# Lecture Notes on Finite Element Methods for Partial Differential Equations

Endre Süli

Mathematical Institute
University of Oxford

August 11, 2020

# Contents

Synopsis:
Finite element methods represent a powerful and general class of techniques for the approximate solution of partial differential equations; the aim of this course is to provide an introduction to their mathematical theory, with special emphasis on theoretical questions such as accuracy, reliability and adaptivity; practical issues concerning the development of efficient finite element algorithms will also be discussed.

Syllabus:
Elements of function spaces. Elliptic boundary value problems: existence, uniqueness and regularity of weak solutions.

Finite element methods: Galerkin orthogonality and Cea's lemma. Piecewise polynomial approximation in Sobolev spaces. The Bramble-Hilbert lemma. Optimal error bounds in the energy norm. Variational crimes.

The Aubin-Nitsche duality argument. Superapproximation properties in mesh-dependent norms. *A posteriori* error analysis by duality: reliability, efficiency and adaptivity.

Finite element approximation of initial boundary value problems. Energy dissipation, conservation and stability. Analysis of finite element methods for evolution problems.

# Reading List

1. *S. Brenner & R. Scott*, The Mathematical Theory of Finite Element Methods. Springer-Verlag, 1994. Corr. 2nd printing 1996. [Chapters 0,1,2,3; Chapter 4: Secs. 4.1–4.4, Chapter 5: Secs. 5.1–5.7].

2. *K. Eriksson, D. Estep, P. Hansbo, & C. Johnson*, Computational Differential Equations. CUP, 1996. [Chapters 5, 6, 8, 14 – 17].

3. *C. Johnson*, Numerical Solution of Partial Differential Equations by the Finite Element Method. CUP, 1990. [Chapters 1–4; Chapter 8: Secs. 8.1–8.4.2; Chapter 9: Secs. 9.1–9.5].

# Chapter 1

# Introduction

Partial differential equations arise in the mathematical modelling of many physical, chemical and biological phenomena and many diverse subject areas such as fluid dynamics, electromagnetism, material science, astrophysics, economy, financial modelling, etc. Very frequently the equations under consideration are so complicated that finding their solutions in closed form or by purely analytical means (e.g. by Laplace and Fourier transform methods, or in the form of a power series) is either impossible or impracticable, and one has to resort to seeking numerical approximations to the unknown analytical solution.

These notes are devoted to a particular class of numerical techniques for the approximate solution of partial differential equations: finite element methods. They were proposed in a seminal work of Richard Courant[1], in 1943; unfortunately, the relevance of this article was not recognised at the time and the idea was forgotten. In the early 1950's the method was rediscovered by engineers, but the mathematical analysis of finite element approximations began much later, in the 1960's, the first important results being due to Miloš Zlámal[2] in 1968. Since then finite element methods have been developed into one of the most general and powerful class of techniques for the numerical solution of partial differential equations and are widely used in engineering design and analysis.

In these notes we shall be concerned with the mathematical aspects of finite element approximation, including stability, accuracy, reliability and adaptivity. We begin by developing some of the theoretical tools: the present chapter is devoted to summarising the elements of the theory of function spaces and reviewing some basic results from the theory of partial differential equations. The concepts and notational conventions introduced here will be used systematically throughout the notes.

[1]R. Courant: *Variational methods for the solution of problems of equilibrium and vibrations.* Bull. Amer. Math. Soc., **49**, pp. 1–23 (1943)

[2]M. Zlámal: *On the finite element method.* Numerische Mathematik, **12**, pp. 394–402 (1968)

## 1.1 Elements of function spaces

As will become apparent in subsequent chapters, the accuracy of finite element approximations to partial differential equations very much depends on the smoothness of the analytical solution to the equation under consideration, and this in turn hinges on the smoothness of the data.

Precise assumptions about the regularity of the solution and the data can be conveniently formulated by considering classes of functions with specific differentiability and integrability properties, called function spaces. In this section we present a brief overview of basic definitions and simple results form the theory of function spaces. For future reference, we remark here that all functions that appear in these notes will be assumed to be real-valued.

### 1.1.1 Spaces of continuous functions

In this section, we describe some simple function spaces which consist of continuously differentiable functions. For the sake of notational convenience, we introduce the concept of a multi-index.

Let $\mathbb{N}$ denote the set of non-negative integers. An $n$-tuple

$$\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$$

is called a **multi–index**. The non-negative integer $|\alpha| := \alpha_1 + \ldots + \alpha_n$ is referred to as the length of the multi–index $\alpha = (\alpha_1, \ldots, \alpha_n)$. We denote $(0, \ldots, 0)$ by $\mathbf{0}$; clearly $|\mathbf{0}| = 0$. Let

$$D^\alpha = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}}.$$

**Example 1** *Suppose that $n = 3$, and $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, $\alpha_j \in \mathbb{N}$, $j = 1, 2, 3$. Then for $u$, a function of three variables $x_1, x_2, x_3$,*

$$\begin{aligned}\sum_{|\alpha|=3} D^\alpha u &= \frac{\partial^3 u}{\partial x_1^3} + \frac{\partial^3 u}{\partial x_1^2 \partial x_2} + \frac{\partial^3 u}{\partial x_1^2 \partial x_3} \\ &+ \frac{\partial^3 u}{\partial x_1 \partial x_2^2} + \frac{\partial^3 u}{\partial x_1 \partial x_3^2} + \frac{\partial^3 u}{\partial x_2^3} \\ &+ \frac{\partial^3 u}{\partial x_1 \partial x_2 \partial x_3} + \frac{\partial^3 u}{\partial x_2^2 \partial x_3} + \frac{\partial^3 u}{\partial x_2 \partial x_3^2} + \frac{\partial^3 u}{\partial x_3^3}. \quad \diamond\end{aligned}$$

This example highlights the importance of multi-index notation: instead of laboriously writing out in detail the ten terms on the right-hand side of the last identity, we can compress the information into a single entity shown on the left.

Let $\Omega$ be an open set in $\mathbb{R}^n$ and let $k \in \mathbb{N}$. We denote by $C^k(\Omega)$ the set of all continuous real-valued functions defined on $\Omega$ such that $D^\alpha u$ is continuous on $\Omega$ for

all $\alpha = (\alpha_1, \ldots, \alpha_n)$ with $|\alpha| \leq k$. Assuming that $\Omega$ is a *bounded* open set, $C^k(\bar{\Omega})$ will denote the set of all $u$ in $C^k(\Omega)$ such that $D^\alpha u$ can be extended from $\Omega$ to a continuous function on $\bar{\Omega}$, the closure of the set $\Omega$, for all $\alpha = (\alpha_1, \ldots, \alpha_n)$, $|\alpha| \leq k$. $C^k(\bar{\Omega})$ can be equipped with the norm

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

In particular when $k = 0$ we shall write $C(\bar{\Omega})$ instead of $C^0(\bar{\Omega})$ to denote the set of all continuous functions defined on $\bar{\Omega}$; in this case,

$$\|u\|_{C(\bar{\Omega})} = \sup_{x \in \Omega} |u(x)| = \max_{x \in \bar{\Omega}} |u(x)|.$$

Similarly, if $k = 1$,

$$\begin{aligned} \|u\|_{C^1(\bar{\Omega})} &= \sum_{|\alpha| \leq 1} \sup_{x \in \Omega} |D^\alpha u(x)| \\ &= \sup_{x \in \Omega} |u(x)| + \sum_{j=1}^{n} \sup_{x \in \Omega} |\frac{\partial u}{\partial x_j}(x)|. \end{aligned}$$

**Example 2** *Consider the open interval $\Omega = (0, 1) \subset \mathbb{R}^1$. The function $u(x) = 1/x$ belongs to $C^k(\Omega)$ for each $k \geq 0$. As $\bar{\Omega} = [0, 1]$ and $\lim_{x \to 0} u(x) = \infty$, it is clear that $u$ is not continuous on $\bar{\Omega}$; the same is true of its derivatives. Therefore $u \notin C^k(\bar{\Omega})$ for any $k \geq 0$.* $\diamond$

The **support** of a continuous function $u$ defined on an open set $\Omega \subset \mathbb{R}^n$ is defined as the closure in $\Omega$ of the set $\{x \in \Omega : u(x) \neq 0\}$. We shall write supp $u$ for the support of $u$. Thus, supp $u$ is the smallest closed subset of $\Omega$ such that $u = 0$ in $\Omega \backslash \text{supp } u$.

**Example 3** *Let $w$ be the function defined on $\mathbb{R}^n$ by*

$$w(x) = \begin{cases} \mathrm{e}^{-\frac{1}{1-|x|^2}} & , \ |x| < 1, \\ 0, & \text{otherwise;} \end{cases}$$

*here $|x| = (x_1^2 + \ldots + x_n^2)^{1/2}$. Clearly, the support of $w$ is the closed unit ball $\{x \in \mathbb{R}^n : |x| \leq 1\}$.* $\diamond$

We denote by $C_0^k(\Omega)$ the set of all $u$ contained in $C^k(\Omega)$ whose support is a bounded subset of $\Omega$. Let

$$C_0^\infty(\Omega) = \bigcap_{k \geq 0} C_0^k(\Omega).$$

**Example 4** *The function $w$ defined in the previous example belongs to the space $C_0^\infty(\mathbb{R}^n)$.* $\diamond$

### 1.1.2 Spaces of integrable functions

Next we consider a class of spaces that consist of (Lebesgue-) integrable functions. Let $p$ be a real number, $p \geq 1$; we denote by $L_p(\Omega)$ the set of all real-valued functions defined on an open subset $\Omega$ of $\mathbb{R}^n$ such that

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

Any two functions which are equal almost everywhere (i.e. equal, except on a set of measure zero) on $\Omega$ are identified with each other. Thus, strictly speaking, $L_p(\Omega)$ consists of equivalence classes of functions; still, we shall not insist on this technicality. $L_p(\Omega)$ is equipped with the norm

$$\|u\|_{L_p(\Omega)} := \left( \int_\Omega |u(x)|^p \, \mathrm{d}x \right)^{1/p}.$$

We shall also consider the space $L_\infty(\Omega)$ consisting of functions $u$ defined on $\Omega$ such that $|u|$ has finite essential supremum on $\Omega$ (namely, there exists a positive constant $M$ such that $|u(x)| \leq M$ for almost every[3] $x$ in $\Omega$; the smallest such number $M$ is called the essential supremum of $|u|$, and we write $M = \mathrm{ess.sup}_{x\in\Omega}|u(x)|$). $L_\infty(\Omega)$ is equipped with the norm

$$\|u\|_{L_\infty(\Omega)} = \mathrm{ess.sup}_{x\in\Omega}|u(x)|.$$

A particularly important case corresponds to taking $p = 2$; then

$$\|u\|_{L_2(\Omega)} = \left( \int_\Omega |u(x)|^2 \, \mathrm{d}x \right)^{1/2}.$$

The space $L_2(\Omega)$ can be equipped with the inner product

$$(u, v) := \int_\Omega u(x)v(x) \, \mathrm{d}x.$$

Clearly $\|u\|_{L_2(\Omega)} = (u, u)^{1/2}$.

**Lemma 1** (The Cauchy–Schwarz inequality) *Let $u$ and $v$ belong to $L_2(\Omega)$; then $u\,v \in L_1(\Omega)$ and*

$$|(u, v)| \leq \|u\|_{L_2(\Omega)} \|v\|_{L_2(\Omega)}.$$

**Proof** Let $\lambda \in \mathbb{R}$; then

$$\begin{aligned} 0 \leq \|u + \lambda v\|^2_{L_2(\Omega)} &= (u + \lambda v, u + \lambda v) \\ &= (u, u) + (u, \lambda v) + (\lambda v, u) + (\lambda v, \lambda v) \\ &= \|u\|^2_{L_2(\Omega)} + 2\lambda(u, v) + \lambda^2 \|v\|^2_{L_2(\Omega)}, \quad \lambda \in \mathbb{R}. \end{aligned}$$

[3] We shall say that a property $P(x)$ is true for almost every $x$ in $\Omega$, if $P(x)$ is true for all $x \in \Omega \backslash \Gamma$ where $\Gamma$ is a subset of $\Omega$ with zero Lebesgue measure.

The right-hand side is a quadratic polynomial in $\lambda$ with real coefficients, and it is non-negative for all $\lambda \in \mathbb{R}$; therefore its discriminant is non-positive, i.e.

$$|2(u,v)|^2 - 4\|u\|^2_{L_2(\Omega)}\|v\|^2_{L_2(\Omega)} \leq 0,$$

and hence the desired inequality. ■

**Corollary 1** (The triangle inequality) *Let $u$ and $v$ belong to $L_2(\Omega)$; then $u+v \in L_2(\Omega)$, and*

$$\|u+v\|_{L_2(\Omega)} \leq \|u\|_{L_2(\Omega)} + \|v\|_{L_2(\Omega)}.$$

**Proof** This is a straightforward consequence of the Cauchy–Schwarz inequality:

$$\begin{aligned} \|u+v\|^2_{L_2(\Omega)} = (u+v, u+v) &= \|u\|^2_{L_2(\Omega)} + 2(u,v) + \|v\|^2_{L_2(\Omega)} \\ &\leq \left(\|u\|_{L_2(\Omega)} + \|v\|_{L_2(\Omega)}\right)^2. \end{aligned}$$

Upon taking the square root of both sides we complete the proof. ■

**Remark 1** *The space $L_p(\Omega)$ with $p \in [1,\infty]$ is a Banach space[4]. In particular, $L_2(\Omega)$ is a Hilbert space: it has an inner product $(\cdot,\cdot)$ and, when equipped with the associated norm $\|\cdot\|_{L_2(\Omega)}$, defined by $\|u\|_{L_2(\Omega)} = (u,u)^{1/2}$, it is a Banach space.* ◇

To conclude this section, we note that a statement analogous to Corollary 1 holds, more generally, in the $L_p$ norm for $1 \leq p \leq \infty$; namely,

$$\|u+v\|_{L_p(\Omega)} \leq \|u\|_{L_p(\Omega)} + \|v\|_{L_p(\Omega)}, \qquad u, v \in L_p(\Omega).$$

Furthermore, the following generalisation of the Cauchy–Schwarz inequality, known as **Hölder's inequality**, is valid for any two functions $u \in L_p(\Omega)$ and $v \in L_{p'}(\Omega)$ with $1/p + 1/p' = 1$:

$$\left| \int_\Omega u(x)v(x)\,\mathrm{d}x \right| \leq \|u\|_{L_p(\Omega)}\|v\|_{L_{p'}(\Omega)}.$$

[4] A normed linear space $X$, with norm $\|\cdot\|_X$, is called a Banach space if, whenever $\{u_m\}_{m=1}^\infty$ is a sequence of elements of $X$ such that

$$\lim_{n,m\to\infty} \|u_n - u_m\|_X = 0, \tag{1.1}$$

there exists $u \in X$ such that $\lim_{m\to\infty}\|u-u_m\|_X = 0$ (i.e. the sequence $\{u_m\}_{m=1}^\infty$ converges to $u$ in $X$). A sequence $\{u_m\}_{m=1}^\infty$ with the property (1.1) is called a Cauchy sequence.

### 1.1.3 Sobolev spaces

In this section we introduce a class of spaces, called Sobolev spaces (after the Russian mathematician S.L. Sobolev), which play an important role in modern differential equation theory. Before we give the precise definition of a Sobolev space, we introduce the concept of weak derivative.

Suppose that $u$ is a smooth function, say $u \in C^k(\Omega)$, with $\Omega$ an open subset of $\mathbb{R}^n$, and let $v \in C_0^\infty(\Omega)$; then the following integration-by-parts formula holds:

$$\int_\Omega D^\alpha u(x) \cdot v(x)\,\mathrm{d}x = (-1)^{|\alpha|} \int_\Omega u(x) \cdot D^\alpha v(x)\,\mathrm{d}x, \quad |\alpha| \le k, \quad \forall v \in C_0^\infty(\Omega).$$

Note that all terms involving integrals over the boundary of $\Omega$, which arise in the course of integrating by parts, have disappeared because $v$ and all of its derivatives are identically zero on the boundary of $\Omega$. This identity represents the starting point for defining the concept of weak derivative.

Now suppose that $u$ is a locally integrable function defined on $\Omega$ (i.e. $u \in L_1(\omega)$ for each bounded open set $\omega$, with $\bar{\omega} \subset \Omega$). Suppose also that there exists a function $w_\alpha$, locally integrable on $\Omega$ and such that

$$\int_\Omega w_\alpha(x) \cdot v(x)\,\mathrm{d}x = (-1)^{|\alpha|} \int_\Omega u(x) \cdot D^\alpha v(x)\,\mathrm{d}x \quad \forall v \in C_0^\infty(\Omega);$$

then we say that $w_\alpha$ is a **weak derivative** of the function $u$ of order $|\alpha| = \alpha_1 + \ldots + \alpha_n$, and we write $w_\alpha = D^\alpha u$. In order to see that this definition is correct it has to be shown that if a locally integrable function has a weak derivative then this must be unique; we remark that this is a straightforward consequence of DuBois Reymond's lemma[5]. Clearly, if $u$ is a sufficiently smooth function, say $u \in C^k(\Omega)$, then its weak derivative $D^\alpha u$ of order $|\alpha| \le k$ coincides with the corresponding partial derivative in the classical pointwise sense,

$$\frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}}.$$

In order to simplify the notation, we shall use the letter $D$ to denote classical as well as weak derivatives; it will always be clear from the context (by considering the smoothness of the function differentiated) which of the two is implied.

**Example 5** *Let $\Omega = \mathbb{R}^1$, and suppose that we wish to determine the weak first derivative of the function $u(x) = (1 - |x|)_+$ defined on $\Omega$. Clearly $u$ is not differentiable at the points $0$ and $\pm 1$. However, because $u$ is locally integrable on $\Omega$, it may,*

[5] **DuBois Reymond's lemma**: Suppose that $w$ is a locally integrable function defined on an open set $\Omega$, $\Omega \subset \mathbb{R}^n$. If

$$\int_\Omega w(x)v(x)\,\mathrm{d}x = 0 \qquad \text{for all } v \text{ in } C_0^\infty(\Omega)$$

then $w(x) = 0$ for almost every $x \in \Omega$.

*nevertheless, have a weak derivative. Indeed, for any* $v \in C_0^\infty(\Omega)$,

$$\begin{aligned}\int_{-\infty}^{+\infty} u(x)v'(x)\,\mathrm{d}x &= \int_{-\infty}^{+\infty}(1-|x|)_+v'(x)\,\mathrm{d}x = \int_{-1}^{1}(1-|x|)v'(x)\,\mathrm{d}x\\ &= \int_{-1}^{0}(1+x)v'(x)\,\mathrm{d}x + \int_{0}^{1}(1-x)v'(x)\,\mathrm{d}x\\ &= -\int_{-1}^{0}v(x)\,\mathrm{d}x + (1+x)v(x)|_{-1}^{0} + \int_{0}^{1}v(x)\,\mathrm{d}x + (1-x)v(x)|_{x=0}^{1}\\ &= \int_{-1}^{0}(-1)v(x)\,\mathrm{d}x + \int_{0}^{1}1\cdot v(x)\,\mathrm{d}x \equiv -\int_{-\infty}^{+\infty} w(x)v(x)\,\mathrm{d}x,\end{aligned}$$

*where*

$$w(x) = \begin{cases} 0, & x < -1,\\ 1, & x \in (-1,0),\\ -1, & x \in (0,1),\\ 0, & x > 1.\end{cases}$$

*Thus, the piecewise constant function* $w$ *is the first (weak) derivative of the continuous piecewise linear function* $u$*, i.e.* $w = u' = Du$. $\diamond$

Now we are ready to give a precise definition of a Sobolev space. Let $k$ be a non-negative integer and suppose that $p \in [1,\infty]$. We define (with $D^\alpha$ denoting a weak derivative of order $|\alpha|$ )

$$W_p^k(\Omega) = \{u \in L_p(\Omega) : D^\alpha u \in L_p(\Omega), |\alpha| \le k\}.$$

$W_p^k(\Omega)$ is called a Sobolev space of order $k$; it is equipped with the (Sobolev) norm

$$\|u\|_{W_p^k(\Omega)} := \left(\sum_{|\alpha|\le k}\|D^\alpha u\|_{L_p(\Omega)}^p\right)^{1/p} \qquad \text{when } 1 \le p < \infty$$

and

$$\|u\|_{W_\infty^k(\Omega)} := \sum_{|\alpha|\le k}\|D^\alpha u\|_{L_\infty(\Omega)} \qquad \text{when } p = \infty.$$

Letting,

$$|u|_{W_p^k(\Omega)} := \left(\sum_{|\alpha|= k}\|D^\alpha u\|_{L_p(\Omega)}^p\right)^{1/p},$$

for $p \in [1,\infty)$, we can write

$$\|u\|_{W_p^k(\Omega)} = \left(\sum_{j=0}^{k}|u|_{W_p^j(\Omega)}^p\right)^{1/p}.$$

Similarly, letting

$$|u|_{W^k_\infty(\Omega)} := \sum_{|\alpha|=k} \|D^\alpha u\|_{L_\infty(\Omega)},$$

we have that

$$\|u\|_{W^k_\infty(\Omega)} = \sum_{j=0}^{k} |u|_{W^j_\infty(\Omega)}.$$

When $k \geq 1$, $|\cdot|_{W^k_p(\Omega)}$ is called the Sobolev semi-norm[6] on $W^k_p(\Omega)$.

An important special case corresponds to taking $p = 2$; the space $W^k_2(\Omega)$ is then a Hilbert space with the inner product

$$(u, v)_{W^k_2(\Omega)} := \sum_{|\alpha|\leq k} (D^\alpha u, D^\alpha v).$$

For this reason, we shall usually write $H^k(\Omega)$ instead of $W^k_2(\Omega)$.

Throughout these notes we shall frequently refer to the Hilbertian Sobolev spaces $H^1(\Omega)$ and $H^2(\Omega)$. Our definitions of $W^k_p(\Omega)$ and its norm and seminorm, for $p = 2$, $k = 1$, give:

$$\begin{aligned}
H^1(\Omega) &= \left\{ u \in L_2(\Omega) : \frac{\partial u}{\partial x_j} \in L_2(\Omega), \quad j = 1, \ldots, n \right\}, \\
\|u\|_{H^1(\Omega)} &= \left\{ \|u\|^2_{L_2(\Omega)} + \sum_{j=1}^{n} \|\frac{\partial u}{\partial x_j}\|^2_{L_2(\Omega)} \right\}^{1/2}, \\
|u|_{H^1(\Omega)} &= \left\{ \sum_{j=1}^{n} \|\frac{\partial u}{\partial x_j}\|^2_{L_2(\Omega)} \right\}^{1/2}.
\end{aligned}$$

Similarly, for $p = 2$ and $k = 2$,

$$\begin{aligned}
H^2(\Omega) &= \left\{ u \in L_2(\Omega) : \frac{\partial u}{\partial x_j} \in L_2(\Omega), \quad j = 1, \ldots, n, \right. \\
&\qquad \left. \frac{\partial^2 u}{\partial x_i \partial x_j} \in L_2(\Omega), \quad i, j = 1, \ldots, n \right\},
\end{aligned}$$

$$\begin{aligned}
\|u\|_{H^2(\Omega)} = \big\{ \|u\|^2_{L_2(\Omega)} &+ \sum_{j=1}^{n} \|\frac{\partial u}{\partial x_j}\|^2_{L_2(\Omega)} \\
&+ \sum_{i,j=1}^{n} \|\frac{\partial^2 u}{\partial x_i \partial x_j}\|^2_{L_2(\Omega)} \Big\}^{1/2},
\end{aligned}$$

[6] When $k \geq 1$, $|\cdot|_{W^k_p(\Omega)}$ is only a semi-norm rather than a norm because if $|u|_{W^k_p(\Omega)} = 0$ for $u \in W^k_p(\Omega)$ it does not necessarily follow that $u(x) = 0$ for almost every $x$ in $\Omega$ (all that is known is that $D^\alpha u(x) = 0$ for almost every $x \in \Omega$, $|\alpha| = k$), so $|\cdot|_{W^k_p(\Omega)}$ does not satisfy the first axiom of norm.

$$|u|_{H^2(\Omega)} = \left\{ \sum_{i,j=1}^{n} \| \frac{\partial^2 u}{\partial x_i \partial x_j} \|^2_{L_2(\Omega)} \right\}^{1/2}.$$

Finally, we define the special Sobolev space $H_0^1(\Omega)$ as the closure of $C_0^\infty(\Omega)$ in the norm of $\| \cdot \|_{H^1(\Omega)}$; in other words, $H_0^1(\Omega)$ is the set of all $u \in H^1(\Omega)$ such that $u$ is the limit in $H^1(\Omega)$ of a sequence $\{u_m\}_{m=1}^\infty$ with $u_m \in C_0^\infty(\Omega)$. It can be shown (assuming that $\partial\Omega$ is sufficiently smooth) that

$$H_0^1(\Omega) = \{u \in H^1(\Omega) : u = 0 \text{ on } \partial\Omega\};$$

i.e. $H_0^1(\Omega)$ is, in fact, the set of all functions $u$ in $H^1(\Omega)$ such that $u = 0$ on $\partial\Omega$, the boundary of the set $\Omega$. We shall use this space when considering a partial differential equation that is coupled with a homogeneous (Dirichlet) boundary condition: $u = 0$ on $\partial\Omega$. We note here that $H_0^1(\Omega)$ is also a Hilbert space, with the same norm and inner product as $H^1(\Omega)$.

We conclude the section with the following useful result.

**Lemma 2** (Poincaré–Friedrichs inequality) *Suppose that $\Omega$ is a bounded open set in $\mathbb{R}^n$ (with a sufficiently smooth boundary[7] $\partial\Omega$) and let $u \in H_0^1(\Omega)$; then there exists a constant $c_\star(\Omega)$, independent of $u$, such that*

$$\int_\Omega |u(x)|^2 \, \mathrm{d}x \le c_\star \sum_{i=1}^{n} \int_\Omega |\frac{\partial u}{\partial x_i}(x)|^2 \, \mathrm{d}x. \tag{1.2}$$

**Proof** As any function $u \in H_0^1(\Omega)$ is the limit in $H^1(\Omega)$ of a sequence $\{u_m\}_{m=1}^\infty \subset C_0^\infty(\Omega)$, it is sufficient to prove this inequality for $u \in C_0^\infty(\Omega)$.

In fact, to simplify matters, we shall restrict ourselves to considering the special case of a rectangular domain $\Omega = (a, b) \times (c, d)$ in $\mathbb{R}^2$. The proof for general $\Omega$ is analogous. Evidently

$$u(x, y) = u(a, y) + \int_a^x \frac{\partial u}{\partial x}(\xi, y) \, \mathrm{d}\xi = \int_a^x \frac{\partial u}{\partial x}(\xi, y) \, \mathrm{d}\xi, \qquad c < y < d.$$

Thence, by the Cauchy–Schwarz inequality,

$$\begin{aligned}
\int_\Omega |u(x, y)|^2 \, \mathrm{d}x \, \mathrm{d}y &= \int_a^b \int_c^d |\int_a^x \frac{\partial u}{\partial x}(\xi, y) \, \mathrm{d}\xi|^2 \, \mathrm{d}y \, \mathrm{d}x \\
&\le \int_a^b \int_c^d (x - a) \left( \int_a^x |\frac{\partial u}{\partial x}(\xi, y)|^2 \, \mathrm{d}\xi \right) \mathrm{d}y \, \mathrm{d}x \\
&\le \int_a^b (x - a) \, \mathrm{d}x \left( \int_c^d \int_a^b |\frac{\partial u}{\partial x}(\xi, y)|^2 \, \mathrm{d}\xi \, \mathrm{d}y \right) \\
&= \frac{1}{2}(b - a)^2 \int_\Omega |\frac{\partial u}{\partial x}(x, y)|^2 \, \mathrm{d}x \, \mathrm{d}y.
\end{aligned}$$

[7] Say, $\Omega$ is a polygonal domain in $\mathbb{R}^2$ or a polyhedron in $\mathbb{R}^3$.

Analogously,

$$\int_\Omega |u(x,y)|^2 \,\mathrm{d}x\,\mathrm{d}y \le \frac{1}{2}(d-c)^2 \int_\Omega |\frac{\partial u}{\partial y}(x,y)|^2 \,\mathrm{d}x\,\mathrm{d}y.$$

By adding the two inequalities, we obtain

$$\int_\Omega |u(x,y)|^2 \,\mathrm{d}x\,\mathrm{d}y \le c_\star \int_\Omega \left( |\frac{\partial u}{\partial x}|^2 + |\frac{\partial u}{\partial y}|^2 \right) \,\mathrm{d}x\,\mathrm{d}y,$$

where $c_\star = \left( \frac{2}{(b-a)^2} + \frac{2}{(d-c)^2} \right)^{-1}$. ∎

For further reference, we note that if $\Omega = (0,1)^2 \subset \mathbb{R}^2$ then $c_\star = \frac{1}{4}$; similarly, if $\Omega = (0,1) \subset \mathbb{R}$ then $c_\star = \frac{1}{2}$.

## 1.2 Weak solutions to elliptic problems

In the first part of this lecture course we shall focus on boundary value problems for elliptic partial differential equations. Elliptic equations are typified by the Laplace equation

$$\Delta u = 0,$$

and its non-homogeneous counterpart, Poisson's equation

$$-\Delta u = f,$$

where we used the notation

$$\Delta = \sum_{i=1}^n \frac{\partial^2}{\partial x_i^2}$$

for the **Laplace operator**.

More generally, let $\Omega$ be a bounded open set in $\mathbb{R}^n$, and consider the linear second-order partial differential equation

$$-\sum_{i,j=1}^n \frac{\partial}{\partial x_j}\left( a_{ij}(x) \frac{\partial u}{\partial x_i} \right) + \sum_{i=1}^n b_i(x) \frac{\partial u}{\partial x_i} + c(x)u = f(x), \quad x \in \Omega, \tag{1.3}$$

where the coefficients $a_{ij}$, $b_i$, $c$ and $f$ satisfy the following conditions:

$$\begin{aligned} a_{ij} &\in C^1(\bar{\Omega}), \quad & i,j = 1,\ldots,n; \\ b_i &\in C(\bar{\Omega}), \quad & i = 1,\ldots,n; \\ c &\in C(\bar{\Omega}), \quad & f \in C(\bar{\Omega}), \end{aligned}$$

and

$$\sum_{i,j=1}^n a_{ij}(x)\xi_i\xi_j \ge \tilde{c} \sum_{i=1}^n \xi_i^2, \;\; \forall \xi = (\xi_1,\ldots,\xi_n) \in \mathbb{R}^n, \quad x \in \bar{\Omega}; \tag{1.4}$$

here $\tilde{c}$ is a positive constant independent of $x$ and $\xi$. The condition (1.4) is usually referred to as uniform ellipticity and (1.3) is called an elliptic equation.

In problems that arise in applications equation (1.3) is usually supplemented by one of the following boundary conditions, with $g$ denoting a given function defined on $\partial\Omega$:

(a) $u = g$ on $\partial\Omega$ (Dirichlet boundary condition);

(b) $\frac{\partial u}{\partial \nu} = g$ on $\partial\Omega$, where $\nu$ denotes the unit outward normal vector to $\partial\Omega$ (Neumann boundary condition);

(c) $\frac{\partial u}{\partial \nu} + \sigma u = g$ on $\partial\Omega$, where $\sigma(x) \geq 0$ on $\partial\Omega$ (Robin boundary condition);

(d) A generalisation of the boundary conditions (b) and (c) is

$$\sum_{i,j=1}^{n} a_{ij} \frac{\partial u}{\partial x_i} \cos\alpha_j + \sigma(x)u = g \quad \text{on} \quad \partial\Omega,$$

where $\alpha_j$ is the angle between the unit outward normal vector $\nu$ to $\partial\Omega$ and the $x_j$ axis (Oblique derivative boundary condition).

In many physical problems more than one type of boundary condition is imposed on $\partial\Omega$ (e.g. $\partial\Omega$ is the union of two disjoint subsets $\partial\Omega_1$ and $\partial\Omega_2$, with a Dirichlet boundary condition on $\partial\Omega_1$ and Neumann boundary condition on $\partial\Omega_2$). The study of such mixed boundary value problems will not be pursued in these notes.

We begin by considering the homogeneous Dirichlet boundary value problem

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial u}{\partial x_i}\right) + \sum_{i=1}^{n} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f(x), \; x \in \Omega, \tag{1.5}$$

$$u = 0 \quad \text{on} \quad \partial\Omega, \tag{1.6}$$

where $a_{ij}$, $b_i$, $c$ and $f$ are as in (1.4).

A function $u \in C^2(\Omega)\cap C(\bar{\Omega})$ satisfying (1.5) and (1.6) is called a **classical solution** of this problem. The theory of partial differential equations tells us that (1.5), (1.6) has a unique classical solution, provided that $a_{ij}$, $b_i$, $c$, $f$ and $\partial\Omega$ are sufficiently smooth. However, in many applications one has to consider equations where these smoothness requirements are violated, and for such problems the classical theory is inappropriate. Take, for example, Poisson's equation with zero Dirichlet boundary condition on $\Omega = (-1,1)^n$ in $\mathbb{R}^n$:

$$\left.\begin{array}{rcll} -\Delta u &=& \operatorname{sgn}\left(\frac{1}{2} - |x|\right), & x \in \Omega, \\ u &=& 0, & x \in \partial\Omega. \end{array}\right\} \tag{$*$}$$

This problem does not have a classical solution, $u \in C^2(\Omega)\cap C(\bar{\Omega})$, for otherwise $\Delta u$ would be a continuous function on $\Omega$, which is not possible because $\operatorname{sgn}(1/2 - |x|)$ is not continuous on $\Omega$.

In order to overcome the limitations of the classical theory and to be able to deal with partial differential equations with "non-smooth" data, we generalise the notion of solution by weakening the differentiability requirements on $u$.

To begin, let us suppose that $u$ is a classical solution of (1.5), (1.6). Then, for any $v \in C_0^1(\Omega)$,

$$\begin{aligned} -\sum_{i,j=1}^{n} \int_\Omega \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial u}{\partial x_i}\right) \cdot v \,\mathrm{d}x &+ \sum_{i=1}^{n} \int_\Omega b_i(x)\frac{\partial u}{\partial x_i} \cdot v \,\mathrm{d}x \\ + \int_\Omega c(x)uv \,\mathrm{d}x &= \int_\Omega f(x)v(x)\,\mathrm{d}x. \end{aligned}$$

Upon integration by parts in the first integral and noting that $v = 0$ on $\partial\Omega$, we obtain:

$$\begin{aligned} \sum_{i,j=1}^{n} \int_\Omega a_{ij}(x)\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}\,\mathrm{d}x &+ \sum_{i=1}^{n} \int_\Omega b_i(x)\frac{\partial u}{\partial x_i} v \,\mathrm{d}x \\ + \int_\Omega c(x)uv \,\mathrm{d}x &= \int_\Omega f(x)v(x)\,\mathrm{d}x \quad \forall v \in C_0^1(\Omega). \end{aligned}$$

In order for this equality to make sense we no longer need to assume that $u \in C^2(\Omega)$: it is sufficient that $u \in L_2(\Omega)$ and $\partial u/\partial x_i \in L_2(\Omega)$, $i = 1, \ldots, n$. Thus, remembering that $u$ has to satisfy a zero Dirichlet boundary condition, it is natural to seek $u$ in the space $H_0^1(\Omega)$, where, as in Section 1.1.3,

$$H_0^1(\Omega) = \{u \in L_2(\Omega) : \frac{\partial u}{\partial x_i} \in L_2(\Omega), \;\; i = 1, \ldots, n, \;\; u = 0 \quad \text{on} \quad \partial\Omega\}.$$

Therefore, we consider the following problem: find $u$ in $H_0^1(\Omega)$ such that

$$\begin{aligned} \sum_{i,j=1}^{n} \int_\Omega a_{ij}(x)\frac{\partial u}{\partial x_i} \cdot \frac{\partial v}{\partial x_j}\,\mathrm{d}x + \sum_{i=1}^{n} \int_\Omega b_i(x)\frac{\partial u}{\partial x_i} v \,\mathrm{d}x \\ + \int_\Omega c(x)uv \,\mathrm{d}x = \int_\Omega f(x)v(x)\,\mathrm{d}x \quad \forall v \in C_0^1(\Omega). \end{aligned} \tag{1.7}$$

We note that $C_0^1(\Omega) \subset H_0^1(\Omega)$, and it is easily seen that when $u \in H_0^1(\Omega)$ and $v \in H_0^1(\Omega)$, (instead of $v \in C_0^1(\Omega)$), the expressions on the left- and right-hand side of (1.7) are still meaningful (in fact, we shall prove this below)[8]. This motivates the following definition.

