An Introduction to the Numerical Analysis of Partial Differential Equations

Endre Süli
University of Oxford

February, 2005

Contents

1 Elements of function spaces **3**

1.1 Spaces of continuous functions 4

1.2 Spaces of integrable functions 5

1.3 Sobolev spaces 7

2 Elliptic boundary value problems: existence and uniqueness of weak solutions **11**

3 Introduction to the theory of finite difference schemes **18**

4 Finite difference approximation of elliptic boundary value problems **27**

5 Finite element methods for elliptic boundary value problems **48**

5.1 Construction of the finite element method: piecewise linear basis functions . 49

5.1.1 One-dimensional problem 49

5.1.2 Two-dimensional problem 51

5.2 Variational formulation of self-adjoint elliptic boundary value problems . . . 55

5.3 Construction of the finite element method: abstract setting 59

5.4 Céa's lemma 60

5.5 Optimal error bounds in the energy norm 62

6 Finite difference approximation of evolutionary problems **69**

6.1 Finite difference methods for parabolic equations 69

6.1.1 Explicit and implicit schemes 71

6.1.2 Stability of explicit and implicit schemes 74

6.1.3 Error analysis of difference schemes for the heat equation 78

6.2 Finite difference methods for hyperbolic equations 81

6.2.1 Explicit finite difference scheme 84

Reading List

R.A. Adams, 1978: *Sobolev Spaces*, Academic Press.

S.C. Brenner and L.R. Scott, 2002: *The Mathematical Theory of Finite Element Methods*, Second edition, Springer.

Ph. Ciarlet, 1978: *The Finite Element Method for Elliptic Problems*, North-Holland.

C. Johnson, 1988: *Numerical Solution of Partial Differential Equations by the Finite Element Method*, CUP.

R.D. Richtmyer and K.W. Morton, 1967: *Difference Methods for Initial-Value Problems*, Wiley Interscience.

J.C. Strikwerda, 1989: *Finite Difference Schemes and Partial Differential Equations*, Wadsworth and Brooks, Mathematical Series.

Introduction

Partial differential equations arise in the mathematical modelling of many physical, chemical and biological phenomena (e.g. dispersion of pollutants in lakes and rivers, spreading of diseases, weather prediction, etc.). Very frequently the equations are so complicated that their solution by analytical means (e.g. by Laplace and Fourier transforms or in a form of a series) is either impossible or impracticable, and one has to resort to numerical techniques instead.

These notes are devoted to the analysis of numerical methods for elliptic, parabolic and hyperbolic partial differential equations, by considering simple model problems. We concentrate on techniques that are most widespread in practice: finite difference and finite element methods, although the analysis of finite volume schemes is also touched on. Preference is given to theoretical results concerning the stability and the accuracy of numerical methods – properties that are of key importance in practical computations.

The material covered in the notes had formed the basis of a 16-lecture introductory course on the analysis of numerical algorithms for partial differential equations at the University of Oxford given over the period 1992–1996. The background material from linear functional analysis and the theory of function spaces discussed herein is intentionally sketchy in order to enable the understanding of some of the key concepts, such as stability and convergence of finite difference and finite element methods, with the bare minimum of analytical prerequisites. Due to the time-constraints imposed by the length of the original lecture course, a significant portion of the theory of numerical algorithms for partial differential equations is not being touched upon; nevertheless, I hope that the notes will serve a helpful purpose as a brief compendium of basic theoretical information about this exciting and practically relevant field of research. For further details, the reader is referred to the numerous excellent books on the subject, some of which appear on the Reading List.

1 Elements of function spaces

The accuracy of numerical methods for the approximate solution of partial differential equations depends on their capabilities to represent the important qualitative features of the (analytical) solution. One such feature that has to be taken into account in the construction and the analysis of numerical methods is the smoothness of the solution, and this depends on the smoothness of the data.

Precise assumptions about the smoothness of the data and of the corresponding solution can be conveniently formulated by considering classes of functions with particular differentiability and integrability properties, called function spaces. In this section we present a brief overview of definitions and basic results form the theory of function spaces which will be used throughout these notes, focusing, in particular, on spaces of continuous functions, spaces of

integrable functions, and Sobolev spaces.

1.1 Spaces of continuous functions

In this section, we describe some simple function spaces that consist of continuous and 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 called 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

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

EXAMPLE. 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,

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

Let Ω 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 Ω such that $D^\alpha u$ is continuous on Ω for all $\alpha = (\alpha_1, \ldots, \alpha_n)$ with $|\alpha| \leq k$. Assuming that Ω 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 Ω to a continuous function on $\bar{\Omega}$, the closure of the set Ω, for all $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\quad |\alpha| \leq k$. $C^k(\bar{\Omega})$ 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 $C^0(\bar{\Omega})$;

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

Similarly, if $k = 1$,

$$\|u\|_{C^1(\bar{\Omega})} = \sum_{|\alpha| \leq 1} \sup_{x \in \Omega} |D^\alpha u(x)|$$

$$= \sup_{x\in\Omega} |u(x)| + \sum_{j=1}^{n} \sup_{x\in\Omega} \left| \frac{\partial u}{\partial x_j}(x) \right| .$$

EXAMPLE. Let $n = 1$, and consider the open interval $\Omega = (0,1) \subset \mathbb{R}^1$. The function $u(x) = 1/x$ belongs to $C^k(\Omega)$ for each $k \geq 0$. Since $\bar{\Omega} = [0,1]$, 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*, supp u, of a continuous function u on Ω is defined as the closure in Ω of the set $\{x \in \Omega : u(x) \neq 0\}$; in other words, supp u is the smallest closed subset of Ω such that $u = 0$ in $\Omega \backslash \text{supp } u$.

EXAMPLE. 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 supp 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 $u \in C^k(\Omega)$ such that supp $u \subset \Omega$ and supp u is bounded. Let

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

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

1.2 Spaces of integrable functions

Next we define 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 Ω such that

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

Functions which are equal almost everywhere (i.e. equal, except on a set of measure zero) on Ω are identified with each other. $L^p(\Omega)$ is equipped with the norm

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

A particularly important case is $p = 2$; 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 1.1 *(The Cauchy–Schwarz inequality). Let $u, v \in L^2(\Omega)$; then,*

$$|(u, v)| \leq \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)}.$$

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

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

The right-hand side is a quadratic polynomial in λ with real coefficients which 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 *(The triangle inequality) Let u, v belong to $L^2(\Omega)$; then, $u + v \in L^2(\Omega)$, and*

$$\|u + v\|_{L^2(\Omega)} \leq \|u\|_{L^2(\Omega)} + \|v\|_{L^2(\Omega)}.$$

Remark The space $L^2(\Omega)$ equipped with the inner product $(\cdot, \cdot)$ (and the associated norm $\|u\|_{L^2(\Omega)} = (u, u)^{1/2}$) is an example of a Hilbert space. In general, a vector space X, equipped with an inner product $(\cdot, \cdot)_X$ (and the associated norm $\|u\|_X = (u, u)_X^{1/2}$) is called a Hilbert space if, whenever $\{u_m\}_{m=1}^\infty$ is a sequence of elements of X such that

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

then, 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).

1.3 Sobolev spaces

In this section we introduce a class of function spaces that play an important role in modern differential equation theory. These spaces, called Sobolev spaces (after the Russian mathematician S.L. Sobolev), consist of functions $u \in L^2(\Omega)$ whose weak derivatives $D^\alpha u$ are also elements of $L^2(\Omega)$. To give a precise definition of a Sobolev space, we shall first explain the meaning of weak derivative.

Suppose u is a smooth function, say $u \in C^k(\Omega)$, and let $v \in C_0^\infty(\Omega)$; then, we have the following integration-by-parts formula:

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

However, in the theory of partial differential equations one often has to consider functions u that do not possess the smoothness hypothesised above, yet they have to be differentiated (in some sense). It is for this purpose that we introduce the idea of a weak derivative.

Suppose that u is locally integrable on Ω (i.e. $u \in L^1(\omega)$ for each bounded open set ω, with $\bar{\omega} \subset \Omega$.) Suppose also that there exists a function w_α, locally integrable on Ω, and such that

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

We then say that w_α is the *weak derivative* of u (of order $|\alpha| = \alpha_1 + \ldots + \alpha_n$) and write $w_\alpha = D^\alpha u$. Clearly, if u is a smooth function then its weak derivatives coincide with those in the classical (pointwise) sense. To simplify the notation, we shall use the letter D to denote both a classical and a weak derivative.

EXAMPLE 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 Ω. Clearly u is not differentiable at the points 0 and ± 1. However, because u is locally integrable on Ω, it may have a weak derivative. Indeed, for any $v \in C_0^\infty(\Omega)$,

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

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. We define (with D^α denoting a weak derivative of order $|\alpha|$)

$$H^k(\Omega) = \{u \in L^2(\Omega) : D^\alpha u \in L^2(\Omega), |\alpha| \le k\}.$$

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

$$\|u\|_{H^k(\Omega)} := \left(\sum_{|\alpha| \le k} \|D^\alpha u\|^2_{L^2(\Omega)} \right)^{1/2}$$

and the inner product

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

With this inner product, $H^k(\Omega)$ is a Hilbert space (for the definition of Hilbert space, see the remark in Section 1.2). Letting

$$|u|_{H^k(\Omega)} := \left(\sum_{|\alpha| = k} \|D^\alpha u\|^2_{L^2(\Omega)} \right)^{1/2},$$

we can write

$$\|u\|_{H^k(\Omega)} = \left(\sum_{j=0}^{k} |u|^2_{H^j(\Omega)} \right)^{1/2}.$$

$|\cdot|_{H^k(\Omega)}$ is called the Sobolev semi-norm (it is only a semi-norm rather than a norm because if $|u|_{H^k(\Omega)} = 0$ for $u \in H^k(\Omega)$ it does not necessarily follow that $u \equiv 0$ on Ω.)

Throughout these notes we shall frequently use $H^1(\Omega)$ and $H^2(\Omega)$.

$$H^1(\Omega) = \left\{ u \in L_2(\Omega) : \frac{\partial u}{\partial x_j} \in L_2(\Omega), \quad j = 1, \ldots, n \right\},$$

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

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

Similarly,

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

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

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

Finally, we define a special Sobolev space,

$$H^1_0(\Omega) = \{ u \in H^1(\Omega) : u = 0 \text{ on } \partial\Omega \},$$

i.e. $H^1_0(\Omega)$ is the set of all functions u in $H^1(\Omega)$ such that $u = 0$ on $\partial\Omega$, the boundary of the set Ω. 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 $H^1_0(\Omega)$ is also a Hilbert space, with the same norm and inner product as $H^1(\Omega)$.

We conclude the section with the following important result.

Lemma 1.2 *(Poincaré–Friedrichs inequality). Suppose that Ω is a bounded open set in $\mathbb{R}^n$ (with a sufficiently smooth boundary $\partial\Omega$) and let $u \in H^1_0(\Omega)$; then, there exists a constant $c_\star(\Omega)$, independent of u, such that*

$$\int_\Omega u^2(x)\,\mathrm{d}x \le c_\star \sum_{i=1}^{n} \int_\Omega \left| \frac{\partial u}{\partial x_i}(x) \right|^2 \mathrm{d}x. \tag{1.1}$$

Proof We shall prove this inequality for the special case of a rectangular domain $\Omega = (a, b) \times (c, d)$. in $\mathbb{R}^2$. The proof for general Ω is analogous.

Evidently

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

Thence, by the Cauchy–Schwarz inequality,

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

Analogously,

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

By adding the two inequalities, we obtain

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

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

2 Elliptic boundary value problems: existence and uniqueness of weak solutions

In the first part of this lecture course we 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

$$-\Delta u = f.$$

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

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left(a_{ij}(x)\frac{\partial u}{\partial x_i}\right) + \sum_{i=1}^{n} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f(x), \quad x \in \Omega, \tag{2.1}$$

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

$$\begin{aligned} a_{ij} &\in C^1(\bar{\Omega}), \quad i,j = 1,\dots,n; \\ b_i &\in C(\bar{\Omega}), \quad i = 1,\dots,n; \\ c &\in C(\bar{\Omega}), \quad f \in C(\bar{\Omega}), \quad \text{and} \\ \sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j &\geq \tilde{c}\sum_{i=1}^{n}\xi_i^2, \quad \forall \xi = (\xi_1,\dots,\xi_n) \in \mathbb{R}^n, \quad x \in \bar{\Omega}; \end{aligned} \tag{2.2}$$

here $\tilde{c}$ is a positive constant independent of x and ξ. The condition (2.2) is usually referred to as uniform ellipticity and (2.1) is called an elliptic equation.

Equation (2.1) is supplemented with one of the following boundary conditions:

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

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

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

(d) A more general version 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 α_j is the angle between the unit outward normal vector n to $\partial\Omega$ and the Ox_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 is imposed on $\partial\Omega_1$ and a Neumann boundary condition on $\partial\Omega_2$). The study of such mixed boundary value problems is beyond the scope of these notes.

We begin by considering the homogeneous Dirichlet boundary value problem

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

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

where a_{ij}, b_i, c and f are as in (2.2).

A function $u\in C^2(\Omega)\cap C(\bar{\Omega})$ satisfying (2.3) and (2.4) is called a classical solution of this problem. The theory of partial differential equations tells us that (2.3), (2.4) has a unique classical solution, provided a_{ij}, b_i, c, f and $\partial\Omega$ are sufficiently smooth. However, in many applications one has to consider boundary value problems 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 the cube $\Omega=(-1,1)^n$ in $\mathbb{R}^n$:

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

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

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 (2.3), (2.4). 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 $H_0^1(\Omega)$ instead, where, as in Section 1.3,

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

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

$$\sum_{i,j=1}^{n} \int_\Omega a_{ij}(x) \frac{\partial u}{\partial x_i} \cdot \frac{\partial v}{\partial x_j} \, \mathrm{d}x + \sum_{i=1}^{n} \int_\Omega b_i(x) \frac{\partial u}{\partial x_i} v \, \mathrm{d}x + \int_\Omega c(x) uv \, \mathrm{d}x = \int_\Omega f(x) v(x) \, \mathrm{d}x \quad \forall v \in C_0^1(\Omega). \tag{2.5}$$

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 (2.5) are still meaningful (in fact, we shall prove this below). This motivates the following definition.

Definition 2.1 *Let $a_{ij} \in C(\bar{\Omega})$, $i, j = 1, \ldots, n$, $b_i \in C(\bar{\Omega})$, $i = 1, \ldots, n$, $c \in C(\bar{\Omega})$, and let $f \in L_2(\Omega)$. A function $u \in H_0^1(\Omega)$ satisfying*

$$\sum_{i,j=1}^{n} \int_\Omega a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} \, \mathrm{d}x + \sum_{i=1}^{n} \int_\Omega b_i(x) \frac{\partial u}{\partial x_i} v \, \mathrm{d}x + \int_\Omega c(x) uv \, \mathrm{d}x = \int_\Omega f(x) v(x) \, \mathrm{d}x \quad \forall v \in H_0^1(\Omega) \tag{2.6}$$

is called a weak solution *of (2.3), (2.4). All partial derivatives in (2.6) should be understood as weak derivatives.*

Clearly if u is a classical solution of (2.3), (2.4), then it is also a weak solution of (2.3), (2.4). However, the converse is not true. If (2.3), (2.4) 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 general problem (2.3), (2.4).

