in Table 6.

Table 6. Solution of $u(x,y)$ at free nodes of R using Gauss Seidel method.

Iterations	Global node numbers								
	7	8	9	12	13	14	17	18	19
1	0	0	0	0	0	0	0	0	0
2	-0.0104	-0.0234	-0.0163	-0.0234	-0.0221	-0.0304	-0.0163	-0.0304	-0.0256
3	-0.0221	-0.036	-0.027	-0.036	-0.0436	-0.0449	-0.027	-0.0449	-0.0329
4	-0.0284	-0.0456	-0.033	-0.0456	-0.0557	-0.0512	-0.033	-0.0512	-0.036
5	-0.0332	-0.0513	-0.0361	-0.0513	-0.0617	-0.0543	-0.0361	-0.0543	-0.0376
6	-0.0361	-0.0543	-0.0376	-0.0543	-0.0647	-0.0558	-0.0376	-0.0558	-0.0383
7	-0.0376	-0.0558	-0.0383	-0.0558	-0.0662	-0.0565	-0.0383	-0.0565	-0.0387
8	-0.0383	-0.0565	-0.0387	-0.0565	-0.067	-0.0569	-0.0387	-0.0569	-0.0389
9	-0.0387	-0.0569	-0.0389	-0.0569	-0.0673	-0.0571	-0.0389	-0.0571	-0.039
10	-0.0389	-0.0571	-0.039	-0.0571	-0.0675	-0.0572	-0.0390	-0.0572	-0.039
11	-0.0390	-0.0572	-0.039	-0.0572	-0.0676	-0.0572	-0.0390	-0.0572	-0.039
12	-0.0390	-0.0572	-0.039	-0.0572	-0.0677	-0.0573	-0.0390	-0.0573	-0.0391
13	-0.0390	-0.0573	-0.0391	-0.0573	-0.0677	-0.0573	-0.0391	-0.0573	-0.0391
14	-0.0391	-0.0573	-0.0391	-0.0573	-0.0677	-0.0573	-0.0391	-0.0573	-0.0391
15	-0.0391	-0.0573	-0.0391	-0.0573	-0.0677	-0.0573	-0.0391	-0.0573	-0.0391

The finite element solution can be described graphically in three dimension as shown in figure 3.8