[8]Note further that since the coefficients $a_{ij}$ no longer appear under derivative signs in (1.7), it is not necessary to assume that $a_{ij} \in C^1(\bar{\Omega})$; $a_{ij} \in L_\infty(\Omega)$ will be seen to be sufficient. Also, the smoothness requirements imposed on the coefficients $b_i$ and $c$ can be relaxed: $b_i \in L_\infty(\Omega)$ for $i = 1, \ldots, n$ and $c \in L_\infty(\Omega)$ will suffice.

**Definition 1** *Let $a_{ij} \in L_\infty(\Omega)$, $i,j = 1,\ldots,n$, $b_i \in L_\infty(\Omega)$, $i = 1,\ldots,n$, $c \in L_\infty(\Omega)$, and let $f \in L_2(\Omega)$. A function $u \in H_0^1(\Omega)$ satisfying*

$$\begin{aligned}\sum_{i,j=1}^{n} \int_\Omega a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} \,\mathrm{d}x \;&+\; \sum_{i=1}^{n} \int_\Omega b_i(x) \frac{\partial u}{\partial x_i} v \,\mathrm{d}x \\ + \int_\Omega c(x) uv \,\mathrm{d}x \;&=\; \int_\Omega f(x) v(x) \,\mathrm{d}x \quad \forall v \in H_0^1(\Omega) \end{aligned} \tag{1.8}$$

*is called a* **weak solution** *of (1.5), (1.6). All partial derivatives in (1.8) should be understood as weak derivatives.*

Clearly if $u$ is a classical solution of (1.5), (1.6), then it is also a weak solution of (1.5), (1.6). However, the converse is not true. If (1.5), (1.6) has a weak solution, this may not be smooth enough to be a classical solution. Indeed, we shall prove below that the boundary value problem $(*)$ has a unique weak solution $u \in H_0^1(\Omega)$, despite the fact that it has no classical solution. Before considering this particular boundary value problem, we look at the wider issue of existence of a unique weak solution to the more general problem (1.5), (1.6).

For the sake of simplicity, we adopt the following notation:

$$\begin{aligned} a(w,v) \;&=\; \sum_{i,j=1}^{n} \int_\Omega a_{ij}(x) \frac{\partial w}{\partial x_i} \frac{\partial v}{\partial x_j} \,\mathrm{d}x \\ &\quad + \sum_{i=1}^{n} \int_\Omega b_i(x) \frac{\partial w}{\partial x_i} v \,\mathrm{d}x + \int_\Omega c(x) wv \,\mathrm{d}x \end{aligned} \tag{1.9}$$

and

$$l(v) = \int_\Omega f(x) v(x) \,\mathrm{d}x. \tag{1.10}$$

With this new notation, problem (1.8) can be written as follows:

$$\text{find } u \in H_0^1(\Omega) \text{ such that } a(u,v) = l(v) \quad \forall v \in H_0^1(\Omega). \tag{1.11}$$

We shall prove the existence of a unique solution to this problem by exploiting the following abstract result from Functional Analysis.

**Theorem 1** (Lax & Milgram theorem) *Suppose that $V$ is a real Hilbert space equipped with norm $\|\cdot\|_V$. Let $a(\cdot,\cdot)$ be a bilinear functional on $V \times V$ such that:*

*(a)* $\exists c_0 > 0 \quad \forall v \in V \quad a(v,v) \geq c_0 \|v\|_V^2$,

*(b)* $\exists c_1 > 0 \quad \forall v, w \in V \quad |a(w,v)| \leq c_1 \|w\|_V \|v\|_V$,
*and let $l(\cdot)$ be a linear functional on $V$ such that*

*(c)* $\exists c_2 > 0 \quad \forall v \in V \quad |l(v)| \leq c_2 \|v\|_V$.

*Then, there exists a unique $u \in V$ such that*

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

For a proof of this result the interested reader is referred to the books: P. Ciarlet: *The Finite Element Method for Elliptic Problems*, North-Holland, 1978; K. Yosida: *Functional Analysis*, Reprint of the 6th ed., Springer-Verlag, 1995.

We apply the Lax-Milgram theorem with $V = H_0^1(\Omega)$ and $\|\cdot\|_V = \|\cdot\|_{H^1(\Omega)}$ to show the existence of a unique weak solution to (1.5), (1.6) (or, equivalently, to (1.11)). Let us recall from Section 1.1.3 that $H_0^1(\Omega)$ is a Hilbert space with the inner product

$$(w, v)_{H^1(\Omega)} = \int_\Omega wv \,\mathrm{d}x + \sum_{i=1}^n \int_\Omega \frac{\partial w}{\partial x_i} \cdot \frac{\partial v}{\partial x_i} \,\mathrm{d}x$$

and the associated norm $\|w\|_{H^1(\Omega)} = (w, w)_{H^1(\Omega)}^{1/2}$. Next we show that $a(\cdot, \cdot)$ and $l(\cdot)$, defined by (1.9) and (1.10), satisfy the hypotheses (a), (b), (c) of the Lax-Milgram theorem.

We begin with (c). The mapping $v \mapsto l(v)$ is linear: indeed, for any $\alpha, \beta \in \mathbb{R}$,

$$\begin{aligned} l(\alpha v_1 + \beta v_2) &= \int_\Omega f(x)(\alpha v_1(x) + \beta v_2(x)) \,\mathrm{d}x \\ &= \alpha \int_\Omega f(x) v_1(x) \,\mathrm{d}x + \beta \int_\Omega f(x) v_2(x) \,\mathrm{d}x \\ &= \alpha l(v_1) + \beta l(v_2), \quad v_1, v_2 \in H_0^1(\Omega); \end{aligned}$$

so $l(\cdot)$ is a linear functional on $H_0^1(\Omega)$. Also, by the Cauchy–Schwarz inequality,

$$\begin{aligned} |l(v)| &= |\int_\Omega f(x) v(x) \,\mathrm{d}x| \le \left( \int_\Omega |f(x)|^2 \,\mathrm{d}x \right)^{1/2} \left( \int_\Omega |v(x)|^2 \,\mathrm{d}x \right)^{1/2} \\ &= \|f\|_{L_2(\Omega)} \|v\|_{L_2(\Omega)} \le \|f\|_{L_2(\Omega)} \|v\|_{H^1(\Omega)}, \end{aligned}$$

for all $v \in H_0^1(\Omega)$, where we have used the obvious inequality $\|v\|_{L_2(\Omega)} \le \|v\|_{H^1(\Omega)}$. Letting $c_2 = \|f\|_{L_2(\Omega)}$, we obtain the required bound.

Next we verify (b). For any fixed $w \in H_0^1(\Omega)$, the mapping $v \mapsto a(v, w)$ is linear. Similarly, for any fixed $v \in H_0^1(\Omega)$, the mapping $w \mapsto a(v, w)$ is linear. Hence $a(\cdot, \cdot)$ is a bilinear functional on $H_0^1(\Omega) \times H_0^1(\Omega)$. Applying the Cauchy–Schwarz inequality, we deduce that

$$\begin{aligned} |a(w, v)| \le & \sum_{i,j=1}^n \max_{x \in \bar{\Omega}} |a_{ij}(x)| \, |\int_\Omega \frac{\partial w}{\partial x_i} \frac{\partial v}{\partial x_j} \,\mathrm{d}x| \\ & + \sum_{i=1}^n \max_{x \in \bar{\Omega}} |b_i(x)| \, |\int_\Omega \frac{\partial w}{\partial x_i} v \,\mathrm{d}x| \end{aligned}$$

$$
\begin{aligned}
&+\max_{x\in\bar\Omega}|c(x)|\,\Big|\int_\Omega w(x)v(x)\,\mathrm{d}x\Big| \\
\le\ & \hat c\left\{\sum_{i,j=1}^n\left(\int_\Omega|\frac{\partial w}{\partial x_i}|^2\,\mathrm{d}x\right)^{1/2}\left(\int_\Omega|\frac{\partial v}{\partial x_j}|^2\,\mathrm{d}x\right)^{1/2}\right. \\
&+\sum_{i=1}^n\left(\int_\Omega|\frac{\partial w}{\partial x_i}|^2\,\mathrm{d}x\right)^{1/2}\left(\int_\Omega|v|^2\,\mathrm{d}x\right)^{1/2} \\
&\left.+\left(\int_\Omega|w|^2\,\mathrm{d}x\right)^{1/2}\left(\int_\Omega|v|^2\,\mathrm{d}x\right)^{1/2}\right\} \\
\le\ & \hat c\left\{\left(\int_\Omega|w|^2\,\mathrm{d}x\right)^{1/2}+\sum_{i=1}^n\left(\int_\Omega|\frac{\partial w}{\partial x_i}|^2\,\mathrm{d}x\right)^{1/2}\right\} \\
&\times\left\{\left(\int_\Omega|v|^2\,\mathrm{d}x\right)^{1/2}+\sum_{j=1}^n\left(\int_\Omega|\frac{\partial v}{\partial x_j}|^2\,\mathrm{d}x\right)^{1/2}\right\} \qquad (1.12)
\end{aligned}
$$

where

$$
\hat c=\max\left\{\max_{1\le i,j\le n}\max_{x\in\bar\Omega}|a_{ij}(x)|,\max_{1\le i\le n}\max_{x\in\bar\Omega}|b_i(x)|,\max_{x\in\bar\Omega}|c(x)|\right\}.
$$

By further majorisation of the right-hand side in (1.12) we deduce that

$$
\begin{aligned}
|a(w,v)|\le 2n\hat c&\left\{\int_\Omega|w|^2\,\mathrm{d}x+\sum_{i=1}^n\int_\Omega|\frac{\partial w}{\partial x_i}|^2\,\mathrm{d}x\right\}^{1/2} \\
&\times\left\{\int_\Omega|v|^2\,\mathrm{d}x+\sum_{j=1}^n\int_\Omega|\frac{\partial v}{\partial x_j}|^2\,\mathrm{d}x\right\}^{1/2},
\end{aligned}
$$

so that, by letting $c_1=2n\hat c$, we obtain inequality (b):

$$
|a(w,v)|\le c_1\|w\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}. \qquad (1.13)
$$

It remains to establish (a). To do so, we shall slightly strengthen the smoothness requirements on the coefficients $b_i$ by demanding that $b_i\in W^1_\infty(\Omega)$ (see, however, Remark 4 at the end of this chapter). Using (1.4), we deduce that

$$
a(v,v)\ge\tilde c\sum_{i=1}^n\int_\Omega|\frac{\partial v}{\partial x_i}|^2\,\mathrm{d}x+\sum_{i=1}^n\int_\Omega b_i(x)\frac12\frac{\partial}{\partial x_i}(v^2)\,\mathrm{d}x+\int_\Omega c(x)|v|^2\,\mathrm{d}x,
$$

where we wrote $\frac{\partial v}{\partial x_i}\cdot v$ as $\frac12\frac{\partial}{\partial x_i}(v^2)$. Integrating by parts in the second term on the right, we obtain

$$
a(v,v)\ge\tilde c\sum_{i=1}^n\int_\Omega|\frac{\partial v}{\partial x_i}|^2\,\mathrm{d}x+\int_\Omega\left(c(x)-\frac12\sum_{i=1}^n\frac{\partial b_i}{\partial x_i}\right)|v|^2\,\mathrm{d}x.
$$

Suppose that $b_i$, $i = 1, \ldots, n$, and $c$ satisfy the inequality

$$c(x) - \frac{1}{2}\sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i} \geq 0, \quad x \in \bar{\Omega}. \tag{1.14}$$

Then

$$a(v, v) \geq \tilde{c} \sum_{i=1}^{n} \int_{\Omega} |\frac{\partial v}{\partial x_i}|^2 \, \mathrm{d}x. \tag{1.15}$$

By virtue of the Poincaré–Friedrichs inequality stated in Lemma 1.2, the right-hand side can be further bounded below to obtain

$$a(v, v) \geq \frac{\tilde{c}}{c_\star} \int_{\Omega} |v|^2 \, \mathrm{d}x. \tag{1.16}$$

Summing (1.15) and (1.16),

$$a(v, v) \geq c_0 \left( \int_{\Omega} |v|^2 \, \mathrm{d}x + \sum_{i=1}^{n} \int_{\Omega} |\frac{\partial v}{\partial x_i}|^2 \, \mathrm{d}x \right), \tag{1.17}$$

where $c_0 = \tilde{c}/(1 + c_\star)$, and hence (a). Having checked all hypotheses of the Lax-Milgram theorem, we deduce the existence of a unique $u \in H_0^1(\Omega)$ satisfying (1.11); consequently, problem (1.5), (1.6) has a unique weak solution. We encapsulate this result in the following theorem.

**Theorem 2** *Suppose that $a_{ij} \in L_\infty(\Omega)$, $i, j = 1, \ldots, n$, $b_i \in W_\infty^1(\Omega)$, $i = 1, \ldots, n$, $c \in L_\infty(\Omega)$, $f \in L_2(\Omega)$, and assume that (1.4) and (1.14) hold; then the boundary value problem (1.5), (1.6) possesses a unique weak solution $u \in H_0^1(\Omega)$. In addition,*

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{c_0} \|f\|_{L_2(\Omega)}. \tag{1.18}$$

**Proof** We only have to show (1.18) as the rest of the theorem has been proved above. By (1.17), (1.11), the Cauchy–Schwarz inequality and recalling the definition of $\|\cdot\|_{H^1(\Omega)}$,

$$\begin{aligned} c_0 \|u\|_{H^1(\Omega)}^2 &\leq a(u, u) = l(u) = (f, u) \\ &\leq |(f, u)| \leq \|f\|_{L_2(\Omega)} \|u\|_{L_2(\Omega)} \\ &\leq \|f\|_{L_2(\Omega)} \|u\|_{H^1(\Omega)}. \end{aligned}$$

Hence the desired inequality. ∎

Now we return to our earlier example $(*)$ which has been shown to have no classical solution. However, applying the above theorem with $a_{ij}(x) \equiv 1$, $i = j$, $a_{ij}(x) \equiv 0$, $i \neq j$, $1 \leq i, j \leq n$, $b_i(x) \equiv 0$, $c(x) \equiv 0$, $f(x) = \mathrm{sgn}(\frac{1}{2} - |x|)$, and $\Omega = (-1, 1)^n$, we see that (1.4) holds with $\tilde{c} = 1$ and (1.14) is trivially fulfilled. Thus $(*)$ has a unique weak solution $u \in H_0^1(\Omega)$ by Theorem 2. Similar results are valid in the case of Neumann, Robin, and oblique derivative boundary value problems, as well as mixed problems.

**Remark 2** *Consider, for example, the following Dirichlet-Neumann mixed boundary value problem:*

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \Gamma_1, \\ \frac{\partial u}{\partial \nu} &= g \qquad \text{on } \Gamma_2, \end{aligned}$$

*where $\Gamma_1$ is a non-empty, relatively open subset of $\partial\Omega$ and $\Gamma_1 \cup \Gamma_2 = \partial\Omega$. We shall suppose that $f \in L_2(\Omega)$ and that $g \in L_2(\Gamma_2)$. Following a similar reasoning as in the case of the Dirichlet boundary value problem, we consider the special Sobolev space*

$$H^1_{0,\Gamma_1}(\Omega) = \{v \in H^1(\Omega) \,:\, v = 0 \quad \text{on } \Gamma_1\},$$

*and define the weak formulation of the mixed problem as follows: find $u \in H^1_{0,\Gamma_1}(\Omega)$ such that*

$$a(u,v) = l(v) \qquad \text{for all } v \text{ in } H^1_{0,\Gamma_1}(\Omega),$$

*where we put*

$$a(u,v) = \int_\Omega \sum_{i=1}^n \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i}\,\mathrm{d}x$$

*and*

$$l(v) = \int_\Omega f(x)v(x)\,\mathrm{d}x + \int_{\Gamma_2} g(s)v(s)\,\mathrm{d}s.$$

*Applying the Lax-Milgram theorem with $V = H^1_{0,\Gamma_1}(\Omega)$, the existence and uniqueness of a weak solution to this mixed problem easily follows.* $\diamond$

**Remark 3** *Theorem 2 implies that the weak formulation of the elliptic boundary value problem (1.5), (1.6) is well-posed in the sense of Hadamard; namely, for each $f \in L_2(\Omega)$ there exists a unique (weak) solution $u \in H^1_0(\Omega)$, and "small" changes in $f$ give rise to "small" changes in the corresponding solution $u$. The latter property follows by noting that if $u_1$ and $u_2$ are weak solutions in $H^1_0(\Omega)$ of (1.5), (1.6) corresponding to right-hand sides $f_1$ and $f_2$ in $L_2(\Omega)$, respectively, then $u_1 - u_2$ is the weak solution in $H^1_0(\Omega)$ of (1.5), (1.6) corresponding to the right-hand side $f_1 - f_2 \in L_2(\Omega)$. Thus, by virtue of (1.18),*

$$\|u_1 - u_2\|_{H^1(\Omega)} \le \frac{1}{c_0}\|f_1 - f_2\|_{L_2(\Omega)}, \tag{1.19}$$

*and hence the required continuous dependence of the solution of the boundary value problem on the right-hand side.* $\diamond$

**Remark 4** *The requirement $b_i \in W^1_\infty(\Omega)$ in Theorem 2 can be relaxed to the original assumption $b_i \in L_\infty(\Omega)$, $i = 1, \ldots, n$. To see this, note that the smoothness requirements on $b_i$ are unrelated to the verification of condition (c) in the Lax-Milgram*

*theorem, and condition (b) can be shown with* $b_i \in L_\infty(\Omega)$, $i = 1, \ldots, n$, *only anyway. Thus, it remains to see how condition (a) may be verified under the hypothesis* $b_i \in L_\infty(\Omega)$, $i = 1, \ldots, n$. *By (1.4) and the Cauchy–Schwarz inequality,*

$$\begin{aligned} a(v,v) &\geq \tilde{c}|v|^2_{H^1(\Omega)} - \left( \sum_{i=1}^n \|b_i\|^2_{L_\infty(\Omega)} \right)^{1/2} |v|_{H^1(\Omega)} \|v\|_{L_2(\Omega)} + \int_\Omega c(x)|v(x)|^2 \, dx \\ &\geq \frac{1}{2}\tilde{c}|v|^2_{H^1(\Omega)} + \int_\Omega \left( c(x) - \frac{2}{\tilde{c}} \sum_{i=1}^n \|b_i\|^2_{L_\infty(\Omega)} \right) |v(x)|^2 \, dx. \end{aligned}$$

*Assuming that*

$$c(x) - \frac{2}{\tilde{c}} \sum_{i=1}^n \|b_i\|^2_{L_\infty(\Omega)} \geq 0 \tag{1.20}$$

*we arrive at the inequality*

$$a(v,v) \geq \frac{1}{2}\tilde{c} \sum_{i=1}^n \int_\Omega |\frac{\partial v}{\partial x_i}|^2 \, dx,$$

*which is analogous to (1.15). Thus, proceeding in the same way as in the transition from (1.15) to (1.17) we arrive at (1.17) with* $c_0 = \tilde{c}/(2 + 2c_*)$*; this verifies condition (a) in the Lax-Milgram theorem, under the assumptions that* $b_i \in L_\infty(\Omega)$, $i = 1, \ldots, n$, *only and (1.4), (1.20) hold.* $\diamond$

# Chapter 2

# Approximation of elliptic problems

In this chapter we describe the construction of finite element methods for elliptic boundary value problems and outline some of their key properties. Unlike finite difference schemes which are constructed in a more-or-less *ad hoc* fashion through replacing the derivatives in the differential equation by divided differences, the derivation of finite element methods is quite systematic.

The first step in the construction of a finite element method for an elliptic boundary value problem (e.g. (1.5), (1.6)) is to convert the problem into its weak formulation:

$$\text{find } u \in V \text{ such that } a(u,v) = l(v) \quad \forall v \in V, \qquad (P)$$

where $V$ is the solution space (e.g. $H_0^1(\Omega)$ for the homogeneous Dirichlet boundary value problem), $a(\cdot,\cdot)$ is a bilinear functional on $V \times V$, and $l(\cdot)$ is a linear functional on $V$ (e.g. (1.9) and (1.10)).

The second step in the construction is to replace $V$ in $(P)$ by a finite-dimensional subspace $V_h \subset V$ which consists of continuous piecewise polynomial functions of a fixed degree associated with a subdivision of the computational domain; then consider the following approximation of $(P)$:

$$\text{find } u_h \in V_h \text{ such that } a(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h. \qquad (P_h)$$

Suppose, for example, that

$$\dim V_h = N(h) \text{ and } V_h = \text{span}\{\phi_1, \ldots, \phi_{N(h)}\},$$

where the (linearly independent) basis functions $\phi_i$, $i = 1, \ldots, N(h)$, have "small" support. Expressing the approximate solution $u_h$ in terms of the basis functions, $\phi_i$, we can write

$$u_h(x) = \sum_{i=1}^{N(h)} U_i \phi_i(x), \qquad (**)$$

where $U_i$, $i = 1, \ldots, N(h)$, are to be determined. Thus $(P_h)$ can be rewritten as follows:

$$\text{find } (U_1, \ldots, U_{N(h)}) \in \mathbb{R}^{N(h)} \text{ such that}$$

$$\sum_{i=1}^{N(h)} a(\phi_i, \phi_j) U_i = l(\phi_j),\ j = 1, \ldots, N(h). \qquad (P_h')$$

This is a system of linear equations for $U = (U_1, \ldots, U_{N(h)})^T$, with the matrix of the system $A = (a(\phi_j, \phi_i))$ of size $N(h) \times N(h)$. Because the $\phi_i$'s have small support, $a(\phi_j, \phi_i) = 0$ for most pairs of $i$ and $j$, so the matrix $A$ is sparse (in the sense that most of its entries are equal to 0); this property is crucial from the point of efficient solution – in particular, fast iterative methods are available for sparse linear systems. Once $(P_h')$ has been solved for $U = (U_1, \ldots, U_{N(h)})^T$, the expansion $(**)$ provides the required approximation to $u$.

After this brief outline of the idea behind the finite element method, we illustrate the construction of this numerical technique by considering some simple examples.

## 2.1 Piecewise linear basis functions

In this section we describe the construction of the finite element method through two simple examples: the first of these is a two-point boundary value problem for a second-order ordinary differential equation; the second model problem is the homogeneous Dirichlet boundary value problem for Poisson's equation on the unit square in the plane. For the time being we shall assume that the finite element space $V_h$ consists of continuous piecewise linear functions. Higher-degree piecewise polynomial approximations will be discussed later on in the notes.

### One-dimensional problem

Let us consider the boundary value problem

$$-(p(x)u')' + q(x)u = f(x), \quad x \in (0,1), \qquad (2.1)$$

$$u(0) = 0, \quad u(1) = 0, \qquad (2.2)$$

where $p \in C[0,1]$, $q \in C[0,1]$, $f \in L_2(0,1)$ with $p(x) \geq \tilde{c} > 0$ and $q(x) \geq 0$ for all $x$ in $[0,1]$. The weak formulation of this problem is:

$$\left.\begin{array}{l} \text{find } u \in H_0^1(0,1) \text{ such that} \\ \displaystyle\int_0^1 p(x)u'(x)v'(x)\,\mathrm{d}x + \int_0^1 q(x)u(x)v(x)\,\mathrm{d}x = \int_0^1 f(x)v(x)\,\mathrm{d}x \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall v \in H_0^1(0,1). \end{array}\right\} (P)$$

In order to construct the finite element approximation of this problem, we subdivide $\bar{\Omega} = [0,1]$ into $N$ subintervals $[x_i, x_{i+1}]$, $i = 0, \ldots, N-1$, by the points $x_i = ih$, $i = 0, \ldots, N$, where $h = 1/N$, $N \geq 2$, as shown in Fig 2.1. We note that in general the mesh points $x_i$ need not be equally spaced: here we have chosen a uniform spacing only to simplify the exposition.

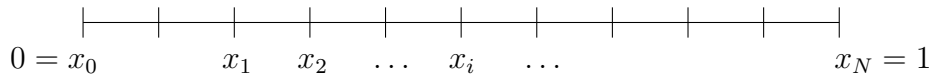


Figure 2.1: Subdivision of $\bar{\Omega} = [0, 1]$.

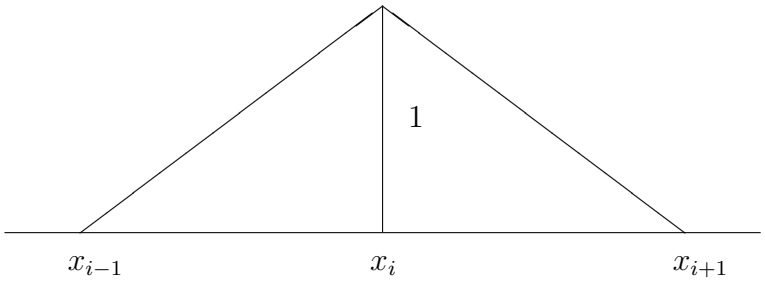


Figure 2.2: The piecewise linear finite element basis function $\phi_i(x)$.

The subintervals $(x_i, x_{i+1})$ are referred to as **element domains** or **elements**, (hence the name **finite element method**). In this example, the weak solution $u \in H_0^1(0,1)$ of problem $(P)$ will be approximated by a continuous piecewise linear function on the subdivision depicted in Figure 2.1. It will be convenient to express our approximation as a linear combination of the finite element basis functions

$$\phi_i(x) = \left(1 - |\frac{x - x_i}{h}|\right)_+, \quad i = 1, \ldots, N-1,$$

shown in Figure 2.2. It is clear that $\phi_i \in H_0^1(0,1)$; furthermore, supp $\phi_i = [x_{i-1}, x_{i+1}]$, $i = 1, \ldots, N-1$, and the functions $\phi_i$, $i = 1, \ldots, N-1$, are linearly independent; therefore

$$V_h := \text{span}\{\phi_1, \ldots, \phi_{N-1}\}$$

is an $(N-1)$-dimensional subspace of $H_0^1(0,1)$.

The finite element approximation of $(P)$ is:

$$\left.\begin{array}{l} \text{find } u_h \in V_h \text{ such that} \\ \displaystyle\int_0^1 p(x)u_h'(x)v_h'(x)\,\mathrm{d}x \;+\; \int_0^1 q(x)u_h(x)v_h(x)\,\mathrm{d}x \\ \qquad\qquad = \displaystyle\int_0^1 f(x)v_h(x)\,\mathrm{d}x \quad \forall v_h \in V_h. \end{array}\right\} (P_h)$$

Since $u_h \in V_h = \text{span}\{\phi_1, \ldots, \phi_{N-1}\}$, it can be written as a linear combination

of the basis functions:

$$u_h(x) = \sum_{i=1}^{N-1} U_i \phi_i(x).$$

Substituting this expansion into $(P_h)$ we obtain the following problem, equivalent to $(P_h)$:

$$\left.\begin{array}{l} \text{find } U = (U_1, \ldots, U_{N-1})^T \in \mathbb{R}^{N-1} \text{ such that} \\ \displaystyle\sum_{i=1}^{N-1} U_i \int_0^1 [p(x)\phi_i'(x)\phi_j'(x) \;+\; q(x)\phi_i(x)\phi_j(x)]\,\mathrm{d}x \\ \qquad\qquad = \displaystyle\int_0^1 f(x)\phi_j(x)\,\mathrm{d}x, \\ \text{for } j = 1, \ldots, N-1. \end{array}\right\} (P_h')$$

Letting

$$a_{ji} := \int_0^1 [p(x)\phi_i'(x)\phi_j'(x) + q(x)\phi_i(x)\phi_j(x)]\,\mathrm{d}x, \quad i, j = 1, \ldots, N-1;$$

$$F_j := \int_0^1 f(x)\phi_j(x)\,\mathrm{d}x, \quad j = 1, \ldots, N-1,$$

$(P_h')$ can be written as a system of linear equations

$$AU = F,$$

where $A = (a_{ji})$, $F = (F_1, \ldots, F_{N-1})^T$. The matrix $A$ is symmetric (i.e. $A^T = A$) and positive definite (i.e. $x^T Ax > 0$, $x \neq 0$). Since supp $\phi_i \cap$ supp $\phi_j$ has empty interior when $|i-j| > 1$, it follows that the matrix $A$ is tri-diagonal (namely, $a_{ji}$ is zero, unless $|i-j| \leq 1$). Having solved the system of linear equations $AU = F$, we substitute the values $U_1, \ldots, U_{N-1}$ into the expansion

$$u_h(x) = \sum_{i=1}^{N-1} U_i \phi_i(x)$$

to obtain $u_h$.

In practice the entries $a_{ji}$ of the matrix $A$ and the entries $F_j$ of the vector $F$ are calculated approximately using numerical integration (quadrature) rules. In the simple case when $p$ and $q$ are constant functions on $[0,1]$, the entries of $A$ can be calculated exactly:

$$\begin{aligned} a_{ij} &= p\int_0^1 \phi_i'(x)\phi_j'(x)\,\mathrm{d}x + q\int_0^1 \phi_i(x)\phi_j(x)\,\mathrm{d}x \\ &= p\begin{cases} 2/h, & i=j, \\ -1/h, & |i-j|=1, \\ 0, & |i-j|>1, \end{cases} + q\begin{cases} 4h/6, & i=j, \\ h/6, & |i-j|=1, \\ 0, & |i-j|>1. \end{cases} \\ &= \begin{cases} 2p/h + 4hq/6, & i=j, \\ -p/h + qh/6, & |i-j|=1, \\ 0, & |i-j|>1. \end{cases} \end{aligned}$$

Figure 2.3: A subdivision (triangulation) of $\bar{\Omega}$.

This gives rise to the following set of linear equations:

$$-p\frac{U_{i-1}-2U_i+U_{i+1}}{h^2}+q\frac{U_{i-1}+4U_i+U_{i+1}}{6}=\frac{1}{h}\int_{x_{i-1}}^{x_{i+1}} f(x)\phi_i(x)\,\mathrm{d}x,$$

$$i=1,\ldots,N-1,$$

with the convention that $U_0 = 0$ and $U_N = 0$ (corresponding to the fact that $u_h(0) = 1$ and $u_h(1) = 0$, respectively). This is a three-point finite difference scheme for the values $U_i$, the values of $u_h(x)$ at the mesh points $x_i$.

**Two-dimensional problem**

Let $\Omega$ be a bounded domain in $\mathbb{R}^2$ with a polygonal boundary $\partial\Omega$; thus $\Omega$ can be exactly covered by a finite number of triangles. It will be assumed that any pair of triangles in a triangulation of $\Omega$ intersect along a complete edge, at a vertex, or not at all, as shown in Fig. 2.3. We shall denote by $h_K$ the diameter (longest side) of triangle $K$, and we define $h = \max_K h_K$.

With each interior node (marked $\odot$ in the figure) we associate a basis function $\phi$ which is equal to 1 at that node and equal to 0 at all the other nodes; $\phi$ is assumed to be a continuous function on $\bar{\Omega}$ and linear in each of the triangles, as shown in Fig. 2.4.

Let us suppose that the interior nodes are labelled $1, 2, \ldots, N(h)$; let $\phi_1(x, y)$, $\ldots, \phi_{N(h)}(x, y)$ be the corresponding basis functions. The functions $\phi_1, \ldots, \phi_{N(h)}$ are linearly independent and they span an $N(h)$-dimensional linear subspace $V_h$ of $H_0^1(\Omega)$.

Let us consider the elliptic boundary value problem

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

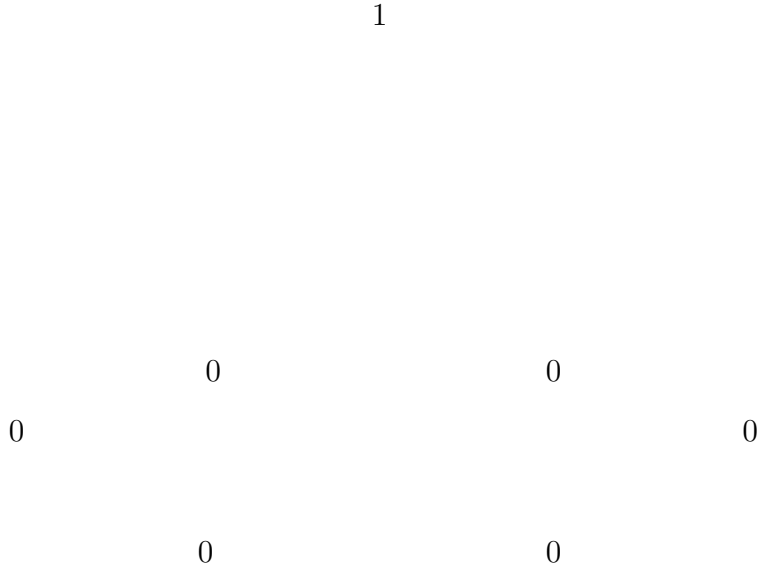


Figure 2.4: A typical finite element basis function $\phi$.

$$u = 0 \quad \text{on} \quad \partial\Omega.$$

In order to construct the finite element approximation of the problem, we begin by considering its weak formulation (see the discussion about weak solutions in Chapter 1 in the special case when $n = 2$, $a_{ij}(x) \equiv 1$ for $i = j$ and $\equiv 0$ for $i \neq j$, $b_i(x) \equiv 0$ for all $i$ and $c(x) \equiv 0$):

$$\begin{aligned} &\text{find } u \in H_0^1(\Omega) \text{ such that} \\ &\int_\Omega \left( \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y = \int_\Omega f v \,\mathrm{d}x\,\mathrm{d}y \quad \forall v \in H_0^1(\Omega). \end{aligned}$$

The finite element approximation of the problem is:

$$\begin{aligned} &\text{find } u_h \in V_h \text{ such that} \\ &\int_\Omega \left( \frac{\partial u_h}{\partial x}\frac{\partial v_h}{\partial x} + \frac{\partial u_h}{\partial y}\frac{\partial v_h}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y = \int_\Omega f v_h \,\mathrm{d}x\,\mathrm{d}y \quad \forall v_h \in V_h. \end{aligned}$$

Writing

$$u_h(x, y) = \sum_{i=1}^{N(h)} U_i \phi_i(x, y),$$

the finite element method can be restated as follows:

$$\text{find } U = (U_1, \ldots, U_{N(h)})^T \in \mathbb{R}^{N(h)} \text{ such that}$$

$$\sum_{i=1}^{N(h)} U_i \left[ \int_\Omega \left( \frac{\partial \phi_i}{\partial x}\frac{\partial \phi_j}{\partial x} + \frac{\partial \phi_i}{\partial y}\frac{\partial \phi_j}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y \right] = \int_\Omega f\phi_j \,\mathrm{d}x\,\mathrm{d}y,$$

for $j = 1, \ldots, N(h)$.

Letting $A = (a_{ij})$, $F = (F_1, \ldots, F_{N(h)})^T$,

$$\begin{aligned} a_{ij} = a_{ji} &= \int_\Omega \left( \frac{\partial \phi_i}{\partial x}\frac{\partial \phi_j}{\partial x} + \frac{\partial \phi_i}{\partial y}\frac{\partial \phi_j}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y, \\ F_j &= \int_\Omega f\phi_j \,\mathrm{d}x\,\mathrm{d}y, \end{aligned}$$

the finite element approximation can be written as a system of linear equations

$$AU = F.$$

Solving this, we obtain $U = (U_1, \ldots, U_{N(h)})^T$, and hence the approximate solution

$$u_h(x, y) = \sum_{i=1}^{N(h)} U_i \phi_i(x, y).$$

The matrix $A$ is called the **stiffness matrix**.