For the sake of simplicity, let us introduce the following notation:

$$a(u, v) = \sum_{i,j=1}^{n} \int_\Omega a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} \, \mathrm{d}x + \sum_{i=1}^{n} \int_\Omega b_i(x) \frac{\partial u}{\partial x_i} v \, \mathrm{d}x + \int_\Omega c(x) uv \, \mathrm{d}x \tag{2.7}$$

and

$$l(v) = \int_\Omega f(x) v(x) \, \mathrm{d}x. \tag{2.8}$$

With this new notation, problem (2.6) can be written as follows:

$$\text{find } u \in H_0^1(\Omega) \text{ such that } a(u,v) = l(v) \quad \forall v \in H_0^1(\Omega). \tag{2.9}$$

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

Theorem 2.2 *(Lax–Milgram theorem) Suppose that V is a real Hilbert space equipped with norm $\|\cdot\|_V$. Let $a(\cdot,\cdot)$ be a bilinear form 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) $\exists c_1 > 0 \quad \forall v, w \in V \quad |a(v,w)| \leq c_1 \|v\|_V \|w\|_V$,

and let $l(\cdot)$ be a linear form 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 book of P. Ciarlet: The Finite Element Method for Elliptic Problems, North-Holland, 1978.

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 (2.3), (2.4) (or, equivalently, to (2.9)). Let us recall from Section 1.3 that $H_0^1(\Omega)$ is a Hilbert space with the inner product

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

and the associated norm $\|u\|_{H^1(\Omega)} = (u,u)_{H^1(\Omega)}^{1/2}$. Next we show that $a(\cdot,\cdot)$ and $l(\cdot)$, defined by (2.7) and (2.8), 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 that $l(\cdot)$ is a linear form on $H_0^1(\Omega)$. Also, by the Cauchy–Schwarz inequality,

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

for all $v \in H_0^1(\Omega)$, where we have used the obvious inequality $\|v\|_{L_2(\Omega)} \leq \|v\|_{H^1(\Omega)}$. 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 form on $H_0^1(\Omega) \times H_0^1(\Omega)$. Employing the Cauchy–Schwarz inequality, we deduce that

$$\begin{aligned}
|a(u,v)| &\leq \sum_{i,j=1}^{n} \max_{x\in\bar{\Omega}} |a_{ij}(x)| \left| \int_\Omega \frac{\partial u}{\partial x_i} \, \frac{\partial v}{\partial x_j} \, \mathrm{d}x \right| \\
&\quad + \sum_{i=1}^{n} \max_{x\in\bar{\Omega}} |b_i(x)| \left| \int_\Omega \frac{\partial u}{\partial x_i} v \, \mathrm{d}x \right| \\
&\quad + \max_{x\in\bar{\Omega}} |c(x)| \left| \int_\Omega u(x) v(x) \, \mathrm{d}x \right| \\
&\leq c \left\{ \sum_{i,j=1}^{n} \left(\int_\Omega \left| \frac{\partial u}{\partial x_i} \right|^2 \mathrm{d}x \right)^{1/2} \left(\int_\Omega \left| \frac{\partial v}{\partial x_j} \right|^2 \mathrm{d}x \right)^{1/2} \right. \\
&\quad + \sum_{i=1}^{n} \left(\int_\Omega \left| \frac{\partial u}{\partial x_i} \right|^2 \mathrm{d}x \right)^{1/2} \left(\int_\Omega |v|^2 \, \mathrm{d}x \right)^{1/2} \\
&\quad \left. + \left(\int_\Omega |u|^2 \, \mathrm{d}x \right)^{1/2} \left(\int_\Omega |v|^2 \, \mathrm{d}x \right)^{1/2} \right\} \\
&\leq c \left\{ \left(\int_\Omega |u|^2 \, \mathrm{d}x \right)^{1/2} + \sum_{i=1}^{n} \left(\int_\Omega \left| \frac{\partial u}{\partial x_i} \right|^2 \mathrm{d}x \right)^{1/2} \right\} \\
&\quad \times \left\{ \left(\int_\Omega |v|^2 \, \mathrm{d}x \right)^{1/2} + \sum_{j=1}^{n} \left(\int_\Omega \left| \frac{\partial v}{\partial x_j} \right|^2 \mathrm{d}x \right)^{1/2} \right\}, \qquad (2.10)
\end{aligned}$$

where

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

By further majorisation of the right-hand side in (2.10),

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

so that, by letting $c_1 = 2nc$, we obtain inequality (b).

It remains to establish (a). Using (2.2), we deduce that

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

where we wrote $\frac{\partial u}{\partial x_i} \cdot u$ as $\frac{1}{2} \frac{\partial}{\partial x_i}(u^2)$. Integrating by parts in the second term on the right, we obtain

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

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

$$c(x) - \frac{1}{2} \sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i} \geq 0, \quad x \in \bar{\Omega}. \tag{2.11}$$

Then,

$$a(u,u) \geq \tilde{c} \sum_{i=1}^{n} \int_{\Omega} \left| \frac{\partial u}{\partial x_i} \right|^2 \, \mathrm{d}x. \tag{2.12}$$

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

$$a(u,u) \geq \frac{\tilde{c}}{c_\star} \int_{\Omega} |u|^2 \, \mathrm{d}x. \tag{2.13}$$

Summing (2.12) and (2.13) multiplied by $c_\star$,

$$a(u,u) \geq c_0 \left(\int_{\Omega} |u|^2 \, \mathrm{d}x + \sum_{i=1}^{n} \int_{\Omega} \left| \frac{\partial u}{\partial x_i} \right|^2 \, \mathrm{d}x \right), \tag{2.14}$$

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 (2.9); thence problem (2.3), (2.4) has a unique weak solution.

We encapsulate this result in the following theorem.

Theorem 2.3 *Suppose that $a_{ij} \in C(\bar{\Omega})$, $i, j = 1, \ldots, n$, $b_i \in C^1(\bar{\Omega})$, $i = 1, \ldots, n$, $c \in C(\bar{\Omega})$, $f \in L_2(\Omega)$, and assume that (2.2) and (2.11) hold; then, the boundary value problem (2.3), (2.4) possesses a unique weak solution $u \in H_0^1(\Omega)$. In addition,*

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{c_0} \|f\|_{L_2(\Omega)}. \tag{2.15}$$

Proof We only have to prove (2.15). By (2.14), (2.9), the Cauchy–Schwarz inequality and recalling the definition of $\|\cdot\|_{H^1(\Omega)}$,

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

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 $a_{ij}(x) \equiv 1$, $i = j$, $a_{ij}(x) \equiv 0$, $i \neq j$, $1 \leq i, j \leq n$, $b_i(x) \equiv 0$, $c(x) \equiv 0$, $f(x) = \operatorname{sgn}(\frac{1}{2} - |x|)$, and $\Omega = (-1,1)^n$, we see that (2.2) holds with $\tilde{c} = 1$ and (2.11) is trivially fulfilled. Thus (∗) has a unique weak solution $u \in H^1_0(\Omega)$.

Remark. The existence and uniqueness of a weak solution to a Neumann, a Robin, or an oblique derivative boundary value problem can be established in a similar fashion, using the Lax–Milgram theorem. ⋄

Remark. Theorem 2.3 implies that the weak formulation of the elliptic boundary value problem (2.3), (2.4) 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 (2.3), (2.4) 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 (2.3), (2.4) corresponding to the right-hand side $f_1 - f_2 \in L^2(\Omega)$. Thus, by virtue of (2.15),

$$\|u_1 - u_2\|_{H^1(\Omega)} \leq \frac{1}{c_0} \|f_1 - f_2\|_{L_2(\Omega)} , \tag{2.16}$$

and hence the required continuous dependence of the solution of the boundary value problem on the right-hand side. ⋄

3 Introduction to the theory of finite difference schemes

Let Ω be a bounded open set in $\mathbb{R}^n$, and suppose we wish to solve the boundary value problem

$$\mathcal{L}u = f \quad \text{in} \quad \Omega, \tag{3.1a}$$
$$lu = g \quad \text{on} \quad \Gamma = \partial\Omega, \tag{3.1b}$$

where $\mathcal{L}$ is a linear partial differential operator, and l is a linear operator which specifies the boundary condition. For example,

$$\mathcal{L}u \equiv -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial u}{\partial x_i}\right) + \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} + cu,$$

and

$$lu \equiv u \quad \text{(Dirichlet boundary condition)},$$

or

$$lu \equiv \frac{\partial u}{\partial \nu} \quad \text{(Neumann boundary condition)},$$

or

$$lu \equiv \sum_{i,j=1}^{n} a_{ij}\frac{\partial u}{\partial x_i}\cos\alpha_j + \sigma(x)u \quad \text{(oblique derivative boundary condition)},$$

or some other appropriate boundary condition.

In general, it is impossible to determine the solution of the boundary value problem (3.1) in closed form. Thus the aim of this chapter is to describe a simple and general numerical technique for the approximate solution of (3.1), called the *finite difference method*. The construction of a finite difference scheme consists of two basic steps: first, the approximation of the computational domain by a finite set of points, and second, the approximation of the derivatives appearing in the differential equation and in the boundary condition by divided differences.

To describe the first of these two steps more precisely, suppose that we have approximated $\bar{\Omega} = \Omega \cup \Gamma$ by a finite set of points

$$\bar{\Omega}_h = \Omega_h \cup \Gamma_h,$$

where $\Omega_h \subset \Omega$ and $\Gamma_h \subset \Gamma$; $\bar{\Omega}_h$ is called a mesh, Ω_h is the set of interior mesh-points and Γ_h the set boundary mesh-points. The parameter $h = (h_1, \ldots, h_n)$ measures the fineness of the mesh (here h_i denotes the mesh-size in the coordinate direction Ox_i): the smaller $|h|$ is, the denser the mesh.

Having constructed the mesh, we proceed by replacing the derivatives in $\mathcal{L}$ by divided differences, and approximate the boundary condition in a similar fashion. This yields the finite difference scheme

$$\mathcal{L}_h U(x) = f_h(x), \quad x \in \Omega_h, \tag{3.2a}$$
$$l_h U(x) = g_h(x), \quad x \in \Gamma_h, \tag{3.2b}$$

where f_h and g_h are suitable approximations of f and g, respectively. Now (3.2) is a system of linear equations involving the values of U at the mesh-points, and can be solved by Gaussian elimination or an iterative method, provided, of course, that it has a unique solution. The sequence $\{U(x) : x \in \bar{\Omega}_h\}$ parametrised by mesh parameter h is an approximation to the sequence $\{u(x) : x \in \bar{\Omega}_h\}$, — the values of the exact solution at the mesh-points.

There are two classes of problems associated with finite difference schemes:

(1) the first, and most fundamental, is the problem of approximation, that is, whether (3.2) approximates the boundary value problem (3.1) in some sense, and whether its solution $\{U(x) : x \in \bar{\Omega}_h\}$ approximates $\{u(x) : x \in \bar{\Omega}_h\}$, the values of the exact solution at the mesh-points.

(2) the second problem concerns the efficient solution of the discrete problem (3.2) using techniques from Numerical Linear Algebra.

In these notes we shall be concerned with the first of these two problems - the question of approximation.

In order to give a simple illustration of the general framework of finite difference approximation, let us consider the following two-point boundary value problem for a second-order linear (ordinary) differential equation:

$$-u'' + c(x)u = f(x), \quad x \in (0,1), \tag{3.3a}$$
$$u(0) = 0, \quad u(1) = 0. \tag{3.3b}$$

The first step in the construction of a finite difference scheme for this boundary value problem is to define the mesh. Let N be an integer, $N \geq 2$, and let $h = 1/N$ be the mesh-size; the mesh-points are $x_i = ih$, $i = 0, \ldots, N$. Formally, $\Omega_h = \{x_i : i = 1, \ldots, N-1\}$, $\Gamma_h = \{x_0, x_N\}$, and $\bar{\Omega}_h = \Omega_h \cup \Gamma_h$. Suppose that u is sufficiently smooth (e.g. $u \in C^4[0,1]$). Then, by Taylor series expansion,

$$\begin{aligned} u(x_{i\pm1}) &= u(x_i \pm h) \\ &= u(x_i) \pm h u'(x_i) + \frac{h^2}{2} u''(x_i) \pm \frac{h^3}{6} u'''(x_i) + \mathcal{O}(h^4), \end{aligned}$$

so that

$$D_x^+ u(x_i) \equiv \frac{u(x_{i+1}) - u(x_i)}{h} = u'(x_i) + \mathcal{O}(h),$$

$$D_x^- u(x_i) \equiv \frac{u(x_i) - u(x_{i-1})}{h} = u'(x_i) + \mathcal{O}(h),$$

and

$$\begin{aligned} D_x^+ D_x^- u(x_i) &= D_x^- D_x^+ u(x_i) \\ &= \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} \\ &= u''(x_i) + \mathcal{O}(h^2). \end{aligned}$$

Thus we replace the second derivative u'' by a second divided difference:

$$-D_x^+ D_x^- u(x_i) + c(x_i)u(x_i) \approx f(x_i), \quad i = 1, \ldots, N-1, \tag{3.4a}$$

$$u(x_0) = 0, \;\; u(x_N) = 0. \tag{3.4b}$$

Now (3.4) indicates that the approximate solution U should be sought as the solution of the system of difference equations:

$$-D_x^+ D_x^- U_i + c(x_i)U_i = f(x_i), \quad i = 1, \ldots, N-1, \tag{3.5a}$$

$$U_0 = 0, \;\; U_N = 0. \tag{3.5b}$$

Using matrix notation, this can be written as

$$\begin{bmatrix} \frac{2}{h^2} + c(x_1) & -\frac{1}{h^2} & & & \bigcirc \\ -\frac{1}{h^2} & \frac{2}{h^2} + c(x_2) & -\frac{1}{h^2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{h^2} & \frac{2}{h^2} + c(x_{N-2}) & -\frac{1}{h^2} \\ \bigcirc & & & -\frac{1}{h^2} & \frac{2}{h^2} + c(x_{N-1}) \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-2} \\ U_{N-1} \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{N-2}) \\ f(x_{N-1}) \end{bmatrix},$$

or, more compactly, $AU = F$, where A is the tri-diagonal $(N-1) \times (N-1)$ matrix displayed above, and U and F are column vectors of size $N-1$.