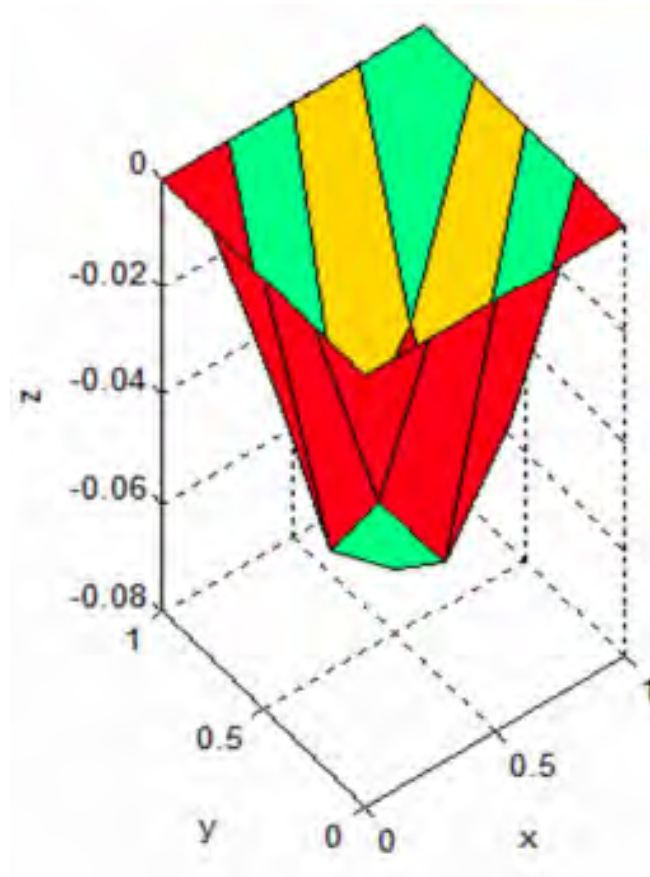


Figure 3.8: finite element solution

Chapter 4

Conculsion

FEM provides greater flexibility to model complex geometries. It can handle general boundary conditions and variable material properties. It has a solid theoretical foundation which gives added reliability and makes it possible to mathematically analyze and estimate the error in the approximate solution. The generalized formulation, its dependence on the boundary conditions, treatment of natural and essential conditions, definitions of the energy space, various norms and the potential energy are fundamental to the finite element method. The approximate solution and hence the error of approximation is determined by the finite element space characterized by the finite element mesh, the polynomial degrees of the elements and the mapping functions. All information generated by the finite element method resides in the standard basis functions, called shape functions, their coefficients and the mapping functions. The errors in the data of interest depend on how the data are computed from the finite element solution. This paper presents the basic understanding of Finite Element Method and the methodology to solve any problem of differential equation. The main idea of Finite Element Method is to choose the appropriate basis functions and then expressing the unknown as combination of the basis functions. Finally a stiffness matrix is generated and solution is obtained. The accuracy of solution increases by using higher order polynomials for the basis function but the calculations become difficult and hence the computation time also increases. Finite Element Method can be executed computationally on Matlab.

Appendix-A

I have tried the MATLAB Code for the Numerical Solution of a Two Dimensional Poisson Equation with Dirichlet Boundary Conditions in the following way. I got good results but still my code has shortcoming specially in the iteration part of finite element solution so I am hopeful those who are referring my project will fill this gap!

```
clc
close all
clear all
%3.6 femcode.m
% [p,t,b]=squaregrid(m,n) % create grid of N = mn nodes to be listed ...
% in p generate mesh of T=2(m-1)(n-1) right triangles in unit square
disp('This MATLAB Program is to Solve 2D Poisson Equation wiz'...
      'Dirchilet Boundry Condntion using Finite Element Method')
m=5; n=5; % includes boundary nodes, mesh spacing 1/(m-1) and 1/(n-1)
[x,y]=ndgrid((0:m-1)/(m-1),(0:n-1)/(n-1)); % matlab forms x and y
%lists
p=[x(:),y(:)]; % N by 2 matrix listing x,y coordinates of all N = mn
%nodes
t=[1,2,m+2; 1,m+2,m+1]; % 3 node numbers for two triangles in first
% square
t=kron(t,ones(m-1,1))+kron(ones(size(t)),(0:m-2)');
% now t lists 3 node numbers of 2(m-1) triangles in the first mesh row
t=kron(t,ones(n-1,1))+kron(ones(size(t)),(0:n-2)*m);
% final t lists 3 node numbers of all triangles in T by 3 matrix
b=[1:m,m+1:m*m*n,2*m:m:m*m*n,m*n-m+2:m*n-1]; % bottom, left, right, top
% b = numbers of all 2m+2n **boundary nodes** preparing for U(b)=0

% [K,F] = assemble(p,t) % K and F for any mesh of triangles: linear
% phi's
N=size(p,1);
disp(['1.The numeber of nodes N='...
      num2str(N)])
T=size(t,1); % number of nodes, number of triangles
disp(['2.The numeber of Triangles T='...
      num2str(T)])
```

```

% p lists x,y coordinates of N nodes, t lists triangles by 3 node
% numbers
K=sparse(N,N); % zero matrix in sparse format: zeros(N) would be
%"dense"
F=zeros(N,1); % load vector F to hold integrals of phi's times load
% f(x,y)
for e=1:T % integration over one triangular element at a time
nodes=t(e,:); % row of t = node numbers of the 3 corners of
% triangle e
Pe=[ones(3,1),p(nodes,:)]; % 3 by 3 matrix with rows=[1 xcorner ycorner]
Area=abs(det(Pe))/2; % area of triangle e = half of parallelogram area
C=inv(Pe); % columns of C are coeffs in a+bx+cy to give phi=1,0,0 at
% nodes
% now compute 3 by 3 Ke and 3 by 1 Fe for element e
grad=C(2:3,:);
Ke=-Area*grad'*grad; % Stiffness matrix representing every
% triangular element of the region R element matrix from
% slopes b,c in grad
Fe=Area/3; % Load vector corresponds to each triangular element
% integral of phi over triangle is volume of pyramid: f(x,y)=1
% multiply Fe by f at centroid for load f(x,y): one-point quadrature!
% centroid would be mean(p(nodes,:)) = average of 3 node coordinates
K(nodes,nodes)=K(nodes,nodes)+Ke; % add Ke to 9 entries of global K
F(nodes)=F(nodes)+Fe; % add Fe to 3 components of load vector F
e=e+1;
end % all T element matrices and vectors now assembled into K and F
Ke;
printmat(Ke, '3. Local Stiffness Matrix', '1 2 3','1 2 3');
Fe;
printmat(Fe, '4. Local load Vectors', '1 2 3','0');
disp([' 5. The three local Load vector corresponds to each triangular...
      element= 'num2str(Fe)])
ff=full(K);
disp(' ----- ')
printmat(ff, '6. Assembled Global Stiffness Matrix',...
'1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25', ...
'1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25' );
disp('5.Assembled global load vectors')
disp(' ----- ')
disp('Global node      Global load vectors.')
disp(' ----- ')
% columnname={'Global node num', 'Global load vectors'};
disp([(1:N)'
F])
% Implement Dirichlet boundary conditions U(b)=0 at nodes in list b
K(b,:)=0; K(:,b)=0; F(b)=0;% put zeros in boundary rows/columns of
% K and F