To simplify matters let us suppose that $\Omega = (0, 1) \times (0, 1)$ and consider the triangulation of $\bar{\Omega}$ shown in Fig. 2.5. The case of a general triangulation will be considered later. Let $\phi_{ij}$ denote the basis function associated with the interior node $(x_i, y_j)$:

$$\phi_{ij}(x, y) = \begin{cases} 1 - \frac{x - x_i}{h} - \frac{y - y_j}{h}, & (x, y) \in 1 \\ 1 - \frac{y - y_j}{h}, & (x, y) \in 2 \\ 1 - \frac{x_i - x}{h}, & (x, y) \in 3 \\ 1 - \frac{x_i - x}{h} - \frac{y_j - y}{h}, & (x, y) \in 4 \\ 1 - \frac{y_j - y}{h}, & (x, y) \in 5 \\ 1 - \frac{x - x_i}{h}, & (x, y) \in 6 \\ 0 & \text{otherwise,} \end{cases}$$

where $1, 2, \ldots, 6$ denote the triangles surrounding the node $(x_i, y_j)$ (see Fig. 2.6.) Thus,

$$\frac{\partial \phi_{ij}}{\partial x} = \begin{cases} -1/h, & (x, y) \in 1 \\ 0, & (x, y) \in 2 \\ 1/h, & (x, y) \in 3 \\ 1/h, & (x, y) \in 4 \\ 0, & (x, y) \in 5 \\ -1/h, & (x, y) \in 6 \\ 0, & \text{otherwise,} \end{cases}$$

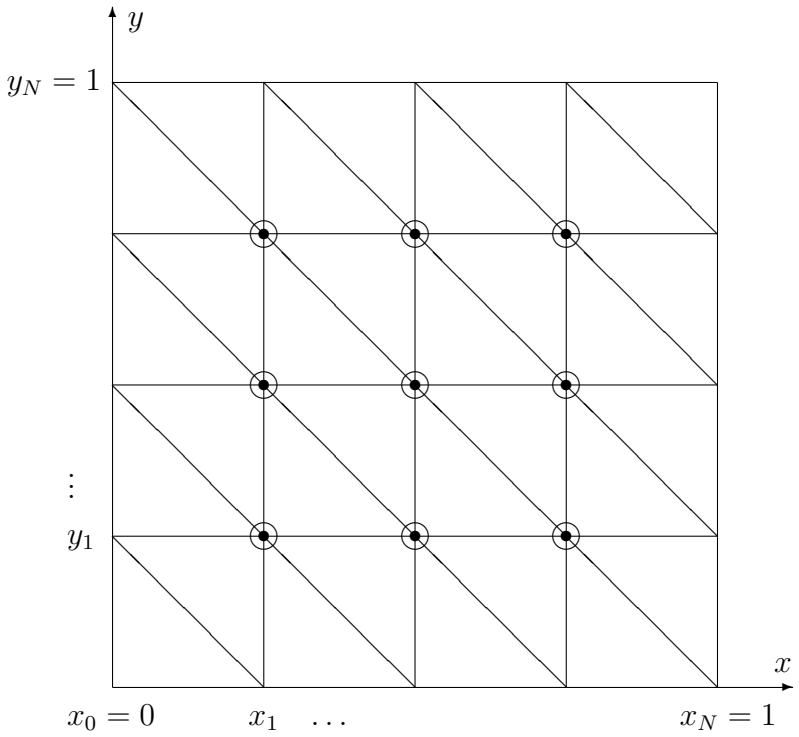


Figure 2.5: Triangulation of $\bar{\Omega} = [0,1] \times [0,1]$.

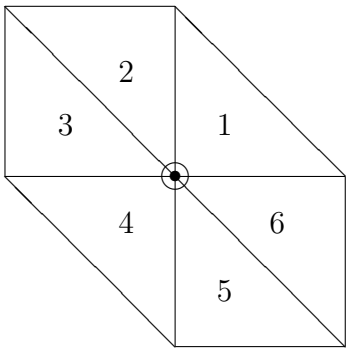


Figure 2.6: Triangles surrounding a node.

and

$$\frac{\partial \phi_{ij}}{\partial y} = \begin{cases} -1/h, & (x,y) \in 1 \\ -1/h, & (x,y) \in 2 \\ 0, & (x,y) \in 3 \\ 1/h, & (x,y) \in 4 \\ 1/h, & (x,y) \in 5 \\ 0, & (x,y) \in 6 \\ 0, & \text{otherwise.} \end{cases}$$

Since

$$\begin{aligned} &\sum_{i=1}^{N-1}\sum_{j=1}^{N-1} U_{ij} \int_\Omega \left( \frac{\partial \phi_{ij}}{\partial x}\frac{\partial \phi_{kl}}{\partial x} + \frac{\partial \phi_{ij}}{\partial y}\frac{\partial \phi_{kl}}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y \\ &= \sum_{i=1}^{N-1}\sum_{j=1}^{N-1} U_{ij} \int_{\text{supp } \phi_{kl}} \left( \frac{\partial \phi_{ij}}{\partial x}\frac{\partial \phi_{kl}}{\partial x} + \frac{\partial \phi_{ij}}{\partial y}\frac{\partial \phi_{kl}}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y \\ &= 4U_{kl} - U_{k-1,l} - U_{k+1,l} - U_{k,l-1} - U_{k,l+1}, \qquad k,l = 1,...,N-1, \end{aligned}$$

the finite element approximation is equivalent to

$$\begin{aligned} &-\frac{U_{k+1,l} - 2U_{k,l} + U_{k-1,l}}{h^2} - \frac{U_{k,l+1} - 2U_{k,l} + U_{k,l-1}}{h^2} \\ &\quad = \frac{1}{h^2} \int\int_{\text{supp } \phi_{kl}} f(x,y)\phi_{kl}(x,y)\,\mathrm{d}x\,\mathrm{d}y, \quad k,l = 1,\ldots,N-1; \\ &U_{kl} = 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Thus, on this special triangulation of $\Omega$, the finite element approximation gives rise to the familiar 5-point finite difference scheme with the forcing function $f$ averaged in a special way.

## 2.2 The self-adjoint elliptic problem

Let us consider, as in Chapter 1, the elliptic boundary value problem

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left( a_{ij}(x)\frac{\partial u}{\partial x_i} \right) + \sum_{i=1}^{n} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f(x), \quad x \in \Omega, \tag{2.3}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{2.4}$$

where $\Omega$ is a bounded open set in $\mathbb{R}^n$, $a_{ij} \in L_\infty(\Omega)$, $i,j = 1,\ldots,n$; $b_i \in W^1_\infty(\Omega)$, $i = 1,\ldots,n$, $c \in L_\infty(\Omega)$, $f \in L_2(\Omega)$, and assume that there exists a positive constant $\tilde{c}$ such that

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \tilde{c}\sum_{i=1}^{n}\xi_i^2 \quad \forall \xi = (\xi_1,\ldots,\xi_n) \in \mathbb{R}^n, \quad \forall x \in \bar{\Omega}. \tag{2.5}$$

We recall from Chapter 1 that the weak formulation of (2.3), (2.4) is:

$$\text{find } u \in H_0^1(\Omega) \text{ such that } a(u,v) = l(v) \quad \forall v \in H_0^1(\Omega), \tag{2.6}$$

where the bilinear functional $a(\cdot,\cdot)$ and the linear functional $l(\cdot)$ are defined by

$$a(u,v) = \sum_{i,j=1}^{n} \int_\Omega a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} \,\mathrm{d}x + \sum_{i=1}^{n} \int_\Omega b_i(x) \frac{\partial u}{\partial x_i} v \,\mathrm{d}x + \int_\Omega c(x) uv \,\mathrm{d}x,$$

and

$$l(v) = \int_\Omega f(x) v(x) \,\mathrm{d}x.$$

We have shown that if

$$c(x) - \frac{1}{2} \sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i} \geq 0, \quad x \in \bar{\Omega},$$

then (2.6) has a unique solution $u$ in $H_0^1(\Omega)$, the weak solution of (2.3), (2.4). In the special case when the boundary value problem is self-adjoint, i.e.

$$a_{ij}(x) = a_{ji}(x), \quad i,j = 1,\ldots,n, \;\; x \in \bar{\Omega},$$

and

$$b_i(x) \equiv 0, \quad i = 1,\ldots,n, \;\; x \in \bar{\Omega},$$

the bilinear functional $a(\cdot,\cdot)$ is symmetric in the sense that

$$a(v,w) = a(w,v) \quad \forall v, w \in H_0^1(\Omega);$$

in the rest of this section this will always be assumed to be the case. Thus we consider

$$\begin{aligned} -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j} \left( a_{ij}(x) \frac{\partial u}{\partial x_i} \right) + c(x) u &= f(x), \quad x \in \Omega, \\ u &= 0, \quad \text{on} \quad \partial\Omega \end{aligned} \tag{2.7}$$

with $a_{ij}(x)$ satisfying the ellipticity condition (2.5); $a_{ij}(x) = a_{ji}(x)$, $c(x) \geq 0$, $x \in \bar{\Omega}$.

It turns out that (2.7) can be restated as a minimisation problem. To be more precise, we define the quadratic functional $J : H_0^1(\Omega) \to \mathbb{R}$ by

$$J(v) = \frac{1}{2} a(v,v) - l(v), \quad v \in H_0^1(\Omega).$$

**Lemma 3** *Let $u$ be the (unique) weak solution to (2.6) in $H_0^1(\Omega)$ and suppose that $a(\cdot,\cdot)$ is a symmetric bilinear functional on $H_0^1(\Omega)$; then $u$ is the unique minimiser of $J(\cdot)$ over $H_0^1(\Omega)$.*

**Proof** Let $u$ be the unique weak solution to (2.6) in $H_0^1(\Omega)$ and, for $v \in H_0^1(\Omega)$, consider $J(v) - J(u)$:

$$\begin{aligned} J(v) - J(u) &= \frac{1}{2}a(v,v) - l(v) - \frac{1}{2}a(u,u) + l(u) \\ &= \frac{1}{2}a(v,v) - \frac{1}{2}a(u,u) - l(v-u) \\ &= \frac{1}{2}a(v,v) - \frac{1}{2}a(u,u) - a(u,v-u) \\ &= \frac{1}{2}[a(v,v) - 2a(u,v) + a(u,u)] \\ &= \frac{1}{2}[a(v,v) - a(u,v) - a(v,u) + a(u,u)] \\ &= \frac{1}{2}a(v-u,v-u). \end{aligned}$$

Thence

$$J(v) - J(u) = \frac{1}{2}a(v-u,v-u).$$

Because of (1.17),

$$a(v-u,v-u) \geq c_0\|v-u\|^2_{H^1(\Omega)},$$

where $c_0$ is a positive constant. Thus

$$J(v) - J(u) \geq \frac{c_0}{2}\|v-u\|^2_{H^1(\Omega)} \quad \forall v \in H_0^1(\Omega), \tag{2.8}$$

and therefore,

$$J(v) \geq J(u) \quad \forall v \in H_0^1(\Omega), \tag{2.9}$$

i.e. $u$ minimises $J(\cdot)$ over $H_0^1(\Omega)$.

In fact, $u$ is the unique minimiser of $J(\cdot)$ in $H_0^1(\Omega)$. Indeed, if $\tilde{u}$ also minimises $J(\cdot)$ on $H_0^1(\Omega)$, then

$$J(v) \geq J(\tilde{u}) \quad \forall v \in H_0^1(\Omega). \tag{2.10}$$

Taking $v = \tilde{u}$ in (2.9) and $v = u$ in (2.10), we deduce that

$$J(u) = J(\tilde{u});$$

but then, by virtue of (2.8),

$$\|\tilde{u} - u\|_{H^1(\Omega)} = 0,$$

and hence $u = \tilde{u}$. ∎

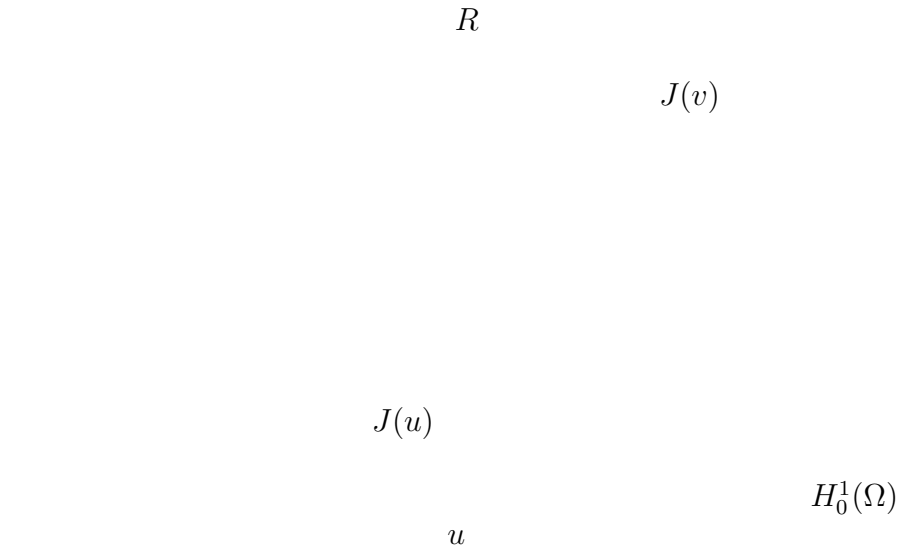


Figure 2.7: The quadratic functional $J(\cdot)$.

It is easily shown that $J(\cdot)$ is convex (down), i.e.

$$J((1-\theta)v + \theta w) \leq (1-\theta)J(v) + \theta J(w) \quad \forall \theta \in [0,1], \quad \forall v, w \in H_0^1(\Omega).$$

This follows from the identity

$$(1-\theta)J(v) + \theta J(w) = J((1-\theta)v + \theta w) + \frac{1}{2}\theta(1-\theta)a(v-w, v-w)$$

and the fact that $a(v-w, v-w) \geq 0$. Moreover, if $u$ minimises $J(\cdot)$ then $J(\cdot)$ has a stationary point at $u$; namely,

$$J'(u)v := \lim_{\lambda \to 0} \frac{J(u+\lambda v) - J(u)}{\lambda} = 0$$

for all $v \in H_0^1(\Omega)$. Since

$$\frac{J(u+\lambda v) - J(u)}{\lambda} = a(u,v) - l(v) + \frac{\lambda}{2}a(v,v),$$

we deduce that if $u$ minimises $J(\cdot)$ then

$$\lim_{\lambda \to 0}[a(u,v) - l(v) + \frac{\lambda}{2}a(v,v)] = a(u,v) - l(v) = 0 \quad \forall v \in H_0^1(\Omega),$$

which proves the following result.

**Lemma 4** *Suppose that $u \in H_0^1(\Omega)$ minimises $J(\cdot)$ over $H_0^1(\Omega)$; then $u$ is the (unique) solution of problem (2.6). The problem (2.6) is called the* **Euler–Lagrange equation** *for this minimisation problem.*

This lemma is precisely the converse of the previous lemma, and the two results together express the equivalence of the weak formulation:

$$\text{find } u \in H_0^1(\Omega) \text{ such that } a(u,v) = l(v) \quad \forall v \in H_0^1(\Omega) \tag{$W$}$$

of the self-adjoint elliptic boundary value problem (2.7) to the associated minimisation problem:

$$\text{find } u \in H_0^1(\Omega) \text{ such that } J(u) \leq J(v) \quad \forall v \in H_0^1(\Omega). \tag{$M$}$$

We shall now use this equivalence to give a variational characterisation of the finite element approximation $u_h$ to $u$ in the self-adjoint case. Given that $V_h$ is a certain finite-dimensional subspace of $H_0^1(\Omega)$ which consists of continuous piecewise polynomials of a fixed degree, the finite element approximation of $(W)$ is:.

$$\text{find } u_h \in V_h \text{ such that } a(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h. \tag{$W_h$}$$

We can repeat the argument presented above (or simply replacing $H_0^1(\Omega)$ by $V_h$ throughout) to show the equivalence of $(W_h)$ to the following minimisation problem:

$$\text{find } u_h \in V_h \text{ such that } J(u_h) \leq J(v_h) \quad \forall v_h \in V_h. \tag{$M_h$}$$

Thus, $u_h$ can be characterised as the unique minimiser of the functional

$$J(v_h) = \frac{1}{2}a(v_h, v_h) - l(v_h)$$

as $v_h$ ranges over the finite element space $V_h$. This means that the finite element solution $u_h$ inherits the energy minimisation property possessed by the weak solution $u \in H_0^1(\Omega)$ in the sense that:

$$J(u_h) = \min_{v_h \in V_h} J(v_h).$$

Of course, in general $J(u) < J(u_h)$.

## 2.3 Calculation and assembly of stiffness matrix

Using the variational characterisation of $u_h$ described at the end of the previous section we return to the construction of the finite element approximation to Poisson's equation $-\Delta u = f$ in $\Omega$ subject to homogeneous Dirichlet boundary condition, $u = 0$ on $\partial\Omega$, in the case of a general triangulation. Rather than restricting ourselves to the special case when $\Omega$ is a square, we now suppose that $\Omega$ is a bounded polygonal

domain in the plane, subdivided into $M$ triangles $K$, so that any pair of (closed) triangles intersect only along a complete edge, at a vertex or not at all. We consider the set of all continuous piecewise linear functions $v_h$ defined on such a triangulation with the property that $v_h = 0$ of $\partial\Omega$; the linear space consisting of all such functions $v_h$ is denoted $V_h$. Thus, $u_h$ is characterised as the unique minimiser of the functional

$$J(v_h) = \frac{1}{2}\int_\Omega |\nabla v_h(x,y)|^2 \,\mathrm{d}x\,\mathrm{d}y - \int_\Omega f(x,y)v_h(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

as $v_h$ ranges over $V_h$. Equivalently, writing

$$v_h(x,y) = \sum_{i=1}^{N} V_i\phi_i(x,y),$$

where $V_i$ is the value of $v_h(x,y)$ at the node $(x_i,y_i)$, $\phi_i$ is the continuous piecewise linear basis function associated with this node, and $N$ is the number of nodes internal to $\Omega$, we can write this minimisation problem in matrix form as follows:

$$\text{find } V \in \mathbb{R}^N \text{ such that } \tfrac{1}{2}V^TAV - V^TF \text{ is minimum,} \tag{2.11}$$

where $V = (V_1,\ldots,V_N)^T$, $A$ is the **(global) stiffness matrix** - an $N\times N$ matrix with $(i,j)$ entry

$$a(\phi_i,\phi_j) = (\nabla\phi_i,\nabla\phi_j) = \int_\Omega \nabla\phi_i(x,y)\cdot\nabla\phi_j(x,y)\,\mathrm{d}x\,\mathrm{d}y,$$

and $F = (F_1,\ldots,F_N)^T$ is the **(global) load vector**, with

$$F_i = (f,\phi_i) = \int_\Omega f(x,y)\phi_i(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Consider any triangle $K$ in the triangulation of $\Omega$, and introduce the position vectors $\mathbf{r}_i = (x_i,y_i)$, $i = 1,2,3$, of its three vertices labelled in the anti-clockwise direction, say. In addition, we consider a so-called local $(\xi,\eta)$ coordinate system and the canonical triangle depicted in Figure 2.8. The coordinate $\mathbf{r} = (x,y)$ of any point in the triangle $K$ can be written as a convex combination of the coordinates of the three vertices:

$$\begin{aligned}\mathbf{r} &= (1-\xi-\eta)\mathbf{r}_1 + \xi\mathbf{r}_2 + \eta\mathbf{r}_3 \\ &\equiv \mathbf{r}_1\psi_1(\xi,\eta) + \mathbf{r_2}\psi_2(\xi,\eta) + \mathbf{r}_3\psi_3(\xi,\eta).\end{aligned} \tag{2.12}$$

The set $\{\psi_1,\psi_2,\psi_3\}$ is called the **nodal basis** (or local basis) for the set of linear polynomials in terms of the local coordinates. Consider the transformation $(\xi,\eta)\mapsto\mathbf{r} = (x,y)$ defined by (2.12) from the canonical triangle to the 'global' $(x,y)$ coordinate system. The Jacobi matrix $J$ of this transformation is given by

$$J = \frac{\partial(x,y)}{\partial(\xi,\eta)} = \begin{bmatrix} x_2-x_1 & y_2-y_1 \\ x_3-x_1 & y_3-y_1 \end{bmatrix}$$

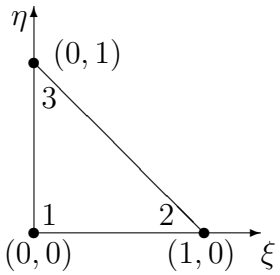


Figure 2.8: Canonical triangle and local coordinates.

from which it follows that the Jacobian is

$$|J| = \det \begin{bmatrix} x_2 - x_1 & y_2 - y_1 \\ x_3 - x_1 & y_3 - y_1 \end{bmatrix} = \det \begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix}, \tag{2.13}$$

namely,

$$|J| = 2A_{123}$$

where $A_{123}$ is the area of the triangle $K = \Delta(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$. Similarly, for any function $v_h \in V_h$,

$$v_h(x, y) = v_h(\mathbf{r}(\xi, \eta)) = V_1\psi_1(\xi, \eta) + V_2\psi_2(\xi, \eta) + V_3\psi_3(\xi, \eta), \tag{2.14}$$

where $V_i$ is the value of $v_h$ at the node of the triangle $K$ with position vector $\mathbf{r}_i$, $i = 1, 2, 3$. In order to determine the entries of the stiffness matrix, we need the gradient of $v_h$ in the global coordinate system; however, from (2.12) and the form of the Jacobi matrix $J$ we have that

$$\begin{bmatrix} \frac{\partial v_h}{\partial \xi} \\ \\ \frac{\partial v_h}{\partial \eta} \end{bmatrix} = J \begin{bmatrix} \frac{\partial v_h}{\partial x} \\ \\ \frac{\partial v_h}{\partial y} \end{bmatrix}, \qquad \begin{bmatrix} \frac{\partial v_h}{\partial x} \\ \\ \frac{\partial v_h}{\partial y} \end{bmatrix} = J^{-1} \begin{bmatrix} \frac{\partial v_h}{\partial \xi} \\ \\ \frac{\partial v_h}{\partial \eta} \end{bmatrix}. \tag{2.15}$$

Consequently,

$$\begin{aligned} \frac{\partial v_h}{\partial x} &= \frac{1}{|J|}\left[(y_3 - y_1)\frac{\partial v_h}{\partial \xi} - (y_2 - y_1)\frac{\partial v_h}{\partial \eta}\right] \\ \frac{\partial v_h}{\partial y} &= \frac{1}{|J|}\left[-(x_3 - x_1)\frac{\partial v_h}{\partial \xi} + (x_2 - x_1)\frac{\partial v_h}{\partial \eta}\right]. \end{aligned} \tag{2.16}$$

Hence

$$\begin{aligned} |J|^2|\nabla v_h|^2 &= |\mathbf{r}_3 - \mathbf{r}_1|^2\left(\frac{\partial v_h}{\partial \xi}\right)^2 + |\mathbf{r}_2 - \mathbf{r}_1|^2\left(\frac{\partial v_h}{\partial \eta}\right)^2 \\ &\quad -2(\mathbf{r}_3 - \mathbf{r}_1)\cdot(\mathbf{r}_2 - \mathbf{r}_1)\frac{\partial v_h}{\partial \xi}\frac{\partial v_h}{\partial \eta} \end{aligned} \tag{2.17}$$

and from (2.14) and (2.12) it follows that

$$\frac{\partial v_h}{\partial \xi} = V_2 - V_1, \qquad \frac{\partial v_h}{\partial \eta} = V_3 - V_1. \tag{2.18}$$

As $v_h(x, y)$ is linear on each triangle $K$ in the triangulation, $\nabla v_h$ is constant on $K$ so the contribution to

$$\int_\Omega |\nabla v_h(x,y)|^2 \,\mathrm{d}x\,\mathrm{d}y = \sum_K \int_K |\nabla v_h(x,y)|^2 \,\mathrm{d}x\,\mathrm{d}y$$

from triangle $K$ is

$$\int_K |\nabla v_h(x,y)|^2 \,\mathrm{d}x\,\mathrm{d}y = A_{123}|\nabla v_h|^2 = \frac{1}{2}|J||\nabla v_h|^2 = \frac{1}{4A_{123}}|J|^2|\nabla v_h|^2.$$

Substitution of (2.17) and (2.18) into this formula yields a quadratic form in the nodal values $V_1$, $V_2$, $V_3$; after a little algebra, we find that the coefficient of $V_1^2$ is

$$|\mathbf{r}_3 - \mathbf{r}_1|^2 + |\mathbf{r}_2 - \mathbf{r}_1|^2 - 2(\mathbf{r}_3 - \mathbf{r}_1)\cdot(\mathbf{r}_2 - \mathbf{r}_1) = |\mathbf{r}_3 - \mathbf{r}_2|^2$$

and the coefficient of $V_1V_2$ is

$$-2|\mathbf{r}_3 - \mathbf{r}_1|^2 + 2(\mathbf{r}_3 - \mathbf{r}_1)\cdot(\mathbf{r}_2 - \mathbf{r}_1) = 2(\mathbf{r}_2 - \mathbf{r}_3)\cdot(\mathbf{r}_3 - \mathbf{r}_1)$$

with similar expressions for the coefficients of $V_2^2$, $V_3^2$ and $V_2V_3$, $V_3V_1$, obtained by cyclic permutations of the indices in these expressions, respectively. Thus we deduce that

$$\int_K |\nabla v_h(x,y)|^2 \,\mathrm{d}x\,\mathrm{d}y = [V_1, V_2, V_3]A^k \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix},$$

where $k \in \{1, \ldots, M\}$ is the number of the triangle $K$ in the global numbering and $A^k$ is the symmetric $3 \times 3$ **element stiffness matrix**:

$$A^k = \frac{1}{4A_{123}} \begin{bmatrix} |\mathbf{r}_2 - \mathbf{r}_3|^2 & (\mathbf{r}_2 - \mathbf{r}_3)\cdot(\mathbf{r}_3 - \mathbf{r}_1) & (\mathbf{r}_2 - \mathbf{r}_3)\cdot(\mathbf{r}_1 - \mathbf{r}_2) \\ & |\mathbf{r}_3 - \mathbf{r}_1|^2 & (\mathbf{r}_3 - \mathbf{r}_1)\cdot(\mathbf{r}_1 - \mathbf{r}_2) \\ \text{symm.} & & |\mathbf{r}_1 - \mathbf{r}_2|^2 \end{bmatrix}.$$

Assembly of the global stiffness matrix entails relating the local numbering of the nodes to the global numbering system. Let us denote by $N$ the number of nodes internal to $\Omega$; as

$$u_h(x,y) = \sum_{i=1}^{N} U_i \phi_i(x,y),$$

$N$ is precisely the number of unknowns: $U_1, \ldots, U_N$. Let us label by $N+1$, $N+2, \ldots N^*$ the nodes that lie on the boundary of $\Omega$ (thus $N^*$ is the total number of nodes of which $N$ are internal and $N^* - N$ are on the boundary). As $u_h = 0$ on the

boundary, we can adopt the notational convention that $U_{N+1} = U_{N+2} = \ldots U_{N^*} = 0$, and write

$$u_h(x,y) = \sum_{i=1}^{N^*} U_i \phi_i(x,y),$$

with the understanding that the coefficients $U_j$, $j = N+1, \ldots, N^*$ are, in fact, known (to be zero) from the boundary condition.

For the $k$th triangle $K$, we consider the Boolean matrix[1] $L^k$ of size $N^* \times 3$ whose entries are defined as follows: if in calculating the matrix $A^k$ the node with position vector $\mathbf{r}_1$ is the $i$th node in the global numbering, $i \in \{1, \ldots, N, \ldots, N^*\}$, then the first column of $L^k$ has unit entry in the $i$th row; similarly, the second and third column depend on the global numbering of the nodes with position vectors $\mathbf{r}_2$ and $\mathbf{r}_3$ appearing in the matrix $A^k$. Then, the so called **full stiffness matrix** $A^*$ is an $N^* \times N^*$ matrix defined as a sum over the elements $K$ in the triangulation of the domain:

$$A^* = \sum_{k=1}^{M} L^k A^k (L^k)^T,$$

where $(L^k)^T$ is the transpose of the matrix $L^k$.

When programming this, instead of working with $M$ Boolean arrays $L^k$, $k = 1, \ldots, M$, it is more economical to store the information contained in the arrays $L^k$ in a single **connectivity array** $LNODS$ which has dimension $M \times 3$, where $M$ is the number of triangles in the triangulation of $\Omega$; $LNODS(k,j) \in \{1, \ldots, N^*\}$ is equal to the global number of the node $r_j$ in the $k$th triangle. By letting $k = 1, \ldots, M$, we loop through all the triangles in the triangulation of $\Omega$, and calculate $A^k_{ij}$ for $i, j = 1, 2, 3$ from the formula for $A^k$ given above; once the value $A^k_{ij}$ has been calculated it is added into the full stiffness matrix $A^*$ at position $(LNODS(k,i), LNODS(k,j))$. The **full load vector** $F^* = (F_1, \ldots, F_N, \ldots F_{N^*})^T$ is built up in the same way.

Once $A^*$ and $F^*$ have been found, we erase the last $N^* - N$ rows and columns of $A^*$ to obtain the **global stiffness matrix** $A$, and the last $N^* - N$ entries of $F^*$ to obtain to **global load vector** $F$, and then solve the linear system

$$AU = F$$

to determine the vector of unknowns $U = (U_1, \ldots, U_N)^T$.

In order to justify more clearly the compression of $A^*$ to $A$ and $F^*$ to $F$, let us note that the minimisation problem (2.11) can be restated in the following equivalent form:

$$\text{find } V^* = (V_1, \ldots, V_N, 0, \ldots, 0)^T \in \mathbb{R}^{N^*} \text{ such that } \tfrac{1}{2} V^{*T} A^* V^* - V^{*T} F^* \text{ is minimum.} \tag{2.19}$$

Since the last $N^* - N$ entries of $V^*$ are equal to 0, the last $N^* - N$ rows and columns of $A^*$ and the last $N^* - N$ entries of $F^*$ can be discarder since they are all multiplied

[1] i.e. a matrix whose entries are 0s and 1s

by entries of $V^*$ that are equal to zero. Even though it may seem that we are doing unnecessary work when computing entries of $A^*$ and $F^*$ which are then thrown away when $A^*$ is compressed to $A$ and $F^*$ is compressed to $F$, the assembly of $A^*$ and $F^*$, followed by compression, is typically a faster process than the direct assembly of $A$ and $F$, since in the latter case special care has to be taken for nodes which belong to triangles with at least one boundary point, leading to a slower assembly process. No such difficulties arise when we work with $A^*$ and $F^*$.

It is worth noting that in practice it is not essential that the first $N$ indices in the set $\{1, \ldots, N^*\}$ correspond to the interior nodes and the last $N^* - N$ to the boundary nodes: indeed, the nodes may be numbered in any order; the only thing that matters is that rows and columns of $A^*$ and entries of $F^*$ corresponding to boundary nodes are discarded when $A$ and $F$ are formed. Here we have chosen the last $N^* - N$ nodes of a total of $N^*$ to be those on the boundary simply for ease of presentation.

## 2.4 Galerkin orthogonality; Céa's lemma

Having described the construction of the finite element method, we now outline the basic tools for its error analysis. Let us consider the elliptic boundary value problem

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left(a_{ij}(x)\frac{\partial u}{\partial x_i}\right) + \sum_{i=1}^{n} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f(x),\ x \in \Omega, \tag{2.20}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{2.21}$$

where $\Omega$ is a bounded open set in $\mathbb{R}^n$, $a_{ij} \in L_\infty(\Omega)$, $i, j = 1, \ldots, n$; $b_i \in W^1_\infty(\Omega)$, $i = 1, \ldots, n$, $c \in L_\infty(\Omega)$, $f \in L_2(\Omega)$, and assume that there exists a positive constant $\tilde{c}$ such that

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \tilde{c}\sum_{i=1}^{n}\xi_i^2 \qquad \forall \xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n,\ \ \forall x \in \bar{\Omega}. \tag{2.22}$$

The weak formulation of (2.20), (2.21) is:

$$\text{find } u \in H_0^1(\Omega) \text{ such that } a(u, v) = l(v) \quad \forall v \in H_0^1(\Omega), \tag{2.23}$$

where the bilinear functional $a(\cdot, \cdot)$ and the linear functional $l(\cdot)$ are defined by

$$a(u, v) = \sum_{i,j=1}^{n} \int_\Omega a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}\,\mathrm{d}x + \sum_{i=1}^{n}\int_\Omega b_i(x)\frac{\partial u}{\partial x_i}v\,\mathrm{d}x + \int_\Omega c(x)uv\,\mathrm{d}x,$$

and

$$l(v) = \int_\Omega f(x)v(x)\,\mathrm{d}x.$$

We have shown that if

$$c(x) - \frac{1}{2}\sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i} \geq 0, \quad x \in \bar{\Omega},$$

then (2.23) has a unique solution $u$ in $H_0^1(\Omega)$, the weak solution of (2.20), (2.21). Moreover,

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{c_0}\|f\|_{L_2(\Omega)},$$

where $c_0$ is as in (1.17).

Now suppose that $V_h$ is a finite-dimensional subspace of $H_0^1(\Omega)$, without making further precise assumptions on the nature of $V_h$ (although we shall implicitly assume that $V_h$ consists of continuous piecewise polynomials defined on a subdivision of "fineness" $h$ of the computational domain $\Omega$). The finite element approximation of (2.23) is:

$$\text{find } u_h \text{ in } V_h \text{ such that } a(u_h, v_h) = l(v_h) \text{ for all } v_h \in V_h. \tag{2.24}$$

As, by hypothesis, $V_h$ is contained in $H_0^1(\Omega)$ it follows from the Lax-Milgram theorem that (2.24) has a unique solution $u_h$ in $V_h$. Moreover, (2.23) holds for any $v = v_h \in V_h$; namely,

$$a(u, v_h) = l(v_h) \qquad \text{for all } v_h \in V_h.$$

Subtracting (2.24) from this identity we deduce that

$$a(u - u_h, v_h) = 0 \qquad \text{for all } v_h \in V_h. \tag{2.25}$$

The property (2.25) is referred to as **Galerkin orthogonality** and will be seen to play a crucial role in the error analysis of finite element methods. Since by (1.17), with $v = u - u_h \in H_0^1(\Omega)$ we have that

$$\|u - u_h\|_{H^1(\Omega)}^2 \leq \frac{1}{c_0} a(u - u_h, u - u_h),$$

it follows from (2.25) that

$$\|u - u_h\|_{H^1(\Omega)}^2 \leq \frac{1}{c_0} a(u - u_h, u - v_h);$$

further, by (1.13),

$$a(u - u_h, u - v_h) \leq c_1 \|u - u_h\|_{H^1(\Omega)} \|u - v_h\|_{H^1(\Omega)}.$$

Combining the last two inequalities, we deduce that

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{c_1}{c_0}\|u - v_h\|_{H^1(\Omega)} \qquad \text{for all } v_h \in V_h.$$

Thus we have proved the following result.