We begin the analysis of the finite difference scheme (3.5) by showing that it has a unique solution. It suffices to show that the matrix A is non-singular. For this purpose, we introduce, for two functions V and W defined at the interior mesh-points x_i, $i = 1, \ldots, N-1$, the inner product

$$(V, W)_h = \sum_{i=1}^{N-1} h V_i W_i$$

(which resembles the L^2-inner product

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

Lemma 3.1 *Suppose that V is a function defined at the mesh-points x_i, $i = 0, \ldots, N$, and let $V_0 = V_N = 0$; then,*

$$(-D_x^+ D_x^- V, V)_h = \sum_{i=1}^{N} h \left| D_x^- V_i \right|^2 . \tag{3.6}$$

Proof Performing summation by parts,

$$\begin{aligned}
(-D_x^+ D_x^- V, V)_h &= -\sum_{i=1}^{N-1} (D_x^+ D_x^- V_i) V_i h \\
&= -\sum_{i=1}^{N-1} \frac{V_{i+1} - V_i}{h} V_i + \sum_{i=1}^{N-1} \frac{V_i - V_{i-1}}{h} V_i \\
&= -\sum_{i=2}^{N} \frac{V_i - V_{i-1}}{h} V_{i-1} + \sum_{i=1}^{N-1} \frac{V_i - V_{i-1}}{h} V_i \\
&= -\sum_{i=1}^{N} \frac{V_i - V_{i-1}}{h} V_{i-1} + \sum_{i=1}^{N} \frac{V_i - V_{i-1}}{h} V_i \\
&= \sum_{i=1}^{N} \frac{V_i - V_{i-1}}{h} (V_i - V_{i-1}) = \sum_{i=1}^{N} h \left| D_x^- V_i \right|^2 ,
\end{aligned}$$

where in the third line we shifted the indices in the first summation, and in the fourth line we made use of the fact that $V_0 = V_N = 0$. □

Returning to the finite difference scheme (3.5), let V be as in the above lemma and note that if $c(x) \geq 0$ then,

$$\begin{aligned}
(AV, V)_h &= (-D_x^+ D_x^- V + cV, V)_h \\
&= (-D_x^+ D_x^- V, V)_h + (cV, V)_h \\
&\geq \sum_{i=1}^{N} h \left| D_x^- V_i \right|^2 .
\end{aligned} \tag{3.7}$$

Thus, if $AV = 0$ for some V, then $D_x^- V_i = 0$, $i = 1, \ldots, N$; because $V_0 = V_N = 0$, this implies that $V_i = 0$, $i = 0, \ldots, N$. Hence $AV = 0$ if and only if $V = 0$. We deduce that A is a non-singular matrix, and (3.5) has a unique solution, $U = A^{-1} F$.

Theorem 3.2 *Suppose that c and f are continuous functions on $[0, 1]$, and $c(x) \geq 0$, $x \in [0, 1]$; then, the finite difference scheme (3.5) possesses a unique solution U.*

We note that, by virtue of Theorem 2.3, the boundary value problem (3.3) has a unique (weak) solution under the same hypotheses on c and f as in Theorem 3.2.

Next, we investigate the approximation properties of the difference scheme (3.5). A key ingredient in our analysis is the fact that the scheme (3.5) is stable (or discretely well-posed) in the sense that "small" perturbations in the data result in "small" perturbations in the corresponding finite difference solution. Effectively, we shall prove the discrete version of the inequality (2.15). For this purpose, we define the *discrete L^2-norm*

$$\|U\|_h = (U,U)_h^{1/2} = \left(\sum_{i=1}^{N-1} h|U_i|^2 \right)^{1/2},$$

and the *discrete Sobolev norm*

$$\|U\|_{1,h} = (\|U\|_h^2 + \left\|D_x^- U\right]\!\right|_h^2)^{1/2},$$

where

$$\|V\!]_h^2 = \sum_{i=1}^{N} h\,|V_i|^2\,.$$

Using this notation, the inequality (3.7) can be written

$$(AV,V)_h \geq \left\|D_x^- V\right]\!\right|_h^2. \tag{3.8}$$

In fact, employing a discrete version of the Poincaré–Friedrichs inequality (1.1), stated in Lemma 3.3 below, we shall prove that

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

where c_0 is a positive constant.

Lemma 3.3 *(Discrete Poincaré–Friedrichs inequality.) Let V be a function defined on the mesh $\{x_i, i = 0,\ldots,N\}$, and such that $V_0 = V_N = 0$; then, there exists a positive constant $c_\star$, independent of V and h, such that*

$$\|V\|_h^2 \leq c_\star \left\|D_x^- V\right]\!\right|_h^2 \tag{3.9}$$

for all such V.

Proof We proceed in the same way as in the proof of (1.1). First note that

$$|V_i|^2 = \left| \sum_{j=1}^{i} (D_x^- V_j) h \right|^2 \leq \left(\sum_{j=1}^{i} h \right) \sum_{j=1}^{i} h\,\left|D_x^- V_j\right|^2.$$

Thence,

$$\begin{aligned}\|V\|_h^2 &= \sum_{i=1}^{N-1} h\,|V_i|^2 \le \sum_{i=1}^{N-1} ih^2 \sum_{j=1}^{i} h\left|D_x^- V_j\right|^2 \\ &\le \frac{(N-1)N}{2} h^2 \sum_{j=1}^{N} h\left|D_x^- V_j\right|^2 \\ &\le \frac{1}{2}\left\|D_x^- V\right]\!\right|_h^2 . \qquad \square\end{aligned}$$

We note that the constant $c_\star = 1/2$ in (3.9).

Using (3.9) to bound the right-hand side of (3.8) from below we obtain

$$(AV, V)_h \ge \frac{1}{c_\star}\|V\|_h^2 . \tag{3.10}$$

Adding (3.8) to (3.10) multiplied by $c_\star$, we deduce that

$$(AV, V)_h \ge (1 + c_\star)^{-1}\left(\|V\|_h^2 + \left\|D_x^- V\right]\!\right|_h^2\right).$$

Letting $c_0 = (1 + c_\star)^{-1}$,

$$(AV, V)_h \ge c_0\|V\|_{1,h}^2 . \tag{3.11}$$

Now the stability of the finite difference scheme (3.5) easily follows.

Theorem 3.4 *The scheme (3.5) is stable in the sense that*

$$\|U\|_{1,h} \le \frac{1}{c_0}\|f\|_h . \tag{3.12}$$

Proof From (3.11) and (3.5) we have that

$$\begin{aligned}c_0\|U\|_{1,h}^2 &\le (AU, U)_h = (f, U)_h \le |(f, U)_h| \\ &\le \|f\|_h\|U\|_h \le \|f\|_h\|U\|_{1,h} ,\end{aligned}$$

and hence (3.12). $\square$

Using this stability result it is easy to derive an estimate of the error between the exact solution u, and its finite difference approximation, U. We define the *global error*, e, by

$$e_i := u(x_i) - U_i, \quad i = 0, \ldots, N.$$

Obviously $e_0 = 0$, $e_N = 0$, and

$$\begin{aligned} Ae_i &= Au(x_i) - AU_i = Au(x_i) - f(x_i) \\ &= -D_x^+ D_x^- u(x_i) + c(x_i)u(x_i) - f(x_i) \\ &= u''(x_i) - D_x^+ D_x^- u(x_i), \quad i = 1, \ldots, N-1. \end{aligned}$$

Thus,

$$Ae_i = \varphi_i, \quad i = 1, \ldots, N-1, \tag{3.13a}$$

$$e_0 = 0, \quad e_N = 0, \tag{3.13b}$$

where $\varphi_i = u''(x_i) - D_x^+ D_x^- u(x_i)$ is the *truncation error*.

Applying (3.12) to the finite difference scheme (3.13), we obtain

$$\|u - U\|_{1,h} = \|e\|_{1,h} \leq \frac{1}{c_0} \|\varphi\|_h . \tag{3.14}$$

It remains to estimate $\|\varphi\|_h$. We have shown on page 19 that, if $u \in C^4[0,1]$, then,

$$\varphi_i = u''(x_i) - D_x^+ D_x^- u(x_i) = \mathcal{O}(h^2),$$

i.e. there is a positive constant C, independent of h, such that

$$|\varphi_i| \leq Ch^2.$$

Consequently,

$$\|\varphi\|_h = \left(\sum_{i=1}^{N-1} h \, |\varphi_i|^2 \right)^{1/2} \leq Ch^2. \tag{3.15}$$

Combining (3.14) and (3.15), it follows that

$$\|u - U\|_{1,h} \leq \frac{C}{c_0} h^2. \tag{3.16}$$

In fact, a more careful treatment of the remainder term in the Taylor series expansion on p. 19 reveals that

$$\varphi_i = u''(x_i) - D_x^+ D_x^- u(x_i) = -\frac{h^2}{12} u^{IV}(\xi_i), \quad \xi_i \in [x_{i-1}, x_{i+1}].$$

Thus

$$|\varphi_i| \leq h^2 \frac{1}{12} \max_{x \in [0,1]} \left| u^{IV}(x) \right| ,$$

and hence

$$C = \frac{1}{12} \max_{x \in [0,1]} \left| u^{IV}(x) \right|$$

in (3.15). Recalling that $c_0 = (1+c_\star)^{-1}$ and $c_\star = 1/2$, we deduce that $c_0 = 2/3$. Substituting the values of the constants C and c_0 into (3.16), it follows that

$$\|u - U\|_{1,h} \leq \frac{1}{8} h^2 \left\|u^{IV}\right\|_{C[0,1]}.$$

Thus we have proved the following result.

Theorem 3.5 *Let $f \in C[0,1]$, $c \in C[0,1]$, with $c(x) \geq 0$, $x \in [0,1]$, and suppose that the corresponding (weak) solution of the boundary value problem (3.3) belongs to $C^4[0,1]$; then,*

$$\|u - U\|_{1,h} \leq \frac{1}{8} h^2 \left\|u^{IV}\right\|_{C[0,1]}. \tag{3.17}$$

The analysis of the finite difference scheme (3.3) contains the key steps of a general error analysis for finite difference approximations of (elliptic) partial differential equations:

(1) The first step is to prove the stability of the scheme in an appropriate mesh-dependent norm (c.f. (3.12), for example). A typical stability result for the general finite difference scheme (3.2) is

$$|||U|||_{\Omega_h} \leq c(\|f_h\|_{\Omega_h} + \|g_h\|_{\Gamma_h}), \tag{3.18}$$

where $|||\cdot|||_{\Omega_h}$, $\|\cdot\|_{\Omega_h}$ and $\|\cdot\|_{\Gamma_h}$ are mesh-dependent norms involving mesh-points of Ω_h (or $\bar{\Omega}_h$) and Γ_h, respectively, and c is a positive constant, independent of h.

(2) The second step is to estimate the size of the *truncation error*,

$$\begin{array}{rcll} \varphi_{\Omega_h} & = & L^h u - f_h, & \text{in} \quad \Omega_h, \\ \varphi_{\Gamma_h} & = & l_h u - g_h, & \text{on} \quad \Gamma_h. \end{array}$$

(in the case of the finite difference scheme (3.3) $\varphi_{\Gamma_h} = 0$, and therefore φ_{Γ_h} never appeared explicitly in our error analysis). If

$$\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h} \to 0 \quad \text{as} \quad h \to 0,$$

for a sufficiently smooth solution u of (3.1), we say that the scheme (3.2) is *consistent.* If p is the largest positive integer such that

$$\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h} \leq C h^p \quad \text{as} \quad h \to 0,$$

(where C is a positive constant independent of h) for all sufficiently smooth u, the scheme is said to have *order of accuracy p.*

The finite difference scheme (3.2) is said to provide a *convergent* approximation to (3.1) in the norm $|||\cdot|||_{\Omega_h}$, if

$$|||u - U|||_{\Omega_h} \to 0 \quad \text{as } h \to 0.$$

If q is the largest positive integer such that

$$|||u - U|||_{\Omega_h} \leq Ch^q \qquad \text{as } h \to 0$$

(where C is a positive constant independent of h), then the scheme is said to have *order of convergence* q.

From these definitions we deduce the following fundamental theorem.

Theorem 3.6 *Suppose that the finite difference scheme (3.2) is stable (i.e. (3.18) holds for all f_h and g_h) and that the scheme is a consistent approximation of (3.1); then, (3.2) is a convergent approximation of (3.1), and the order of convergence is not smaller then the order of accuracy.*

Proof We define the *global error* $e = u - U$. Then,

$$L^h e = L^h(u - U) = L^h u - L^h U = L^h u - f_h.$$

Thus

$$L^h e = \varphi_{\Omega_h},$$

and similarly,

$$l_h e = \varphi_{\Gamma_h}.$$

By stability,

$$|||u - U|||_{\Omega_h} = |||e|||_{\Omega_h} \leq c(\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h}),$$

and hence the stated result. □

Thus, paraphrasing Theorem 3.6, *stability* and *consistency* imply *convergence*. This abstract result is at the heart of the error analysis of finite difference approximations of differential equations.

4 Finite difference approximation of elliptic boundary value problems

In Section 3 we presented a detailed error analysis for a finite difference approximation of a two-point boundary value problem. Here we shall carry out a similar analysis for the model problem

$$-\Delta u + c(x)u = f(x) \quad \text{in } \Omega, \tag{4.1a}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{4.1b}$$

where $\Omega = (0,1) \times (0,1)$, c is a continuous function on $\bar{\Omega}$ and $c(x) \geq 0$. As far as the smoothness of the function f is concerned, we shall consider two separate cases:

(a) First we shall assume that f is a continuous function on $\bar{\Omega}$. In this case, the error analysis will proceed along the same lines as in Section 3.

(b) We shall then consider the case when f is only in $L^2(\Omega)$. In this instance the boundary value problem (4.1) does not have a classical solution – only a weak solution exists. This lack of smoothness gives rise to some technical difficulties: in particular, we cannot use a Taylor series expansion to estimate the size of the truncation error. We shall bypass the problem by employing a different technique, instead.

(a) ($f \in C(\bar{\Omega})$) The first step in the construction of the finite difference approximation of (4.1) is to define the mesh. Let N be an integer, $N \geq 2$, and let $h = 1/N$; the mesh-points are (x_i, y_j), $i, j = 0, \ldots, N$, where $x_i = ih$, $y_j = jh$. These mesh-points form the mesh

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

Similarly as in Section 3, 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 $\Gamma_h = \bar{\Omega}_h \setminus \Omega_h$. Analogously to (3.5), the difference scheme is:

$$-(D_x^+ D_x^- U_{ij} + D_y^+ D_y^- U_{ij}) + c(x_i, y_j)U_{ij} = f(x_i, y_j), \quad (x_i, y_j) \in \Omega_h, \tag{4.2a}$$

$$U = 0 \quad \text{on } \Gamma_h. \tag{4.2b}$$

In an expanded form, this can be written

$$-\left\{\frac{U_{i+1,j} - 2U_{ij} + U_{i-1,j}}{h^2} + \frac{U_{i,j+1} - 2U_{ij} + U_{i,j-1}}{h^2}\right\} + c(x_i, y_j)U_{ij} = f(x_i, y_j),$$

$$i, j = 1, \ldots, N-1, \tag{4.3}$$

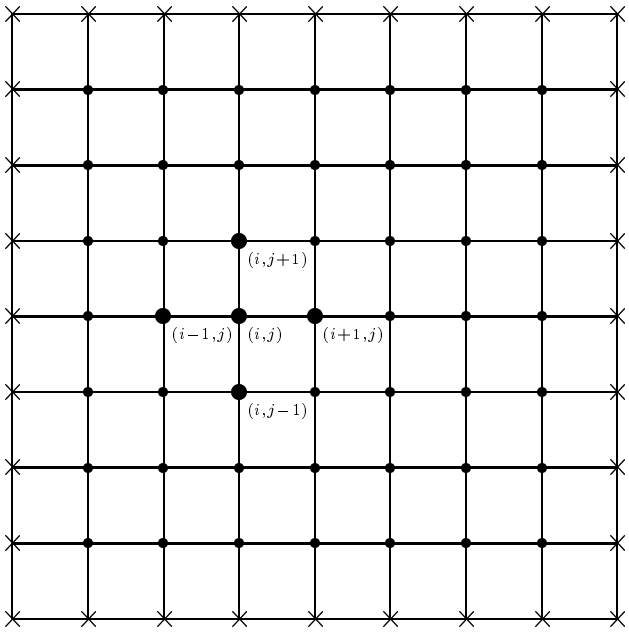


Figure 1: The mesh $\Omega_h(\cdot)$, the boundary mesh $\Gamma_h(\times)$, and a typical 5-point difference stencil.

$$U_{ij} = 0, \text{ if } i = 0, \ i = N \text{ or if } j = 0, \ j = N. \tag{4.4}$$

For each i and j, $1 \leq i, j \leq N-1$, the finite difference equation (4.3) involves five values of the approximate solution U: $U_{i,j}$, $U_{i-1,j}$, $U_{i+1,j}$, $U_{i,j-1}$, $U_{i,j+1}$. It is again possible to write (4.3), (4.4) as a system of linear equations

$$AU = F, \tag{4.5}$$

where

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

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

and A is an $(N-1)^2 \times (N-1)^2$ sparse matrix of banded 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. 1, while the sparsity structure of A is depicted in Fig. 2.