```

```

K(b,b)=speye(length(b),length(b)); % put I into boundary submatrix
% of K
Kb=K;% Stiffness matrix Kb (sparse format)
Kbf=full(Kb);
Fb=F; %load vector Fb
% Solving for the vector U will produce U(b)=0 at boundary nodes
U=Kb\Fb; % The FEM approximation is  $U_1 \phi_1 + \dots + U_N \phi_N$ 
c=[7 8 9 12 13 14 17 18 19];
UU=U(c');
disp('          7. Finite Element Solutions ')
disp(' ----- ')
disp('      Node      Finite Element Solution ')
disp(' ----- ')
disp([c'                                UU])
% Plot the FEM approximation U(x,y) with values U_1 to U_N at the
% nodes
plot(c',UU,'r+')
xlabel('Global Nodes numbers')
ylabel('Solutions ')
title('Finite Element solution ')
U3d=[U(1:5,:) U(6:10,:) U(11:15,:) U(16:20,:) U(21:25,:)];
figure
surf(x,y,U3d)
xlabel('x')
ylabel('y')
zlabel('Z')
title('Finite Element solution ')

```

References

- [1] Atkinson K, Han W. *Theoretical Numerical Analysis*. Berlin: Springer; 2005.
- [2] Kreyszig E. *Introductory Functional Analysis with Applications*. New York: wiley; 1989 Feb.
- [3] Hutson V, Pym J, Cloud M. *Applications of Functional Analysis and Operator Theory*. Elsevier; 2005 Feb 8.
- [4] Stavroulakis IP, Tersian SA. *Partial Differential Equations: an introduction with Mathematica and MAPLE*. World Scientific; 2004.
- [5] Aubin JP. *Applied Functional Analysis*. John Wiley & Sons; 2011 Sep 30.
- [6] Cioranescu D, Donato P. *An Introduction to Homogenization*.
- [7] Benyam Mebrate, Purnachandra Rao Koya, *Numerical Solution of a Two Dimensional Poisson Equation with Dirichlet Boundary Conditions*, American Journal of Applied Mathematics. Vol. 3, No. 6, 2015, pp. 297-304