**Lemma 5** (Céa's lemma) *The finite element approximation $u_h$ to $u \in H_0^1(\Omega)$, the weak solution to the problem (2.20), (2.21), is the near-best fit to $u$ in the norm $\|\cdot\|_{H^1(\Omega)}$; i.e.,*

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{c_1}{c_0} \min_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}.$$

**Remark 5** *We shall prove in the next chapter that, for a typical finite element space $V_h$,*

$$\min_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \leq C(u) h^s$$

*where $C(u)$ is a positive constant, dependent on the smoothness of $u$, $h$ is the mesh-size parameter (the maximum diameter of elements in the subdivision of the computational domain) and $s$ is a positive real number, dependent on the smoothness of $u$ and the degree of piecewise polynomials comprising the space $V_h$. Hence, with the aid of Céa's lemma we shall be able to deduce that*

$$\|u - u_h\|_{H^1(\Omega)} \leq C(u) \left( \frac{c_1}{c_0} \right) h^s \tag{2.26}$$

*which is a bound of the global error $e_h = u - u_h$ in terms of the mesh-size parameter $h$. Such a bound on the global error is called an* a priori *error bound (the terminology stems from the fact that (2.26) can be stated* prior *to computing $u_h$). It shows, in particular, that as $h \to 0$ when refining the subdivision further and further, the sequence of finite element solutions $\{u_h\}_h$ converges to $u$ in the $H^1(\Omega)$ norm. While this result is reassuring from the theoretical point of view, it is of little practical relevance as the constant $C(u)$ involved in (2.26) is difficult to quantify (given that it depends on the unknown analytical solution $u$). Later on we shall discuss* a posteriori *error bounds which make explicit use of the computed solution $u_h$ and provide computable bounds on the global error.* $\diamond$

**Example 6** *In this example we highlight a further point concerning the* a priori *error bound (2.26): for certain elliptic problems the ratio $c_1/c_0$ can be very large, and then the mesh-size $h$ has to be taken extremely small before any reduction in the size of the global error is observed. Suppose that $\Omega$ is a bounded open set in $\mathbb{R}^n$. Consider the following boundary value problem:*

$$\begin{aligned} -\varepsilon \Delta u + \mathbf{b} \cdot \nabla u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned}$$

*where $\varepsilon > 0$, $\mathbf{b} = (b_1, \ldots, b_n)^T$, with $b_i \in W_\infty^1(\Omega)$ for $i = 1, \ldots, n$. For the sake of simplicity, we shall suppose that* div $b \leq 0$ *almost everywhere on $\Omega$. Such problems arise in the mathematical modelling of advection-diffusion phenomena. When advection dominates diffusion the so-called Péclet number*

$$Pe = \frac{\left( \sum_{i=1}^n \|b_i\|_{L_\infty(\Omega)}^2 \right)^{1/2}}{\varepsilon}$$

*is very large (say, of the order* $10^6$ *to* $10^8$*).*

*A simple calculation shows that for the present problem*

$$c_1 = \left( \varepsilon^2 + \sum_{i=1}^{n} \|b_i\|^2_{L_\infty(\Omega)} \right)^{1/2}$$

*and*

$$c_0 = \frac{\varepsilon}{(1 + c_\star^2)^{1/2}}.$$

*Therefore*

$$\frac{c_1}{c_0} = (1 + c_\star^2)^{1/2}(1 + Pe^2)^{1/2},$$

*and (2.26) gives*

$$\|u - u_h\|_{H^1(\Omega)} \leq (1 + c_\star^2)^{1/2}(1 + Pe^2)^{1/2} C(u) h^s. \tag{2.27}$$

*Thus, when* $\varepsilon << 1$*, the constant on the right-hand side in this error bound is made very large through the presence of the Peclet number; in fact, things are even worse: the constant* $C(u)$ *also depends on* $\varepsilon$ *through* $u$ *(typically* $C(u) >> 1$ *when* $\varepsilon << 1$*).*

*We shall not consider the finite element approximation of advection-dominated diffusion problems any further. The point that we wish to make is merely that care should be taken when attempting to draw practically relevant conclusions from theoretical results of the kind (2.26). As it happens, the poor quality of the* a priori *error bound (2.27) when* $Pe >> 1$ *is merely a reflection of the fact that for advection-dominated diffusion equations conventional finite element methods are genuinely badly behaved: on coarse meshes the numerical solution exhibits large unphysical oscillations which can only be eliminated by severely reducing the mesh-size* $h$*.* $\diamond$

In order to put this example into perspective, we now discuss the other extreme case, when $\mathbf{b} \equiv 0$ on $\Omega$: then $c_1 = c_0 = \varepsilon$, so Céa's lemma implies that

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

In fact, since the left-hand side of this inequality cannot be strictly less than the right-hand side (this can be seen by choosing $v_h = u_h$ on the right), it follows that

$$\|u - u_h\|_{H^1(\Omega)} = \min_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)},$$

so that $u_h$ is the best approximation to $u$ from $V_h$ in the $H^1(\Omega)$ norm. We shall show that a result of this kind holds in a slightly more general setting, when the problem is self-adjoint, namely $a_{ij}(x) \equiv a_{ji}(x)$ for all $i, j = 1, \ldots, n$, $b_i(x) \equiv 0$ for $i = 1, \ldots, n$. Let us define

$$(v, w)_a := a(v, w), \quad v, w \in H_0^1(\Omega).$$

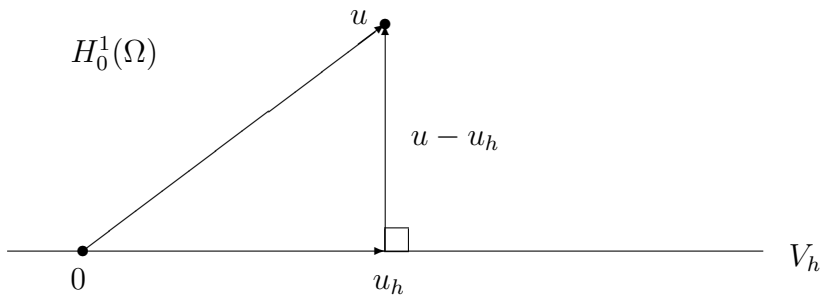


Figure 2.9: The error $u - u_h$ is orthogonal to $V_h$.

Because $a(\cdot, \cdot)$ is a symmetric bilinear functional on $H_0^1(\Omega) \times H_0^1(\Omega)$ and

$$a(v, v) \geq c_0 \|v\|^2_{H^1(\Omega)} \quad \forall v \in H_0^1(\Omega),$$

it is easily seen that $(\cdot, \cdot)_a$ satisfies all axioms of an inner product. Let $\| \cdot \|_a$ denote the associated **energy norm** defined by:

$$\|v\|_a := [a(v, v)]^{1/2}.$$

Since $V_h \subset H_0^1(\Omega)$, taking $v = v_h \in V_h$ in the statement of $(W)$, we deduce that

$$a(u, v_h) = l(v_h), \quad v_h \in V_h; \tag{2.28}$$

also by, $(W_h)$,

$$a(u_h, v_h) = l(v_h), \quad v_h \in V_h. \tag{2.29}$$

Subtracting (2.29) from (2.28) and using the fact that $a(\cdot, \cdot)$ is a bilinear functional, we deduce the Galerkin orthogonality property

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in V_h,$$

i.e.

$$(u - u_h, v_h)_a = 0 \quad \forall v_h \in V_h. \tag{2.30}$$

Thus, in the self-adjoint case, the error $u - u_h$ between the exact solution $u$ and its finite element approximation $u_h$ is orthogonal to $V_h$ in the inner product $(\cdot, \cdot)_a$ (see Figure 2.9). By virtue of the orthogonality property (2.30),

$$\begin{aligned} \|u - u_h\|_a^2 &= (u - u_h, u - u_h)_a \\ &= (u - u_h, u)_a - (u - u_h, u_h)_a \\ &= (u - u_h, u)_a \\ &= (u - u_h, u)_a - (u - u_h, v_h)_a \\ &= (u - u_h, u - v_h)_a \quad \forall v_h \in V_h. \end{aligned}$$

Thence, by the Cauchy–Schwarz inequality,

$$\begin{aligned} \|u-u_h\|_a^2 &= (u-u_h, u-v_h)_a \\ &\leq \|u-u_h\|_a \|u-v_h\|_a \quad \forall v_h \in V_h; \end{aligned}$$

therefore

$$\|u-u_h\|_a \leq \|u-v_h\|_a \quad \forall v_h \in V_h.$$

Consequently,

$$\|u-u_h\|_a = \min_{v_h \in V_h} \|u-v_h\|_a.$$

Thus we have proved the following refinement of Céa's lemma in the self-adjoint case.

**Lemma 6** *The finite element approximation $u_h \in V_h$ of $u \in H_0^1(\Omega)$ is the best fit to $u$ from $V_h$ in the energy norm $\|\cdot\|_a$, i.e.*

$$\|u-u_h\|_a = \min_{v_h \in V_h} \|u-v_h\|_a.$$

Céa's lemma is the key to the error analysis of the finite element method for elliptic boundary value problems. In the next section we describe how such an analysis proceeds in the self-adjoint case, for a particularly simple finite element space $V_h$ consisting of continuous piecewise linear functions on $\Omega$. The general case is very similar and will be considered later on in the notes.

## 2.5 Optimal error bound in the energy norm

In this section, we shall employ Céa's lemma to derive an optimal error bound for the finite element approximation $(W_h)$ of problem $(W)$ in the case of piecewise linear basis functions. We shall consider two examples: a one-dimensional model problem – a two-point boundary value problem, and a two-dimensional model problem – Poisson's equation subject to homogeneous Dirichlet boundary condition.

### One-dimensional problem

Consider, for $f \in L_2(0,1)$, the boundary value problem

$$\begin{aligned} -u'' + u &= f(x), \qquad 0 < x < 1, \\ u(0) &= 0, \quad u(1) = 0. \end{aligned}$$

Its weak formulation is:

$$\text{find } u \text{ in } H_0^1(0,1) \text{ such that } a(u,v) = l(v) \quad \forall v \in H_0^1(0,1),$$

where

$$a(u,v) = \int_0^1 (u'(x)v'(x) + u(x)v(x))\,\mathrm{d}x$$

and

$$l(v) = \int_0^1 f(x)v(x)\,\mathrm{d}x.$$

The symmetric bilinear functional $a(\cdot,\cdot)$ induces the energy norm $\|\cdot\|_a$ defined by

$$\|w\|_a = (a(w,w))^{1/2} = \left[\int_0^1 \left(|w'(x)|^2 + |w(x)|^2\right)\,\mathrm{d}x\right]^{1/2} = \|w\|_{H^1(0,1)}.$$

The finite element approximation of this problem, using piecewise linear basis functions, has been described in Section 2.1 (take $p(x) \equiv 1$ and $q(x) \equiv 1$ there to obtain the present problem). Here, instead of restricting ourselves to uniform subdivisions of $[0,1]$, we consider a general nonuniform subdivision:

$$0 = x_0 < x_1 < \ldots < x_{N-1} < x_N = 1,$$

where the mesh-points $x_i$, $i = 0,\ldots,N$, are not necessarily equally spaced. It will be supposed that $N \geq 2$ so that we have at least one mesh-point inside $(0,1)$. We put $h_i = x_i - x_{i-1}$ and define the mesh parameter $h = \max_i h_i$. For such a subdivision, we consider the finite element basis function

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

for $i = 1,\ldots,N-1$. We put

$$V_h = \mathrm{span}\{\phi_1,\ldots,\phi_{N-1}\}.$$

Clearly $V_h$ is an $(N-1)$-dimensional subspace of $H_0^1(0,1)$. We approximate the boundary value problem by the finite element method

$$\text{find } u_h \text{ in } V_h \text{ such that } a(u_h,v_h) = l(v_h) \quad \forall v_h \in V_h.$$

Now since the bilinear functional $a(\cdot,\cdot)$ is symmetric it follows from Céa's lemma that

$$\|u - u_h\|_{H^1(0,1)} = \|u - u_h\|_a = \min_{v_h \in V_h} \|u - v_h\|_a = \min_{v_h \in V_h} \|u - v_h\|_{H^1(0,1)}. \tag{2.31}$$

Let $\mathcal{I}_h u \in V_h$ denote the continuous piecewise linear function on the subdivision $\{x_0, x_1, \ldots, x_N\}$ which coincides with $u$ at the mesh-points $x_i$, $i = 0,\ldots,N$. Thus,

$$\mathcal{I}_h u(x) = \sum_{i=1}^{N-1} u(x_i)\phi_i(x).$$

The function $\mathcal{I}_h u$ is called the **interpolant** of $u$ from the finite element space $V_h$. Choosing $v_h = \mathcal{I}_h u$ in (2.31), we see that

$$\|u - u_h\|_{H^1(0,1)} \leq \|u - \mathcal{I}_h u\|_{H^1(0,1)}. \tag{2.32}$$

Thus, to derive a bound on the global error $u - u_h$ in the $H^1(0,1)$ norm, we shall now seek a bound on the interpolation error $u - \mathcal{I}_h u$ in the same norm. The rest of this subsection is devoted to the proof of the following estimate:

$$\|u - \mathcal{I}_h u\|_{H^1(0,1)} \leq \frac{h}{\pi}\left(1 + \frac{h^2}{\pi^2}\right)^{1/2} \|u''\|_{L_2(0,1)}. \tag{2.33}$$

**Theorem 3** *Suppose that $u \in H^2(0,1)$ and let $\mathcal{I}_h u$ be the interpolant of $u$ from the finite element space $V_h$ defined above; then the following error bounds hold:*

$$\begin{aligned} \|u - \mathcal{I}_h u\|_{L_2(0,1)} &\leq \left(\frac{h}{\pi}\right)^2 \|u''\|_{L_2(0,1)}, \\ \|u' - (\mathcal{I}_h u)'\|_{L_2(0,1)} &\leq \frac{h}{\pi}\|u''\|_{L_2(0,1)}. \end{aligned}$$

**Proof** Consider a subinterval $[x_{i-1}, x_i]$, $1 \leq i \leq N$, and define $\zeta(x) = u(x) - \mathcal{I}_h u(x)$ for $x \in [x_{i-1}, x_i]$. Then $\zeta \in H^2(x_{i-1}, x_i)$ and $\zeta(x_{i-1}) = \zeta(x_i) = 0$. Therefore $\zeta$ can be expanded into a convergent Fourier sine-series,

$$\zeta(x) = \sum_{k=1}^{\infty} a_k \sin \frac{k\pi(x - x_{i-1})}{h_i}, \quad x \in [x_{i-1}, x_i].$$

Hence,

$$\int_{x_{i-1}}^{x_i} [\zeta(x)]^2 \, \mathrm{d}x = \frac{h_i}{2} \sum_{k=1}^{\infty} |a_k|^2.$$

Differentiating the Fourier sine-series for $\zeta$ twice, we deduce that the Fourier coefficients of $\zeta'$ are $(k\pi/h_i)a_k$, while those of $\zeta''$ are $-(k\pi/h_i)^2 a_k$. Thus,

$$\begin{aligned} \int_{x_{i-1}}^{x_i} [\zeta'(x)]^2 \, \mathrm{d}x &= \frac{h_i}{2} \sum_{k=1}^{\infty} \left(\frac{k\pi}{h_i}\right)^2 |a_k|^2, \\ \int_{x_{i-1}}^{x_i} [\zeta''(x)]^2 \, \mathrm{d}x &= \frac{h_i}{2} \sum_{k=1}^{\infty} \left(\frac{k\pi}{h_i}\right)^4 |a_k|^2. \end{aligned}$$

Because $k^4 \geq k^2 \geq 1$, it follows that

$$\begin{aligned} \int_{x_{i-1}}^{x_i} [\zeta(x)]^2 \, \mathrm{d}x &\leq \left(\frac{h_i}{\pi}\right)^4 \int_{x_{i-1}}^{x_i} [\zeta''(x)]^2 \, \mathrm{d}x, \\ \int_{x_{i-1}}^{x_i} [\zeta'(x)]^2 \, \mathrm{d}x &\leq \left(\frac{h_i}{\pi}\right)^2 \int_{x_{i-1}}^{x_i} [\zeta''(x)]^2 \, \mathrm{d}x. \end{aligned}$$

However $\zeta''(x) = u''(x) - (\mathcal{I}_h u)''(x) = u''(x)$ for $x \in (x_{i-1}, x_i)$ because $\mathcal{I}_h u$ is a linear function on this interval. Therefore, upon summation over $i = 1, \ldots, N$ and letting $h = \max_i h_i$, we obtain

$$\begin{aligned} \|\zeta\|^2_{L_2(0,1)} &\leq \left(\frac{h}{\pi}\right)^4 \|u''\|^2_{L_2(0,1)}, \\ \|\zeta'\|^2_{L_2(0,1)} &\leq \left(\frac{h}{\pi}\right)^2 \|u''\|^2_{L_2(0,1)}. \end{aligned}$$

After taking the square root and recalling that $\zeta = u - (\mathcal{I}_h u)$ these yield the desired bounds on the interpolation error. ∎

Now (2.33) follows directly from this theorem by noting that

$$\begin{aligned} \|u - \mathcal{I}_h u\|^2_{H^1(0,1)} &= \|u - \mathcal{I}_h u\|^2_{L_2(0,1)} + \|(u - \mathcal{I}_h u)'\|^2_{L_2(0,1)} \\ &\leq \frac{h^2}{\pi^2}\left(1 + \frac{h^2}{\pi^2}\right) \|u''\|^2_{L_2(0,1)}. \end{aligned}$$

Having established the bound (2.33) on the interpolation error, we arrive at the following *a priori* error bound by inserting (2.33) into the inequality (2.32):

$$\|u - u_h\|_{H^1(0,1)} \leq \frac{h}{\pi}\left(1 + \frac{h^2}{\pi^2}\right)^{1/2} \|u''\|_{L_2(0,1)}. \tag{2.34}$$

This shows that, provided $u'' \in L_2(0,1)$, the error in the finite element solution, measured in the $H^1(0,1)$ norm, converges to 0 at the rate $\mathcal{O}(h)$ as $h \to 0$.

As a final note concerning this example, we remark that our hypothesis on $f$ (namely that $f \in L_2(0,1)$) implies that $u'' \in L_2(0,1)$. Indeed, choosing $v = u$ in the weak formulation of the boundary value problem gives

$$\begin{aligned} \int_0^1 |u'(x)|^2 \,\mathrm{d}x + \int_0^1 |u(x)|^2 \,\mathrm{d}x &= \int_0^1 f(x)u(x) \,\mathrm{d}x \\ &\leq \left(\int_0^1 |f(x)|^2 \,\mathrm{d}x\right)^{1/2} \left(\int_0^1 |u(x)|^2 \,\mathrm{d}x\right)^{1/2}. \end{aligned} \tag{2.35}$$

Hence,

$$\left(\int_0^1 |u(x)|^2 \,\mathrm{d}x\right)^{1/2} \leq \left(\int_0^1 |f(x)|^2 \,\mathrm{d}x\right)^{1/2},$$

i.e.

$$\|u\|_{L_2(0,1)} \leq \|f\|_{L_2(0,1)}.$$

Thereby, from (2.35),

$$\|u'\|_{L_2(0,1)} \leq \|f\|_{L_2(0,1)}.$$

Finally, as $u'' = u - f$ from the differential equation, we have that

$$\|u''\|_{L_2(0,1)} = \|u - f\|_{L_2(0,1)} \leq \|u\|_{L_2(0,1)} + \|f\|_{L_2(0,1)} \leq 2\|f\|_{L_2(0,1)}.$$

Thus we have proved that $u'' \in L_2(0,1)$ (in fact, as we also know that $u$ and $u'$ belong to $L_2(0,1)$, we have proved more: $u \in H^2(0,1)$). Substituting this bound on $\|u''\|_{L_2(0,1)}$ into (2.34) gives

$$\|u - u_h\|_{H^1(0,1)} \leq \frac{2h}{\pi}\left(1 + \frac{h^2}{\pi^2}\right)^{1/2} \|f\|_{L_2(0,1)}.$$

This now provides a computable upper bound on the global error $u - u_h$ in the $H^1(0,1)$ norm, since $f$ is a given function and $h = \max_i h_i$ can be easily calculated for any given subdivision of $[0,1]$.

The argument presented in this example is representative of a general finite element (*a priori*) error analysis. In a nutshell, it consisted of using:

a) Céa's lemma, together with

b) an interpolation error bound.

These two ingredients then lead us to the error bound (2.34). Finally, if we are fortunate enough to have a bound of the type $\|u''\|_{L_2(0,1)} \leq C_* \|f\|_{L_2(0,1)}$ (or, in other words, $|u|_{H^2(0,1)} \leq C_* \|f\|_{L_2(0,1)}$), which is called

c) an elliptic regularity estimate,

then, at least in principle, we obtain a computable bound on the global error. Unfortunately, for (multi-dimensional) elliptic boundary value problems proving an elliptic regularity estimate of the form

$$|u|_{H^2(\Omega)} \leq C_* \|f\|_{L_2(\Omega)} \tag{2.36}$$

is a highly non-trivial task (this issue will be touched on in the next section, in discussion about the Aubin–Nitsche duality argument). In fact, for multi-dimensional problems (2.36) will not hold unless the boundary $\partial\Omega$ and the coefficients $a_{ij}$, $b_i$ and $c$ are sufficiently smooth. Worse still, even when (2.36) holds precise estimates of the size of the constant $C_*$ are only available in rare circumstances. The upshot is that an *a priori* error bound will usually not provide a computable estimate of the global error. This is a serious drawback from the point of view of practical computations where one would like to have precise information about the size of the error between the analytical solution and its finite element approximation. Later on in the notes we shall discuss an alternative approach, *a posteriori* error analysis, which resolves this difficulty and provides computable bounds on the error in terms of $u_h$.

**Two-dimensional problem**

Let $\Omega = (0,1) \times (0,1)$, and consider the elliptic boundary value problem

$$-\Delta u = f \quad \text{in } \Omega, \tag{2.37}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{2.38}$$

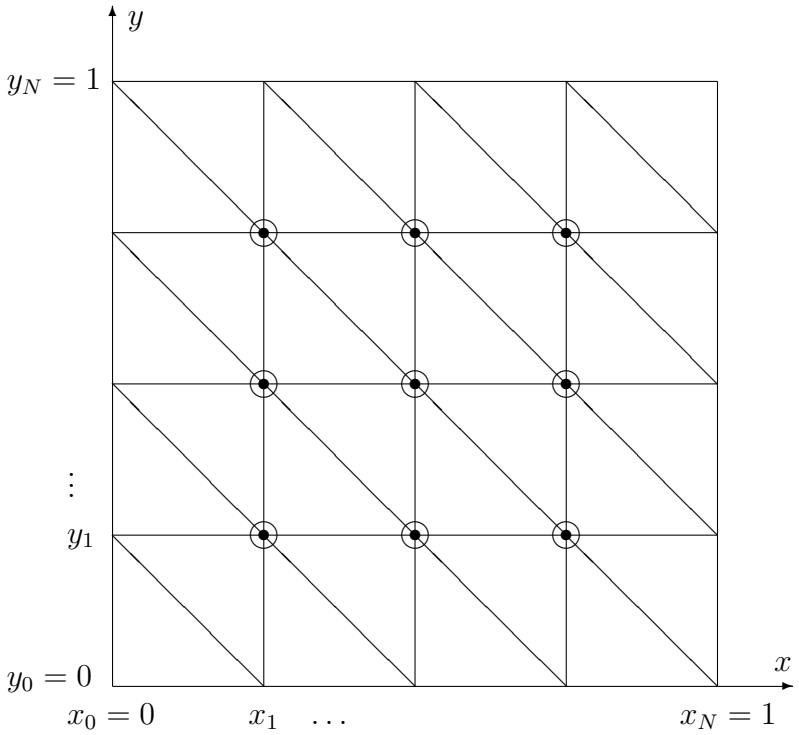


Figure 2.10: Triangulation of $\bar{\Omega} = [0,1] \times [0,1]$.

We recall that the weak formulation of this problem is:

$$\begin{aligned} &\text{find } u \in H_0^1(\Omega) \text{ such that} \\ &\int_\Omega \left( \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y \;=\; \int_\Omega f v \,\mathrm{d}x\,\mathrm{d}y \quad \forall v \in H_0^1(\Omega). \end{aligned} \tag{2.39}$$

In order to construct the finite element approximation, we triangulate the domain as shown in the Fig 2.10. Let $h = 1/N$, and define $x_i = ih$, $i = 0, \ldots, N$, $y_j = jh$, $j = 0, \ldots, N$. With each node, $(x_i, y_j)$, contained in the interior of $\Omega$ (labelled $\odot$ in the figure), we associate a basis function $\phi_{ij}$, $i, j = 1, \ldots, N-1$, defined by

$$\phi_{ij}(x,y) \;=\; \begin{cases} 1 - \frac{x-x_i}{h} - \frac{y-y_j}{h}, & (x,y) \in 1 \\ 1 - \frac{y-y_j}{h}, & (x,y) \in 2 \\ 1 - \frac{x_i-x}{h}, & (x,y) \in 3 \\ 1 - \frac{x_i-x}{h} - \frac{y_j-y}{h}, & (x,y) \in 4 \\ 1 - \frac{y_j-y}{h}, & (x,y) \in 5 \\ 1 - \frac{x-x_i}{h}, & (x,y) \in 6 \\ 0 & \text{otherwise.} \end{cases}$$

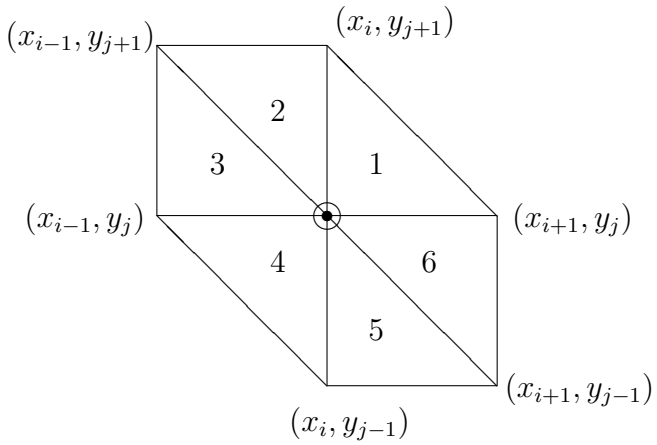


Figure 2.11: Triangles surrounding the node $(x_i, y_j)$.

Let $V_h = \text{span}\{\phi_{ij},\ i = 1, \ldots, N-1;\ j = 1, \ldots, N-1\}$. The finite element approximation of (2.37) (and (2.39)) is:

$$\begin{gathered}\text{find } u_h \in V_h \text{ such that} \\ \int_\Omega \left( \frac{\partial u_h}{\partial x}\frac{\partial v_h}{\partial x} + \frac{\partial u_h}{\partial y}\frac{\partial v_h}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y = \int_\Omega f v_h \,\mathrm{d}x\,\mathrm{d}y \quad \forall v_h \in V_h. \end{gathered} \tag{2.40}$$

Letting

$$\begin{aligned} l(v) &= \int_\Omega f(x)v(x)\,\mathrm{d}x \quad \text{and} \\ (v,w)_a = a(v,w) &= \int_\Omega \left( \frac{\partial v}{\partial x}\frac{\partial w}{\partial x} + \frac{\partial v}{\partial y}\frac{\partial w}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y, \end{aligned}$$

(2.39) and the finite element method (2.40) can be written, respectively, as follows:

$$\text{find } u \in H_0^1(\Omega) \text{ such that } a(u,v) = l(v) \quad \forall v \in H_0^1(\Omega), \tag{5.13$'$}$$

and

$$\text{find } u_h \in V_h \text{ such that } a(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h. \tag{5.14$'$}$$

According to Céa's lemma,

$$\|u - u_h\|_a = \min_{v_h \in V_h} \|u - v_h\|_a \le \|u - \mathcal{I}_h u\|_a, \tag{2.41}$$

where $\mathcal{I}_h u$ denotes the continuous piecewise linear interpolant of the function $u$ on

the set $\bar{\Omega} = [0,1] \times [0,1]$:

$$(\mathcal{I}_h u)(x,y) = \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} u(x_i, y_j) \phi_{ij}(x,y).$$

Clearly $(\mathcal{I}_h u)(x_k, y_l) = u(x_k, y_l)$. Let us estimate $\|u - \mathcal{I}_h u\|_a$:

$$\begin{aligned} \|u - \mathcal{I}_h u\|_a^2 &= \int_\Omega |\frac{\partial}{\partial x}(u - \mathcal{I}_h u)|^2 \, \mathrm{d}x \, \mathrm{d}y + \int_\Omega |\frac{\partial}{\partial y}(u - \mathcal{I}_h u)|^2 \, \mathrm{d}x \, \mathrm{d}y \\ &= \sum_{\triangle} \left\{ \int_\triangle |\frac{\partial}{\partial x}(u - \mathcal{I}_h u)|^2 \, \mathrm{d}x \, \mathrm{d}y + \int_\triangle |\frac{\partial}{\partial y}(u - \mathcal{I}_h u)|^2 \, \mathrm{d}x \, \mathrm{d}y \right\} \end{aligned} \tag{2.42}$$

where $\triangle$ is a triangle in the subdivision of $\Omega$. Suppose, for example, that

$$\triangle = \{(x,y) : x_i \le x \le x_{i+1};\ y_j \le y \le y_{j+1} + x_i - x\}.$$

In order to estimate

$$\int_\triangle |\frac{\partial}{\partial x}(u - \mathcal{I}_h u)|^2 \, \mathrm{d}x \, \mathrm{d}y + \int_\triangle |\frac{\partial}{\partial y}(u - \mathcal{I}_h u)|^2 \, \mathrm{d}x \, \mathrm{d}y,$$

we define the canonical triangle

$$K = \{(s,t) : 0 \le s \le 1,\ 0 \le t \le 1 - s\}$$

and the affine mapping $(x,y) \mapsto (s,t)$ from $\triangle$ to $K$, by

$$\begin{aligned} x &= x_i + sh, \qquad 0 \le s \le 1, \\ y &= y_j + th, \qquad 0 \le t \le 1. \end{aligned}$$

Let $\bar{u}(s,t) := u(x,y)$. Then,

$$\begin{aligned} \frac{\partial u}{\partial x} = \frac{\partial \bar{u}}{\partial s} \cdot \frac{\partial s}{\partial x} + \frac{\partial \bar{u}}{\partial t} \cdot \frac{\partial t}{\partial x} &= \frac{1}{h} \cdot \frac{\partial \bar{u}}{\partial s}, \\ \frac{\partial u}{\partial y} = \frac{\partial \bar{u}}{\partial s} \cdot \frac{\partial s}{\partial y} + \frac{\partial \bar{u}}{\partial t} \cdot \frac{\partial t}{\partial y} &= \frac{1}{h} \cdot \frac{\partial \bar{u}}{\partial t}. \end{aligned}$$

The Jacobian of the mapping $(s,t) \mapsto (x,y)$ is

$$|J| = \left| \frac{\partial(x,y)}{\partial(s,t)} \right| = \begin{vmatrix} x_s & x_t \\ y_s & y_t \end{vmatrix} = h^2.$$

Thus

$$\int_\triangle |\frac{\partial}{\partial x}(u - \mathcal{I}_h u)|^2 \, \mathrm{d}x \, \mathrm{d}y = \qquad \text{(P.T.O.)}$$

$$
\begin{aligned}
&= \int_K |\frac{\partial}{\partial s}\left(\bar{u}(s,t) - [(1-s-t)\bar{u}(0,0) + s\bar{u}(1,0) + t\bar{u}(0,1)]\right)|^2 \,\mathrm{d}s\,\mathrm{d}t \\
&= \int_0^1 \int_0^{1-s} |\frac{\partial \bar{u}}{\partial s}(s,t) - [\bar{u}(1,0) - \bar{u}(0,0)]|^2 \,\mathrm{d}s\,\mathrm{d}t \\
&= \int_0^1 \int_0^{1-s} |\frac{\partial \bar{u}}{\partial s}(s,t) - \int_0^1 \frac{\partial \bar{u}}{\partial s}(\sigma,0)\,\mathrm{d}\sigma|^2 \,\mathrm{d}s\,\mathrm{d}t \\
&= \int_0^1 \int_0^{1-s} |\int_0^1 \left(\frac{\partial \bar{u}}{\partial s}(s,t) - \frac{\partial \bar{u}}{\partial s}(\sigma,t)\right)\mathrm{d}\sigma + \int_0^1 \left(\frac{\partial \bar{u}}{\partial s}(\sigma,t) - \frac{\partial \bar{u}}{\partial s}(\sigma,0)\right)\mathrm{d}\sigma|^2 \,\mathrm{d}s\,\mathrm{d}t \\
&= \int_0^1 \int_0^{1-s} |\int_0^1 \int_\sigma^s \frac{\partial^2 \bar{u}}{\partial s^2}(\theta,t)\,\mathrm{d}\theta\,\mathrm{d}\sigma + \int_0^1 \int_0^t \frac{\partial^2 \bar{u}}{\partial s\,\partial t}(\sigma,\eta)\,\mathrm{d}\eta\,\mathrm{d}\sigma|^2 \,\mathrm{d}s\,\mathrm{d}t \\
&\le 2\int_0^1 \int_0^{1-s} \int_0^1 \int_0^1 |\frac{\partial^2 \bar{u}}{\partial s^2}(\theta,t)|^2 \,\mathrm{d}\theta\,\mathrm{d}\sigma\,\mathrm{d}s\,\mathrm{d}t \\
&\quad +2\int_0^1 \int_0^{1-s} \int_0^1 \int_0^1 |\frac{\partial^2 \bar{u}}{\partial s\,\partial t}(\sigma,\eta)|^2 \,\mathrm{d}\eta\,\mathrm{d}\sigma\,\mathrm{d}s\,\mathrm{d}t \\
&\le 2\int_0^1 \int_0^1 |\frac{\partial^2 \bar{u}}{\partial s^2}(\theta,t)|^2 \,\mathrm{d}\theta\,\mathrm{d}t + \int_0^1 \int_0^1 |\frac{\partial^2 \bar{u}}{\partial s\,\partial t}(\sigma,\eta)|^2 \,\mathrm{d}\sigma\,\mathrm{d}\eta \\
&= 2\int_{x_i}^{x_{i+1}} \int_{y_j}^{y_{j+1}} |\frac{\partial^2 u}{\partial x^2}(x,y)|^2 \cdot |h^2|^2 \cdot h^{-2} \,\mathrm{d}x\,\mathrm{d}y \\
&\quad + \int_{x_i}^{x_{i+1}} \int_{y_j}^{y_{j+1}} |\frac{\partial^2 u}{\partial x\,\partial y}(x,y)|^2 \cdot |h^2|^2 \cdot h^{-2} \,\mathrm{d}x\,\mathrm{d}y.
\end{aligned}
$$

Therefore,

$$
\int_\triangle |\frac{\partial}{\partial x}(u - \mathcal{I}_h u)|^2 \,\mathrm{d}x\,\mathrm{d}y \le 2h^2 \int_{x_i}^{x_{i+1}} \int_{y_j}^{y_{j+1}} \left(|\frac{\partial^2 u}{\partial x^2}|^2 + \frac{1}{2}|\frac{\partial^2 u}{\partial x\,\partial y}|^2\right) \mathrm{d}x\,\mathrm{d}y. \tag{2.43}
$$

Similarly,

$$
\int_\triangle |\frac{\partial}{\partial y}(u - \mathcal{I}_h u)|^2 \,\mathrm{d}x\,\mathrm{d}y \le 2h^2 \int_{x_i}^{x_{i+1}} \int_{y_j}^{y_{j+1}} \left(|\frac{\partial^2 u}{\partial y^2}|^2 + \frac{1}{2}|\frac{\partial^2 u}{\partial x\,\partial y}|^2\right) \mathrm{d}x\,\mathrm{d}y. \tag{2.44}
$$

Substituting (2.43) and (2.44) into (2.42),

$$
\|u - \mathcal{I}_h u\|_a^2 \le 4h^2 \int_\Omega \left(|\frac{\partial^2 u}{\partial x^2}|^2 + |\frac{\partial^2 u}{\partial x\,\partial y}|^2 + |\frac{\partial^2 u}{\partial y^2}|^2\right) \mathrm{d}x\,\mathrm{d}y. \tag{2.45}
$$

Finally by (2.41) and (2.45),

$$
\|u - u_h\|_a \le 2h|u|_{H^2(\Omega)}. \tag{2.46}
$$

Thus we have proved the following result.

**Theorem 4** *Let $u$ be the weak solution of the boundary value problem (2.37), and let $u_h$ be its piecewise linear finite element approximation defined by (2.40). Suppose that $u \in H^2(\Omega) \cap H_0^1(\Omega)$; then*

$$\|u - u_h\|_a \leq 2h|u|_{H^2(\Omega)}.$$

**Corollary 2** *Under the hypotheses of Theorem 4,*

$$\|u - u_h\|_{H^1(\Omega)} \leq \sqrt{5}h|u|_{H^2(\Omega)}.$$

**Proof** According to Theorem 4,

$$\|u - u_h\|_a^2 \;=\; |u - u_h|_{H^1(\Omega)}^2 \leq 4h^2|u|_{H^2(\Omega)}^2.$$

Since $u \in H_0^1(\Omega)$, $u_h \in V_h \subset H_0^1(\Omega)$, it follows that $u - u_h \in H_0^1(\Omega)$. By the Poincaré–Friedrichs inequality,

$$\|u - u_h\|_{L_2(\Omega)}^2 \leq \frac{1}{4}|u - u_h|_{H^1(\Omega)}^2; \tag{2.47}$$

thus,

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)}^2 &= \|u - u_h\|_{L_2(\Omega)}^2 + |u - u_h|_{H^1(\Omega)}^2 \\ &\leq \frac{5}{4}|u - u_h|_{H^1(\Omega)}^2 \leq 5h^2|u|_{H^2(\Omega)}^2, \end{aligned}$$

and that completes the proof. ■

From (2.47) and (2.46) we also see that

$$\|u - u_h\|_{L_2(\Omega)} \leq h\,|u|_{H^2(\Omega)}. \tag{2.48}$$

**The Aubin–Nitsche duality argument.** The error estimate (2.48) indicates that the error in the $L_2$-norm between $u$ and its finite element approximation $u_h$ is of the size $\mathcal{O}(h)$. It turns out, however, that this bound is quite pessimistic and can be improved to $\mathcal{O}(h^2)$; the proof of this is presented below.