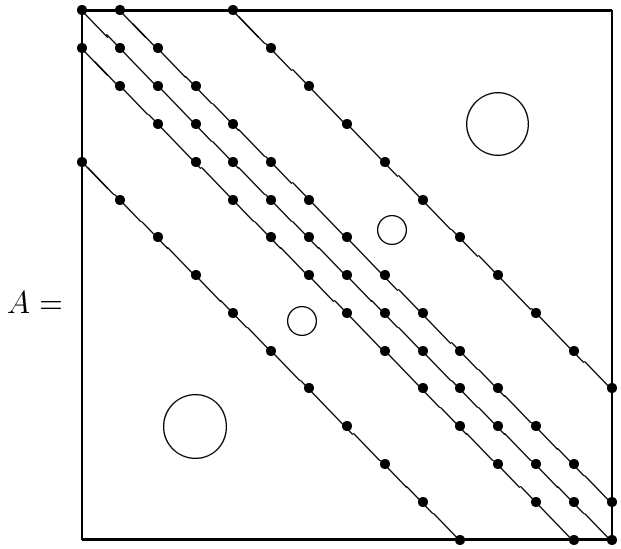


Figure 2: The sparsity structure of the banded matrix A.

Next we show that (4.2) has a unique solution. We proceed in the same way as in Section 3. For two functions, V and W, defined on Ω_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 4.1 *Suppose that V is a function defined on $\bar{\Omega}_h$ and that $V = 0$ on Γ_h; then,*

$$(-D_x^+ D_x^- V, V)_h + (-D_y^+ D_y^- V, V)_h = \sum_{i=1}^{N} \sum_{j=1}^{N-1} h^2 |D_x^- V_{ij}|^2 + \sum_{i=1}^{N-1} \sum_{j=1}^{N} h^2 |D_y^- V_{ij}|^2. \quad (4.6)$$

Proof (4.6) is a straightforward consequence of (3.6) and the analogous identity for $-D_y^+ D_y^-$. □

Returning to the analysis of the finite difference scheme (4.2), we note that, since $c(x, y) \geq 0$ on $\bar{\Omega}$, by (4.6) we have

$$\begin{aligned}
(AV, V)_h &= (-D_x^+ D_x^- V - D_y^+ D_y^- V + cV, V)_h \\
&= (-D_x^+ D_x^- V, V)_h + (-D_y^+ D_y^- V, V)_h + (cV, V)_h \\
&\geq \sum_{i=1}^{N} \sum_{j=1}^{N-1} h^2 |D_x^- V_{ij}|^2 + \sum_{i=1}^{N-1} \sum_{j=1}^{N} h^2 |D_y^- V_{ij}|^2, \qquad (4.7)
\end{aligned}$$

for any V defined on $\bar{\Omega}_h$ such that $V = 0$ on Γ_h. Now this implies, just as in the one-dimensional analysis presented in Section 3, that A is a non-singular matrix. Indeed if $AV = 0$, then (4.7) 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 Γ_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 (4.2). Thus the solution of the finite difference scheme (4.2) may be found by solving the system of linear equations (4.5).

In order to prove the stability of the finite difference scheme (4.2), we introduce (similarly as in one dimension) the mesh–dependent norms

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

and

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

where

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

and

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

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

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

With this new notation, the inequality (4.7) takes the following form:

$$(AV, V)_h \geq \left\|D_x^- V\right]\right|_x^2 + \left\|D_y^- V\right]\right|_y^2. \tag{4.8}$$

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 4.2 *(Discrete Poincaré–Friedrichs inequality.)*

Let V be a function defined on $\bar{\Omega}_h$ and such that $V = 0$ on Γ_h; then, there exists a constant c_*, independent of V and h, such that

$$\|V\|_h^2 \leq c_* \left(\left\| D_x^- V \right] \right|_x^2 + \left| \left\| D_y^- V \right] \right|_y^2 \right) \tag{4.9}$$

for all such V.

Proof (4.9) is a straightforward consequence of its one-dimensional counterpart (3.9). It follows from (3.9) that, for each fixed j, $1 \leq j \leq N-1$,

$$\sum_{i=1}^{N-1} h|V_{ij}|^2 \leq \frac{1}{2} \sum_{i=1}^{N} h|D_x^- V_{ij}|^2. \tag{4.10}$$

Analogously, for each fixed i, $1 \leq i \leq N-1$,

$$\sum_{j=1}^{N-1} h|V_{ij}|^2 \leq \frac{1}{2} \sum_{j=1}^{N} h|D_y^- V_{ij}|^2. \tag{4.11}$$

We multiply (4.10) by h and sum through j, $1 \leq j \leq N-1$, multiply (4.11) by h and sum through i, $1 \leq i \leq N-1$, and add these two inequalities to obtain

$$2\|V\|_h^2 \leq \frac{1}{2} \left(\left\| D_x^- V \right] \right|_x^2 + \left| \left\| D_y^- V \right] \right|_y^2 \right).$$

Hence (4.9) with $c_* = \frac{1}{4}$. □

Now (4.8) and (4.9) imply that

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

Finally, combining this with (4.8) and recalling the definition of the norm $\|\cdot\|_{1,h}$, we obtain

$$(AV, V)_h \geq c_0 \|V\|_{1,h}^2 , \tag{4.12}$$

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

Theorem 4.3 *The scheme (4.2) is stable in the sense that*

$$\|U\|_{1,h} \leq \frac{1}{c_0} \|f\|_h . \tag{4.13}$$

Proof Identical to the proof of (3.12) $\square$.

Having established stability, we turn to the question of accuracy. We define the global error, e, by

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

Then, assuming that $u \in C^4(\bar{\Omega})$, and employing Taylor series expansions,

$$\begin{aligned} Ae_{ij} &= \Delta u(x_i, y_j) - (D_x^+ D_x^- u(x_i, y_j) + D_y^+ D_y^- u(x_i, y_j)) \\ &= \left[\frac{\partial^2 u}{\partial x^2}(x_i, y_j) - D_x^+ D_x^- u(x_i, y_j)\right] + \left[\frac{\partial^2 u}{\partial y^2}(x_i, y_j) - D_y^+ D_y^- u(x_i, y_j)\right] \\ &= -\frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_i, y_j) - \frac{h^2}{12}\frac{\partial^4 u}{\partial y^4}(x_i, \eta_j), \qquad 1 \le i, j \le N-1, \end{aligned}$$

where $\xi_i \in [x_{i-1}, x_{i+1}]$, $\eta_j \in [y_{j-1}, y_{j+1}]$.

Let

$$\varphi_{ij} = -\frac{h^2}{12}\left(\frac{\partial^4 u}{\partial x^4}(\xi_i, y_j) + \frac{\partial^4 u}{\partial y^4}(x_i, \eta_j)\right), \quad 1 \le i, j \le N-1;$$

then,

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

By virtue of (4.13),

$$\|u - U\|_{1,h} = \|e\|_{1,h} \le \frac{1}{c_0}\|\varphi\|_h . \tag{4.14}$$

Noting that

$$|\varphi_{ij}| \le \frac{h^2}{12}\left(\left\|\frac{\partial^4 u}{\partial x^4}\right\|_{C(\bar{\Omega})} + \left\|\frac{\partial^4 u}{\partial y^4}\right\|_{C(\bar{\Omega})}\right),$$

we deduce that the truncation error, φ, satisfies

$$\|\varphi\|_h \le \frac{h^2}{12}\left(\left\|\frac{\partial^4 u}{\partial x^4}\right\|_{C(\bar{\Omega})} + \left\|\frac{\partial^4 u}{\partial y^4}\right\|_{C(\bar{\Omega})}\right). \tag{4.15}$$

Finally (4.14) and (4.15) yield the following result.

Theorem 4.4 *Let $f \in C(\bar{\Omega})$, $c \in C(\bar{\Omega})$, with $c(x, y) \ge 0$, $(x, y) \in \bar{\Omega}$, and suppose that the corresponding weak solution of the boundary value problem (4.1) belongs to $C^4(\bar{\Omega})$; then,*

$$\|u - U\|_{1,h} \le \frac{5h^2}{48}\left(\left\|\frac{\partial^4 u}{\partial x^4}\right\|_{C(\bar{\Omega})} + \left\|\frac{\partial^4 u}{\partial y^4}\right\|_{C(\bar{\Omega})}\right). \tag{4.16}$$

Proof Recall that $c_0 = (1 + c_*)^{-1}, c_* = \frac{1}{4}$, so that $1/c_0 = \frac{5}{4}$, and combine (4.14) and (4.15). □

According to this result, the five-point difference scheme (4.2) for the boundary value problem (4.1) is second-order convergent, provided that u is sufficiently smooth.

In general, however, even if f and c are smooth functions, the corresponding solution, u, of (4.1) will not be a smooth function because the boundary, Γ, of the domain, Ω, is a non-smooth curve. Thus, the hypothesis $u \in C^4(\bar{\Omega})$ is unrealistic.

Our analysis has another limitation: it has been performed under the assumption that $f \in C(\bar{\Omega})$ which was required in order to ensure that the values of f are well defined at the mesh-points. However, in physical applications one often has to consider differential equations with f discontinuous (e.g. piecewise continuous), or, more generally, $f \in L^2(\Omega)$. We know that in this case Theorem 2.3 still implies that the problem has a unique weak solution, so it is natural to ask whether one can construct an accurate finite difference approximation of the weak solution. This brings us to case (b), formulated on page 26.

(b) ($f \in L^2(\Omega)$). We retain the same finite difference mesh as in case (a), but we modify the difference scheme (4.3) to cater for the fact that f is not necessarily continuous on $\bar{\Omega}$.

The idea is to replace $f(x_i, y_j)$ in (4.3) by a cell-average of f,

$$Tf_{ij} = \frac{1}{h^2} \int_{K_{ij}} f(x, y) \, \mathrm{d}x \, \mathrm{d}y,$$

where

$$K_{ij} = \left[x_i - \frac{h}{2}, x_i + \frac{h}{2} \right] \times \left[y_j - \frac{h}{2}, y_j + \frac{h}{2} \right].$$

This, seemingly ad hoc approach, has the following justification. Integrating the partial differential equation $-\Delta u + cu = f$ over the cell K_{ij}, and using Gauss' theorem, we have

$$-\int_{\partial K_{ij}} \frac{\partial u}{\partial \nu} \, \mathrm{d}l + \int_{K_{ij}} cu \, \mathrm{d}x \, \mathrm{d}y = \int_{K_{ij}} f \, \mathrm{d}x \, \mathrm{d}y \qquad (**)$$

where ∂K_{ij} is the boundary of K_{ij}, and ν the unit outward normal to ∂K_{ij}. The normal vectors to ∂K_{ij} point in the coordinate directions, so the normal derivative $\partial u/\partial \nu$ can be approximated by divided differences using the values of u at the five mesh-points marked "•" on Fig. 3. Approximating the second integral on the left by mid-point quadrature, and dividing both sides by $\mathrm{meas}(K_{ij}) = h^2$, we obtain

$$-(D_x^+ D_x^- u(x_i, y_j) + D_y^+ D_y^- u(x_i, y_j)) + c(x_i, y_j) u(x_i, y_j) \approx \frac{1}{h^2} \int_{K_{ij}} f(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$

REMARK Finite difference schemes which arise from integral formulations of a differential equation, such as (**), are called finite volume methods. ◇

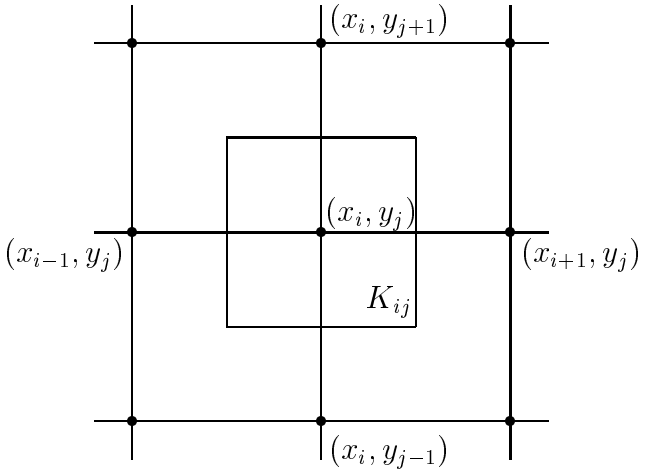


Figure 3: The cell K_{ij}

Clearly, Tf_{ij} is well defined for f in $L^2(\Omega)$ (and, in fact, even for $f \in L^1(\Omega)$); this follows by noting that

$$\begin{aligned} |Tf_{ij}| &= \frac{1}{h^2} \left| \int_{K_{ij}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \right| \\ &\le \frac{1}{h^2} \left(\int_{K_{ij}} 1^2\,\mathrm{d}x\,\mathrm{d}y \right)^{1/2} \left(\int_{K_{ij}} |f(x,y)|^2\,\mathrm{d}x\,\mathrm{d}y \right)^{1/2} \\ &= \frac{1}{h} \|f\|_{L^2(K_{ij})}\,, \end{aligned} \tag{4.17}$$

which, in turn, is bounded by $h^{-1}\|f\|_{L^2(\Omega)}$. Thus we define our finite difference (or, more precisely, finite volume) approximation of (4.1) by

$$-(D_x^+ D_x^- U_{ij} + D_y^+ D_y^- U_{ij}) + c(x_i, y_j) U_{ij} = Tf_{ij}, \quad (x_i, y_j) \in \Omega_h, \tag{4.18a}$$

$$U = 0 \qquad \text{on } \Gamma_h. \tag{4.18b}$$

Since we have not changed the difference operator on the left-hand side, the argument presented on page 28 still applies, and therefore (4.18) has a unique solution, U.

Theorem 4.5 *The scheme (4.18) is stable in the sense that*

$$\|U\|_{1,h} \le \frac{1}{c_0} \|f\|_{L^2(\Omega)}\,. \tag{4.19}$$

Proof According to (4.12) and (4.17),

$$\begin{aligned} c_0 \|U\|_{1,h}^2 &\leq (AU, U)_h = (Tf, U)_h \\ &\leq \|Tf\|_h \|U\|_h \leq \|Tf\|_h \|U\|_{1,h} \\ &\leq \|f\|_{L^2(\Omega)} \|U\|_{1,h}, \end{aligned}$$

and hence (4.19). □

Having established the stability of the scheme (4.18), we consider the question of its accuracy. Let us define the global error, e, as before,

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

Clearly,

$$\begin{aligned} Ae_{ij} &= Au(x_i, y_j) - AU_{ij} \\ &= Au(x_i, y_j) - Tf_{ij} \\ &= -(D_x^+ D_x^- u(x_i, y_j) + D_y^+ D_y^- u(x_i, y_j)) + c(x_i, y_j)u(x_i, y_j) \\ &\quad + \left(T\left(\frac{\partial^2 u}{\partial x^2}\right)(x_i, y_j) + T\left(\frac{\partial^2 u}{\partial y^2}\right)(x_i, y_j) - T(cu)(x_i, y_j) \right). \end{aligned} \tag{4.20}$$

Noting that

$$\begin{aligned} T\left(\frac{\partial^2 u}{\partial x^2}\right)(x_i, y_j) &= \frac{1}{h}\int_{y_j - h/2}^{y_j + h/2} \frac{\frac{\partial u}{\partial x}(x_i + h/2, y) - \frac{\partial u}{\partial x}(x_i - h/2, y)}{h} \, \mathrm{d}y \\ &= \frac{1}{h}\int_{y_j - h/2}^{y_j + h/2} D_x^+ \frac{\partial u}{\partial x}(x_i - h/2, y) \, \mathrm{d}y \\ &= D_x^+ \left[\frac{1}{h}\int_{y_j - h/2}^{y_j + h/2} \frac{\partial u}{\partial x}(x_i - h/2, y) \, \mathrm{d}y \right], \end{aligned}$$

and similarly,

$$T\left(\frac{\partial^2 u}{\partial y^2}\right)(x_i, y_j) = D_y^+ \left[\frac{1}{h}\int_{x_i - h/2}^{x_i + h/2} \frac{\partial u}{\partial y}(x, y_j - h/2) \, \mathrm{d}x \right],$$

(4.20) can be rewritten as

$$Ae = D_x^+ \varphi_1 + D_y^+ \varphi_2 + \psi,$$

where

$$\begin{aligned} \varphi_1(x_i, y_j) &= \frac{1}{h}\int_{y_j - h/2}^{y_j + h/2} \frac{\partial u}{\partial x}(x_i - h/2, y) \, \mathrm{d}y - D_x^- u(x_i, y_j), \\ \varphi_2(x_i, y_j) &= \frac{1}{h}\int_{x_i - h/2}^{x_i + h/2} \frac{\partial u}{\partial y}(x, y_j - h/2) \, \mathrm{d}x - D_y^- u(x_i, y_j), \\ \psi(x_i, y_j) &= (cu)(x_i, y_j) - T(cu)(x_i, y_j). \end{aligned}$$

Thus,

$$Ae = D_x^+ \varphi_1 + D_y^+ \varphi_2 + \psi \qquad \text{in } \Omega_h, \tag{4.21a}$$

$$e = 0 \qquad \text{on } \Gamma_h. \tag{4.21b}$$

As the stability of the difference scheme would only imply the crude bound

$$\|e\|_{1,h} \le \frac{1}{c_0} \left\| D_x^+ \varphi_1 + D_y^+ \varphi_2 + \psi \right\|_h$$

which makes no use of the special form of the truncation error

$$\varphi = D_x^+ \varphi_1 + D_y^+ \varphi_2 + \psi,$$

we shall proceed in a different way. According to (4.12),

$$\begin{aligned} c_0 \|e\|_{1,h}^2 &\le (Ae, e)_h \\ &= (D_x^+ \varphi_1, e)_h + (D_y^+ \varphi_2, e)_h + (\psi, e)_h. \end{aligned} \tag{4.22}$$

Using summation by parts, we shall pass the difference operators D_x^+ and D_y^+ from φ_1 and φ_2, respectively, onto e. Recalling that $e = 0$ on Γ_h,

$$\begin{aligned} (D_x^+ \varphi_1, e)_h &= \sum_{j=1}^{N-1} h \left(\sum_{i=1}^{N-1} h \frac{\varphi_1(x_{i+1}, y_j) - \varphi_1(x_i, y_j)}{h} e_{ij} \right) \\ &= -\sum_{j=1}^{N-1} h \left(\sum_{i=1}^{N} h \varphi_1(x_i, y_j) \frac{e_{ij} - e_{i-1,j}}{h} \right) \\ &= -\sum_{j=1}^{N-1} h \left(\sum_{i=1}^{N} h \varphi_1(x_i, y_j) D_x^- e_{ij} \right) \\ &= -\sum_{i=1}^{N} \sum_{j=1}^{N-1} h^2 \varphi_1(x_i, y_j) D_x^- e_{ij} \\ &\le \left(\sum_{i=1}^{N} \sum_{j=1}^{N-1} h^2 |\varphi_1(x_i, y_j)|^2 \right)^{1/2} \left(\sum_{i=1}^{N} \sum_{j=1}^{N-1} h^2 |D_x^- e_{ij}|^2 \right)^{1/2} \\ &= \|\varphi_1\|_x \left\| D_x^- e \right\|_x . \end{aligned}$$

Thus,

$$(D_x^+ \varphi_1, e)_h \le \|\varphi_1\|_x \left\| D_x^- e \right\|_x . \tag{4.23}$$

Similarly,

$$(D_y^+ \varphi_2, e)_h \le \|\varphi_2\|_y \left\| D_y^- e \right\|_y \tag{4.24}$$

(see page 29 for the definition of the mesh-dependent norms $\|\cdot\|\!\|_x$ and $\|\cdot\|\!\|_y$.) By the Cauchy–Schwarz inequality we also have that

$$(\psi, e)_h \leq \|\psi\|_h \, \|e\|_h \, . \tag{4.25}$$

Upon substituting (4.23) – (4.25) into (4.22) we obtain

$$\begin{aligned} c_0 \, \|e\|_{1,h}^2 &\leq \|\varphi_1\|\!\|_x \left\|D_x^- e\right\|\!\|_x + \|\varphi_2\|\!\|_y \left\|D_y^- e\right\|\!\|_y + \|\psi\|_h \, \|e\|_h \\ &\leq \left(\|\varphi_1\|\!\|_x^2 + \|\varphi_2\|\!\|_y^2 + \|\psi\|_h^2 \right)^{1/2} \left(\left\|D_x^- e\right\|\!\|_x^2 + \left\|D_y^- e\right\|\!\|_y^2 + \|e\|_h^2 \right)^{1/2} \\ &= \left(\|\varphi_1\|\!\|_x^2 + \|\varphi_2\|\!\|_y^2 + \|\psi\|_h^2 \right)^{1/2} \|e\|_{1,h} \, . \end{aligned}$$

Dividing both sides by $\|e\|_{1,h}$ yields the following result.

Lemma 4.6 *The global error, e, of the finite difference scheme (4.18) satisfies*

$$\|e\|_{1,h} \leq \frac{1}{c_0} (\|\varphi_1\|\!\|_x^2 + \|\varphi_2\|\!\|_y^2 + \|\psi\|_h^2)^{1/2}, \tag{4.26}$$

where φ_1, φ_2, and ψ are defined by

$$\varphi_1(x_i, y_j) = \frac{1}{h} \int_{y_j - h/2}^{y_j + h/2} \frac{\partial u}{\partial x}(x_i - h/2, y) \, \mathrm{d}y - D_x^- u(x_i, y_j), \tag{4.27}$$

$$\varphi_2(x_i, y_j) = \frac{1}{h} \int_{x_i - h/2}^{x_i + h/2} \frac{\partial u}{\partial y}(x, y_j - h/2) \, \mathrm{d}x - D_y^- u(x_i, y_j), \tag{4.28}$$

$$\begin{aligned} \psi(x_i, y_j) = (cu)(x_i, y_j) - \frac{1}{h^2} \int_{x_i - h/2}^{x_i + h/2} \int_{y_j - h/2}^{y_j + h/2} (cu)(x, y) \, \mathrm{d}x \, \mathrm{d}y, \\ i = 1, \ldots, N-1, \quad j = 1, \ldots, N. \end{aligned} \tag{4.29}$$

To complete the error analysis, it remains to estimate φ_1, φ_2 and ψ. Using Taylor series expansions it is easily seen that

$$|\varphi_1(x_i, y_j)| \leq \frac{h^2}{24} \left(\left\| \frac{\partial^3 u}{\partial x \partial y^2} \right\|_{C(\bar{\Omega})} + \left\| \frac{\partial^3 u}{\partial x^3} \right\|_{C(\bar{\Omega})} \right), \tag{4.30}$$

$$|\varphi_2(x_i, y_j)| \leq \frac{h^2}{24} \left(\left\| \frac{\partial^3 u}{\partial x^2 \partial y} \right\|_{C(\bar{\Omega})} + \left\| \frac{\partial^3 u}{\partial y^3} \right\|_{C(\bar{\Omega})} \right), \tag{4.31}$$

$$|\psi(x_i, y_j)| \leq \frac{h^2}{24} \left(\left\| \frac{\partial^2 (cu)}{\partial x^2} \right\|_{C(\bar{\Omega})} + \left\| \frac{\partial^2 (cu)}{\partial y^2} \right\|_{C(\bar{\Omega})} \right), \tag{4.32}$$

and hence the bounds for $\|\varphi_1\|\!\|_x$, $\|\varphi_2\|\!\|_y$ and $\|\psi\|\!\|_h$. We have the following theorem.

Theorem 4.7 *Let $f \in L^2(\Omega)$, $c \in C^2(\bar{\Omega})$ with $c(x,y) \geq 0$, $(x,y) \in \bar{\Omega}$, and suppose that the corresponding weak solution of the boundary value problem (4.1) belongs to $C^3(\bar{\Omega})$. Then,*

$$\|u - U\|_{1,h} \leq \frac{5}{96} h^2 M_3, \tag{4.33}$$

where

$$\begin{aligned} M_3 = \Bigg\{ & \left(\left\| \frac{\partial^3 u}{\partial x \partial y^2} \right\|_{C(\bar{\Omega})} + \left\| \frac{\partial^3 u}{\partial x^3} \right\|_{C(\bar{\Omega})} \right)^2 \\ & + \left(\left\| \frac{\partial^3 u}{\partial x^2 y} \right\|_{C(\bar{\Omega})} + \left\| \frac{\partial^3 u}{\partial y^3} \right\|_{C(\bar{\Omega})} \right)^2 \\ & + \left(\left\| \frac{\partial^2 (cu)}{\partial x^2} \right\|_{C(\bar{\Omega})} + \left\| \frac{\partial^2 (cu)}{\partial y^2} \right\|_{C(\bar{\Omega})} \right)^2 \Bigg\}^{1/2}. \end{aligned}$$

Proof Recalling that $1/c_0 = 5/4$ and substituting (4.30) - (4.32) into the right-hand side of (4.26), (4.33) immediately follows. □

Comparing (4.33) with (4.16), we see that while the smoothness requirement on the solution has been relaxed from $u \in C^4(\bar{\Omega})$ to $u \in C^3(\bar{\Omega})$, second-order convergence has been retained.

The hypothesis $u \in C^3(\bar{\Omega})$ can be further relaxed by using integral representations of φ_1, φ_2 and ψ instead of Taylor series expansions. The key idea is to use the Newton-Leibniz formula

$$w(b) - w(a) = \int_a^b w'(x)\,\mathrm{d}x.$$

Thus, denoting $x_{i\pm 1/2} = x_i \pm h/2$ and $y_{j\pm 1/2} = y_j \pm h/2$, we have

$$\begin{aligned} \varphi_1(x_i, y_j) &= \frac{1}{h^2} \int_{x_{i-1}}^{x_i} \int_{y_{j-1/2}}^{y_{j+1/2}} \left[\frac{\partial u}{\partial x}(x_{i-1/2}, y) - \frac{\partial u}{\partial x}(x, y_j) \right] \mathrm{d}x\,\mathrm{d}y \\ &= \frac{1}{h^2} \int_{x_{i-1}}^{x_i} \int_{y_{j-1/2}}^{y_{j+1/2}} \left[\frac{\partial u}{\partial x}(x_{i-1/2}, y) - \frac{\partial u}{\partial x}(x, y) \right] \mathrm{d}x\,\mathrm{d}y \\ &\quad + \frac{1}{h^2} \int_{x_{i-1}}^{x_i} \int_{y_{j-1/2}}^{y_{j+1/2}} \left[\frac{\partial u}{\partial x}(x, y) - \frac{\partial u}{\partial x}(x, y_j) \right] \mathrm{d}x\,\mathrm{d}y \\ &= \frac{1}{h^2} \int_{y_{j-1/2}}^{y_{j+1/2}} \left[\int_{x_{i-1}}^{x_i} 1 \cdot \int_x^{x_{i-1/2}} \frac{\partial^2 u}{\partial x^2}(\xi, y)\,\mathrm{d}\xi \right] \mathrm{d}x\,\mathrm{d}y \\ &\quad + \frac{1}{h^2} \int_{x_{i-1}}^{x_i} \left[\int_{y_{j-1/2}}^{y_{j+1/2}} 1 \cdot \int_{y_j}^{y} \frac{\partial^2 u}{\partial x \partial y}(x, \eta)\,\mathrm{d}\eta \right] \mathrm{d}x\,\mathrm{d}y \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{h^2}\int_{y_{j-1/2}}^{y_{j+1/2}} \left[x \int_{x}^{x_{i-1/2}} \frac{\partial^2 u}{\partial x^2}(\xi, y)\,\mathrm{d}\xi \Big|_{x_{i-1}}^{x_i} + \int_{x_{i-1}}^{x_i} x \frac{\partial^2 u}{\partial x^2}(x, y)\,\mathrm{d}x \right] \mathrm{d}y \\
&\quad + \frac{1}{h^2}\int_{x_{i-1}}^{x_i} \left[y \int_{y_j}^{y} \frac{\partial^2 u}{\partial x \partial y}(x, \eta)\,\mathrm{d}\eta \Big|_{y_{j-1/2}}^{y_{j+1/2}} - \int_{y_{j-1/2}}^{y_{j+1/2}} y \frac{\partial^2 u}{\partial x \partial y}(x, y)\,\mathrm{d}y \right] \mathrm{d}x \\
&= \frac{1}{h^2}\int_{y_{j-1/2}}^{y_{j+1/2}} \left[\int_{x_{i-1}}^{x_{i-1/2}} (x - x_{i-1}) \frac{\partial^2 u}{\partial x^2}(x, y)\,\mathrm{d}x + \int_{x_{i-1/2}}^{x_i} (x - x_i) \frac{\partial^2 u}{\partial x^2}(x, y)\,\mathrm{d}x \right] \mathrm{d}y \\
&\quad - \frac{1}{h^2}\int_{x_{i-1}}^{x_i} \left[\int_{y_{j-1/2}}^{y_j} (y - y_{j-1/2}) \frac{\partial^2 u}{\partial x \partial y}(x, y)\,\mathrm{d}y + \int_{y_j}^{y_{j+1/2}} (y - y_{j+1/2}) \frac{\partial^2 u}{\partial x \partial y}(x, y)\,\mathrm{d}y \right] \mathrm{d}x.
\end{aligned}
$$

We define the functions

$$
A(x) = \begin{cases} \frac{1}{2}(x - x_{i-1})^2, & x \in [x_{i-1}, x_{i-1/2}], \\ \frac{1}{2}(x - x_i)^2, & x \in [x_{i-1/2}, x_i], \end{cases}
$$

$$
B(y) = \begin{cases} \frac{1}{2}(y - y_{j-1/2})^2, & y \in [y_{j-1/2}, y_j], \\ \frac{1}{2}(y - y_{j+1/2})^2, & y \in [y_j, y_{j+1/2}]. \end{cases}
$$