Let us first observe that if $w \in H^2(\Omega) \cap H_0^1(\Omega)$, $\Omega = (0,1) \times (0,1)$, then

$$\begin{aligned} \|\Delta w\|_{L_2(\Omega)}^2 &= \int_\Omega \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2}\right)^2 \mathrm{d}x\,\mathrm{d}y \\ &= \int_\Omega \left(\frac{\partial^2 w}{\partial x^2}\right)^2 \mathrm{d}x\,\mathrm{d}y + 2\int_\Omega \frac{\partial^2 w}{\partial x^2}\cdot\frac{\partial^2 w}{\partial y^2}\,\mathrm{d}x\,\mathrm{d}y + \int_\Omega \left(\frac{\partial^2 w}{\partial y^2}\right)^2 \mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

Performing integration by parts and using the fact that $w = 0$ on $\partial\Omega$,

$$\begin{aligned} \int_\Omega \frac{\partial^2 w}{\partial x^2}\cdot\frac{\partial^2 w}{\partial y^2}\,\mathrm{d}x\,\mathrm{d}y &= \int_\Omega \frac{\partial^2 w}{\partial x\,\partial y}\cdot\frac{\partial^2 w}{\partial x\,\partial y}\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_\Omega \left|\frac{\partial^2 w}{\partial x\,\partial y}\right|^2 \mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

Thus

$$\begin{aligned} \|\Delta w\|^2_{L_2(\Omega)} &= \int_\Omega \left( |\frac{\partial^2 w}{\partial x^2}|^2 + 2|\frac{\partial^2 w}{\partial x\,\partial y}|^2 + |\frac{\partial^2 w}{\partial y^2}|^2 \right) \mathrm{d}x\,\mathrm{d}y \\ &= |w|^2_{H^2(\Omega)}. \end{aligned}$$

Given $g \in L_2(\Omega)$, let $w_g \in H_0^1(\Omega)$ be the weak solution of the boundary value problem

$$-\Delta w_g = g \quad \text{in } \Omega, \tag{2.49}$$

$$w_g = 0 \quad \text{on } \partial\Omega; \tag{2.50}$$

then $w_g \in H^2(\Omega) \cap H_0^1(\Omega)$, and

$$|w_g|_{H^2(\Omega)} = \|\Delta w_g\|_{L_2(\Omega)} = \|g\|_{L_2(\Omega)}. \tag{2.51}$$

After this brief preparation, we turn to the derivation of the optimal error bound in the $L_2$-norm.

According to the Cauchy–Schwarz inequality for the $L_2$-inner product $(\cdot,\cdot)$,

$$(u - u_h, g) \le \|u - u_h\|_{L_2(\Omega)} \|g\|_{L_2(\Omega)} \quad \forall g \in L_2(\Omega).$$

Therefore,

$$\|u - u_h\|_{L_2(\Omega)} = \sup_{g \in L_2(\Omega)} \frac{(u - u_h, g)}{\|g\|_{L_2(\Omega)}}. \tag{2.52}$$

For $g \in L_2(\Omega)$, the function $w_g \in H_0^1(\Omega)$ is the weak solution of the problem (2.49), so it satisfies

$$a(w_g, v) = l_g(v) \quad \forall v \in H_0^1(\Omega), \tag{2.53}$$

where

$$\begin{aligned} l_g(v) &= \int_\Omega gv \,\mathrm{d}x\,\mathrm{d}y = (g, v), \\ a(w_g, v) &= \int_\Omega \left( \frac{\partial w_g}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial w_g}{\partial y}\frac{\partial v}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

Consider the finite element approximation of (2.53):

$$\text{find } w_{gh} \in V_h \text{ such that } a(w_{gh}, v_h) = l_g(v_h) \quad \forall v_h \in V_h. \tag{2.54}$$

From (2.53), (2.54) and the error bound (2.46), we deduce that

$$\|w_g - w_{gh}\|_a \le 2h|w_g|_{H^2(\Omega)},$$

and therefore, by (2.51),

$$\|w_g - w_{gh}\|_a \leq 2h\|g\|_{L_2(\Omega)}. \tag{2.55}$$

Now

$$\begin{aligned}(u - u_h, g) &= (g, u - u_h) = l_g(u - u_h) \\ &= a(w_g, u - u_h) = a(u - u_h, w_g). \end{aligned} \tag{2.56}$$

Because $w_{gh} \in V_h$, (2.30) implies that

$$a(u - u_h, w_{gh}) = 0,$$

and therefore, by (2.56),

$$\begin{aligned}(u - u_h, g) &= a(u - u_h, w_g) - a(u - u_h, w_{gh}) \\ &= a(u - u_h, w_g - w_{gh}) \\ &= (u - u_h, w_g - w_{gh})_a.\end{aligned}$$

Applying the Cauchy–Schwarz inequality on the right,

$$(u - u_h, g) \leq \|u - u_h\|_a \|w_g - w_{gh}\|_a,$$

and thence by (2.46) and (2.55)

$$(u - u_h, g) \leq 4h^2 |u|_{H^2(\Omega)} \cdot \|g\|_{L_2(\Omega)}. \tag{2.57}$$

Substituting (2.57) into the right-hand side of (2.52), we obtain

$$\|u - u_h\|_{L_2(\Omega)} \leq 4h^2 |u|_{H^2(\Omega)},$$

which is our improved error bound in the $L_2$-norm. The proof presented above is called the **Aubin–Nitsche duality argument**.

## 2.6 Superapproximation in mesh-dependent norms

We have shown that the piecewise linear finite element approximation $u_h$ to the solution $u$ of the homogeneous Dirichlet boundary value problem for Poisson's equation obeys the following error bounds:

$$\begin{aligned}\|u - u_h\|_{H^1(\Omega)} &\leq Ch|u|_{H^2(\Omega)}, \\ \|u - u_h\|_{L_2(\Omega)} &\leq Ch^2|u|_{H^2(\Omega)},\end{aligned}$$

where $C$ denotes a generic positive constant and $h$ is the maximum element size in the subdivision. It is possible to show that these error bounds are sharp in the sense they cannot in general be improved. However, it was observed by engineers

that, when sampled at certain special points, the finite element approximation $u_h$ is more accurate than these error bounds might indicate. Indeed, we shall prove here that when measured in a discrete counterpart of the Sobolev $H^1(\Omega)$ norm, based on sampling at the mesh points, the error $u - u_h$ is $\mathcal{O}(h^2)$. A result of this kind is usually referred to as a superapproximation property.

We consider the model problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, && (2.58)\\ u &= 0 \qquad \text{on } \partial\Omega, && (2.59) \end{aligned}$$

where $\Omega = (0,1) \times (0,1)$. We showed in Section 2.1 that when using continuous piecewise linear finite elements on the uniform triangulation shown in Figure 2.5, the finite element solution $u_h(x,y)$ can be expressed in terms of the finite element basis functions $\phi_{ij}(x,y)$ as

$$u_h(x,y) = \sum_{i=1}^{N-1}\sum_{j=1}^{N-1} U_{ij}\phi_{ij}(x,y),$$

where the $U_{ij}(= u_h(x_i,y_j))$ are obtained by solving the set of difference equations

$$\begin{aligned} &-\frac{U_{i+1,j} - 2U_{ij} + U_{i-1,j}}{h^2} - \frac{U_{i,j+1} - 2U_{ij} + U_{i,j-1}}{h^2} \\ &\quad = \frac{1}{h^2}\int\int_{\text{supp } \phi_{ij}} f(x,y)\phi_{ij}(x,y)\,\mathrm{d}x\,\mathrm{d}y, \quad i,j = 1,\ldots,N-1; \\ &U_{ij} = 0 \quad \text{when } i = 0 \text{ or } i = N \text{ or } j = 0 \text{ or } j = N. \end{aligned}$$

Since $u_h(x,y) = 0$ when $(x,y) \in \partial\Omega$, we have adopted the convention that $U_{ij} = 0$ when $i = 0$ or $i = N$ or $j = 0$ or $j = N$. For simplicity, we shall write

$$F_{ij} = \frac{1}{h^2}\int\int_{\text{supp } \phi_{ij}} f(x,y)\phi_{ij}(x,y)\,\mathrm{d}x\,\mathrm{d}y, \qquad \text{for } i,j = 1,\ldots,N-1.$$

Here $N$ is an integer, $N \geq 2$, and $h = 1/N$; the mesh-points are $(x_i,y_j)$, $i,j = 0,\ldots,N$, where $x_i = ih$, $y_j = jh$. These form the finite difference mesh

$$\bar{\Omega}_h = \{(x_i,y_j) : i,j = 0,\ldots,N\}.$$

We consider the set of interior mesh points

$$\Omega_h = \{(x_i,y_j) : i,j = 1,...,N-1\},$$

and the set of boundary mesh points $\Gamma_h = \bar{\Omega}_h \setminus \Omega_h$. In more compact notation, the difference scheme can be written as follows:

$$\begin{aligned} -(D_x^+D_x^-U_{ij} + D_y^+D_y^-U_{ij}) &= F_{ij}, \quad (x_i,y_j) \in \Omega_h, && (2.60)\\ U &= 0 \qquad\quad \text{on } \Gamma_h, && (2.61) \end{aligned}$$

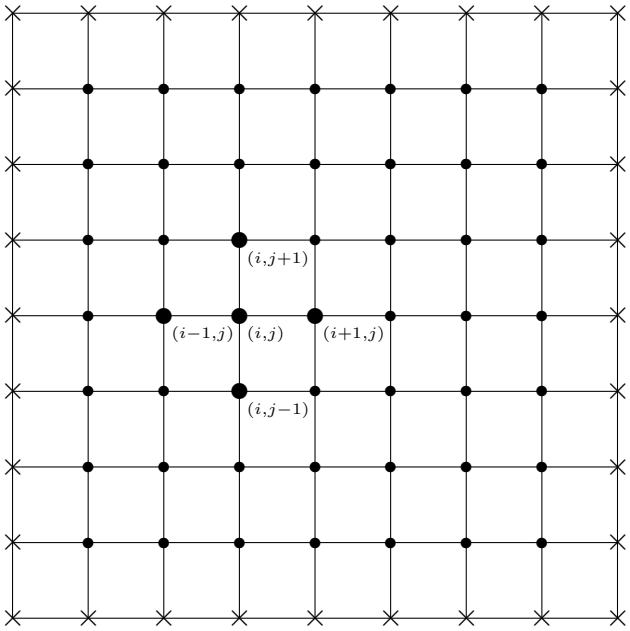


Figure 2.12: The mesh $\Omega_h(\cdot)$, the boundary mesh $\Gamma_h(\times)$, and a typical 5-point difference stencil.

where $D_x^+$ and $D_x^-$ denote the forward and backward divided difference operators in the $x$ direction, respectively, defined by

$$D_x^+ V_{ij} = \frac{V_{i+1,j} - V_{ij}}{h}, \qquad D_x^- V_{ij} = \frac{V_{ij} - V_{i-1,j}}{h},$$

and

$$D_x^+ D_x^- V_{ij} = D_x^+ (D_x^- V_{ij}) = \frac{V_{i+1,j} - 2V_{ij} + V_{i-1,j}}{h^2}$$

is the second divided difference operator in the $x$ direction. Similar definitions apply in the $y$ direction.

For each $i$ and $j$, $1 \le i, j \le N-1$, the finite difference equation (2.60) involves five values of $U$: $U_{i,j}$, $U_{i-1,j}$, $U_{i+1,j}$, $U_{i,j-1}$, $U_{i,j+1}$. It is possible to write (2.60) as a system of linear equations

$$AU = F, \tag{2.62}$$

where

$$\begin{aligned} U \;=\; & (U_{11}, U_{12}, \ldots, U_{1,N-1}, U_{21}, U_{22}, \ldots, U_{2,N-1}, \ldots, \\ & \ldots, U_{i1}, U_{i2}, \ldots, U_{i,N-1}, \ldots, U_{N-1,1}, U_{N-1,2}, \ldots, U_{N-1,N-1})^T, \end{aligned}$$

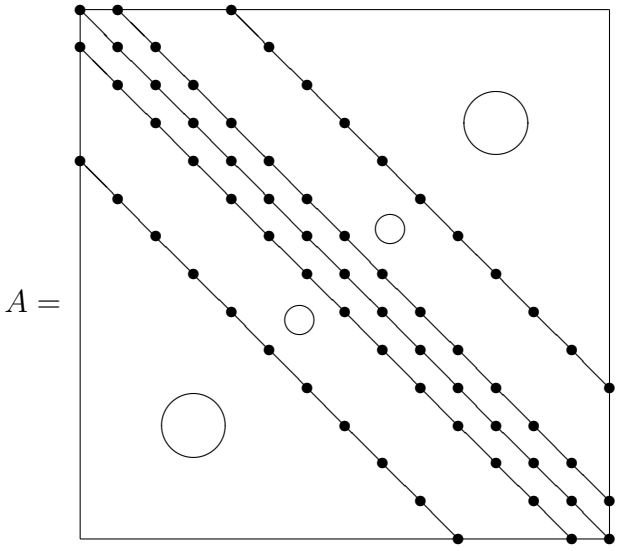


Figure 2.13: The sparsity structure of the band matrix $A$.

$$\begin{aligned} F &= (F_{11}, F_{12}, \ldots, F_{1,N-1}, F_{21}, F_{22}, \ldots, F_{2,N-1}, \ldots, \\ &\quad \ldots, F_{i1}, F_{i2}, \ldots, F_{i,N-1}, \ldots, F_{N-1,1}, F_{N-1,2}, \ldots, F_{N-1,N-1})^T, \end{aligned}$$

and $A$ is an $(N-1)^2 \times (N-1)^2$ sparse matrix of band structure. A typical row of the matrix contains five non-zero entries, corresponding to the five values of $U$ in the finite difference stencil shown in Fig. 2.12, while the sparsity structure of $A$ is depicted in Fig. 2.13.

Next we show that (2.62) has a unique solution[2]. For two functions, $V$ and $W$, defined on $\Omega_h$, we introduce the inner product

$$(V, W)_h = \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} h^2 V_{ij} W_{ij}$$

(which resembles the $L_2$-inner product $(v, w) = \int_\Omega v(x, y) w(x, y)\, \mathrm{d}x\, \mathrm{d}y$.)

**Lemma 7** *Suppose that $V$ is a function defined on $\bar{\Omega}_h$ and that $V = 0$ on $\Gamma_h$; then*

$$\begin{aligned} &(-D_x^+ D_x^- V, V)_h + (-D_y^+ D_y^- V, V)_h \\ &\quad = \sum_{i=1}^{N} \sum_{j=1}^{N-1} h^2 |D_x^- V_{ij}|^2 + \sum_{i=1}^{N-1} \sum_{j=1}^{N} h^2 |D_y^- V_{ij}|^2. \end{aligned} \tag{2.63}$$

[2]The uniqueness of the solution to the linear system (2.62) is a trivial consequence of the uniqueness of solution $u_h$ to the finite element method. The argument that follows is an alternative way of verifying uniqueness; we present it here since some of its ingredients will be exploited in the course of the proof of the superapproximation property.

**Proof** We shall prove that the first term on the left is equal to the first term on the right, and the second term on the left to the second term on the right.

$$\begin{aligned}
(-D_x^+ D_x^- V, V)_h &= -\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}(V_{i+1,j} - 2V_{ij} + V_{i-1,j})V_{ij} \\
&= -\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}(V_{i+1,j} - V_{ij})V_{ij} + \sum_{i=1}^{N-1}\sum_{j=1}^{N-1}(V_{ij} - V_{i-1,j})V_{ij} \\
&= -\sum_{i=2}^{N}\sum_{j=1}^{N-1}(V_{ij} - V_{i-1,j})V_{i-1,j} + \sum_{i=1}^{N-1}\sum_{j=1}^{N-1}(V_{ij} - V_{i-1,j})V_{ij} \\
&= -\sum_{i=1}^{N}\sum_{j=1}^{N-1}(V_{ij} - V_{i-1,j})V_{i-1,j} + \sum_{i=1}^{N}\sum_{j=1}^{N-1}(V_{ij} - V_{i-1,j})V_{ij} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N-1}(V_{ij} - V_{i-1,j})^2 = \sum_{i=1}^{N}\sum_{j=1}^{N-1} h^2 |D_x^- V_{ij}|^2.
\end{aligned}$$

Similarly,

$$(-D_y^+ D_y^- V, V)_h = \sum_{i=1}^{N-1}\sum_{j=1}^{N} h^2 |D_y^- V_{ij}|^2,$$

and that completes the proof. ∎

Returning to the analysis of the finite difference scheme (2.60), we note that by (2.63) we have

$$\begin{aligned}
(AV, V)_h &= (-D_x^+ D_x^- V - D_y^+ D_y^- V, V)_h \\
&= (-D_x^+ D_x^- V, V)_h + (-D_y^+ D_y^- V, V)_h \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N-1} h^2 |D_x^- V_{ij}|^2 + \sum_{i=1}^{N-1}\sum_{j=1}^{N} h^2 |D_y^- V_{ij}|^2, \qquad (2.64)
\end{aligned}$$

for any $V$ defined on $\bar{\Omega}_h$ such that $V = 0$ on $\Gamma_h$. Now this implies that $A$ is a non-singular matrix. Indeed if $AV = 0$, then (2.64) yields:

$$D_x^- V_{ij} = \frac{V_{ij} - V_{i-1,j}}{h} = 0, \qquad \begin{array}{l} i = 1, \ldots, N, \\ j = 1, \ldots, N-1; \end{array}$$

$$D_y^- V_{ij} = \frac{V_{ij} - V_{i,j-1}}{h} = 0, \qquad \begin{array}{l} i = 1, \ldots, N-1, \\ j = 1, \ldots, N. \end{array}$$

Since $V = 0$ on $\Gamma_h$, these imply that $V \equiv 0$. Thus $AV = 0$ if and only if $V = 0$. Hence $A$ is non-singular, and $U = A^{-1}F$ is the unique solution of (2.60). In summary then, the (unique) solution of the finite difference scheme (2.60) may be found by solving the system of linear equations (2.62).

In order to prove the stability of the finite difference scheme (2.60), we introduce the mesh–dependent norms

$$\|U\|_h = (U,U)_h^{1/2},$$

and

$$\|U\|_{1,h} = (\|U\|_h^2 + \|D_x^- U]|_x^2 + \|D_y^- U]|_y^2)^{1/2},$$

where

$$\|D_x^- U]|_x = \left( \sum_{i=1}^{N} \sum_{j=1}^{N-1} h^2 |D_x^- U_{ij}|^2 \right)^{1/2}$$

and

$$\|D_y^- U]|_y = \left( \sum_{i=1}^{N-1} \sum_{j=1}^{N} h^2 |D_y^- U_{ij}|^2 \right)^{1/2}.$$

The norm $\|\cdot\|_{1,h}$ is the discrete version of the Sobolev norm $\|\cdot\|_{H^1(\Omega)}$,

$$\|u\|_{H^1(\Omega)} = \left( \|u\|_{L^2(\Omega)}^2 + \|\frac{\partial u}{\partial x}\|_{L^2(\Omega)}^2 + \|\frac{\partial u}{\partial y}\|_{L^2(\Omega)}^2 \right)^{1/2}.$$

With this new notation, the inequality (2.64) takes the following form:

$$(AV,V)_h \geq \|D_x^- V]|_x^2 + \|D_y^- V]|_y^2. \tag{2.65}$$

Using the discrete Poincaré-Friedrichs inequality stated in the next lemma, we shall be able to deduce that

$$(AV,V)_h \geq c_0 \|V\|_{1,h}^2,$$

where $c_0$ is a positive constant.

**Lemma 8** (Discrete Poincaré–Friedrichs inequality) *Let $V$ be a function defined on $\bar{\Omega}_h$ and such that $V = 0$ on $\Gamma_h$; then there exists a constant $c_*$, independent of $V$ and $h$, such that*

$$\|V\|_h^2 \leq c_* \left( \|D_x^- V]|_x^2 + \|D_y^- V]|_y^2 \right) \tag{2.66}$$

*for all such $V$.*

**Proof** Writing

$$|V_{ij}|^2 = \left( \sum_{k=1}^{i} h D_x^- V_{kj} \right)^2,$$

we deduce that

$$|V_{ij}|^2 \leq \left(\sum_{k=1}^{i} h\right) \left(\sum_{k=1}^{i} h|D_x^- V_{kj}|^2\right) \leq i \sum_{k=1}^{N} h^2 |D_x^- V_{kj}|^2.$$

Multiplying both sides by $h^2$ and summing through $i = 1, \ldots, N-1$ and $j = 1, \ldots, N-1$, on noting that

$$h^2 \sum_{i=1}^{N-1} i = h^2 \frac{(N-1)N}{2} \leq \frac{1}{2},$$

we deduce that

$$\|V\|_h^2 \leq \frac{1}{2} \|D_x^- V]|_x^2.$$

Analogously,

$$\|V\|_h^2 \leq \frac{1}{2} \|D_y^- V]|_y^2.$$

Adding these two inequalities we complete the proof of (2.66) with $c_* = \frac{1}{4}$. ■

Now (2.65) and (2.66) imply that

$$(AV, V)_h \geq \frac{1}{c_*} \|V\|_h^2.$$

Finally, combining this with (2.65) and recalling the definition of the norm $\|\cdot\|_{1,h}$, we obtain

$$(AV, V)_h \geq c_0 \|V\|_{1,h}^2, \tag{2.67}$$

where $c_0 = (1 + c_*)^{-1}$.

**Theorem 5** *The scheme (2.60) is stable in the sense that*

$$\|U\|_{1,h} \leq \frac{1}{c_0} \|F\|_h. \tag{2.68}$$

**Proof** The proof is simple: it follows from (2.67) and the Cauchy–Schwarz inequality that

$$c_0 \|V\|_{1,h}^2 \leq (AV, V)_h = (F, V)_h \leq \|F\|_h \|V\|_h \leq \|F\|_h \|V\|_{1,h},$$

and hence the result. ■

Having established stability, we turn to the question of accuracy. We define the global error $e_h$ by $e_h(x, y) = u(x, y) - u_h(x, y)$ and note that $u_h(x_i, y_j) = U_{ij}$ for $i, j = 1, \ldots, N-1$. Since $u_h(x, y) = 0$ when $(x, y) \in \partial\Omega$, we have adopted the convention that $U_{ij} = 0$ when $i = 0$ or $i = N$ or $j = 0$ or $j = N$. Thus, writing $e_{ij} = e_h(x_i, y_j)$, we have that

$$e_{ij} = u(x_i, y_j) - U_{ij}, \quad 0 \leq i, j \leq N,$$

with $e_{ij} = 0$ when $i = 0$ or $i = N$ or $j = 0$ or $j = N$.

Now,

$$\begin{aligned} Ae_{ij} &= Au(x_i,y_j) - AU_{ij} = Au(x_i,y_j) - F_{ij} \\ &= Au(x_i,y_j) - \frac{1}{h^2}\int\int_{\text{supp }\phi_{ij}} \phi_{ij}(x,y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \left[\frac{1}{h^2}\int\int_{\text{supp }\phi_{ij}} \phi_{ij}(x,y)\frac{\partial^2 u}{\partial x^2}(x,y)\,\mathrm{d}x\,\mathrm{d}y - D_x^+ D_x^- u(x_i,y_j)\right] \\ &\quad + \left[\frac{1}{h^2}\int\int_{\text{supp }\phi_{ij}} \phi_{ij}(x,y)\frac{\partial^2 u}{\partial y^2}(x,y)\,\mathrm{d}x\,\mathrm{d}y - D_y^+ D_y^- u(x_i,y_j)\right] \\ &\equiv \varphi_{ij}. \end{aligned}$$

Thus,

$$\begin{aligned} Ae_{ij} &= \varphi_{ij}, \quad 1 \le i,j \le N-1, \\ e &= 0 \qquad \text{on } \Gamma_h. \end{aligned}$$

By virtue of (2.68),

$$\|u-U\|_{1,h} = \|e\|_{1,h} \le \frac{1}{c_0}\|\varphi\|_h. \tag{2.69}$$

Assuming that $u \in C^4(\bar{\Omega})$ and employing a Taylor series expansion of $u(x,y)$ about $(x_i,y_j)$, we deduce that

$$|\varphi_{ij}| \le K_0 h^2 \left( \|\frac{\partial^4 u}{\partial x^4}\|_{C(\bar{\Omega})} + \|\frac{\partial^4 u}{\partial y^4}\|_{C(\bar{\Omega})} \right),$$

where $K_0$ is a positive constant independent of $h$. Thus,

$$\|\varphi\|_h \le K_0 h^2 \left( \|\frac{\partial^4 u}{\partial x^4}\|_{C(\bar{\Omega})} + \|\frac{\partial^4 u}{\partial y^4}\|_{C(\bar{\Omega})} \right). \tag{2.70}$$

Finally (2.69) and (2.70) yield the following result.

**Theorem 6** *Let $f \in L_2(\Omega)$ and suppose that the corresponding weak solution $u \in H_0^1(\Omega)$ belongs to $C^4(\bar{\Omega})$; then*

$$\|u-u_h\|_{1,h} \le \frac{5}{4} K_0 h^2 \left( \|\frac{\partial^4 u}{\partial x^4}\|_{C(\bar{\Omega})} + \|\frac{\partial^4 u}{\partial y^4}\|_{C(\bar{\Omega})} \right). \tag{2.71}$$

**Proof** Recall that $c_0 = (1+c_*)^{-1}, c_* = \frac{1}{4}$, so that $1/c_0 = \frac{5}{4}$, and combine (2.69) and (2.70). ■

According to this result, the piecewise linear finite element approximation of the homogeneous Dirichlet boundary value problem on uniform triangular subdivision of size $h$ is $\mathcal{O}(h^2)$ convergent to the weak solution in the discrete Sobolev $H^1$ norm,

$\|\cdot\|_{1,h}$, provided that $u \in C^4(\bar{\Omega})$. Since this exceeds the first order of convergence of the global error observed in the Sobolev norm $\|\cdot\|_{H^1(\Omega)}$, the result encapsulated in Theorem 6 is referred to as a superapproximation property. In fact the smoothness requirement $u \in H_0^1(\Omega) \cap C^4(\bar{\Omega})$ can be relaxed to $u \in H_0^1(\Omega) \cap H^3(\Omega)$ while retaining the superapproximation property $\|u - u_h\|_{1,h} = \mathcal{O}(h^2)$; the proof of this is more technical and relies on the Bramble-Hilbert lemma (See Chapter 3).

# Chapter 3

# Piecewise polynomial approximation

In the previous chapter we discussed finite element approximations to elliptic boundary value problems using piecewise polynomials of degree 1. The purpose of this chapter is to develop, in a more general setting, the construction of finite element spaces and to formalise the concepts introduced in Chapter 2.

## 3.1 Construction of finite element spaces

Let us consider an elliptic boundary value problem written in its weak formulation:

$$\text{find } u \text{ in } V \text{ such that } a(u,v) = l(v) \quad \forall v \in V,$$

where $H_0^1(\Omega) \subset V \subset H^1(\Omega)$; in the case of a homogeneous Dirichlet boundary value problem $V = H_0^1(\Omega)$, in the case of a Neumann, Robin or oblique derivative boundary value problem, $V = H^1(\Omega)$. In order to define a finite element approximation to this problem we need to construct a finite-dimensional subspace $V_h$ of $V$ consisting of continuous piecewise polynomial functions of a certain degree defined on a subdivision of the computational domain $\Omega$. We have already discussed the special case when $V_h$ consists of continuous piecewise linear functions; here we shall put this construction into a general context.

### 3.1.1 The finite element

We begin by giving a formal definition of a finite element.

**Definition 2** *Let us suppose that*

*(i)* $K \subset \mathbb{R}^n$ *is a simply connected bounded open set with piecewise smooth boundary (***the element domain***);*

*(ii)* $\mathcal{P}$ *is a finite-dimensional space of functions defined on* $K$ *(***the space of shape functions***);*

*(iii)* $\mathcal{N} = \{N_1, \ldots, N_k\}$ *is a basis for* $\mathcal{P}'$ *(***the set of nodal variables***).*

*Then* $(K, \mathcal{P}, \mathcal{N})$ *is called a* **finite element***.*

In this definition $\mathcal{P}'$ denotes the algebraic dual of the linear space $\mathcal{P}$.

**Definition 3** *Let* $(K, \mathcal{P}, \mathcal{N})$ *be a finite element, and let* $\{\psi_1, \psi_2, \ldots, \psi_k\}$ *be a basis for* $\mathcal{P}$*, dual to* $\mathcal{N}$*; namely,*

$$N_i(\psi_j) = \delta_{ij}, \qquad 1 \leq i, j \leq k.$$

*Such a basis is called a* **nodal basis** *for* $\mathcal{P}$*.*

We give a simple example to illustrate these definitions.

**Example 7** (The one-dimensional Lagrange element) *Let* $K = (0, 1)$*,* $\mathcal{P}$ *the set of linear polynomials, and* $\mathcal{N} = \{N_1, N_2\}$*, where* $N_1(v) = v(0)$ *and* $N_2(v) = v(1)$ *for all* $v \in \mathcal{P}$*. Then* $(K, \mathcal{P}, \mathcal{N})$ *is a finite element, with nodal basis* $\{\psi_1, \psi_2\}$ *where* $\psi_1(x) = 1 - x$ *and* $\psi_2(x) = x$*.* $\diamond$

Next we give an equivalent characterisation of condition (iii) in Definition 2.

**Lemma 9** *Let* $\mathcal{P}$ *be a* $k$*-dimensional linear space of functions on* $\mathbb{R}^n$*, and suppose that* $\{N_1, N_2, \ldots, N_k\}$ *is a subset of the dual space* $\mathcal{P}'$*. Then the following two statements are equivalent:*

*(a)* $\{N_1, N_2, \ldots, N_k\}$ *is a basis for* $\mathcal{P}'$*;*

*(b) Given that* $v \in \mathcal{P}$ *and* $N_i(v) = 0$ *for* $i = 1, \ldots, k$*, then* $v \equiv 0$*.*

**Proof** Let $\{\psi_1, \ldots, \psi_k\}$ be a basis for $\mathcal{P}$. Now $\{N_1, \ldots, N_k\}$ is a basis for $\mathcal{P}'$ if and only if any $L \in \mathcal{P}'$ can written in a unique fashion as a linear combination of the $N_i$'s:

$$L = \alpha_1 N_1 + \ldots + \alpha_k N_k.$$

This is equivalent to demanding that, for each $i = 1, \ldots, k$, $L(\psi_i)$ can be written in a unique fashion as a linear combination

$$L(\psi_i) = \alpha_1 N_1(\psi_i) + \ldots + \alpha_k N_k(\psi_i).$$

Let us define the matrix $B = (N_j(\psi_i))_{i,j=1,\ldots,k}$ and the vectors

$$\mathbf{y} = (L(\psi_1), \ldots, L(\psi_k))^T, \qquad \mathbf{a} = (\alpha_1, \ldots, \alpha_k)^T.$$

Then the last condition is equivalent to requiring that the system of linear equations $B\mathbf{a} = \mathbf{y}$ has a unique solution, which, in turn, is equivalent to demanding that the matrix $B$ be invertible. Given any $v \in \mathcal{P}$, we can write

$$v = \beta_1 \psi_1 + \ldots + \beta_k \psi_k.$$

Now $N_i(v) = 0$ for all $i = 1, \ldots, k$ if and only if

$$\beta_1 N_i(\psi_1) + \ldots + \beta_k N_i(\psi_k) = 0, \quad i = 1, \ldots, k. \tag{3.1}$$

Thus (b) is equivalent to requiring that (3.1) implies $\beta_1 = \ldots = \beta_k = 0$. Let $C = (N_i(\psi_j))_{i,j=1,\ldots,k}$; then (b) holds if and only if $C\mathbf{b} = \mathbf{0}$, with $\mathbf{b} = (\beta_1, \ldots, \beta_k)^T$, implies that $\mathbf{b} = \mathbf{0}$, which is equivalent to demanding that $C$ be invertible. However $C^t = B$ and therefore (a) and (b) are equivalent. ∎

Motivated by this result, we introduce the following definition.

**Definition 4** *We say that* $\mathcal{N}$ **determines** $\mathcal{P}$ *if* $\psi \in \mathcal{P}$ *with* $N(\psi) = 0$ *for all* $N \in \mathcal{N}$ *implies that* $\psi = 0$.

We shall need the following Lemma.

**Lemma 10** *Suppose that* $P$ *is a polynomial of degree* $d \geq 1$ *that vanishes on the hyperplane* $\{x : L(x) = 0\}$ *where* $L$ *is a non-degenerate linear function. Then we can write* $P$ *in the factorised form* $P = LQ$ *where* $Q$ *is a polynomial of degree* $(d-1)$.

**Proof** Let us write $\hat{x} = (x_1, \ldots, x_{n-1})$. Suppose that we have carried out an affine change of variables such that $L(\hat{x}, x_n) = x_n$; so, the hyperplane $L(\hat{x}, x_n) = 0$ is the hyperplane $x_n = 0$; then, by hypothesis, $P(\hat{x}, 0) \equiv 0$. Since $P$ is of degree $d$, we have that

$$P(\hat{x}, x_n) = \sum_{j=0}^{d} \sum_{|\alpha| \leq d-j} c_{\alpha j} \hat{x}^\alpha x_n^j,$$

where $\alpha = (i_1, \ldots, i_{n-1})$ and $\hat{x}^\alpha = x_1^{i_1} \cdots x_{n-1}^{i_{n-1}}$. Letting $x_n = 0$ we get

$$0 \equiv P(\hat{x}, 0) = \sum_{|\alpha| \leq d} c_{\alpha 0} \hat{x}^\alpha,$$

which implies that $c_{\alpha 0} = 0$ for $|\alpha| \leq d$. Hence

$$P(\hat{x}, x_n) = \sum_{j=1}^{d} \sum_{|\alpha| \leq d-j} c_{\alpha j} \hat{x}^\alpha x_n^j = x_n \sum_{j=1}^{d} \sum_{|\alpha| \leq d-j} c_{\alpha j} \hat{x}^\alpha x_n^{j-1} = x_n Q = LQ$$

where $Q$ is of degree $(d-1)$. ∎

### 3.1.2 Examples of triangular finite elements

Let $K$ be a triangle and let $\mathcal{P}_k$ denote the set of all polynomials of degree $\leq k$ in two variables. The dimension of the linear space $\mathcal{P}_k$ is displayed in Table 3.1.2.

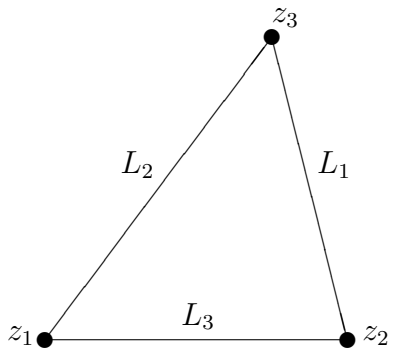


Figure 3.1: Linear Lagrange triangle with edges $L_1$, $L_2$, $L_3$ and vertices $z_1$, $z_2$, $z_3$.

| $k$ | $\dim \mathcal{P}_k$ |
|---|---|
| 1 | 3 |
| 2 | 6 |
| 3 | 10 |
| ... | ... |
| $k$ | $\frac{1}{2}(k+1)(k+2)$ |

Table 3.1.2: The dimension of the linear space $\mathcal{P}_k$.

**Lagrange elements**

**Example 8** *Let $k = 1$ and take $\mathcal{P} = \mathcal{P}_1$, $\mathcal{N} = \mathcal{N}_1 = \{N_1, N_2, N_3\}$ (and therefore the dimension of $\mathcal{P}_1$ is 3), where $N_i(v) = v(z_i)$ and $z_1$, $z_2$, $z_3$ are the vertices of the triangle $K$, as shown in Figure 3.1.*

*In the figure $\bullet$ indicates function evaluation at the point where the dot is placed. We verify condition (iii) of Definition 2 using part (b) of Lemma 9; namely, we prove that $\mathcal{N}_1$ determines $\mathcal{P}_1$. Indeed, let $L_1$, $L_2$ and $L_3$ be non-trivial linear functions which define the lines that contain the three edges of the triangle. Suppose that a polynomial $P \in \mathcal{P}_1$ vanishes at $z_1$, $z_2$ and $z_3$. Since $P|_{L_1}$ is a linear function of one variable that vanishes at two points, it follows that $P \equiv 0$ on $L_1$. By virtue of Lemma 10, we can write $P = cL_1$ where $c$ is a constant (i.e. a polynomial of degree $1 - 1 = 0$). However, because $L_1(z_1) \neq 0$,*

$$0 = P(z_1) = cL_1(z_1)$$

*implies that $c = 0$; thus $P \equiv 0$. Hence, according to Lemma 9, $\mathcal{N}_1$ determines $\mathcal{P}_1$.* $\diamond$

**Example 9** *Now take $k = 2$, let $\mathcal{P} = \mathcal{P}_2$ and $\mathcal{N} = \mathcal{N}_2 = \{N_1, N_2, \ldots, N_6\}$ (so we have that dim $\mathcal{P}_2 = 6$), where*

$$N_i(v) = \begin{cases} v(\textit{at the } i\textit{th vertex of the triangle}), & i = 1, 2, 3 \\ v(\textit{at the midpoint of the } (i-3)\textit{rd edge}), & i = 4, 5, 6. \end{cases}$$

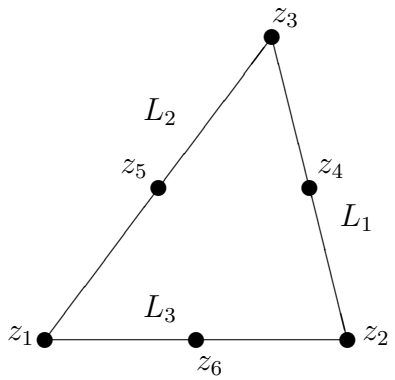


Figure 3.2: Quadratic Lagrange triangle with edges $L_1$, $L_2$, $L_3$, vertices $z_1$, $z_2$, $z_3$, and $z_4$, $z_5$ and $z_6$ denoting the midpoints of $L_1$, $L_2$ and $L_3$, respectively.