Note that A and B are continuous functions, $A(x_{i-1}) = A(x_i) = 0$, and $B(y_{j-1/2}) = B(y_{j+1/2}) = 0$. Thus, upon integration by parts,

$$
\begin{aligned}
\varphi_1(x_i, y_j) &= \frac{1}{h^2}\int_{y_{j-1/2}}^{y_{j+1/2}} \left[\int_{x_{i-1}}^{x_i} A'(x) \frac{\partial^2 u}{\partial x^2}(x, y)\,\mathrm{d}x \right] \mathrm{d}y \\
&\quad - \frac{1}{h^2}\int_{x_{i-1}}^{x_i} \left[\int_{y_{j-1/2}}^{y_{j+1/2}} B'(y) \frac{\partial^2 u}{\partial x \partial y}(x, y)\,\mathrm{d}y \right] \mathrm{d}x \\
&= -\frac{1}{h^2}\int_{y_{j-1/2}}^{y_{j+1/2}} \left[\int_{x_{i-1}}^{x_i} A(x) \frac{\partial^3 u}{\partial x^3}(x, y)\,\mathrm{d}x \right] \mathrm{d}y \\
&\quad + \frac{1}{h^2}\int_{x_{i-1}}^{x_i} \left[\int_{y_{j-1/2}}^{y_{j+1/2}} B(y) \frac{\partial^3 u}{\partial x \partial y^2}(x, y)\,\mathrm{d}y \right] \mathrm{d}x.
\end{aligned}
$$

But

$$
|A(x)| \le \frac{h^2}{8}, \quad |B(y)| \le \frac{h^2}{8},
$$

and therefore,

$$
\begin{aligned}
|\varphi_1(x_i, y_j)| &\le \frac{1}{8}\int_{x_{i-1}}^{x_i}\int_{y_{j-1/2}}^{y_{j+1/2}} \left| \frac{\partial^3 u}{\partial x^3}(x, y) \right| \mathrm{d}x\,\mathrm{d}y \\
&\quad + \frac{1}{8}\int_{x_{i-1}}^{x_i}\int_{y_{j-1/2}}^{y_{j+1/2}} \left| \frac{\partial^3 u}{\partial x \partial y^2}(x, y) \right| \mathrm{d}x\,\mathrm{d}y.
\end{aligned}
$$

Consequently,

$$\|\varphi_1\|_x^2 \le \frac{h^4}{32}\left(\left\|\frac{\partial^3 u}{\partial x^3}\right\|_{L_2(\Omega)}^2 + \left\|\frac{\partial^3 u}{\partial x \partial y^2}\right\|_{L_2(\Omega)}^2\right). \tag{4.34}$$

Analogously,

$$\|\varphi_2\|_y^2 \le \frac{h^4}{32}\left(\left\|\frac{\partial^3 u}{\partial y^3}\right\|_{L_2(\Omega)}^2 + \left\|\frac{\partial^3 u}{\partial x^2 \partial y}\right\|_{L_2(\Omega)}^2\right). \tag{4.35}$$

In order to estimate ψ, we note that

$$\begin{aligned}
\psi(x_i, y_j) = \frac{1}{h^2}\int_{x_{i-1/2}}^{x_{i+1/2}}\int_{y_{j-1/2}}^{y_{j+1/2}} \Bigg(\int_x^{x_i}\frac{\partial w}{\partial x}(s,y)\,\mathrm{d}s + \\
+ \int_y^{y_j}\frac{\partial w}{\partial y}(x,t)\,\mathrm{d}t + \int_x^{x_i}\int_y^{y_j}\frac{\partial^2 w}{\partial x \partial y}(s,t)\,\mathrm{d}s\,\mathrm{d}t\Bigg)\,\mathrm{d}x\,\mathrm{d}y \\
= -\frac{1}{h^2}\int_{x_{i-1/2}}^{x_{i+1/2}}\int_{y_{j-1/2}}^{y_{j+1/2}} C(x)\frac{\partial^2 w}{\partial x^2}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
-\frac{1}{h^2}\int_{x_{i-1/2}}^{x_{i+1/2}}\int_{y_{j-1/2}}^{y_{j+1/2}} D(y)\frac{\partial^2 w}{\partial y^2}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
+\frac{1}{h^2}\int_{x_{i-1/2}}^{x_{i+1/2}}\int_{y_{j-1/2}}^{y_{j+1/2}} \left(\int_x^{x_i}\int_y^{y_j}\frac{\partial^2 w}{\partial x \partial y}(s,t)\,\mathrm{d}s\,\mathrm{d}t\right)\,\mathrm{d}x\,\mathrm{d}y,
\end{aligned}$$

where $w(x,y) = c(x,y)u(x,y)$,

$$C(x) = \begin{cases} \frac{1}{2}(x - x_{i-1/2})^2, & x \in [x_{i-1/2}, x_i], \\ \frac{1}{2}(x - x_{i+1/2})^2, & x \in [x_i, x_{i+1/2}], \end{cases}$$

and

$$D(y) = \begin{cases} \frac{1}{2}(y - y_{j-1/2})^2, & y \in [y_{j-1/2}, y_j], \\ \frac{1}{2}(y - y_{j+1/2})^2, & y \in [y_j, y_{j+1/2}]. \end{cases}$$

Thence,

$$\begin{aligned}
|\psi(x_i, y_j)| \le \frac{1}{8}\Bigg(\int_{x_{i-1/2}}^{x_{i+1/2}}\int_{y_{j-1/2}}^{y_{j+1/2}}\left|\frac{\partial^2 w}{\partial x^2}(x,y)\right|\,\mathrm{d}x\,\mathrm{d}y \\
+ \int_{x_{i-1/2}}^{x_{i+1/2}}\int_{y_{j-1/2}}^{y_{j+1/2}}\left|\frac{\partial^2 w}{\partial y^2}(x,y)\right|\,\mathrm{d}x\,\mathrm{d}y \\
+ 2\int_{x_{i-1/2}}^{x_{i+1/2}}\int_{y_{j-1/2}}^{y_{j+1/2}}\left|\frac{\partial^2 w}{\partial x \partial y}\right|\,\mathrm{d}x\,\mathrm{d}y\Bigg),
\end{aligned}$$

so that, with $w = cu$, we have

$$\|\psi\|_h^2 \leq \frac{3h^4}{64}\left(\left\|\frac{\partial^2 w}{\partial x^2}\right\|_{L_2(\Omega)}^2 + \left\|\frac{\partial^2 w}{\partial y^2}\right\|_{L_2(\Omega)}^2 + 4\left\|\frac{\partial^2 w}{\partial x \partial y}\right\|_{L_2(\Omega)}^2\right). \tag{4.36}$$

Substituting (4.34)–(4.36) into the right-hand side of (4.26) and recalling that $1/c_0 = 4/5$, we obtain the following result.

Theorem 4.8 *Let $f \in L_2(\Omega)$, $c \in C^2(\bar{\Omega})$, with $c(x,y) \geq 0$, $(x,y) \in \bar{\Omega}$, and suppose that the corresponding weak solution of the boundary value problem (4.1) belongs to $H^3(\Omega)$. Then,*

$$\|u - U\|_{1,h} \leq Ch^2 \|u\|_{H^3(\Omega)}, \tag{4.37}$$

where C is a positive constant (computable from (4.34)–(4.36)).

It can be shown that the error estimate (4.37) is best possible in the sense that further relaxation of the regularity hypothesis on u leads to a loss of second-order convergence. Error estimates of this type, where the highest possible accuracy has been attained with the minimum hypotheses on the smoothness of the solution are called optimal error estimates. Thus, for example, (4.37) is an optimal error estimate for the difference scheme (4.18), but (4.33) is not.

We have used integral representations of differences to show the bounds (4.34)–(4.36). Alternatively one can use the following abstract device.

Lemma 4.9 *(The Bramble-Hilbert Lemma) Suppose $\Phi : H^k(\Omega) \to \mathbb{R}$ is a linear form, i.e. for all $u, v \in H^k(\Omega)$, and all $\alpha, \beta \in \mathbb{R}$,*

$$\Phi(\alpha u + \beta v) = \alpha\Phi(u) + \beta\Phi(v),$$

and assume that:

(a) $\Phi(p) = 0$ for every polynomial p of degree $\leq k-1$, and

(b) there exists a positive constant C such that

$$|\Phi(u)| \leq C\|u\|_{H^k(\Omega)} \quad \forall u \in H^k(\Omega).$$

Then, there exists a constant $C_1 = C_1(\Omega, C, k)$ such that

$$|\Phi(u)| \leq C_1 |u|_{H^k(\Omega)} \quad \forall u \in H^k(\Omega).$$

Proof See P. Ciarlet: The Finite Element Method for Elliptic Problems, North-Holland, 1979.

We shall use the Bramble-Hilbert lemma to re-derive the bound (4.34) for φ_1. Let $K = [-1/2, 1/2] \times [-1/2, 1/2]$, and consider the affine mapping

$$\begin{cases} x = x_i - h/2 + sh, & -1/2 \le s \le 1/2, \\ y = y_j + th, & -1/2 \le t \le 1/2, \end{cases}$$

of K onto $K_{ij}^- = [x_{i-1}, x_i] \times [y_{j-1/2}, y_{j+1/2}]$. We define

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

In terms of $\bar{u}$, φ_1 can be rewritten as follows:

$$\varphi_1(x_i, y_j) = \frac{1}{h}\Phi(\bar{u}),$$

where

$$\Phi(\bar{u}) = \int_{-1/2}^{1/2} \frac{\partial \bar{u}}{\partial s}(0,t)\,\mathrm{d}t - \{\bar{u}(\tfrac{1}{2}, 0) - \bar{u}(-\tfrac{1}{2}, 0)\}.$$

Clearly $\Phi : \bar{u} \mapsto \Phi(\bar{u})$ is a linear form, and $\Phi(p) = 0$ for every polynomial of the form

$$p = a_0 + a_1 s + a_2 t + a_3 s^2 + a_4 st + a_5 t^2$$

(i.e. $\Phi(p) = 0$ if p is a polynomial of degree ≤ 2). In addition,

$$|\Phi(\bar{u})| \le \int_{-1/2}^{1/2} \left| \frac{\partial \bar{u}}{\partial s}(0,t) \right| \mathrm{d}t + 2 \max_{(s,t)\in K} |\bar{u}(s,t)|\,. \tag{4.38}$$

Lemma 4.10 *Let $v \in H^2(K)$; then,*

(a) $\displaystyle\int_{-1/2}^{1/2} \left| \frac{\partial v}{\partial s}(0,t) \right| \mathrm{d}t \le \sqrt{2}\,\|v\|_{H^2(K)}\,,$

(b) $\displaystyle\max_{(s,t)\in K} |v(s,t)| \le 2\,\|v\|_{H^2(K)}\,.$

Proof

(a) Note that, for any $s \in [-1/2, 1/2]$,

$$\left| \frac{\partial v}{\partial s}(0,t) \right| \le \left| \frac{\partial v}{\partial s}(s,t) \right| + \left| \int_s^0 \frac{\partial^2 v}{\partial s^2}(\sigma,t)\,\mathrm{d}\sigma \right|.$$

Thus,

$$\left|\frac{\partial v}{\partial s}(0,t)\right| \le \left|\frac{\partial v}{\partial s}(s,t)\right| + \int_{-1/2}^{1/2} \left|\frac{\partial^2 v}{\partial s^2}(\sigma,t)\right| \,\mathrm{d}\sigma.$$

Integrating both sides in s and t,

$$\begin{aligned}
\int_{-1/2}^{1/2} \left|\frac{\partial v}{\partial s}(0,t)\right| \,\mathrm{d}t &\le \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial v}{\partial s}(s,t)\right| \,\mathrm{d}s\,\mathrm{d}t + \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial^2 v}{\partial s^2}(\sigma,t)\right| \,\mathrm{d}\sigma\,\mathrm{d}t, \\
&\le \left(\int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial v}{\partial s}(s,t)\right|^2 \,\mathrm{d}s\,\mathrm{d}t\right)^{1/2} + \left(\int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial^2 v}{\partial s^2}(\sigma,t)\right|^2 \,\mathrm{d}\sigma\,\mathrm{d}t\right)^{1/2} \\
&= \left\|\frac{\partial v}{\partial s}\right\|_{L^2(K)} + \left\|\frac{\partial^2 v}{\partial s^2}\right\|_{L^2(K)}.
\end{aligned}$$

Finally, using the inequality

$$a + b \le \sqrt{2}(a^2 + b^2)^{1/2}, \quad a, b \ge 0,$$

and the definition of $\|\cdot\|_{H^2(K)}$, we get (a).

(b) Let $(x,y) \in K$ and $(s,t) \in K$. Then,

$$\begin{aligned}
v(x,y) = v(s,t) &+ \int_s^x \frac{\partial v}{\partial s}(\sigma,t)\,\mathrm{d}\sigma + \int_t^y \frac{\partial v}{\partial t}(s,\tau)\,\mathrm{d}\tau \\
&+ \int_s^x \int_t^y \frac{\partial^2 v}{\partial s \partial t}(\sigma,\tau)\,\mathrm{d}\sigma\,\mathrm{d}\tau,
\end{aligned}$$

and therefore

$$\begin{aligned}
|v(x,y)| \le |v(s,t)| &+ \int_{-1/2}^{1/2} \left|\frac{\partial v}{\partial s}(\sigma,t)\right| \,\mathrm{d}\sigma + \int_{-1/2}^{1/2} \left|\frac{\partial v}{\partial t}(s,\tau)\right| \,\mathrm{d}\tau \\
&+ \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial^2 v}{\partial s \partial t}(\sigma,\tau)\right| \,\mathrm{d}\sigma\,\mathrm{d}\tau.
\end{aligned}$$

Integrating both sides in s and t, we obtain

$$\begin{aligned}
|v(x,y)| &= \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} |v(s,t)| \,\mathrm{d}s\,\mathrm{d}t + \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial v}{\partial s}(\sigma,t)\right| \,\mathrm{d}\sigma\,\mathrm{d}t \\
&+ \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial v}{\partial t}(s,\tau)\right| \,\mathrm{d}s\,\mathrm{d}\tau + \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial^2 v}{\partial s \partial t}(\sigma,\tau)\right| \,\mathrm{d}\sigma\,\mathrm{d}\tau \\
&\le \|v\|_{L^2(K)} + \left\|\frac{\partial v}{\partial s}\right\|_{L^2(K)} + \left\|\frac{\partial v}{\partial t}\right\|_{L^2(K)} + \left\|\frac{\partial^2 v}{\partial s \partial t}\right\|_{L^2(K)} \\
&\le 2\,\|v\|_{H^2(K)} \quad \forall (x,y) \in K.
\end{aligned}$$

Taking the maximum over all (x,y) in K, we obtain (b). □

Equipped with the inequalities (a) and (b), we return to (4.38). It follows that

$$|\Phi(\bar{u})| \leq (\sqrt{2}+4)\,\|\bar{u}\|_{H^2(K)}\,.$$

Since $\|\bar{u}\|_{H^2(K)} \leq \|\bar{u}\|_{H^3(K)}$, we also have

$$|\Phi(\bar{u})| \leq (\sqrt{2}+4)\,\|\bar{u}\|_{H^3(K)}\,.$$

Thus we have shown that the mapping Φ satisfies the hypotheses of the Bramble-Hilbert lemma with $k=3$ and $\Omega = K$.