*We have to show that $\mathcal{N}_2$ determines $\mathcal{P}_2$. Let $L_1$, $L_2$ and $L_3$ be non-trivial linear functions which define the lines containing the edges of the triangle. Let $P \in \mathcal{P}_2$ be such that $P(z_i) = 0$ for $i = 1, \ldots, 6$. As $P|_{L_1}$ is a quadratic function of one variable that vanishes at three points, it follows that $P \equiv 0$ on $L_1$. By Lemma 10, $P = L_1 Q_1$, where the degree of $Q_1$ is one less than the degree of $P$, so $Q_1$ is of degree 1. However, by an analogous argument $P$ also vanishes along $L_2$. Therefore, $L_1 Q_1|_{L_2} \equiv 0$. Thus, on $L_2$, either $L_1 \equiv 0$ or $Q_1 \equiv 0$. But $L_1$ can be equal to zero only at one point of $L_2$ because the triangle is non-degenerate. Thus $Q_1 \equiv 0$ on $L_2$, except possibly at one point. However, by continuity of $Q_1$, we then have that $Q_1 \equiv 0$ along the whole of $L_2$.*

*Now applying again Lemma 10, we deduce that $Q_1 = L_2 Q_2$, where the degree of $Q_2$ is one less than the degree of $Q_1$, so $Q_2$ is of degree 0. Thus, $Q_2 \equiv c$, where $c$ is a constant. Hence, $P = cL_1L_2$. However $P(z_6) = 0$ and $z_6$ does not lie on either $L_1$ or $L_2$. Consequently,*

$$0 = P(z_6) = cL_1(z_6)L_2(z_6).$$

*Therefore, $c = 0$ since $L_1(z_6) \neq 0$ and $L_2(z_6) \neq 0$. This finally implies that $P \equiv 0$, so $\mathcal{N}_2$ determines $\mathcal{P}_2$.* $\diamond$

### Hermite elements

**Example 10** *Let us suppose that $k = 3$, and let $\mathcal{P} = \mathcal{P}_3$. Let $\bullet$ denote evaluation at a point and let $\bigcirc$ signify evaluation of the gradient at the centre point of the circle. We shall prove that $\mathcal{N} = \mathcal{N}_3 = \{N_1, N_2, \ldots, N_{10}\}$, as depicted in Figure 3.3, determines $\mathcal{P}_3$ (whose dimension is precisely 10). Let, as before, $L_1$, $L_2$ and $L_3$ be the lines corresponding to the three sides of the triangle and suppose that $P \in \mathcal{P}_3$ and $N_i(P) = 0$ for the $i = 1, 2 \ldots, 10$. The restriction of $P$ to $L_1$ is a cubic polynomial of one variable with double roots at $z_2$ and $z_3$. Hence $P \equiv 0$ along $L_1$. Similarly, $P \equiv 0$ along the edges $L_2$ and $L_3$. By Lemma 10, we can write $P = cL_1L_2L_3$, where $c$ is a constant. However,*

$$0 = P(z_4) = cL_1(z_4)L_2(z_4)L_3(z_4),$$

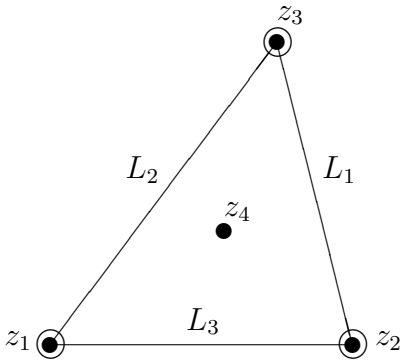


Figure 3.3: Cubic Hermite triangle with edges $L_1$, $L_2$, $L_3$ and vertices $z_1$, $z_2$ and $z_3$, and centroid $z_3$.

*and so $c = 0$, since $L_i(z_4) \neq 0$ for $i = 1, 2, 3$. Thus $P \equiv 0$ and we deduce that $\mathcal{N}$ uniquely determines $\mathcal{P}_3$.* $\diamond$

### 3.1.3 The interpolant

Having described a number of finite elements, we now wish to piece them together to construct finite-dimensional subspaces of Sobolev spaces.

**Definition 5** *Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element and let the set $\{\psi_i : i = 1, \ldots, k\}$, be a basis for $\mathcal{P}$ dual to $\mathcal{N}$. Given that $v$ is a function for which all $N_i \in \mathcal{N}$, $i = 1, \ldots, k$, are defined, we introduce the* **local interpolant** *$\mathcal{I}_K v$ by*

$$\mathcal{I}_K v = \sum_{i=1}^{k} N_i(v)\psi_i.$$

In order to illustrate the idea of local interpolant, we give a simple example.

**Example 11** *Consider the triangle $K$ shown in Figure 3.4, let $\mathcal{P} = \mathcal{P}_1$, $\mathcal{N} = \{N_1, N_2, N_3\}$ as in the case of the linear Lagrange element ($k = 1$), and suppose that we wish to find the local interpolant $\mathcal{I}_K v$ of the function $v$ defined by $v(x, y) = (1 + x^2 + y^2)^{-1}$.*

*By definition,*

$$\mathcal{I}_K v = N_1(v)\psi_1 + N_2(v)\psi_2 + N_3(v)\psi_3.$$

*Thus we must determine $\psi_i$, $i = 1, 2, 3$, to be able to write down the local interpolant. This we do, using Definition 3, as follows. The line $L_1$ has equation $y = 1 - x$; as $\psi_1$ vanishes at $z_2$ and $z_3$, and thereby along the whole of $L_1$, it follows that $\psi_1 = cL_1 = c(1 - x - y)$, where $c$ is a constant to be determined. Also, $N_1\psi_1 = 1$, so $c = \psi_1(z_1) = 1$. Hence, $\psi_1 = 1 - x - y$. Similarly, $\psi_2 = L_2(x, y)/L_2(z_2) = x$ and $\psi_3 = L_3(x, y)/L_3(z_3) = y$.*

*Having found $\psi_1$, $\psi_2$ and $\psi_3$, we have that*

$$\mathcal{I}_K(v) = N_1(v)(1 - x - y) + N_2(v)x + N_3(v)y.$$

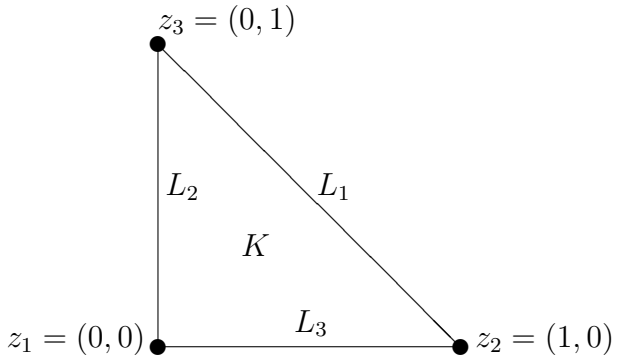


Figure 3.4: Linear Lagrange triangle with edges $L_1$, $L_2$, $L_3$, and the vertices $z_1$, $z_2$ and $z_3$ where the local interpolant is evaluated.

*In fact, in our case* $N_1(v) = v(z_1) = 1$, $N_2(v) = v(z_2) = \frac{1}{2}$ *and* $N_3(v) = v(z_3) = \frac{1}{2}$, *and so*

$$\mathcal{I}_K(v) = 1 - \frac{1}{2}(x+y). \qquad \diamond$$

The next lemma summarises the key properties of the local interpolant.

**Lemma 11** *The local interpolant has the following properties:*

*a) The mapping* $v \mapsto \mathcal{I}_K v$ *is linear.*

*b)* $N_i(\mathcal{I}_K(v)) = N_i(v)$, $i = 1, \ldots, k$.

*c)* $\mathcal{I}_K(v) = v$ *for* $v \in \mathcal{P}$; *consequently* $\mathcal{I}_K$ *is idempotent on* $\mathcal{P}$, *that is,* $\mathcal{I}_K^2 = \mathcal{I}_K$.

**Proof**

a) Since each $N_i : v \mapsto N_i(v)$, $i = 1, \ldots, k$, is a linear functional, $v \mapsto \mathcal{I}_K v$ has the same property.

b) Clearly

$$\begin{aligned} N_i(\mathcal{I}_K(v)) &= N_i\left(\sum_{j=1}^{k} N_j(v)\psi_j\right) = \sum_{j=1}^{k} N_j(v) N_i(\psi_j) \\ &= \sum_{j=1}^{k} N_j(v)\delta_{ij} = N_i(v), \end{aligned}$$

for $i = 1, \ldots, k$, where $\delta_{ij} = 1$ when $i = j$ and $= 0$ when $i \neq j$.

c) It follows from b) that $N_i(v - \mathcal{I}_K(v)) = 0$, $i = 1, \ldots, k$, which implies that $\mathcal{I}_K(v) = v$ for all $v \in \mathcal{P}$. The second assertion follows from this; indeed, $\mathcal{I}_K^2 v = \mathcal{I}_K(\mathcal{I}_K v) = \mathcal{I}_K v$ since $\mathcal{I}_K v \in \mathcal{P}$.

That completes the proof of the lemma. ■

We can now glue together the element domains to obtain a subdivision of the computational domain, and merge the local interpolants to obtain a global interpolant.

**Definition 6** *A* **subdivision** *of the computational domain $\Omega$ is a finite collection of open sets $\{K_i\}$ such that*

*(1) $K_i \cap K_j = \emptyset$ if $i \neq j$, and*

*(2) $\bigcup_i \bar{K}_i = \bar{\Omega}$.*

**Definition 7** *Suppose that $\Omega$ is a bounded open set in $\mathbb{R}^n$ with subdivision $\mathcal{T}$. Assume that each element domain $K$ in the subdivision is equipped with some type of shape functions $\mathcal{P}$ and nodal variables $\mathcal{N}$, such that $(K, \mathcal{P}, \mathcal{N})$ forms a finite element. Let $m$ be the order of the highest partial derivative involved in the nodal variables. For $v \in C^m(\bar{\Omega})$ the* **global interpolant** *$\mathcal{I}_h v$ is defined on $\bar{\Omega}$ by*

$$\mathcal{I}_h v|_{K_i} = \mathcal{I}_{K_i} v \qquad \forall K_i \in \mathcal{T}.$$

In the absence of further conditions on the subdivision it is not possible to assert the continuity of the global interpolant. Next we shall formulate a simple condition which ensures that the global interpolant is a continuous function on $\bar{\Omega}$. To keep the presentation simple, we shall restrict ourselves to the case of two space dimensions, namely when $\Omega \subset \mathbb{R}^2$, although an analogous definition can be introduced in $\mathbb{R}^n$.

**Definition 8** *A* **triangulation** *of a polygonal domain $\Omega$ is a subdivision of $\Omega$ consisting of triangles which have the property that*

*(3) No vertex of any triangle lies in the interior of an edge of another triangle.*

From now on, we shall use the word *triangulation* without necessarily implying that $\Omega \subset \mathbb{R}^2$: when $\Omega \subset \mathbb{R}^n$ and $n = 2$ we shall mean that the condition of this definition is satisfied; when $n > 2$, the obvious generalisation of this condition to $n$ dimensions will be meant to hold.

**Definition 9** *We say that an interpolant has* **continuity of order** *$r$ (or, briefly, that it is $C^r$) if $\mathcal{I}_h v \in C^r(\bar{\Omega})$ for all $v \in C^m(\bar{\Omega})$. The space*

$$\{\mathcal{I}_h v \,:\, v \in C^m(\bar{\Omega})\}$$

*is called a $C^r$ finite element space.*

For simplicity, again, the next result is stated and proved in the case of $n = 2$; an analogous result holds for $n > 2$.

**Theorem 7** *The Lagrange and Hermite elements on triangles are all $C^0$ elements. More precisely, given a triangulation $\mathcal{T}$ of $\Omega$, it is possible to choose edge nodes for the corresponding elements $(K, \mathcal{P}, \mathcal{N})$, $K \in \mathcal{T}$, such that the global interpolant $\mathcal{I}_h v$ belongs to $C^0(\bar{\Omega})$ for all $v$ in $C^m(\bar{\Omega})$, where $m = 0$ for Lagrange and $m = 1$ for Hermite elements.*

**Proof** It suffices to show that continuity holds across each edge. Let $K_i$, $i = 1, 2$, be two triangles in the triangulation $\mathcal{T}$ with common edge $e$. Assuming that we choose nodes interior to $e$ in a symmetric way, it follows that the edge nodes on $e$ for the elements on both $K_1$ and $K_2$ are at the same location in space.

Let $w = \mathcal{I}_{K_1} v - \mathcal{I}_{K_2} v$, where we interpret $\mathcal{I}_{K_1} v$ and $\mathcal{I}_{K_2} v$ to be defined everywhere by extension outside $K_1$ and $K_2$, respectively, as polynomials. Now $w$ is a polynomial of degree $k$ and its restriction to the edge $e$ vanishes at the one-dimensional Lagrange (or Hermite) nodes. Therefore $w|_e \equiv 0$. Hence $\mathcal{I}_{K_1}|_e v = \mathcal{I}_{K_2} v|_e$, i.e. the global interpolant is continuous across each edge. ∎

In order to be able to compare global interpolation operators on different elements, we introduce the following definition (now for $K \subset \Omega \subset \mathbb{R}^n$.)

**Definition 10** *Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element and suppose that $F(x) = Ax + b$ where $A$ is a non-singular $n \times n$ matrix and $x$ and $b$ are $n$-component column vectors. The finite element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ is* **affine equivalent** *to $(K, \mathcal{P}, \mathcal{N})$ if:*

*(a) $F(K) = \hat{K}$;*

*(b) $F^* \hat{\mathcal{P}} = \mathcal{P}$ and*

*(c) $F_* \mathcal{N} = \hat{\mathcal{N}}$.*

*Here $F^*$ is the pull-back of $F$ defined by $F^*(\hat{v}) = \hat{v} \circ F$, and $F_*$ is the push-forward of $F$ defined by $(F_* N)(\hat{v}) = N(F^*(\hat{v})) = N(\hat{v} \circ F)$.*

**Example 12** *Lagrange elements on triangles with appropriate choice of edge and interior nodes are affine equivalent. The same is true of Hermite elements on triangles.* ◇

### 3.1.4 Examples of rectangular elements

To conclude this section we consider finite elements defined on rectangles. Let

$$Q_k = \left\{ \sum_j c_j p_j(x) q_j(y) \ : \ p_j \text{ and } q_j \text{ are polynomials of degree} \leq k \right\}.$$

It can be shown that $Q_k$ is a linear space of dimension $(\dim \mathcal{P}_k^1)^2$, where $\mathcal{P}_k^1$ denotes the set of all polynomials of a single variable of degree $\leq k$ and $\dim \mathcal{P}_k^1$ signifies its dimension.

We give two examples, without going into details.

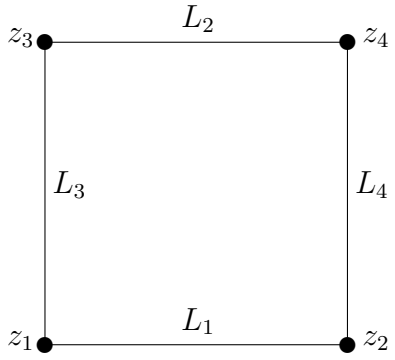


Figure 3.5: Bilinear Lagrange rectangle with edges $L_1$, $L_2$, $L_3$, $L_4$ and the vertices $z_1$, $z_2$, $z_3$ and $z_4$.

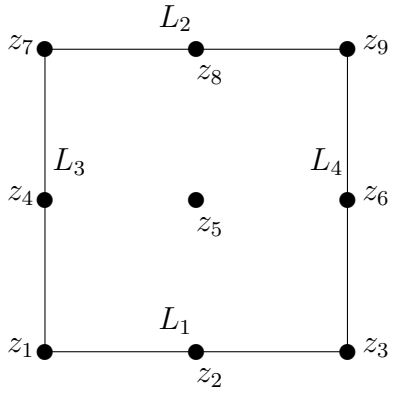


Figure 3.6: Biquadratic Lagrange rectangle with edges $L_1$, $L_2$, $L_3$, $L_4$ and the vertices $z_1$, $z_3$, $z_7$, $z_9$, midpoints of edges $z_2$, $z_4$, $z_6$, $z_8$, and centroid $z_5$.

**Example 13** (Bilinear Lagrange rectangle) *Let $k = 1$ and suppose that $K$ is a rectangle. Further, let $\mathcal{P} = Q_1$ and let $\mathcal{N} = \{N_1, \ldots, N_4\}$ with $N_i(v) = v(z_i)$ with $z_i$, $i = 1, \ldots, 4$, as in Figure 3.5. We leave it as an exercise to the reader to show, using Lemmas 9 and 10 that $\mathcal{N}$ determines $\mathcal{P} = Q_1$ (the dimension of $Q_1$ is equal to 4).* $\diamond$

**Example 14** (Biquadratic Lagrange rectangle) *Let $k = 2$ and suppose that $K$ is a rectangle. We let $\mathcal{P} = Q_2$ and put $\mathcal{N} = \{N_1, \ldots, N_9\}$ with $N_i(v) = v(z_i)$ with $z_i$, $i = 1, \ldots, 9$, as in Figure 3.6. It is left as an exercise to show that $\mathcal{N}$ determines $\mathcal{P} = Q_2$ (the dimension of $Q_2$ is equal to 9).* $\diamond$

## 3.2 Polynomial approximation in Sobolev spaces

In this section we shall develop the approximation theory for the finite element spaces described in the previous section. We shall adopt a constructive approach which will enable us to calculate the constants in the error estimates explicitly. The technique is based on the use of the Hardy-Littlewood maximal function, following

the work of Ricardo Durán[1]. An alternative approach which exploits the theory of Riesz potentials is presented in the Brenner-Scott monograph cited in the reading list.

### 3.2.1 The Bramble-Hilbert lemma

A key device in finite element error analysis is the following result.

**Lemma 12** (Bramble-Hilbert lemma) *Suppose that $\Omega$ is a bounded open set in $\mathbb{R}^n$ and assume that $\Omega$ is star-shaped with respect to every point in a set $B$ of positive measure contained in $\Omega$ (i.e. for all $x \in \Omega$ the closed convex hull of $\{x\} \cup B$ is a subset of $\Omega$). Let $l$ be a bounded linear functional on the Sobolev space $W_p^m(\Omega)$, $m \geq 1$, $1 < p < \infty$, such that $l(Q) = 0$ for any polynomial $Q$ of degree $\leq m-1$. Then there exists a constant $C_1 > 0$ such that*

$$|l(v)| \leq C_1 |v|_{W_p^m(\Omega)} \qquad \text{for all } v \in W_p^m(\Omega).$$

**Proof** By hypothesis, there exists $C_0 > 0$ such that

$$|l(v)| \leq C_0 \|v\|_{W_p^m(\Omega)} \qquad \forall v \in W_p^m(\Omega).$$

Since $l(Q) = 0$ for all $Q \in \mathcal{P}_{m-1}$, we have by the linearity of $l$ that

$$\begin{aligned} |l(v)| = |l(v-Q)| &\leq C_0 \|v - Q\|_{W_p^m(\Omega)} \\ &= C_0 \left( \sum_{j=0}^{m} |v-Q|^p_{W_p^j(\Omega)} \right)^{1/p} \\ &= C_0 \left( \sum_{j=0}^{m-1} |v-Q|^p_{W_p^j(\Omega)} + |v|^p_{W_p^m(\Omega)} \right)^{1/p} \\ &\leq C_0 \left( \sum_{j=0}^{m-1} |v-Q|_{W_p^j(\Omega)} + |v|_{W_p^m(\Omega)} \right). \end{aligned}$$

In order to complete the proof it remains to prove that

$$\begin{aligned} &\exists K_j > 0 \ \ \forall v \in W_p^j(\Omega) \ \ \exists Q \in \mathcal{P}_{m-1} \qquad \text{such that} \\ &\qquad |v-Q|_{W_p^j(\Omega)} \leq K_j |v|_{W_p^m(\Omega)}, \quad j = 0, \ldots, m-1. \end{aligned} \tag{3.2}$$

This will be done in the rest of the section. Once (3.2) has been verified, we shall have that

$$|l(v)| \leq C_0 \left( 1 + \sum_{j=0}^{m-1} K_j \right) |v|_{W_p^m(\Omega)},$$

[1] R. Durán: On polynomial approximation in Sobolev spaces. SIAM Journal of Numerical Analysis, **20**, No. 5., (1983), pp. 985–988.

and the proof will be complete, with the constant

$$C_1 = C_0 \left( 1 + \sum_{j=0}^{m-1} K_j \right).$$

■

The original proof of (3.2) given by Bramble and Hilbert was based on the use of the Hahn-Banach theorem and was non-constructive in nature in the sense that it did not provide computable constants $K_j$, $j = 0, \ldots, m-1$; only the existence of such constants was proved. The remainder of this section is devoted to the (constructive) proof of (3.2). Our main tool is the following lemma.

**Lemma 13** *Let $g \in L_p(\mathbb{R}^n)$, $1 < p < \infty$. Given $\nu \in \mathbb{R}^n$ such that $|\nu| = 1$, we define*

$$g_1(x, \nu) = \sup_{t>0} \frac{1}{t} \int_0^t |g(x + s\nu)| \, ds$$

*and*

$$g^*(x) = \left( \int_{|\nu|=1} g_1(x, \nu)^p \, d\sigma_\nu \right)^{1/p}.$$

*Then*

$$\|g^*\|_{L_p(\mathbb{R}^n)} \leq \frac{p}{p-1} \omega_n^{1/p} \|g\|_{L_p(\mathbb{R}^n)},$$

*where $\omega_n$ is the measure of the unit sphere in $\mathbb{R}^n$.*

**Proof** Since $g_1(\cdot, \nu)$ is the Hardy-Littlewood maximal function of $g(\cdot)$ in the direction $\nu$, it follows that[2]

$$\int_{\mathbb{R}^n} g_1(x, \nu)^p \, \mathrm{d}x \leq \left( \frac{p}{p-1} \right)^p \int_{\mathbb{R}^n} |g(x)|^p \, \mathrm{d}x,$$

and therefore

$$\begin{aligned} \int_{\mathbb{R}^n} |g^*(x)|^p \, \mathrm{d}x &= \int_{\mathbb{R}^n} \left( \int_{|\nu|=1} g_1(x, \nu)^p \, \mathrm{d}\sigma_\nu \right) \mathrm{d}x \\ &= \int_{|\nu|=1} \left( \int_{\mathbb{R}^n} g_1(x, \nu)^p \, \mathrm{d}x \right) \mathrm{d}\sigma_\nu \\ &\leq \omega_n \left( \frac{p}{p-1} \right)^p \int_{\mathbb{R}^n} |g(x)|^p \, \mathrm{d}x. \end{aligned}$$

Upon taking the $p$th root of the two sides in this inequality we arrive at the desired result.

■

Now we are ready to prove (3.2).

[2]See E.M. Stein and T.S. Murphy: Harmonic Analysis: Real Variable Methods, Orthogonality and Oscillatory Integrals. Princeton University Press, 1993; or A.P. Calderón: Estimates for singular integral operators in terms of maximal functions. Stud. Math. **44**, (1972), pp.563–582.

**Theorem 8** *Let $\Omega \subset \mathbb{R}^n$ be a bounded open set which is star-shaped with respect to each point in a set of positive measure $B \subset \Omega$. Let $1 < p < \infty$, $0 \le j < m$, and let $d$ be the diameter of $\Omega$. If $v \in W_p^m(\Omega)$ then*

$$\inf_{Q \in \mathcal{P}_{m-1}} |v - Q|_{W_p^j(\Omega)} \le C \frac{d^{m-j+(n/p)}}{|B|^{1/p}} |v|_{W_p^m(\Omega)},$$

*where $|B|$ denotes the measure of $B$ and*

$$C = (\sharp\{\alpha \,:\, |\alpha| = j\}) \, \frac{m-j}{n^{1/p}} \, \frac{p}{p-1} \, \omega_n^{1/p} \left( \sum_{|\beta|=m-j} (\beta!)^{-p'} \right)^{1/p'},$$

*with $1/p + 1/p' = 1$. Here, for a set $A$, $\sharp A$ denotes the number of elements in $A$ and, for a multi-index $\beta = (\beta_1, \ldots, \beta_n)$, $\beta! = \beta_1! \cdot \ldots \cdot \beta_n!$.*

**Proof** Because $C^\infty(\bar{\Omega})$ is dense[3] in $W_p^m(\Omega)$, it suffices to prove the theorem for $v \in C^\infty(\bar{\Omega})$. Given $x \in B$, we define

$$P_m(v)(x,y) = \sum_{|\beta|<m} D^\beta v(x) \frac{(y-x)^\beta}{\beta!}$$

and

$$Q_m(v)(y) = \frac{1}{|B|} \int_B P_m(v)(x,y) \, \mathrm{d}x.$$

Here we used the multi-index notation $(y-x)^\beta = (y_1 - x_1)^{\beta_1} \cdot \ldots \cdot (y_n - x_n)^{\beta_n}$. It is easy to prove by induction that

$$D^\alpha Q_m(v)(y) = Q_{m-|\alpha|}(D^\alpha v)(y).$$

Thus,

$$\begin{aligned} |v - Q_m(v)|_{W_p^j(\Omega)} &= \left( \sum_{|\alpha|=j} \|D^\alpha (v - Q_m(v))\|_{L_p(\Omega)}^p \right)^{1/p} \\ &\le \sum_{|\alpha|=j} \|D^\alpha (v - Q_m(v))\|_{L_p(\Omega)} \\ &= \sum_{|\alpha|=j} \|D^\alpha v - Q_{m-j}(D^\alpha v)\|_{L_p(\Omega)}. \end{aligned}$$

Now let us estimate $\|D^\alpha v - Q_{m-j}(D^\alpha v)\|_{L_p(\Omega)}$ for each $\alpha$, $|\alpha| = j$. As

$$(D^\alpha v - Q_{m-j}(D^\alpha v))(y) = \frac{1}{|B|} \int_B [D^\alpha v(y) - P_{m-j}(D^\alpha v)(x,y)] \, \mathrm{d}x,$$

[3] See R.A. Adams: Sobolev Spaces. Academic Press, 1975.

it follows, by applying Minkowski's inequality for integrals[4] that

$$\begin{aligned}&\|D^\alpha v - Q_{m-j}(D^\alpha v)\|_{L_p(\Omega)}\\ &\qquad \le \frac{1}{|B|}\int_B \left(\int_\Omega |D^\alpha v(y) - P_{m-j}(D^\alpha v)(x,y)|^p \,\mathrm{d}y\right)^{1/p} \mathrm{d}x. \end{aligned} \tag{3.3}$$

Now recalling the integral-remainder for Taylor series, for $x \in B$, $y \in \Omega$, we have that

$$\begin{aligned}&|D^\alpha v(y) - P_{m-j}(D^\alpha v)(x,y)|\\ &= \left|(m-j)\sum_{|\beta|=m-j}\frac{(y-x)^\beta}{\beta!}\int_0^1 D^\beta D^\alpha v(x+t(y-x))(1-t)^{m-j-1}\,\mathrm{d}t\right|\\ &\le (m-j)d^{m-j}\int_0^1 \sum_{|\beta|=m-j}\frac{1}{\beta!}|D^\beta D^\alpha v(x+t(y-x))|\,\mathrm{d}t\\ &= (m-j)d^{m-j}\frac{1}{|y-x|}\int_0^{|y-x|}\sum_{|\beta|=m-j}\frac{1}{\beta!}\left|D^\beta D^\alpha v\left(x+s\frac{y-x}{|y-x|}\right)\right|\mathrm{d}s.\end{aligned}$$

Let $g$ be the function that coincides with $\sum_{|\beta|=m-j}\frac{1}{\beta!}|D^\beta D^\alpha v|$ in $\Omega$ and is identically zero outside $\Omega$. If $g_1$ and $g^*$ are the functions associated with $g$, as defined in Lemma 13, we have that

$$|D^\alpha v(y) - P_{m-j}(D^\alpha v)(x,y)| \le (m-j)d^{m-j} g_1\left(x, \frac{y-x}{|y-x|}\right),$$

and therefore

$$|D^\alpha v(y) - P_{m-j}(D^\alpha v)(x,y)|^p \le (m-j)^p d^{(m-j)p} g_1^p\left(x, \frac{y-x}{|y-x|}\right).$$

Noting that $B \subset \Omega$, it follows for each $x \in B$ that

$$\int_\Omega |D^\alpha v(y) - P_{m-j}(D^\alpha v)(x,y)|^p \,\mathrm{d}y \le (m-j)^p d^{(m-j)p}\int_{|y-x|\le d} g_1^p\left(x, \frac{y-x}{|y-x|}\right)\mathrm{d}y.$$

Thus,

$$\begin{aligned}&\left(\int_\Omega |D^\alpha v(y) - P_{m-j}(D^\alpha v)(x,y)|^p \,\mathrm{d}y\right)^{1/p}\\ &\qquad \le (m-j)d^{m-j}\left(\int_0^d\int_{|\nu|=1} g_1^p(x,\nu)\,\mathrm{d}\sigma_\nu r^{n-1}\,\mathrm{d}r\right)^{1/p}\\ &\qquad = (m-j)d^{m-j}\left(\frac{d^n}{n}\right)^{1/p}\left(\int_{|\nu|=1} g_1^p(x,\nu)\,\mathrm{d}\sigma_\nu\right)^{1/p}\\ &\qquad = (m-j)d^{m-j}\left(\frac{d^n}{n}\right)^{1/p} g^*(x).\end{aligned}$$

[4]Minkowski's integral inequality states that, for a function $u \in C(\bar{B}\times\bar{\Omega})$,

$$\|\int_B u(x,\cdot)\,\mathrm{d}x\|_{L_p(\Omega)} \le \int_B \|u(x,\cdot)\|_{L_p(\Omega)}\,\mathrm{d}x.$$

Inserting this into (3.3) we get, by Hölder's inequality (see Ch.1, Sec.1.1.2).

$$\begin{aligned}\|D^\alpha v - Q_{m-j}(D^\alpha v)\|_{L_p(\Omega)} &\leq \frac{1}{|B|}\,(m-j)\,d^{m-j}\left(\frac{d^n}{n}\right)^{1/p}\int_B g^*(x)\,\mathrm{d}x \\ &\leq \frac{1}{|B|}\,(m-j)\,d^{m-j}|B|^{1-(1/p)}\left(\frac{d^n}{n}\right)^{1/p}\|g^*\|_{L_p(\mathbb{R}^n)}\end{aligned}$$

and hence, by Lemma 13,

$$\begin{aligned}\|D^\alpha v - Q_{m-j}(D^\alpha v)\|_{L_p(\Omega)} &\leq \frac{m-j}{n^{1/p}}\,\frac{d^{m-j+(n/p)}}{|B|^{1/p}}\,\frac{p}{p-1}\,\omega_n^{1/p}\|g\|_{L_p(\mathbb{R}^n)} \\ &= \frac{m-j}{n^{1/p}}\,\frac{d^{m-j+(n/p)}}{|B|^{1/p}}\,\frac{p}{p-1}\,\omega_n^{1/p}\|g\|_{L_p(\Omega)}.\end{aligned}$$

However,

$$\begin{aligned}\|g\|_{L_p(\Omega)} &= \|\sum_{|\beta|=m-j}\frac{1}{\beta!}|D^\beta D^\alpha v|\,\|_{L_p(\Omega)} \\ &\leq \sum_{|\beta|=m-j}\frac{1}{\beta!}\|D^\beta D^\alpha v\|_{L_p(\Omega)} \\ &\leq \left(\sum_{|\beta|=m-j}(\beta!)^{-p'}\right)^{1/p'}\left(\sum_{|\beta|=m-j}\|D^\beta D^\alpha v\|^p_{L_p(\Omega)}\right)^{1/p},\end{aligned}$$

where $1/p + 1/p' = 1$. Therefore,

$$\begin{aligned}\|D^\alpha v - Q_{m-j}(D^\alpha v)\|_{L_p(\Omega)} \leq \frac{m-j}{n^{1/p}}\,\frac{d^{m-j+(n/p)}}{|B|^{1/p}}\,\frac{p}{p-1}\,\omega_n^{1/p}\left(\sum_{|\beta|=m-j}(\beta!)^{-p'}\right)^{1/p'} \\ \times\left(\sum_{|\beta|=m-j}\|D^\beta D^\alpha v\|^p_{L_p(\Omega)}\right)^{1/p}.\end{aligned}$$

Recalling that $|\alpha| = j$, we deduce that

$$|v - Q_m(v)|_{W_p^j(\Omega)} \leq K_j |v|_{W_p^m(\Omega)},$$

where

$$K_j = C\frac{d^{m-j+(n/p)}}{|B|^{1/p}}.$$

Since $Q_m \in \mathcal{P}_{m-1}$, this completes the proof of Theorem 8, and thereby also the proof of the Bramble-Hilbert lemma (Lemma 12). ■

**Corollary 3** *Let $\Omega \subset \mathbb{R}^n$ be a bounded open set of diameter $d$ which is star-shaped with respect to every point of an open ball $B \subset \Omega$ of diameter $\mu d$, $\mu \in (0,1]$. Suppose that $1 < p < \infty$ and $0 \leq j < m$, $m \geq 1$. If $v \in W_p^m(\Omega)$ then*

$$\inf_{Q\in\mathcal{P}_{m-1}}|v-Q|_{W_p^j(\Omega)} \leq C(m,n,p,j,\mu)d^{m-j}|v|_{W_p^m(\Omega)},$$

*where*

$$C(m,n,p,j,\mu) = \mu^{-n/p}\left(\sharp\{\alpha\,:\,|\alpha|=j\}\right)\frac{p(m-j)}{p-1}\left(\sum_{|\beta|=m-j}(\beta!)^{-p'}\right)^{1/p},$$

*with* $1/p + 1/p' = 1$.

**Proof** Note that

$$|B| = \frac{(\mu d)^n \omega_n}{n}.$$

■

In order to minimise the size of the constant $C(m,n,p,j,\mu)$, $\mu$ should be taken as large as possible a number in the interval $(0,1]$ such that $\Omega$ is star-shaped with respect to each point in a ball $B \subset \Omega$ of radius $\mu d$.

### 3.2.2 Error bounds on the interpolation error

We shall apply Corollary 3 to derive a bound on the error between a function and its finite element interpolant. We begin by estimating the norm of the local interpolation operator.

**Lemma 14** *Let $(K,\mathcal{P},\mathcal{N})$ be a finite element such that the diameter of $K$ is equal to 1, $\mathcal{P} \subset W^m_\infty(K)$ and $\mathcal{N} \subset (C^l(\bar{K}))'$ (i.e. the nodal variables in $\mathcal{N}$ involve derivatives up to order $l$, and each element of the set $\mathcal{N}$ is a bounded linear functional on $C^l(\bar{K})$). Then the local interpolation operator is bounded from $C^l(\bar{K})$ into $W^m_p(K)$ for $1 < p < \infty$.*

**Proof** Let $\mathcal{N} = \{N_1,\ldots,N_k\}$, and let $\{\psi_1,\ldots,\psi_k\} \subset \mathcal{P}$ be the basis dual to $\mathcal{N}$. The local interpolant of a function $u$ is defined by the formula

$$\mathcal{I}_K u = \sum_{i=1}^k N_i(u)\psi_i,$$

where each $\psi_i \in W^m_\infty(K) \subset W^m_p(K)$, $1 < p < \infty$, by hypothesis. Thus

$$\begin{aligned}
\|\mathcal{I}_K u\|_{W^m_p(K)} &\le \sum_{i=1}^k |N_i(u)|\,\|\psi_i\|_{W^m_p(K)} \\
&\le \left(\sum_{i=1}^k \|N_i\|_{(C^l(\bar{K})'}\|\psi_i\|_{W^m_p(K)}\right)\|u\|_{C^l(\bar{K})} \\
&= Const.\,\|u\|_{C^l(\bar{K})}.
\end{aligned}$$

and that completes the proof. ■

We define

$$\sigma(K) = \sup_{v\in C^l(\bar{K})}\frac{\|\mathcal{I}_K v\|_{W^m_p(K)}}{\|v\|_{C^l(\bar{K})}},$$

the **norm of the local interpolation operator** $\mathcal{I}_K : C^l(\bar{K}) \to W^m_p(K)$.

**Theorem 9** *Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element satisfying the following conditions:*

*(i) $K$ is star-shaped with respect to some ball contained in $K$;*

*(ii) $\mathcal{P}_{m-1} \subset \mathcal{P} \subset W^m_\infty(K)$;*

*(iii) $\mathcal{N} \subset (C^l(\bar{K}))'$.*

*Suppose that $1 < p < \infty$ and $m - l - (n/p) > 0$. Then, for $0 \leq j \leq m$ and $v \in W^m_p(K)$ we have that*

$$|v - \mathcal{I}_K v|_{W^j_p(K)} \leq C(m, n, p, \mu, \sigma(\hat{K}))h_K^{m-j}|v|_{W^m_p(K)},$$

*where $h_K$ is the diameter of $K$, $\hat{K} = \{x/h_K : x \in K\}$ and $\mu$ is the largest real number in the interval $(0, 1]$ such that a ball of diameter $\mu h_K$ is contained in $K$.*

**Proof** It suffices to take $K$ with diameter equal to 1, in which case $K = \hat{K}$; the general case follows by a simple scaling argument. Also, note that the local interpolation operator is well defined on $W^m_p(K)$ by the Sobolev embedding theorem[5] and there exists a constant $C = C_{m,n,p}$ such that, for all $v \in W^m_p(K)$,

$$\|v\|_{C^l(\bar{K})} \leq C_{m,n,p}\|v\|_{W^m_p(K)}.$$

Let $Q_m v$ be as in the proof of Theorem 8. Since $\mathcal{I}_K f = f$ for any $f \in \mathcal{P}$, we have that

$$\mathcal{I}_K Q_m v = Q_m v,$$

because $Q_m v \in \mathcal{P}_{m-1} \subset \mathcal{P}$. Thus,

$$\begin{aligned}
\|v - \mathcal{I}_K v\|_{W^m_p(K)} &\leq \|v - Q_m v\|_{W^m_p(K)} + \|Q_m v - \mathcal{I}_K v\|_{W^m_p(K)} \\
&= \|v - Q_m v\|_{W^m_p(K)} + \|\mathcal{I}_K(Q_m v - \mathcal{I}_K v)\|_{W^m_p(K)} \\
&\leq \|v - Q_m v\|_{W^m_p(K)} + \sigma(K)\,\|Q_m v - \mathcal{I}_K v\|_{C^l(\bar{K})} \\
&\leq (1 + C_{m,n,p}\sigma(K))\,\|v - Q_m v\|_{W^m_p(K)},
\end{aligned}$$

by the Sobolev embedding theorem. Finally, by (3.2) we deduce that

$$\|v - \mathcal{I}_K v\|_{W^m_p(K)} \leq C(m, n, p, \mu, \sigma(K))\,|v|_{W^m_p(K)},$$

and hence, for $0 \leq j \leq m$, we have that

$$|v - \mathcal{I}_K v|_{W^j_p(K)} \leq C(m, n, p, \mu, \sigma(K))\,|v|_{W^m_p(K)}.$$

That completes the proof. ■

Next we show that, under a certain regularity condition on the subdivision $\mathcal{T} = \{K\}$ of the computational domain $\Omega$, the constant $C(m, n, p, \mu, \sigma(\hat{K}))$ can be made independent of $\sigma(\hat{K})$.

[5]The Sobolev embedding theorem asserts that $W^m_p(K) \subset C^l(\bar{K})$ for $m - l > \frac{n}{p}$, $1 \leq p < \infty$ and the identity operator $Id : v \in W^m_p(K) \mapsto v \in C^l(\bar{K})$ is a bounded linear operator.

Let us suppose that each $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ is affine equivalent to a single reference element $(K, \mathcal{P}, \mathcal{N})$ through an affine transformation

$$x \mapsto \hat{x} = Ax \equiv ax + b,$$

where $a = (a_{ij})$ is an invertible $n \times n$ matrix and $b$ is a column vector of size $n$, of the same length as the column vector $x$.[6] We shall denote the entries of the matrix $a^{-1}$ by $(a^{-1})_{ij}$. The definition of affine equivalence yields:

$$\hat{\mathcal{I}}_{\hat{K}} \hat{v}(\hat{x}) = \sum_{N \in \mathcal{N}} (A_* N) \hat{v} (A^{-1})^* \psi_N(\hat{x}),$$

where

$$(A_* N)(\hat{v}) = N(A^* \hat{v}), \quad (A^* \hat{v}) x = \hat{v}(Ax).$$

Thus,

$$\begin{aligned} |(A_* N)(\hat{v})| &= |N(A^* \hat{v})| \leq C_N \|A^* \hat{v}\|_{C^l(\bar{K})} \\ &\leq C_{N,n,l} \left(1 + \max_{1 \leq i,j \leq n} |a_{ij}|\right)^l \|\hat{v}\|_{C^l(\bar{\hat{K}})}. \end{aligned}$$

Also,

$$\|(A^{-1})^* \psi_N\|_{W_p^m(\hat{K})} \leq C'_{N,n,m} \left(1 + \max_{1 \leq i,j \leq n} |(a^{-1})_{ij}|\right)^m |\det a|^{1/p} \|\psi_N\|_{W_p^m(K)}.$$

Since $\|\psi_N\|_{W_p^m(K)}$ is a fixed constant on the reference element $K$, we have that

$$\begin{aligned} \|\hat{\mathcal{I}}_{\hat{K}} \hat{v}\|_{W_p^m(\hat{K})} \leq C_{ref} &\left(1 + \max_{1 \leq i,j \leq n} |a_{ij}|\right)^l \\ &\times \left(1 + \max_{1 \leq i,j \leq n} |(a^{-1})_{ij}|\right)^m |\det a|^{1/p} \|\hat{v}\|_{C^l(\bar{\hat{K}})}, \end{aligned}$$

where

$$C_{ref} = |\mathcal{N}| \max_{N \in \mathcal{N}} \{C_{N,n,l}\} \max_{N \in \mathcal{N}} \{C'_{N,n,m}\} \max_{N \in \mathcal{N}} \{\|\psi_N\|_{W_p^m(K)}\},$$

and $|\mathcal{N}|$ denotes the number of nodal variables (i.e. the dimension of $\mathcal{P}$). Thus, we have shown that

$$\sigma(\hat{K}) \leq C_{ref} \left(1 + \max_{1 \leq i,j \leq n} |a_{ij}|\right)^l \times \left(1 + \max_{1 \leq i,j \leq n} |(a^{-1})_{ij}|\right)^m |\det a|^{1/p}.$$

[6] To avoid confusion between the finite element $(K, \mathcal{P}, \mathcal{N})$, associated with an element domain $K$ in the triangulation, and the affine image of $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ considered here, it would have been better to use a new symbol $(\tilde{K}, \tilde{\mathcal{P}}, \tilde{\mathcal{N}})$, say, instead of $(K, \mathcal{P}, \mathcal{N})$ and $\tilde{x}$ instead of $x$, to denote the affine image; but this would have complicated the notation. Here and in the next 17 lines we shall, temporarily, adopt this sloppy notation. Thereafter, $(K, \mathcal{P}, \mathcal{N})$ will, again, signify a finite element on an element domain $K$ in the triangulation.

Assuming that the subdivision $\mathcal{T} = \{K\}$ is **regular** in the sense that

$$\exists \mu > 0 \ \forall K \in \mathcal{T} \quad \mu h_K \leq \rho_K (\leq h_K),$$

where $h_K$ is the diameter of $K$ and $\rho_K$ is the radius of the largest sphere (largest circle for $n = 2$) contained in $K$, it is a straightforward exercise in geometry to show that $\sigma(\hat{K}) \leq C_\mu$, where $C_\mu$ is a fixed constant dependent on $\mu$, but independent of $\hat{K} \in \mathcal{T}$. Consequently,

$$|v - \mathcal{I}_K v|_{W_p^j(K)} \leq C(m, p, n, \mu) h_K^{m-j} |v|_{W_p^m(K)} \tag{3.4}$$

for each $K \in \mathcal{T}$ provided that $\mathcal{T}$ is a regular subdivision, $1 < p < \infty$, $m - l - (n/p) > 0$ and $0 \leq j \leq m$; from this, and recalling the definition of the global interpolant of $v \in W_p^m(\Omega)$ it follows that

$$\left( \sum_{K \in \mathcal{T}} h_K^{(j-m)p} |v - \mathcal{I}_h v|^p_{W_p^j(K)} \right)^{1/p} \leq C(m, p, n, \mu) |v|_{W_p^m(\Omega)}. \tag{3.5}$$

We shall also need the following somewhat cruder statement which is a straightforward consequence of (3.4); still supposing that the triangulation $\mathcal{T}$ is regular, $1 < p < \infty$, $m - l - (n/p) > 0$ and $0 \leq j \leq m$, and $v \in W_p^m(\Omega)$, we have that

$$|v - \mathcal{I}_h v|_{W_p^j(\Omega)} \leq C(m, p, n, \mu) h^{m-j} |v|_{W_p^m(\Omega)}, \tag{3.6}$$

where $h = \max_{K \in \mathcal{T}} h_K$. These interpolation error estimates are of crucial importance in finite element error analysis.

## 3.3 Optimal error bounds in the $H^1(\Omega)$ norm – revisited

In this section we return to the discussion of error estimation in the $H^1(\Omega)$ norm. In Chapter 2 we showed, for the finite element approximation $u_h \in V_h$ to the weak solution $u$ of the homogeneous Dirichlet boundary value problem for a second-order elliptic equation, that

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{c_1}{c_0} \inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}. \tag{3.7}$$

(c.f. Céa's lemma 5). Thus, restricting ourselves to the case of Poisson's equation and a continuous piecewise linear approximation $u_h$ defined on a uniform triangulation of $\Omega = (0, 1)^2$, we proved that, whenever $u \in H^2(\Omega) \cap H_0^1(\Omega)$, we have

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

Now, equipped with the interpolation error estimate (3.6) we can derive an analogous error bound in a more general setting; also, we can generalise to higher degree piecewise polynomial approximations.

Suppose that $\Omega \subset \mathbb{R}^n$ and that it can be represented as a union of element domains $K$ such that conditions (i), (ii) and (iii) of Theorem 9 hold with $p = 2$. Given such a triangulation $\mathcal{T}$ of $\Omega$ we shall suppose that it is regular in the sense introduced in the previous section and we put

$$V_h = \mathcal{I}_h(H^m(\Omega) \cap H_0^1(\Omega)).$$

Then,

$$\inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \leq \|u - \mathcal{I}_h u\|_{H^1(\Omega)} \leq C(m, n, \mu) h^{m-1} |u|_{H^m(\Omega)}.$$

Substituting this into (3.7), we arrive at the following error bound:

$$\|u - u_h\|_{H^1(\Omega)} \leq C(m, n, \mu, c_1, c_0) h^{m-1} |u|_{H^m(\Omega)}, \tag{3.8}$$

provided that $u \in H^m(\Omega) \cap H_0^1(\Omega)$. In particular, this will be the case if we use continuous piecewise polynomials of degree $m - 1$ on a regular triangulation of $\Omega$, with $m = 2$ corresponding to our earlier result with piecewise linear basis functions. The inequality (3.8) is usually referred to as an **optimal error bound**, since for a given $m$ the smallest possible error that can be, in general, achieved in the $H^1(\Omega)$ norm is of size $\mathcal{O}(h^{m-1})$.

## 3.4 Variational crimes

To conclude this chapter we briefly comment on a further issue which arises in the implementation of finite element methods. Let us consider the weak formulation of the second-order elliptic partial differential equation (1.5) on a bounded open set $\Omega \subset \mathbb{R}^n$, in the case of a homogeneous Dirichlet boundary condition (1.6):

$$\text{find } u \in H^1(\Omega) \text{ such that } a(u, v) = l(v) \text{ for all } v \in H_0^1(\Omega),$$

where, as before,

$$\begin{aligned} a(w, v) &= \sum_{i,j=1}^{n} \int_\Omega a_{ij}(x) \frac{\partial w}{\partial x_i} \frac{\partial v}{\partial x_j} \, \mathrm{d}x \\ &\quad + \sum_{i=1}^{n} \int_\Omega b_i(x) \frac{\partial w}{\partial x_i} v \, \mathrm{d}x + \int_\Omega c(x) w v \, \mathrm{d}x \end{aligned} \tag{3.9}$$

and

$$l(v) = \int_\Omega f(x) v(x) \, \mathrm{d}x. \tag{3.10}$$

The associated finite element method is based on choosing a finite element subspace $V_h \subset H_0^1(\Omega)$ consisting of continuous piecewise polynomials of a certain degree defined on a subdivision of the computational domain $\Omega$, and considering the approximate problem

$$\text{find } u_h \in V_h \text{ such that } a(u_h, v_h) = l(v_h) \text{ for all } v_h \in V_h.$$

Unfortunately, unless the coefficients $a_{ij}$, $b_i$ and $c$ and the right-hand side $f$ are exceptionally simple functions, the integrals which appear in the definitions of $a(\cdot,\cdot)$ and $l(\cdot)$ will not be possible to evaluate exactly, and numerical integration rules (such as the trapezium rule, Simpson's rule, Gauss-type rules and their multi-dimensional counterparts) will have to be used to calculate $a(\cdot,\cdot)$ and $l(\cdot)$ approximately. Without focusing on any particular quadrature rule, we attempt to analyse the effects of this quadrature-induced perturbation on the accuracy of the exactly-integrated finite element method. To keep the discussion simple, let us suppose that the bilinear form $a(\cdot,\cdot)$ is still calculated exactly, but that $l(\cdot)$ has been replaced by an approximation $l_h(\cdot)$, thereby leading to the following definition of $u_h$:

$$\text{find } u_h \in V_h \text{ such that } a(u_h, v_h) = l_h(v_h) \text{ for all } v_h \in V_h.$$

We recall that a key step in developing the finite element error analysis was the presence of the Galerkin orthogonality property. With this new definition of $u_h$, however, we have that

$$a(u - u_h, v_h) = a(u, v_h) - a(u_h, v_h) = l(v_h) - l_h(v_h) \neq 0, \quad v_h \in V_h,$$

and Galerkin orthogonality no longer holds. We say that we have committed a **variational crime** by replacing $l(\cdot)$ by $l_h(\cdot)$. We wish to study the extent to which the accuracy of the basic finite element approximation is disturbed by this variational crime.

Assuming that

$$c(x) - \frac{1}{2}\sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i} \geq 0,$$

we have that

$$a(v, v) \geq c_0 \|v\|^2_{H^1(\Omega)},$$

with $c_0$ a positive constant (as in Section 1.2). Thus,

$$\begin{aligned} c_0\|u - u_h\|^2_{H^1(\Omega)} &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\ &= a(u - u_h, u - v_h) + l(v_h - u_h) - l_h(v_h - u_h) \\ &\leq c_1\|u - u_h\|_{H^1(\Omega)}\|u - v_h\|_{H^1(\Omega)} \\ &\qquad + \sup_{w_h \in V_h} \frac{|l(w_h) - l_h(w_h)|}{\|w_h\|_{H^1(\Omega)}} \|v_h - u_h\|_{H^1(\Omega)}, \end{aligned}$$

with $c_1$ a positive constant (as in Section 1.2). To simplify writing, we define

$$\|l - l_h\|_{-1,h} = \sup_{w_h \in V_h} \frac{|l(w_h) - l_h(w_h)|}{\|w_h\|_{H^1(\Omega)}}.$$

Hence,

$$\begin{aligned} c_0\|u-u_h\|^2_{H^1(\Omega)} &\le c_1\|u-u_h\|_{H^1(\Omega)}\|u-v_h\|_{H^1(\Omega)} \\ &\quad +\|l-l_h\|_{-1,h}\left(\|u-u_h\|_{H^1(\Omega)} + \|u-v_h\|_{H^1(\Omega)}\right) \\ &= \left(c_1\|u-v_h\|_{H^1(\Omega)} + \|l-l_h\|_{-1,h}\right)\|u-u_h\|_{H^1(\Omega)} \\ &\quad +\|l-l_h\|_{-1,h}\|u-v_h\|_{H^1(\Omega)}. \end{aligned} \tag{3.11}$$

Now applying the elementary inequality

$$ab \le \frac{1}{2c_0}a^2 + \frac{c_0}{2}b^2, \quad a, b \ge 0,$$

we have that

$$\begin{aligned} &\left(c_1\|u-v_h\|_{H^1(\Omega)} + \|l-l_h\|_{-1,h}\right)\|u-u_h\|_{H^1(\Omega)} \\ &\quad \le \frac{1}{2c_0}\left(c_1\|u-v_h\|_{H^1(\Omega)} + \|l-l_h\|_{-1,h}\right)^2 + \frac{c_0}{2}\|u-u_h\|^2_{H^1(\Omega)}. \end{aligned}$$

Substituting this into (3.11) gives

$$\begin{aligned} c_0^2\|u-u_h\|^2_{H^1(\Omega)} \;\le\; &\left(c_1\|u-v_h\|_{H^1(\Omega)} + \|l-l_h\|_{-1,h}\right)^2 \\ &+2c_0\|l-l_h\|_{-1,h}\|u-v_h\|_{H^1(\Omega)}. \end{aligned}$$

Noting that $c_0 \le c_1$, this yields

$$c_0^2\|u-u_h\|^2_{H^1(\Omega)} \le 2\left(c_1\|u-v_h\|_{H^1(\Omega)} + \|l-l_h\|_{-1,h}\right)^2,$$

and therefore,

$$\|u-u_h\|_{H^1(\Omega)} \le \frac{c_1\sqrt{2}}{c_0}\|u-v_h\|_{H^1(\Omega)} + \frac{\sqrt{2}}{c_0}\|l-l_h\|_{-1,h}.$$

Equivalently,

$$\|u-u_h\|_{H^1(\Omega)} \le \frac{c_1\sqrt{2}}{c_0}\|u-v_h\|_{H^1(\Omega)} + \frac{\sqrt{2}}{c_0}\sup_{w_h \in V_h}\frac{|l(w_h) - l_h(w_h)|}{\|w_h\|_{H^1(\Omega)}}. \tag{3.12}$$

Since $v_h \in V_h$ is arbitrary, it follows that we have proved the following perturbed version of Céa's lemma:

$$\begin{aligned} \|u-u_h\|_{H^1(\Omega)} \;\le\; &\frac{c_1\sqrt{2}}{c_0}\min_{v_h \in V_h}\|u-v_h\|_{H^1(\Omega)} \\ &+\frac{\sqrt{2}}{c_0}\sup_{w_h \in V_h}\frac{|l(w_h) - l_h(w_h)|}{\|w_h\|_{H^1(\Omega)}}. \end{aligned}$$

The second term on the right-hand side of (3.12) quantifies the extent to which the accuracy of the exactly integrated finite element method is affected by the failure of Galerkin orthogonality. Indeed, arguing in the same manner as in Section 3.3 will lead to the error bound

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} &\leq C(m, n, \mu, c_1, c_0) h^{m-1} |u|_{H^m(\Omega)} \\ &\quad + \frac{\sqrt{2}}{c_0} \sup_{w_h \in V_h} \frac{|l(w_h) - l_h(w_h)|}{\|w_h\|_{H^1(\Omega)}}. \end{aligned}$$

Thus, in order to retain the accuracy of the exactly integrated method, the numerical quadrature rule has to be selected so that the second term on the right is also of size $\mathcal{O}(h^{m-1})$; in the case of continuous piecewise linear basis functions ($m = 2$) this means that the additional error should be at most $\mathcal{O}(h)$.

The situation when $a(\cdot, \cdot)$ is perturbed to $a_h(\cdot, \cdot)$ is analysed in a similar manner. We shall not discuss variational crimes which arise from replacing the computational domain $\Omega$ by a "conveniently chosen close-by domain" $\Omega_h$; for the details of the analysis the reader is referred to the books on the reading list and references therein.

# Chapter 4

# A posteriori error analysis by duality

In this chapter we shall derive a computable bound on the global error and indicate the implementation of this result into an adaptive algorithm with reliable error control.

## 4.1 The one-dimensional model problem

In order to illuminate the key ideas and avoid technical difficulties, we shall consider the two-point boundary value problem

$$\begin{aligned} -u'' + b(x)u' + c(x)u &= f(x), \qquad 0 < x < 1, \\ u(0) &= 0, \quad u(1) = 0, \end{aligned}$$

where $b \in W^1_\infty(0,1)$, $c \in L_\infty(0,1)$ and $f \in L_2(0,1)$. Letting

$$a(w,v) = \int_0^1 [w'(x)v'(x) + b(x)w'(x)v(x) + c(x)w(x)v(x)]\,\mathrm{d}x$$

and

$$l(v) = \int_0^1 f(x)v(x)\,\mathrm{d}x,$$

the weak formulation of this problem can be stated as follows:

$$\text{find } u \in H^1_0(0,1) \text{ such that } a(u,v) = l(v) \text{ for all } v \in H^1_0(0,1).$$

Assuming that

$$c(x) - \frac{1}{2}b'(x) \geq 0, \quad \text{for } x \in (0,1), \tag{4.1}$$

there exists a unique weak solution, $u \in H^1_0(0,1)$.

The finite element approximation of this problem is constructed by considering a (possibly non-uniform) subdivision of the interval $[0,1]$ by the points $0 = x_0 < x_1 <$

$\ldots < x_{N-1} < x_N = 1$ and defining the finite element space $V_h \subset H_0^1(0,1)$ consisting of continuous piecewise polynomials of a certain degree on this subdivision. To keep matters simple, let us suppose that $V_h$ consists of continuous piecewise linear functions, as described in Chapter 2. The finite element approximation of the boundary value problem is:

$$\text{find } u_h \in V_h \text{ such that } a(u_h, v_h) = l(v_h) \text{ for all } v_h \in V_h.$$

We let $h_i = x_i - x_{i-1}$, $i = 1, \ldots, N$, and put $h = \max_i h_i$.

We wish to derive an *a posteriori* error bound; that is, we aim to quantify the size of the global error $u - u_h$ in terms of the mesh parameter $h$ and the computed solution $u_h$ (rather then the analytical solution $u$, as in an *a priori* error analysis). To do so, we consider the following auxiliary boundary value problem

$$\begin{aligned} -z'' - (b(x)z)' + c(x)z = (u - u_h)(x), \qquad & 0 < x < 1, \\ z(0) = 0, \quad z(1) = 0, & \end{aligned}$$

called the **dual** or **adjoint** problem.

We begin our error analysis by noting that the definition of the dual problem and a straightforward integration by parts yield (recall that $(u-u_h)(0) = 0$, $(u-u_h)(1) = 0$):

$$\begin{aligned} \|u - u_h\|^2_{L_2(0,1)} &= (u - u_h, u - u_h) = (u - u_h, -z'' - (bz)' + cz) \\ &= a(u - u_h, z). \end{aligned}$$

By virtue of the Galerkin orthogonality property,

$$a(u - u_h, z_h) = 0 \qquad \forall z_h \in V_h.$$

In particular, choosing $z_h = \mathcal{I}_h z \in V_h$, the continuous piecewise linear interpolant of the function $z$, associated with the subdivision $0 = x_0 < x_1 < \ldots < x_{N-1} < x_N = 1$, we have that

$$a(u - u_h, \mathcal{I}_h z) = 0.$$

Thus,

$$\begin{aligned} \|u - u_h\|^2_{L_2(0,1)} &= a(u - u_h, z - \mathcal{I}_h z) = a(u, z - \mathcal{I}_h z) - a(u_h, z - \mathcal{I}_h z) \\ &= (f, z - \mathcal{I}_h z) - a(u_h, z - \mathcal{I}_h z). \end{aligned} \tag{4.2}$$

We observe that by this stage the right-hand side no longer involves the unknown analytical solution $u$. Now,

$$\begin{aligned} a(u_h, z - \mathcal{I}_h z) &= \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} u_h'(x)\,(z - \mathcal{I}_h z)'(x)\,\mathrm{d}x \\ &\quad + \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} b(x)\,u_h'(x)\,(z - \mathcal{I}_h z)(x)\,\mathrm{d}x \\ &\quad + \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} c(x)\,u_h(x)\,(z - \mathcal{I}_h z)(x)\,\mathrm{d}x. \end{aligned}$$

Integrating by parts in each of the $(N-1)$ integrals in the first sum on the right-hand side, noting that $(z - \mathcal{I}_h z)(x_i) = 0$, $i = 0, \ldots, N$, we deduce that

$$a(u_h, z - \mathcal{I}_h z) = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \left[-u_h''(x) + b(x)u_h'(x) + c(x)u_h(x)\right] (z - \mathcal{I}_h z)(x)\, \mathrm{d}x.$$

Further

$$(f, z - \mathcal{I}_h z) = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} f(x)\, (z - \mathcal{I}_h z)(x)\, \mathrm{d}x.$$

Substituting these two identities into (4.2), we deduce that

$$\|u - u_h\|_{L_2(0,1)}^2 = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} R(u_h)(x)\, (z - \mathcal{I}_h z)(x)\, \mathrm{d}x, \tag{4.3}$$

where, for $i = 1, \ldots, N$,

$$R(u_h)(x) = f(x) + u_h''(x) - b(x)u_h'(x) - c(x)u_h(x), \qquad x \in (x_{i-1}, x_i).$$

The function $R(u_h)$ is called **the finite element residual**; it measures the extent to which $u_h$ fails to satisfy the differential equation $-u'' + b(x)u' + c(x)u = f(x)$ on the interval $(0, 1)$. Now, applying the Cauchy–Schwarz inequality on the right-hand side of (4.3) yields

$$\|u - u_h\|_{L_2(0,1)}^2 \leq \sum_{i=1}^{N} \|R(u_h)\|_{L_2(x_{i-1},x_i)} \|z - \mathcal{I}_h z\|_{L_2(x_{i-1},x_i)}.$$

Recalling from the proof of Theorem 3 (with $\zeta = z - \mathcal{I}_h z$ and noting that $\zeta''(x) = z''(x)$ for all $x$ in $(x_{i-1}, x_i)$, since $\mathcal{I}_h z$ is a linear function on $(x_{i-1}, x_i)$, $i = 1, \ldots, N$) that

$$\|z - \mathcal{I}_h z\|_{L_2(x_{i-1},x_i)} \leq \left(\frac{h_i}{\pi}\right)^2 \|z''\|_{L_2(x_{i-1},x_i)}, \quad i = 1, \ldots, N,$$

we deduce that

$$\|u - u_h\|_{L_2(0,1)}^2 \leq \frac{1}{\pi^2} \sum_{i=1}^{N} h_i^2 \|R(u_h)\|_{L_2(x_{i-1},x_i)} \|z''\|_{L_2(x_{i-1},x_i)}$$

and consequently,

$$\|u - u_h\|_{L_2(0,1)}^2 \leq \frac{1}{\pi^2} \left( \sum_{i=1}^{N} h_i^4 \|R(u_h)\|_{L_2(x_{i-1},x_i)}^2 \right)^{1/2} \|z''\|_{L_2(0,1)}. \tag{4.4}$$

The rest of the analysis is aimed at eliminating $z''$ from the right-hand side of (4.4). We recall that

$$z'' = u_h - u - (b\, z)' + c\, z = u_h - u - b\, z' + (c - b')\, z,$$

and therefore,

$$\begin{aligned}\|z''\|_{L_2(0,1)} &\leq \|u-u_h\|_{L_2(0,1)} + \|b\|_{L_\infty(0,1)}\|z'\|_{L_2(0,1)} \\ &\quad + \|c-b'\|_{L_\infty(0,1)}\|z\|_{L_2(0,1)}. \end{aligned} \tag{4.5}$$

We shall show that both $\|z'\|_{L_2(0,1)}$ and $\|z\|_{L_2(0,1)}$ can be bounded in terms of $\|u-u_h\|_{L_2(0,1)}$ and then, by virtue of (4.5), we shall deduce that the same is true of $\|z''\|_{L_2(0,1)}$. Let us observe that

$$(-z'' - (bz)' + cz, z) = (u-u_h, z).$$

Integrating by parts and noting that $z(0)=0$ and $z(1)=0$ yields

$$\begin{aligned}(-z'' - (bz)' + cz, z) &= (z', z') + (bz, z') + (cz, z) \\ &= \|z'\|^2_{L_2(0,1)} + \frac{1}{2}\int_0^1 b(x)[z^2(x)]'\,\mathrm{d}x + \int_0^1 c(x)[z(x)]^2\,\mathrm{d}x.\end{aligned}$$

Integrating by parts, again, in the second term on the right gives

$$(-z'' - (bz)' + cz, z) = \|z'\|^2_{L_2(0,1)} - \frac{1}{2}\int_0^1 b'(x)[z^2(x)]\,\mathrm{d}x + \int_0^1 c(x)[z(x)]^2\,\mathrm{d}x.$$

Hence,

$$\|z'\|^2_{L_2(0,1)} + \int_0^1 \left(c(x) - \frac{1}{2}b'(x)\right)[z(x)]^2\,\mathrm{d}x = (u-u_h, z),$$

and thereby, noting (4.1),

$$\|z'\|^2_{L_2(0,1)} \leq (u-u_h, z) \leq \|u-u_h\|_{L_2(0,1)}\|z\|_{L_2(0,1)}. \tag{4.6}$$

By the Poincaré–Friedrichs inequality,

$$\|z\|^2_{L_2(0,1)} \leq \frac{1}{2}\|z'\|^2_{L_2(0,1)}.$$

Thus, (4.6) gives

$$\|z\|_{L_2(0,1)} \leq \frac{1}{2}\|u-u_h\|_{L_2(0,1)}. \tag{4.7}$$

Inserting this into the right-hand side of (4.6) yields

$$\|z'\|_{L_2(0,1)} \leq \frac{1}{\sqrt{2}}\|u-u_h\|_{L_2(0,1)}. \tag{4.8}$$

Now we substitute (4.7) and (4.8) into (4.5) to deduce that

$$\|z''\|_{L_2(0,1)} \leq K\|u-u_h\|_{L_2(0,1)}. \tag{4.9}$$

Where

$$K = 1 + \frac{1}{\sqrt{2}}\|b\|_{L_\infty(0,1)} + \frac{1}{2}\|c-b'\|_{L_\infty(0,1)}.$$

It is important to note here that $K_0$ involves only known quantities, namely the coefficients in the differential equation under consideration, and therefore it can be computed without difficulty. Inserting (4.9) into (4.4), we arrive at our final result, the computable *a posteriori* error bound,

$$\|u - u_h\|_{L_2(0,1)} \leq K_0 \left( \sum_{i=1}^{N} h_i^4 \|R(u_h)\|^2_{L_2(x_{i-1},x_i)} \right)^{1/2}, \tag{4.10}$$

where $K_0 = K/\pi^2$.

The name *a posteriori* stems from the fact that (4.10) can only be employed to quantify the size of the approximation error that has been committed in the course of the computation after $u_h$ has been computed. In the next section we shall describe the construction of an adaptive mesh refinement algorithm based on the bound (4.10).

## 4.2 An adaptive algorithm

Suppose that $TOL$ is a prescribed tolerance and that our aim is to compute a finite element approximation $u_h$ to the unknown solution $u$ (with the same definition of $u$ and $u_h$ as in the previous section) so that

$$\|u - u_h\|_{L_2(0,1)} \leq TOL.$$

We shall use the *a posteriori* error bound (4.10) to achieve this goal by successively refining the subdivision, and computing a succession of numerical solutions $u_h$ on these subdivisions, until the stopping criterion

$$K_0 \left( \sum_{i=1}^{N} h_i^4 \|R(u_h)\|^2_{L_2(x_{i-1},x_i)} \right)^{1/2} \leq TOL$$

is satisfied. The algorithm proceeds as follows:

1. Choose an initial subdivision

   $$\mathcal{T}_0 : \quad 0 = x_0^{(0)} < x_1^{(0)} < \ldots < x_{N_0-1}^{(0)} < x_{N_0}^{(0)} = 1$$

   of the interval $[0,1]$, with $h_i^{(0)} = x_i^{(0)} - x_{i-1}^{(0)}$ for $i = 1, \ldots, N_0$, and $h^{(0)} = \max_i h_i^{(0)}$, and consider the associated finite element space $V_{h^{(0)}}$ (of dimension $N_0 - 1$).

2. Compute the corresponding solution $u_{h^{(0)}} \in V_{h^{(0)}}$.

3. Given a computed solution $u_{h^{(m)}} \in V_{h^{(m)}}$ for some $m \geq 0$, defined on a subdivision $\mathcal{T}_m$, stop if

   $$K_0 \left( \sum_{i=1}^{N_m} \left( h_i^{(m)} \right)^4 \|R(u_{h^{(m)}})\|^2_{L_2(x_{i-1}^{(m)},x_i^{(m)})} \right)^{1/2} \leq TOL. \tag{4.11}$$

4. If not, then determine a new subdivision

$$\mathcal{T}_{m+1} : \quad 0 = x_0^{(m+1)} < x_1^{(m+1)} < \ldots < x_{N_{m+1}-1}^{(m+1)} < x_{N_{m+1}}^{(m+1)} = 1$$

of the interval $[0,1]$, with $h_i^{(m+1)} = x_i^{(m+1)} - x_{i-1}^{(m+1)}$ for $i = 1, \ldots, N_{m+1}$ and $h^{(m+1)} = \max_i h_i^{(m+1)}$, and an associated finite element space $V_{h^{(m+1)}}$ (of dimension $N_{m+1} - 1$), with $h^{(m+1)}$ as large as possible (and consequently $N_{m+1}$ as small as possible), such that

$$K_0 \left( \sum_{i=1}^{N_{m+1}} \left( h_i^{(m+1)} \right)^4 \| R(u_{h^{(m)}}) \|^2_{L_2(x_{i-1}^{(m+1)}, x_i^{(m+1)})} \right)^{1/2} = TOL, \tag{4.12}$$

and continue.

Here (4.11) is the stopping criterion and (4.12) is the mesh modification strategy. According to the *a posteriori* error bound (4.10), when the algorithm terminates the global error $\|u - u_h\|_{L_2(0,1)}$ is controlled to within the prescribed tolerance $TOL$. The relation (4.12) defines the new mesh-size by maximality. The natural condition for maximality is **equidistribution**; this means that the residual contributions from individual elements in the subdivision are required to be equal:

$$\left( h_i^{(m+1)} \right)^4 \| R(u_{h^{(m)}}) \|^2_{L_2(x_{i-1}^{(m+1)}, x_i^{(m+1)})} = \frac{TOL^2}{K_0^2 N_{m+1}}$$

for each $i = 1, \ldots, N_{m+1}$; the implementation can be simplified by replacing $N_{m+1}$ on the right-hand side by $N_m$. Then, we have a simple formula for $h_i^{(m+1)}$:

$$h_i^{(m+1)} = \left( \frac{TOL^2}{K_0^2 N_m \| R(u_{h^{(m)}}) \|^2_{L_2(x_{i-1}^{(m+1)}, x_i^{(m+1)})}} \right)^{1/4}, \quad i = 1, \ldots, N_{m+1},$$

from which the $h_i^{(m+1)}$ can be found by treating this as an equation in $h_i^{(m+1)}$, and solving it numerically, for $m$ and $i$ fixed, by some root-finding algorithm (e.g. successive bisection or fixed-point iteration), starting from $i = 1$.

**Reliability** means that the computational error is controlled in a given norm on a given tolerance level. Thus what we have described above is a reliable computational algorithm. **Efficiency** means that the computational effort required to achieve reliability is minimal. It is unclear from the present discussion whether the adaptive algorithm described above is efficient in this sense: although we have minimised the computational effort required to ensure that the right-hand side in the error bound (4.10) is below the given tolerance, the extent to which this implies that we have also minimised the amount of computational effort required to ensure that the left-hand side in (4.10) is less than TOL depends on the sharpness of the inequality (4.10), and this will vary from case to case, depending very much on the choice of the functions $b$, $c$ and $f$.