Hence, there exists a constant C_1 such that

$$|\Phi(\bar{u})| \leq C_1\,|\bar{u}|_{H^3(K)} \qquad \forall \bar{u} \in H^3(K).$$

Returning from $(s,t) \in K$ to our original variables $(x,y) \in K_{ij}^-$, we deduce that

$$|\Phi(\bar{u})| \leq C_1 h^{3-1}\,|u|_{H^3(K_{ij}^-)}\,,$$

and therefore,

$$|\varphi_1(x_i,y_j)| = \frac{1}{h}\,|\Phi(\bar{u})| \leq C_1 h\,|u|_{H^3(K_{ij}^-)}\,.$$

Consequently,

$$\begin{aligned}
\|\varphi_1\|_x^2 &= \sum_{i=1}^{N}\sum_{j=1}^{N-1} h^2\,|\varphi_1(x_i,y_j)|^2 \\
&\leq C_1^2 h^4 \sum_{i=1}^{N}\sum_{j=1}^{N-1} |u|^2_{H^3(K_{ij}^-)} \\
&\leq C_1^2 h^4\,|u|^2_{H^3(\Omega)}\,.
\end{aligned}$$

Therefore,

$$\|\varphi_1\|_x \leq C_1 h^2\,|u|_{H^3(\Omega)}\,. \tag{4.39}$$

Similarly,

$$\|\varphi_2\|_y \leq C_2 h^2\,|u|_{H^3(\Omega)} \tag{4.40}$$

and

$$\|\psi\|_h \leq C_3 h^2\,|u|_{H^2(\Omega)}\,. \tag{4.41}$$

The bounds (4.39)–(4.41) derived by using the Bramble-Hilbert lemma are essentially the same as those obtained earlier by integral representations, and stated in (4.34)–(4.36). There

is, however, an important practical difference: while the constants involved in (4.34)–(4.36) are known, those which appear in (4.39)–(4.41) (namely, C_1, C_2, C_3) are unknown because the Bramble-Hilbert lemma does not tell us what these are, so the constant in the resulting error estimate is not computable. We note, however, that in recent years several constructive proofs of the Bramble-Hilbert lemma have been derived for restricted classes of Ω. (e.g. Ω convex or star-shaped). These constructive proofs give an explicit expression for C_1 (see the statement of the Bramble-Hilbert lemma) in terms of C, k and the area (volume) of Ω.

Concluding remarks. We have carried out an error analysis of finite difference schemes for the partial differential equation

$$-\Delta u + c(x,y)u = f(x,y)$$

on a square domain Ω. The error analysis of difference schemes for more general elliptic equations would proceed along similar lines. Consider, for example,

$$\begin{aligned} &- \left[\frac{\partial}{\partial x}\left(a_1(x,y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(a_2(x,y)\frac{\partial u}{\partial y}\right)\right] \\ &+ b_1(x,y)\frac{\partial u}{\partial x} + b_2(x,y)\frac{\partial u}{\partial y} + c(x,y)u = f(x,y) \end{aligned}$$

on the unit square Ω in $\mathbb{R}^2$. We approximate the equation by

$$\begin{aligned} &- \frac{1}{h}\left[a_1(x_{i+1/2},y_j)\frac{U_{i+1,j}-U_{i,j}}{h} - a_1(x_{i-1/2},y_j)\frac{U_{i,j}-U_{i-1,j}}{h}\right] \\ &- \frac{1}{h}\left[a_2(x_i,y_{j+1/2})\frac{U_{i,j+1}-U_{i,j}}{h} - a_2(x_i,y_{j-1/2})\frac{U_{i,j}-U_{i,j-1}}{h}\right] \\ &+ b_1(x_i,y_j)\frac{U_{i+1,j}-U_{i-1,j}}{2h} + b_1(x_i,y_j)\frac{U_{i,j+1}-U_{i,j-1}}{2h} \\ &+ c(x_i,y_j)U_{ij} = \frac{1}{h^2}\int_{x_{i-1/2}}^{x_{i+1/2}}\int_{y_{i-1/2}}^{y_{i+1/2}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

This is still a five-point difference scheme. Provided $u \in H^3(\Omega) \cap H_0^1(\Omega)$, the scheme is second-order convergent in the $\|\cdot\|_{1,h}$ norm (i.e. (4.38) holds).

When Ω has a curved boundary, a non-uniform mesh has to be used near $\partial\Omega$ to avoid a loss of accuracy. To be more precise, let us introduce the following notation: let $h_{i+1} = x_{i+1} - x_i$, $h_i = x_i - x_{i-1}$, and let $\hbar_i = \frac{1}{2}(h_{i+1} + h_i)$. We define

$$\begin{aligned} D_x^+ U_i &= \frac{U_{i+1}-U_i}{\hbar_i}, \qquad D_x^- U_i = \frac{U_i - U_{i-1}}{h_i}, \\ D_x^+ D_x^- U_i &= \frac{1}{\hbar_i}\left(\frac{U_{i+1}-U_i}{h_{i+1}} - \frac{U_i - U_{i-1}}{h_i}\right). \end{aligned}$$

Similarly, let $k_{j+1} = y_{j+1} - y_j$, $k_j = y_j - y_{j-1}$, and let

$$\hbar_i = \frac{1}{2}(k_{j+1} + k_j).$$

Let

$$D_y^+ U_j = \frac{U_{j+1} - U_j}{k_j}, \quad D_y^- U_j = \frac{U_j - U_{j-1}}{k_j},$$
$$D_y^+ D_y^- U_j = \frac{1}{k_j} \left(\frac{U_{j+1} - U_j}{k_{j+1}} - \frac{U_j - U_{j-1}}{k_j} \right).$$

So, on a general non-uniform mesh

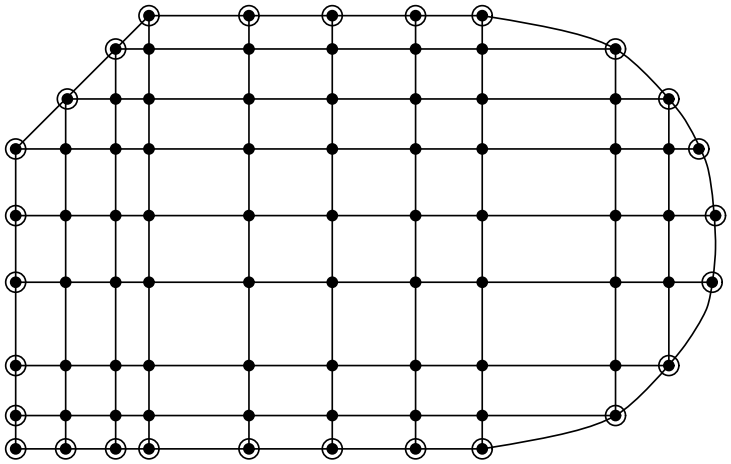
$$\bar{\Omega}_h = \{(x_i, y_j) : x_{i+1} - x_i = h_i,\ y_{j+1} - y_j = k_j\},$$

the Laplace operator, Δ, can be approximated by $D_x^+ D_x^- + D_y^+ D_y^-$, with the difference operators $D_x^+ D_x^-$, $D_y^+ D_y^-$ defined above.

Consider, for example, the Dirichlet problem

$$-\Delta u = f(x,y) \quad \text{in } \Omega,$$
$$u = 0 \quad \text{on } \partial\Omega,$$

where Ω and the non-uniform mesh $\bar{\Omega}_h$ are depicted in Fig. 4.



• Ω_h; $\odot$ Γ_h, $\bar{\Omega}_h = \Omega_h \cap \Gamma_h$.

Figure 4: Non-uniform mesh $\bar{\Omega}_h$.

The finite difference approximation of this boundary value problem is

$$-(D_x^+ D_x^- U_{ij} + D_y^+ D_y^- U_{ij}) = f(x_i, y_j) \quad \text{in } \Omega_h,$$
$$U_{ij} = 0 \quad \text{on } \Gamma_h.$$

Equivalently,

$$-\frac{1}{h_i} \left(\frac{U_{i+1,j} - U_{ij}}{h_{i+1}} - \frac{U_{ij} - U_{i-1,j}}{h_i} \right) - \frac{1}{k_j} \left(\frac{U_{i,j+1} - U_{ij}}{k_{j+1}} - \frac{U_{ij} - U_{i,j-1}}{k_j} \right) = f(x_i, y_j) \quad \text{in } \Omega_h,$$
$$U_{ij} = 0 \quad \text{on } \Gamma_h.$$

A typical difference stencil is shown in Fig. 5; clearly we still have a five-point difference scheme.

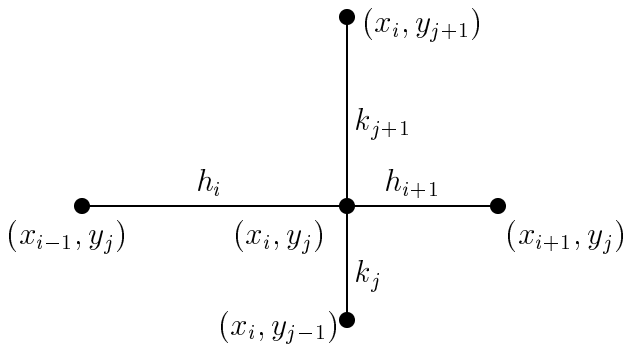


Figure 5: Five-point stencil on a non-uniform mesh.

5 Finite element methods for elliptic boundary value problems

In sections 3 and 4 we described the construction of finite difference methods for elliptic boundary value problems and outlined some simple techniques for their analysis. There, because of the very nature of finite difference schemes, the emphasis was placed on approximating the values of the exact solution at a finite number of mesh-points. In this section we concentrate on an alternative approach which is based on the approximation of the exact solution by continuous piecewise polynomial functions. Numerical methods of this type are called finite element methods.

Finite element methods were proposed by Courant in 1943, but the importance of his contribution was not recognised at the time and the idea was forgotten. The method was rediscovered by engineers in the early 1950's, though the mathematical analysis of finite element schemes only began in the 1960's, the first important theoretical results being those of Zlámal in 1968.

In this section we present some of the basic properties of finite element methods for elliptic boundary value problems. Unlike finite difference schemes which are constructed in a more-or-less ad hoc fashion by replacing the derivatives in the differential equation by divided differences, the derivation of finite element methods is much more systematic.

The first step in the construction of a finite element method for an elliptic boundary value problem (e.g. (2.3), (2.4)) is to convert the problem into its weak formulation:

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

where V is the solution space (e.g. $H_0^1(\Omega)$ for a homogeneous Dirichlet boundary value problem), $a(\cdot,\cdot)$ is a bilinear form on $V \times V$, and $l(\cdot)$ is a linear form on V (e.g. (2.7) and (2.8)).

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, and to consider the following approximation of (P):

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

Suppose, for example, that $\dim V_h = N(h)$ and $V_h = \text{span}\{\phi_1, \ldots, \phi_{N(h)}\}$, where the linearly independent basis functions ϕ_i, $i = 1, \ldots, N(h)$, have "small" support. Expressing the approximate solution u_h in terms of the basis functions, ϕ_i, we can write

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

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

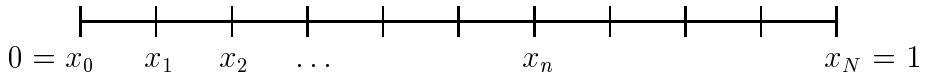


Figure 6: Subdivision of $\bar{\Omega} = [0, 1]$.

$$\text{find } (U_1, \ldots, U_{N(h)}) \in \mathbb{R}^{N(h)} \text{ such that } \sum_{i=1}^{N(h)} a(\phi_i, \phi_j) U_i = l(\phi_j),\ j = 1, \ldots, N(h). \qquad (P_h')$$

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 ϕ_i's have small support, $a(\phi_j, \phi_i) = 0$ for most i and j, so the matrix A is sparse. Once the system of linear equations (P_h') has been solved for $U = (U_1, \ldots, U_{N(h)})^T$, $(*)$ provides the required approximation of u.

After this brief outline of the finite element method, we illustrate the construction of this numerical technique through some simple examples.

5.1 Construction of the finite element method: piecewise linear basis functions

In this section we describe two specific examples of finite element methods for boundary value problems.

5.1.1 One-dimensional problem

Let us consider the boundary value problem

$$-(p(x)u')' + q(x)u = f(x), \quad x \in (0, 1), \qquad (5.1a)$$

$$u(0) = 0, \ \ u(1) = 0, \qquad (5.1b)$$

where $p \in C[0, 1]$, $q \in C[0, 1]$, $f \in L^2(0, 1)$, $p(x) \geq \tilde{c} > 0$, $q(x) \geq 0$, $x \in [0, 1]$. The weak formulation of this problem is:

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

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 $h = 1/N$, $N \geq 2$ (see Fig. 6).

The subintervals are called "elements". The solution, $u \in H_0^1(0,1)$, of (P) will be approximated by a continuous piecewise linear function on this subdivison. For this purpose we define the finite element basis functions

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

Here, for $z \in \mathbf{R}$, we used the notation $z_+ = \max\{0, z\}$. Clearly $\phi_i \in H_0^1(0,1)$, and $\operatorname{supp} \phi_i = [x_{i-1}, x_{i+1}]$, $i = 1, \ldots, N-1$. The functions ϕ_i, $i = 1, \ldots, N-1$, are linearly independent and therefore

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

is an $(N-1)$-dimensional subspace of $H_0^1(0,1)$. The finite element approximation of (P) is:

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

Since $u_h \in V_h = \operatorname{span}\{\phi_1, \ldots, \phi_{N-1}\}$, 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 into (P_h) we obtain the following problem, equivalent to (P_h):

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

Letting

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

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

(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. $x^T A x > 0$, $x \neq 0$). Since $\operatorname{supp}\phi_i \cup \operatorname{supp}\phi_j$ has empty interior when $|i-j| > 1$, it follows that the matrix A is tri-diagonal. Having solved the system of linear equations $AU = F$, we substitute the values $U_1, \ldots, U_{N-1}$ into

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

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 quadrature rules. In the simple case when p and q are constant functions on $[0,1]$, the entries of A can be calculated exactly:

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

5.1.2 Two-dimensional problem

Let Ω be a bounded domain in $\mathbb{R}^2$ with a polygonal boundary $\partial\Omega$, so that Ω can be exactly covered by a finite number of triangles. We shall suppose that a family of such sets of triangles is parametrised by h, where h is the maximum diameter of triangles in the set. We shall assume that any pair of triangles in a triangulation of Ω intersect along a complete edge, at a vertex, or not at all, as shown in Fig. 7.