# Chapter 5

# Evolution problems

In previous chapters we considered the finite element approximation of elliptic boundary value problems. This chapter is devoted to finite element methods for time-dependent problems; in particular, we shall be concerned with the finite element approximation of parabolic equations. Hyperbolic equations will not be discussed in these notes.

## 5.1 The parabolic model problem

Let $\Omega$ be a bounded open set in $\mathbb{R}^n$, $n \geq 1$, with boundary $\Gamma = \partial\Omega$, and let $T > 0$. In $Q = \Omega \times (0, T]$, we consider the initial boundary value problem for the unknown function $u(x,t)$, $x \in \Omega$, $t \in (0,T]$ :

$$\frac{\partial u}{\partial t} - \sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}(a_{ij}(x,t)\frac{\partial u}{\partial x_i}) + \sum_{i=1}^{n} b_i(x,t)\frac{\partial u}{\partial x_i} + c(x,t)u = f(x,t),$$
$$x \in \Omega, \quad t \in (0,T], \tag{5.1}$$
$$u(x,t) = 0, \qquad x \in \Gamma, \quad t \in [0,T], \tag{5.2}$$
$$u(x,0) = u_0(x), \qquad x \in \bar{\Omega}. \tag{5.3}$$

Suppose that $u_0 \in L_2(\Omega)$, and that there exists a positive constant $\tilde{c}$ such that

$$\sum_{i,j=1}^{n} a_{ij}(x,t)\xi_i\xi_j \geq \tilde{c}\sum_{i=1}^{n}\xi_i^2,$$
$$\forall \xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n, \ \forall x \in \bar{\Omega}, \ t \in [0,T]. \tag{5.4}$$

We shall also assume that

$$a_{ij} \in L_\infty(Q), \qquad b_i \in W^1_\infty(Q), \quad i,j = 1, \ldots, n,$$
$$c \in L_\infty(Q), \qquad f \in L_2(Q),$$

and that

$$c(x,t) - \frac{1}{2}\sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i}(x,t) \geq 0, \quad (x,t) \in \bar{Q}, \tag{5.5}$$

as in the elliptic case.

A partial differential equation of the form (5.1) is called a parabolic equation (of second order). Simple examples of parabolic equations are the heat equation

$$\frac{\partial u}{\partial t} = \Delta u$$

and the unsteady advection-diffusion equation

$$\frac{\partial u}{\partial t} - \Delta u + \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} = 0.$$

The proof of the existence of a unique solution to a parabolic initial boundary value problem is more technical than for an elliptic boundary value problem and it is omitted here. Instead, we shall simply assume that (5.1)–(5.3) has a unique solution and investigate its decay in $t$ ($t$ typically signifies time), and discuss the question of continuous dependence of the solution on the initial datum $u_0$ and the forcing function $f$.

We recall that, for $v, w \in L_2(\Omega)$, the inner product $(u, v)$ and the norm $\|v\|_{L_2(\Omega)}$ are defined by

$$\begin{aligned} (v,w) &= \int_\Omega v(x)w(x)\,\mathrm{d}x, \\ \|v\|_{L_2(\Omega)} &= (v,v)^{1/2}. \end{aligned}$$

Taking the inner product of (5.1) with $u$, noting that $u(x,t) = 0$, $x \in \Gamma$, integrating by parts, and applying (5.4) and (5.5), we get

$$\left(\frac{\partial u}{\partial t}(\cdot,t), u(\cdot,t)\right) + \tilde{c}\sum_{i=1}^{n} \|\frac{\partial u}{\partial x_i}(\cdot,t)\|^2_{L_2(\Omega)} \leq (f(\cdot,t), u(\cdot,t)).$$

Noting that

$$\left(\frac{\partial u}{\partial t}(\cdot,t), u(\cdot,t)\right) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2_{L_2(\Omega)},$$

and using the Poincaré–Friedrichs inequality (1.2), we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2_{L_2(\Omega)} + \frac{\tilde{c}}{c_\star}\|u(\cdot,t)\|^2_{L_2(\Omega)} \leq (f(\cdot,t), u(\cdot,t)).$$

Let $K = \tilde{c}/c_\star$; then, by the Cauchy–Schwarz inequality,

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2_{L_2(\Omega)} + K\|u(\cdot,t)\|^2_{L_2(\Omega)} &\le \|f(\cdot,t)\|_{L_2(\Omega)}\|u(\cdot,t)\|_{L_2(\Omega)} \\ &\le \frac{1}{2K}\|f(\cdot,t)\|^2_{L_2(\Omega)} + \frac{K}{2}\|u(\cdot,t)\|^2_{L_2(\Omega)}.\end{aligned}$$

Thence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2_{L_2(\Omega)} + K\|u(\cdot,t)\|^2_{L_2(\Omega)} \le \frac{1}{K}\|f(\cdot,t)\|^2_{L_2(\Omega)}.$$

Multiplying both sides by $\mathrm{e}^{Kt}$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{Kt}\|u(\cdot,t)\|^2_{L_2(\Omega)}\right) \le \frac{\mathrm{e}^{Kt}}{K}\|f(\cdot,t)\|^2_{L_2(\Omega)}.$$

Integrating from 0 to $t$,

$$\mathrm{e}^{Kt}\|u(\cdot,t)\|^2_{L_2(\Omega)} - \|u_0\|^2_{L_2(\Omega)} \le \frac{1}{K}\int_0^t \mathrm{e}^{K\tau}\|f(\cdot,\tau)\|^2_{L_2(\Omega)}\,\mathrm{d}\tau.$$

Hence

$$\|u(\cdot,t)\|^2_{L_2(\Omega)} \le \mathrm{e}^{-Kt}\|u_0\|^2_{L_2(\Omega)} + \frac{1}{K}\int_0^t \mathrm{e}^{-K(t-\tau)}\|f(\cdot,\tau)\|^2_{L_2(\Omega)}\,\mathrm{d}\tau. \tag{5.6}$$

Assuming that (5.1)–(5.3) has a solution, (5.6) implies that the solution is unique. Indeed, if $u_1$ and $u_2$ are solutions to (5.1)–(5.3), then $u = u_1 - u_2$ satisfies (5.1)–(5.3) with $f \equiv 0$ and $u_0 \equiv 0$; therefore, by (5.6), $u \equiv 0$, i.e. $u_1 \equiv u_2$.

Let us also look at the special case when $f \equiv 0$ in (5.1). This corresponds to considering the evolution of the solution from the initial datum $u_0$ in the absence of external forces. In this case (5.6) yields

$$\|u(\cdot,t)\|^2_{L_2(\Omega)} \le \mathrm{e}^{-Kt}\|u_0\|^2_{L_2(\Omega)}, \quad t \ge 0; \tag{5.7}$$

in physical terms, the energy $\frac{1}{2}\|u(\cdot,t)\|^2_{L_2(\Omega)}$ dissipates exponentially. Since $K = \tilde{c}/c_\star$, we have

$$\|u(\cdot,t)\|^2_{L_2(\Omega)} \le \mathrm{e}^{-\tilde{c}t/c_\star}\|u_0\|^2_{L_2(\Omega)}, \quad t \ge 0, \tag{5.8}$$

and we deduce that the rate of **dissipation** depends on the lower bound, $\tilde{c}$, on the "diffusion coefficients" $a_{ij}$ (i.e. the smaller $\tilde{c}$, the slower the decay of the energy). Conservation of energy would correspond to

$$\|u(\cdot,t)\|^2_{L_2(\Omega)} = \|u_0\|^2_{L_2(\Omega)};$$

this will only occur by formally setting $\tilde{c} = 0$, however since $\tilde{c} > 0$ by hypothesis, **conservation** of energy will not be observed for a physical process modelled by a second-order parabolic equation.

In the next section we consider some simple finite element methods for the numerical solution of parabolic initial boundary value problems. In order to simplify the presentation, we restrict ourselves to the heat equation in one space dimension, but the analysis that we shall present also applies in the general setting.

## 5.2 Forward and backward Euler schemes

We consider the following simple model problem for the heat equation in one space dimension. Let $Q = \Omega \times (0,T]$, where $\Omega = (0,1)$, $T > 0$;

$$\begin{aligned} &\text{find } u(x,t) \text{ such that} \\ &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t), \quad x \in (0,1), \;\; t \in (0,T], \\ &u(0,t) = 0, \;\; u(1,t) = 0, \;\; t \in [0,T], \\ &u(x,0) = u_0(x), \quad x \in [0,1]. \end{aligned} \tag{5.9}$$

We describe two schemes for the numerical solution of (5.9). They both use the same discretisation in the $x$ variable but while the first scheme (called the forward Euler scheme) employs a forward divided difference in $t$ to approximate $\partial u/\partial t$, the second (called the backward Euler scheme) uses a backward difference in $t$.

**The forward Euler scheme.** We begin by constructing a mesh on $\bar{Q} = [0,1] \times [0,T]$. Let $h = 1/N$ be the mesh-size in the $x$-direction and let $\Delta t = T/M$ be the mesh-size in the $t$-direction; here $N$ and $M$ are two integers, $N \geq 2$, $M \geq 1$. We define the uniform mesh $\bar{Q}_h^{\Delta t}$ on $\bar{Q}$ by

$$\bar{Q}_h^{\Delta t} = \{(x_j, t^m) : x_j = jh, \; 0 \leq j \leq N; \; t^m = m \cdot \Delta t, \; 0 \leq m \leq M\}.$$

Let $V_h \subset H_0^1(0,1)$ denote the set of all continuous piecewise linear functions defined on the $x$-mesh

$$0 = x_0 < x_1 < \ldots < x_{N-1} < x_N = 1$$

which vanish at the end-points, $x = 0$ and $x = 1$.

We approximate (5.9) by the finite element method, referred to as the **forward Euler scheme**:

$$\begin{aligned} &\text{find } u_h^m \in V_h, \quad 0 \leq m \leq M, \text{ such that} \\ &\left(\frac{u_h^{m+1} - u_h^m}{\Delta t}, v_h\right) + a(u_h^m, v_h) = (f(\cdot, t^m), v_h) \;\; \forall v_h \in V_h, \\ &(u_h^0 - u_0, v_h) = 0 \quad \forall v_h \in V_h, \end{aligned} \tag{5.10}$$

where $u_h^m$ represents the approximation of $u(\cdot, t^m)$, and $a(\cdot,\cdot)$ is defined by

$$a(w,v) = \int_0^1 w'(x)v'(x)\,\mathrm{d}x.$$

Clearly, (5.10) can be rewritten as follows:

$$\begin{aligned} (u_h^{m+1}, v_h) = (u_h^m, v_h) - \Delta t\, a(u_h^m, v_h) + \Delta t\, (f(\cdot, t^m), v_h) \\ \forall v_h \in V_h, 0 \leq m \leq M-1, \end{aligned}$$

with

$$(u_h^0, v_h) = (u_0, v_h) \quad \forall v_h \in V_h.$$

Thus, given $u_h^m$, to find $u_h^{m+1}$ at time level $t^{m+1}$ we have to solve a system of linear equations with symmetric positive definite matrix $M$ of size $(N-1)\times(N-1)$, with entries $(\phi_i, \phi_j)$ where $\phi_i$ denotes the one-dimensional piecewise linear finite element basis function associated with the $x$-mesh point $x_i$; the same matrix arises when determining $u_h^0$. It is a simple matter to show that this matrix is tridiagonal and has the form

$$M = \frac{h}{6}\begin{pmatrix} 4 & 1 & 0 & 0 & \dots & 0 \\ 1 & 4 & 1 & 0 & \dots & 0 \\ 0 & 1 & 4 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 1 & 4 \end{pmatrix}.$$

The matrix $M$ is usually referred to as the mass matrix.

**The backward Euler scheme.** Alternatively, one can approximate the time derivative by a backward difference, which gives rise to the following **backward Euler scheme**:

$$\begin{aligned} &\text{find } u_h^m \in V_h, \quad 0 \le m \le M, \text{ such that} \\ &\left(\frac{u_h^{m+1} - u_h^m}{\Delta t}, v_h\right) + a(u_h^{m+1}, v_h) = (f(\cdot, t^{m+1}), v_h) \ \ \forall v_h \in V_h, \\ &(u_h^0 - u_0, v_h) = 0 \quad \forall v_h \in V_h, \end{aligned} \tag{5.11}$$

where $u_h^m$ represents the approximation of $u(\cdot, t^m)$. Equivalently, (5.11) can be written

$$\begin{aligned} (u_h^{m+1}, v_h) + \Delta t\, a(u_h^{m+1}, v_h) = (u_h^m, v_h) + \Delta t\, (f(\cdot, t^{m+1}), v_h) \\ \forall v_h \in V_h, 0 \le m \le M-1, \end{aligned}$$

with

$$(u_h^0, v_h) = (u_0, v_h) \quad \forall v_h \in V_h.$$

Thus, given $u_h^m$, to find $u_h^{m+1}$ at time level $t^{m+1}$ we have to solve a system of linear equations with symmetric positive definite matrix $A$ of size $(N-1)\times(N-1)$, with entries $(\phi_i, \phi_j) + \Delta t\, (\phi_i', \phi_j')$ where $\phi_i$ denotes the one-dimensional piecewise linear finite element basis function associated with the $x$-mesh point $x_i$; finding $u_h^0$ still only involves inverting the mass matrix $M$. It is clear that $A = M + \Delta t\, K$, where

$$K = \frac{1}{h}\begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & -1 & 2 \end{pmatrix}$$

is the so-called **stiffness matrix**.

## 5.3 Stability of $\theta$-schemes

We shall study the stability of the schemes (5.10) and (5.11) simultaneously, by embedding them into a one-parameter family of finite element schemes:

$$\begin{aligned}
&\text{find } u_h^m \in V_h, \quad 0 \le m \le M, \text{ such that} \\
&\left(\frac{u_h^{m+1} - u_h^m}{\Delta t}, v_h\right) + a(u_h^{m+\theta}, v_h) = (f^{m+\theta}, v_h) \ \forall v_h \in V_h \\
&(u_h^0 - u_0, v_h) = 0 \quad \forall v_h \in V_h,
\end{aligned} \tag{5.12}$$

where $0 \le \theta \le 1$, and for the sake of notational simplicity, we wrote

$$f^{m+\theta}(x) = \theta f(x, t^{m+1}) + (1-\theta) f(x, t^m)$$

and

$$u_h^{m+\theta}(x) = \theta u_h^{m+1}(x) + (1-\theta) u_h^m(x).$$

For $\theta = 0$ this gives the forward Euler scheme, for $\theta = 1$ the backward Euler scheme. The method corresponding to $\theta = \frac{1}{2}$ is known as the **Crank-Nicolson scheme**. Recall that

$$\begin{aligned}
(w, v) &= \int_0^1 w(x) v(x) \, \mathrm{d}x, \\
\|v\|_{L_2(\Omega)} &= (v, v)^{1/2}.
\end{aligned}$$

Taking the inner product of (5.12) with $u_h^{m+\theta}$ we get

$$\left(\frac{u_h^{m+1} - u_h^m}{\Delta t}, u_h^{m+\theta}\right) + a(u_h^{m+\theta}, u_h^{m+\theta}) = (f^{m+\theta}, u_h^{m+\theta}).$$

Equivalently,

$$\left(\frac{u_h^{m+1} - u_h^m}{\Delta t}, u_h^{m+\theta}\right) + |u_h^{m+\theta}|^2_{H^1(\Omega)} = (f^{m+\theta}, u_h^{m+\theta}).$$

Since

$$u_h^{m+\theta} = \Delta t \left(\theta - \frac{1}{2}\right) \frac{u_h^{m+1} - u_h^m}{\Delta t} + \frac{u_h^{m+1} + u_h^m}{2},$$

it follows that

$$\begin{aligned}
\Delta t \left(\theta - \frac{1}{2}\right) \left\| \frac{u_h^{m+1} - u_h^m}{\Delta t} \right\|^2_{L_2(\Omega)} + \frac{\|u_h^{m+1}\|^2_{L_2(\Omega)} - \|u_h^m\|^2_{L_2(\Omega)}}{2\Delta t} \\
+ |u_h^{m+\theta}|^2_{H^1(\Omega)} = (f^{m+\theta}, u_h^{m+\theta}).
\end{aligned} \tag{5.13}$$

Suppose that $\theta \in [1/2, 1]$; then $\theta - 1/2 \geq 0$, and therefore

$$\begin{aligned}\frac{\|u_h^{m+1}\|^2_{L_2(\Omega)} - \|u_h^m\|^2_{L_2(\Omega)}}{2\Delta t} + |u_h^{m+\theta}|^2_{H^1(\Omega)} &\leq (f^{m+\theta}, u_h^{m+\theta}) \\ &\leq \|f^{m+\theta}\|_{L_2(\Omega)}\|u_h^{m+\theta}\|_{L_2(\Omega)}.\end{aligned}$$

According to the Poincaré–Friedrichs inequality,

$$\|u_h^{m+\theta}\|^2_{L_2(\Omega)} \leq \frac{1}{2}|u_h^{m+\theta}|^2_{H^1(\Omega)}.$$

Thus

$$\frac{\|u_h^{m+1}\|^2_{L_2(\Omega)} - \|u_h^m\|^2_{L_2(\Omega)}}{2\Delta t} + 2\|u_h^{m+\theta}\|^2_{L_2(\Omega)} \leq \frac{1}{2}\|f^{m+\theta}\|^2_{L_2(\Omega)} + \frac{1}{2}\|u_h^{m+\theta}\|^2_{L_2(\Omega)},$$

so that

$$\|u_h^{m+1}\|^2_{L_2(\Omega)} \leq \|u_h^m\|^2_{L_2(\Omega)} + \Delta t\|f^{m+\theta}\|^2_{L_2(\Omega)}.$$

Summing through $m$, $m = 0, \ldots, k$, we get that

$$\|u_h^k\|^2_{L_2(\Omega)} \leq \|u_h^0\|^2_{L_2(\Omega)} + \sum_{m=0}^{k-1} \Delta t\|f^{m+\theta}\|^2_{L_2(\Omega)}, \tag{5.14}$$

for all $k$, $1 \leq k \leq M$.

The inequality (5.14) can be thought of as the discrete version of (5.6). If follows from (5.14) that

$$\max_{1\leq k\leq M} \|u_h^k\|^2_{L_2(\Omega)} \leq \|u_h^0\|^2_{L_2(\Omega)} + \sum_{m=0}^{M-1} \Delta t\|f^{m+\theta}\|^2_{L_2(\Omega)},$$

i.e.

$$\max_{1\leq k\leq M} \|u_h^k\|_{L_2(\Omega)} \leq \left[\|u_h^0\|^2_{L_2(\Omega)} + \sum_{m=0}^{M-1} \Delta t\|f^{m+\theta}\|^2_{L_2(\Omega)}\right]^{1/2}, \tag{5.15}$$

which expresses the continuous dependence of the solution to the finite element scheme (5.12) on the initial data and the right-hand side. This property is called stability.

Thus we have proved that for $\theta \in [1/2, 1]$, the scheme (5.12) is stable, without any limitations on the time step in terms of $h$. In other words, the scheme (5.12) is **unconditionally stable** for $\theta \in [1/2, 1]$.

Now let us consider the case $\theta \in [0, 1/2)$. According to (5.13),

$$\begin{aligned}&\frac{\|u_h^{m+1}\|^2_{L_2(\Omega)} - \|u_h^m\|^2_{L_2(\Omega)}}{2\Delta t} + |u_h^{m+\theta}|^2_{H^1(\Omega)} \\ &\quad = \Delta t(\frac{1}{2} - \theta)\|\frac{u_h^{m+1} - u_h^m}{\Delta t}\|^2_{L_2(\Omega)} + (f^{m+\theta}, u^{m+\theta}).\end{aligned} \tag{5.16}$$

Recalling (5.12) with $v_h = (u_h^{m+1} - u_h^m)/\Delta t$, we have that

$$\|\frac{u_h^{m+1} - u_h^m}{\Delta t}\|^2_{L_2(\Omega)} = \left(f^{m+\theta}, \frac{u_h^{m+1} - u_h^m}{\Delta t}\right) - a\left(u_h^{m+\theta}, \frac{u_h^{m+1} - u_h^m}{\Delta t}\right).$$

Therefore,

$$\begin{aligned}\|\frac{u_h^{m+1} - u_h^m}{\Delta t}\|^2_{L_2(\Omega)} &\leq \|f^{m+\theta}\|_{L_2(\Omega)}\|\frac{u_h^{m+1} - u_h^m}{\Delta t}\|_{L_2(\Omega)} \\ &\quad + |u_h^{m+\theta}|_{H^1(\Omega)}\,|\frac{u_h^{m+1} - u_h^m}{\Delta t}|_{H^1(\Omega)}. \end{aligned} \tag{5.17}$$

Next we shall prove that, for each $w_h \in V_h$,

$$|w_h|_{H^1(\Omega)} \leq \frac{\sqrt{12}}{h}\|w_h\|_{L_2(\Omega)}. \tag{5.18}$$

We shall then use this inequality to estimate the terms appearing on the right-hand side of (5.17). Let $W_i$ denote the value of the piecewise linear function $w_h \in V_h$ at the mesh-point $x_i$, $i = 0, \ldots, N$, and note that $W_0 = W_N = 0$. A simple calculation reveals that

$$|w_h|^2_{H^1(\Omega)} = \sum_{i=1}^{N} h|\frac{W_i - W_{i-1}}{h}|^2 \leq \frac{4}{h^2}\sum_{i=1}^{N-1} h|W_i|^2. \tag{5.19}$$

On the other hand,

$$\begin{aligned}\|w_h\|^2_{L_2(\Omega)} &= \frac{h}{6}\sum_{i=1}^{N-1}(W_iW_{i-1} + 4W_i^2 + W_iW_{i+1}) \\ &\geq \frac{h}{6}\sum_{i=1}^{N-1}\left[-\frac{1}{2}W_i^2 - \frac{1}{2}W_{i-1}^2 + 4W_i^2 - \frac{1}{2}W_i^2 - \frac{1}{2}W_{i+1}^2\right] \\ &\geq \frac{1}{3}\sum_{i=1}^{N-1} h|W_i|^2. \end{aligned} \tag{5.20}$$

From (5.20) and (5.19) we deduce (5.18).

Now, equipped with the inequality (5.18), we continue the stability analysis. Applying (5.18) with $w_h = (u_h^{m+1} - u_h^m)/\Delta t$, we deduce that

$$\begin{aligned}\|\frac{u_h^{m+1} - u_h^m}{\Delta t}\|^2_{L_2(\Omega)} &\leq \|f^{m+\theta}\|_{L_2(\Omega)}\|\frac{u_h^{m+1} - u_h^m}{\Delta t}\|_{L_2(\Omega)} \\ &\quad + \frac{\sqrt{12}}{h}|u_h^{m+\theta}|_{H^1(\Omega)}\,\|\frac{u_h^{m+1} - u_h^m}{\Delta t}\|_{L_2(\Omega)}\end{aligned}$$

and hence

$$\|\frac{u_h^{m+1} - u_h^m}{\Delta t}\|_{L_2(\Omega)} \leq \|f^{m+\theta}\|_{L_2(\Omega)} + \frac{\sqrt{12}}{h}|u_h^{m+\theta}|_{H^1(\Omega)} \tag{5.21}$$

By (5.21), for any $\epsilon \in (0,1)$,

$$\begin{aligned}\|\frac{u_h^{m+1}-u_h^m}{\Delta t}\|^2_{L_2(\Omega)} &\leq \left(\frac{\sqrt{12}}{h}|u_h^{m+\theta}|_{H^1(\Omega)} + \|f^{m+\theta}\|_{L_2(\Omega)}\right)^2 \\ &\leq (1+\epsilon)\frac{12}{h^2}|u_h^{m+\theta}|^2_{H^1(\Omega)} + (1+\epsilon^{-1})\|f^{m+\theta}\|^2_{L_2(\Omega)},\end{aligned}$$

where the inequality $(a+b)^2 \leq (1+\epsilon)a^2 + (1+\frac{1}{\epsilon})b^2$, $a,b \geq 0$, $\epsilon > 0$, has been applied. Substituting into (5.16),

$$\begin{aligned}&\frac{\|u_h^{m+1}\|^2_{L_2(\Omega)} - \|u_h^m\|^2_{L_2(\Omega)}}{2\Delta t} + \left(1 - \Delta t(\frac{1}{2}-\theta)\cdot\frac{12(1+\epsilon)}{h^2}\right)|u_h^{m+\theta}|^2_{H^1(\Omega)} \\ &\leq \|f^{m+\theta}\|_{L_2(\Omega)}\|u_h^{m+\theta}\|_{L_2(\Omega)} + \Delta t(\frac{1}{2}-\theta)(1+\epsilon^{-1})\|f^{m+\theta}\|^2_{L_2(\Omega)}.\end{aligned} \tag{5.22}$$

According to the Poincaré–Friedrichs inequality,

$$\|u_h^{m+\theta}\|^2_{L_2(\Omega)} \leq \frac{1}{2}|u_h^{m+\theta}|^2_{H^1(\Omega)},$$

and therefore,

$$\begin{aligned}\|f^{m+\theta}\|_{L_2(\Omega)}\|u_h^{m+\theta}\|_{L_2(\Omega)} &\leq \frac{1}{8\epsilon^2}\|f^{m+\theta}\|^2_{L_2(\Omega)} + 2\epsilon^2\|u_h^{m+\theta}\|^2_{L_2(\Omega)} \\ &\leq \frac{1}{8\epsilon^2}\|f^{m+\theta}\|^2_{L_2(\Omega)} + \epsilon^2|u_h^{m+\theta}|^2_{H^1(\Omega)}.\end{aligned} \tag{5.23}$$

Substituting (5.23) into (5.22),

$$\begin{aligned}&\frac{\|u_h^{m+1}\|^2_{L_2(\Omega)} - \|u_h^m\|^2_{L_2(\Omega)}}{2\Delta t} + \left(1 - \Delta t\frac{6(1-2\theta)(1+\epsilon)}{h^2} - \epsilon^2\right)|u_h^{m+\theta}|^2_{H^1(\Omega)} \\ &\leq \frac{1}{8\epsilon^2}\|f^{m+\theta}\|^2_{L_2(\Omega)} + \Delta t(\frac{1}{2}-\theta)(1+\epsilon^{-1})\|f^{m+\theta}\|^2_{L_2(\Omega)}.\end{aligned}$$

Let us suppose that

$$\Delta t \leq \frac{h^2}{6(1-2\theta)}(1-\epsilon), \quad \theta \in [0, 1/2),$$

where $\epsilon$ is a fixed real number, $\epsilon \in (0,1)$. Then

$$1 - \Delta t\,\frac{6(1-2\theta)(1+\epsilon)}{h^2} - \epsilon^2 \geq 0,$$

so that

$$\|u_h^{m+1}\|^2_{L_2(\Omega)} \leq \|u_h^m\|^2_{L_2(\Omega)} + \frac{\Delta t}{4\epsilon^2}\|f^{m+\theta}\|^2_{L_2(\Omega)} + \Delta t^2(1-2\theta)(1+\epsilon^{-1})\|f^{m+\theta}\|^2_{L_2(\Omega)}.$$

Letting $c_\epsilon = 1/(4\epsilon^2) + \Delta t(1-2\theta)(1+\epsilon^{-1})$, upon summation through all $m$ this implies that

$$\max_{1\le k\le M} \|u_h^k\|^2_{L_2(\Omega)} \le \|u_h^0\|^2_{L_2(\Omega)} + c_\epsilon \sum_{m=0}^{M-1} \Delta t \|f^{m+\theta}\|^2_{L_2(\Omega)}.$$

Taking the square root of both sides, we deduce that for $\theta \in [0, 1/2)$ the scheme (5.12) is conditionally stable in the sense that

$$\max_{1\le k\le M} \|u_h^k\|_h \le \left[ \|u_h^0\|^2_{L_2(\Omega)} + c_\epsilon \sum_{m=0}^{M-1} \Delta t \|f^{m+\theta}\|^2_{L_2(\Omega)} \right]^{1/2}, \tag{5.24}$$

provided that

$$\Delta t \le \frac{h^2}{6(1-2\theta)}(1-\epsilon), \quad 0 < \epsilon < 1. \tag{5.25}$$

To summarise: when $\theta \in [1/2, 1]$, the method (5.12) is unconditionally stable. In particular, the backward Euler scheme, corresponding to $\theta = 1$, and the Crank-Nicolson scheme, corresponding to $\theta = 1/2$, are unconditionally stable, and (5.15) holds. When $\theta \in [0, 1/2)$, the scheme (5.12) is conditionally stable, subject to the time step limitation (5.25). The forward Euler scheme, corresponding to $\theta = 0$, is only conditionally stable.

## 5.4 Error analysis in the $L_2$ norm

In this section we investigate the accuracy of the finite element method (5.12) for the numerical solution of the initial boundary value problem (5.9). For simplicity, we shall restrict ourselves to the backward Euler scheme ($\theta = 1$); for other values of $\theta \in [0, 1]$ the analysis is completely analogous.

We decompose the global error $e_h$ as follows:

$$e_h^m = u(\cdot, t^m) - u_h^m = \eta^m + \xi^m,$$

where

$$\eta^m = u(\cdot, t^m) - Pu(\cdot, t^m), \quad \xi^m = Pu(\cdot, t^m) - u_h^m,$$

and for $t \in [0, T]$, $Pu(\cdot, t) \in V_h$ denotes the **Dirichlet projection** of $u(\cdot, t)$ defined by

$$a(Pu(\cdot, t), v_h) = a(u(\cdot, t), v_h) \quad \forall v_h \in V_h.$$

The existence and uniqueness of $Pu(\cdot, t) \in V_h$ follows by the Lax-Milgram theorem. Hence,

$$a(\eta^m, v_h) = 0 \quad \forall v_h \in V_h,$$

and therefore, by Céa's lemma,

$$|\eta^m|_{H^1(\Omega)} \leq |u(\cdot,t^m) - \mathcal{I}_h u(\cdot,t^m)|_{H^1(\Omega)} \leq \frac{h}{\pi}|u(\cdot,t^m)|_{H^2(\Omega)},$$

where $\mathcal{I}_h u(\cdot,t^m) \in V_h$ denotes the continuous piecewise linear interpolant of $u(\cdot,t^m)$ from $V_h$. By the Aubin–Nitsche duality argument,

$$\|\eta^m\|_{L_2(\Omega)} \leq \frac{h^2}{\pi^2}|u(\cdot,t^m)|_{H^2(\Omega)}. \tag{5.26}$$

Since also,

$$a\left(\frac{\eta^{m+1}-\eta^m}{\Delta t}, v_h\right) = 0 \quad \forall v_h \in V_h,$$

by an identical argument we deduce that

$$\left\|\frac{\eta^{m+1}-\eta^m}{\Delta t}\right\|_{L_2(\Omega)} \leq \frac{h^2}{\pi^2}\left|\frac{u(\cdot,t^{m+1})-u(\cdot,t^m)}{\Delta t}\right|_{H^2(\Omega)}. \tag{5.27}$$

For $m = 0$,

$$(\xi^0, v_h) = (e_h^0, v_h) - (\eta^0, v_h) = -(\eta^0, v_h)$$

and therefore, choosing $v_h = \xi^0$ and applying the Cauchy–Schwarz inequality on the right,

$$\|\xi^0\|_{L_2(\Omega)} \leq \|\eta^0\|_{L_2(\Omega)} \leq \frac{h^2}{\pi^2}|u_0|_{H^2(\Omega)}. \tag{5.28}$$

It is easily seen that $\xi^m \in V_h$ satisfies the following identity:

$$\begin{aligned}\left(\frac{\xi^{m+1}-\xi^m}{\Delta t}, v_h\right) &+ a(\xi^{m+1}, v_h) \\ &= \left(\frac{u(\cdot,t^{m+1})-u(\cdot,t^m)}{\Delta t} - \frac{\partial u}{\partial t}(\cdot,t^{m+1}) - \frac{\eta^{m+1}-\eta^m}{\Delta t}, v_h\right)\end{aligned}$$

According to the stability result proved earlier on,

$$\max_{1\leq m\leq M}\|\xi^m\|_{L_2(\Omega)} \leq \left[\|\xi^0\|^2_{L_2(\Omega)} + \sum_{m=0}^{M-1}\Delta t\|\varphi^{m+1}\|^2_{L_2(\Omega)}\right]^{1/2}, \tag{5.29}$$

where

$$\varphi^{m+1} = \frac{u(\cdot,t^{m+1})-u(\cdot,t^m)}{\Delta t} - \frac{\partial u}{\partial t}(\cdot,t^{m+1}) - \frac{\eta^{m+1}-\eta^m}{\Delta t}.$$

By (5.28),

$$\|\xi^0\|_{L_2(\Omega)} \leq \frac{h^2}{\pi^2}|u_0|_{H^2(\Omega)}. \tag{5.30}$$

It remains to estimate $\|\varphi^{m+1}\|_{L_2(\Omega)}$. Now

$$\begin{aligned}\|\varphi^{m+1}\|_{L_2(\Omega)} &\leq \|\frac{u(\cdot,t^{m+1})-u(\cdot,t^m)}{\Delta t}-\frac{\partial u}{\partial t}(\cdot,t^{m+1})\|_{L_2(\Omega)} \\ &\quad +\|\frac{\eta^{m+1}-\eta^m}{\Delta t}\|_{L_2(\Omega)} \equiv I+II. \end{aligned} \tag{5.31}$$

For term $I$, Taylor's formula with integral remainder yields that

$$\frac{u(x,t^{m+1})-u(x,t^m)}{\Delta t}-\frac{\partial u}{\partial t}(x,t^{m+1})=-\frac{1}{\Delta t}\int_{t^m}^{t^{m+1}}(t-t^m)\frac{\partial^2 u}{\partial t^2}(x,t)\,\mathrm{d}t,$$

and therefore

$$I\leq\sqrt{\Delta t}\left(\int_{t^m}^{t^{m+1}}\|\frac{\partial^2 u}{\partial t^2}(\cdot,t)\|^2_{L_2(\Omega)}\,\mathrm{d}t\right)^{1/2}.$$

Further, by (5.27),

$$\begin{aligned} II &\leq \frac{h^2}{\pi^2}\left|\frac{u(\cdot,t^{m+1})-u(\cdot,t^m)}{\Delta t}\right|_{H^2(\Omega)}=\frac{h^2}{\pi^2}\left|\frac{1}{\Delta t}\int_{t^m}^{t^{m+1}}\frac{\partial u}{\partial t}(\cdot,t)\,\mathrm{d}t\right|_{H^2(\Omega)} \\ &\leq\frac{h^2}{\pi^2\sqrt{\Delta t}}\left(\int_{t^m}^{t^{m+1}}\left|\frac{\partial u}{\partial t}(\cdot,t)\right|^2_{H^2(\Omega)}\,\mathrm{d}t\right)^{1/2}. \end{aligned}$$

Substituting the bounds on terms $I$ and $II$ onto (5.31) and inserting the resulting inequality and (5.30) into (5.29), we obtain the following error bound:

$$\max_{1\leq m\leq M}\|\xi^m\|_{L_2(\Omega)}\leq C_1(h^2+\Delta t),\quad \theta\in(1/2,1], \tag{5.32}$$

where $C_1$ is a positive constant, independent of $h$ and $\Delta t$, and depending only on norms of the analytical solution $u$. But,

$$\max_{1\leq m\leq M}\|u(\cdot,t^m)-u_h^m\|_{L_2(\Omega)}\leq\max_{1\leq m\leq M}\|\xi^m\|_{L_2(\Omega)}+\max_{1\leq m\leq M}\|\eta^m\|_{L_2(\Omega)}.$$

Thus, by (5.32) and (5.26), we deduce that

$$\max_{1\leq m\leq M}\|u(\cdot,t^m)-u_h^m\|_{L_2(\Omega)}\leq C_2(h^2+\Delta t),$$

where $C_2$ is a positive constant independent of $h$ and $\Delta t$.

The Crank-Nicolson scheme ($\theta=1/2$) can be shown to converge in the norm $\|\cdot\|_{L_2(\Omega)}$ unconditionally, with error $\mathcal{O}(h^2+(\Delta t)^2)$. For $\theta\in(1/2,1]$ the scheme converges unconditionally with error $\mathcal{O}(h^2+\Delta t)$. For $\theta\in[0,1/2)$ the scheme converges with error $\mathcal{O}(h^2+\Delta t)$, but only conditionally. The stability and convergence results presented here can be extended to parabolic equations in more than one space dimension, but the exposition of that theory, while very similar to the one-dimensional case, is beyond the scope of these notes.