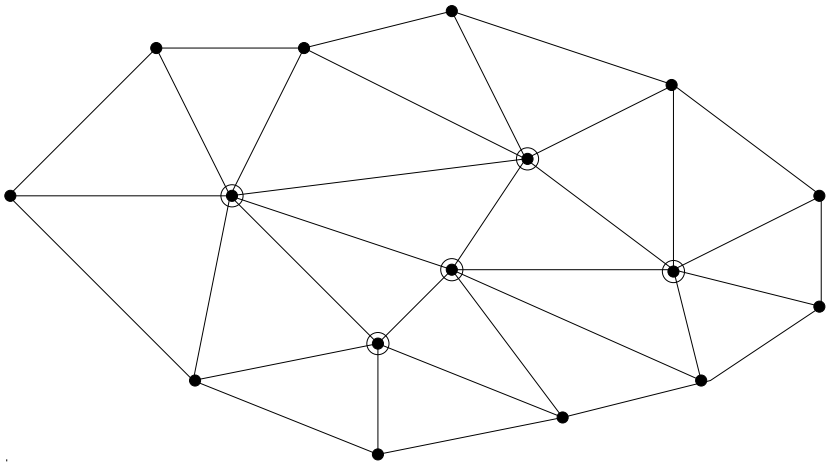


Figure 7: A subdivision (triangulation) of $\bar{\Omega}$.

With each interior node (marked $\odot$ in the figure) we associate a basis function ϕ which is equal to 1 at that node and to 0 at all the other nodes; ϕ is assumed to be continuous and piecewise linear on the triangulation, as shown in Fig. 8.

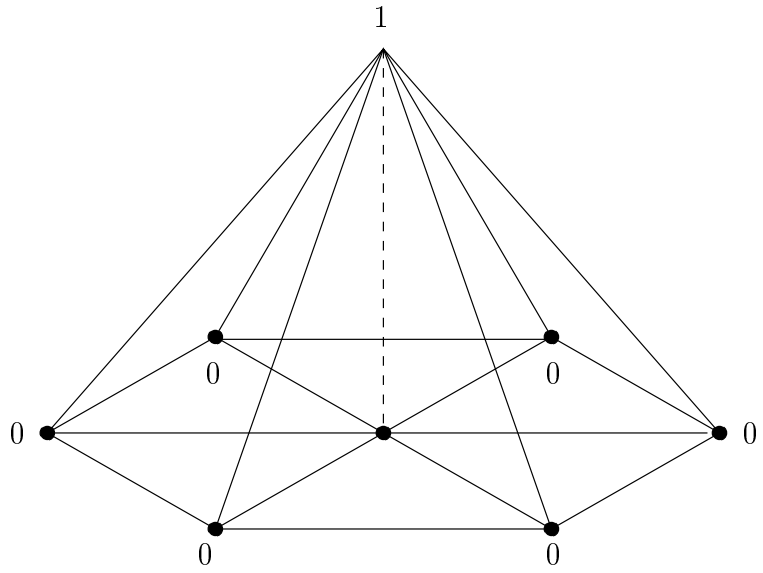


Figure 8: A typical finite element basis function.

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

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

The weak formulation of this problem is:

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

The finite element approximation of the problem is:

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

Writing

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

the finite element approximation can be restated as follows:

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

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

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 restated 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

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

To simplify matters let us suppose that $\Omega = (0,1) \times (0,1)$ and consider the triangulation of $\bar{\Omega}$ shown in Fig. 9.

Let ϕ_{ij} denote the basis function associated with the interior node (x_i, y_j):

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

where $1, 2, \ldots, 6$ denote the triangles surrounding the node (x_i, y_j) (see Fig. 10.)

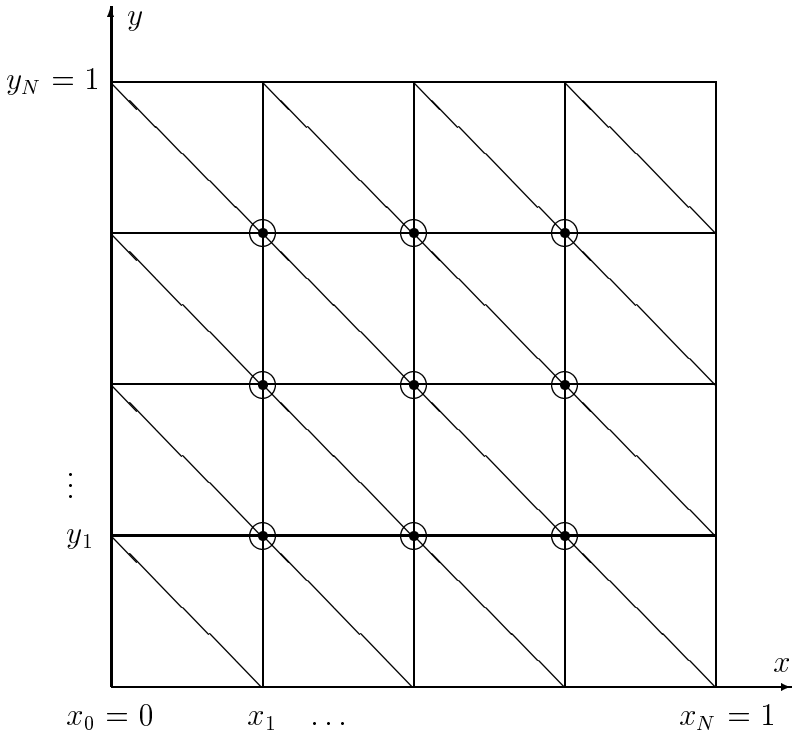


Figure 9: Subdivision (triangulation) of $\bar{\Omega} = [0,1] \times [0,1]$.

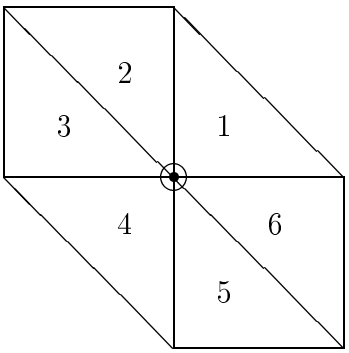


Figure 10: Triangles surrounding a node.

Thus

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

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

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

the finite element approximation is equivalent to

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

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

5.2 Variational formulation of self-adjoint elliptic boundary value problems

Let us consider, as in Section 2, the elliptic boundary value problem

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j} \left(a_{ij}(x) \frac{\partial u}{\partial x_i} \right) + \sum_{i=1}^{n} b_i(x) \frac{\partial u}{\partial x_i} + c(x)u = f(x), \qquad x \in \Omega, \tag{5.2a}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{5.2b}$$

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

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \tilde{c}\sum_{i=1}^{n}\xi_i^2 \quad \forall \xi = (\xi_1,\ldots,\xi_n) \in \mathbb{R}^n, \quad \forall x \in \bar{\Omega}. \tag{5.3}$$

We recall from Section 2 that the weak formulation of (5.2) is:

$$\text{find } u \in H_0^1(\Omega) \text{ such that } a(u,v) = l(v) \quad \forall v \in H_0^1(\Omega), \tag{5.4}$$

where the bilinear form $a(\cdot,\cdot)$ and the linear form $l(\cdot)$ 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 (5.4) has a unique solution u in $H_0^1(\Omega)$, — the weak solution of (5.2).

In the special case when the boundary value problem is self-adjoint, i.e.

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

and

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

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

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

in the following this will always be assumed to be the case. Thus we consider

$$-\sum_{i,j=1}^{n}\frac{\partial}{\partial x_j}\left(a_{ij}(x)\frac{\partial u}{\partial x_i}\right) + c(x)u = f(x), \quad x \in \Omega, \tag{5.5a}$$

$$u = 0, \quad \text{on} \quad \partial\Omega \tag{5.5b}$$

with $a_{ij}(x)$ satisfying the ellipticity condition (5.3); $a_{ij}(x) = a_{ji}(x)$, $c(x) \geq 0$, $x \in \bar{\Omega}$.

It turns out that (5.5) can be restated as a minimisation problem. To be more precise, let us 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 5.1 *Let u be the (unique) solution of (5.4) and suppose that $a(\cdot,\cdot)$ is a symmetric bilinear form 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 solution of (5.4) 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 (2.14),

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

where c_0 is a positive constant. Thus

$$J(v) - J(u) \geq \frac{c_0}{2}\left\|v-u\right\|_{H^1(\Omega)}^2 \quad \forall v \in H_0^1(\Omega), \tag{5.6}$$

and therefore,

$$J(v) \geq J(u) \quad \forall v \in H_0^1(\Omega), \tag{5.7}$$

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

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

$$J(v) \geq J(\tilde{u}) \quad \forall v \in H_0^1(\Omega). \tag{5.8}$$

Taking $v = \tilde{u}$ in (5.7) and $v = u$ in (5.8), we deduce that

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

but then, by virtue of (5.6),

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

and hence $u = \tilde{u}$. $\square$

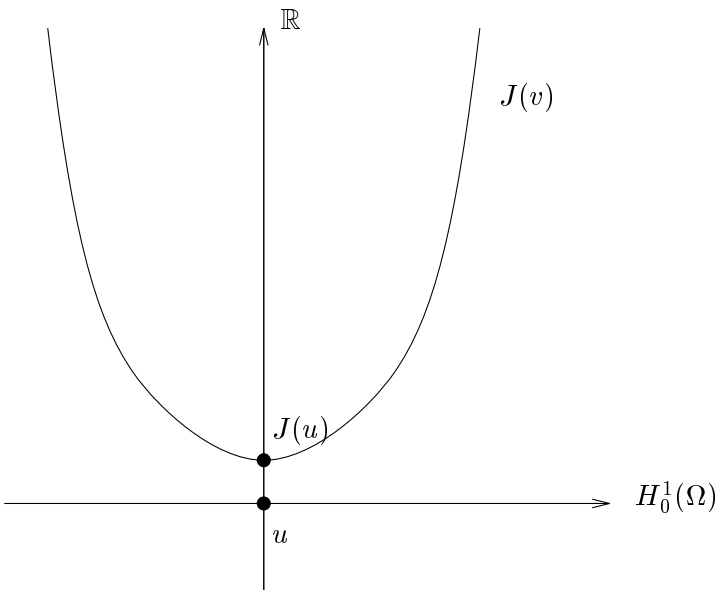


Figure 11: The quadratic functional $J(\cdot)$.

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 $a(v-w, v-w) \geq 0$ on noting that $\theta \in [0,1]$.

Moreover, if u minimises $J(\cdot)$ then the Gateaux derivative $J'(u)$ of $J(\cdot)$ at u,

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

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 5.2 *Suppose that $u \in H_0^1(\Omega)$ minimises $J(\cdot)$ over $H_0^1(\Omega)$; then, u is the (unique) solution of problem (5.4).*

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 (5.5) 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 use of this equivalence to perform an error analysis of the finite element method.

5.3 Construction of the finite element method: abstract setting

Let us consider the self-adjoint elliptic boundary value problem (5.5), and recall that its weak formulation is

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

where

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

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

we suppose that $a_{ij}(x) = a_{ji}(x)$, $i,j = 1,\ldots,n$, $x \in \bar{\Omega}$, $c(x) \geq 0$, $x \in \bar{\Omega}$, a_{ij}, $c \in C(\bar{\Omega})$, $f \in L_2(\Omega)$, and the ellipticity condition (5.3) holds. Recall also that (W) is equivalent to the minimisation problem

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

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

We can derive the finite element approximation of (5.5) by replacing the space $H_0^1(\Omega)$ in (W) by a certain finite-dimensional subspace $V_h \subset H_0^1(\Omega)$ which consists of continuous piecewise polynomials of a fixed degree k, $k \geq 1$.

Leaving aside for a moment the question of the actual construction of V_h, we consider, instead, some general questions concerning finite element methods which do not depend on the particular properties of V_h.

In its most general form, the finite element approximation of (W) is:

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

As $V_h \subset V = H_0^1(\Omega)$, the existence of a unique solution $u_h \in V_h$ is a straightforward consequence of the Lax–Milgram theorem (see, Section 2). In addition, we can repeat the argument presented in the previous section to show the equivalence of (W_h) to the following minimisation problem:

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

Next we study the approximation properties of (W_h).

5.4 Céa's lemma

Céa's lemma expresses the fact that, in a certain sense, the finite element solution $u_h \in V_h$ is the best approximation to $u \in V = H_0^1(\Omega)$ from V_h. To be more precise, we define

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

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

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

(cf. Section 2), it is easily seen that $(\cdot, \cdot)_a$ satisfies all axioms of an inner product. Let $\|\cdot\|_a$ denote the associated "energy norm":

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

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

$$a(u, v_h) = l(v_h), \quad v_h \in V_h; \tag{5.9}$$

also by, (W_h),

$$a(u_h, v_h) = l(v_h), \quad v_h \in V_h. \tag{5.10}$$

Subtracting (5.10) from (5.9) and using the fact that $a(\cdot, \cdot)$ is a bilinear form, we deduce that

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

i.e.

$$(u - u_h, v_h)_a = 0 \quad \forall v_h \in V_h. \tag{5.11}$$

Thus, the error between the exact solution u and its finite element approximation u_h is orthogonal to V_h.

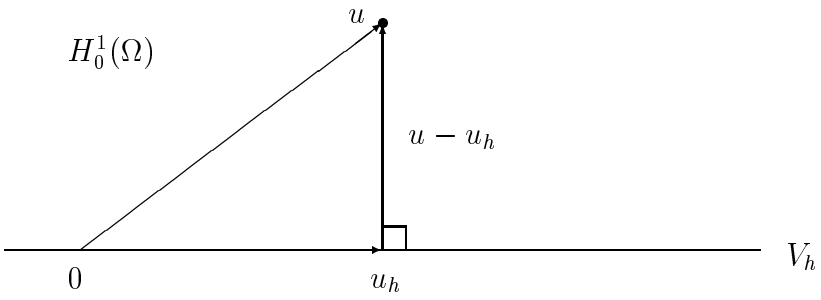


Figure 12: The error $u - u_h$ is orthogonal to V_h.

By virtue of the orthogonality property (5.11) (see Figure 12),

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

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,

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

the minimum being achieved when $v_h = u_h$. Thus we have proved the following result

Lemma 5.3 *(Céa's lemma) 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.*

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

This result is the key to the error analysis of the finite element method for self-adjoint elliptic boundary value problems. In the next section we describe how such an analysis proceeds for a particularly simple finite element space, V_h, consisting of continuous piecewise linear functions on Ω.

5.5 Optimal error bounds 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.

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

$$-\Delta u = f \quad \text{in } \Omega, \tag{5.12a}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{5.12b}$$

We recall that the weak formulation of this problem is:

$$\begin{gathered}\text{find } u \in H_0^1(\Omega) \text{ such that} \\ \int_\Omega \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y = \int_\Omega f v \,\mathrm{d}x\,\mathrm{d}y \quad \forall v \in H_0^1(\Omega).\end{gathered} \tag{5.13}$$

In order to construct the finite element approximation, we triangulate the domain as shown in the Fig. 13. 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 Ω (labelled $\odot$ in the figure), we associate a basis-function ϕ_{ij}, $i, j = 1, \ldots, N-1$, defined by

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

Let $V_h = \operatorname{span}\{\phi_{ij},\ i = 1, \ldots, N-1;\ j = 1, \ldots, N-1\}$. The finite element approximation of (5.12) (and (5.13)) is:

$$\begin{gathered}\text{find } u_h \in V_h \text{ such that} \\ \int_\Omega \left(\frac{\partial u_h}{\partial x}\frac{\partial v_h}{\partial x} + \frac{\partial u_h}{\partial y}\frac{\partial v_h}{\partial y} \right) \mathrm{d}x\,\mathrm{d}y = \int_\Omega f v_h \,\mathrm{d}x\,\mathrm{d}y \quad \forall v_h \in V_h.\end{gathered} \tag{5.14}$$

Letting

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

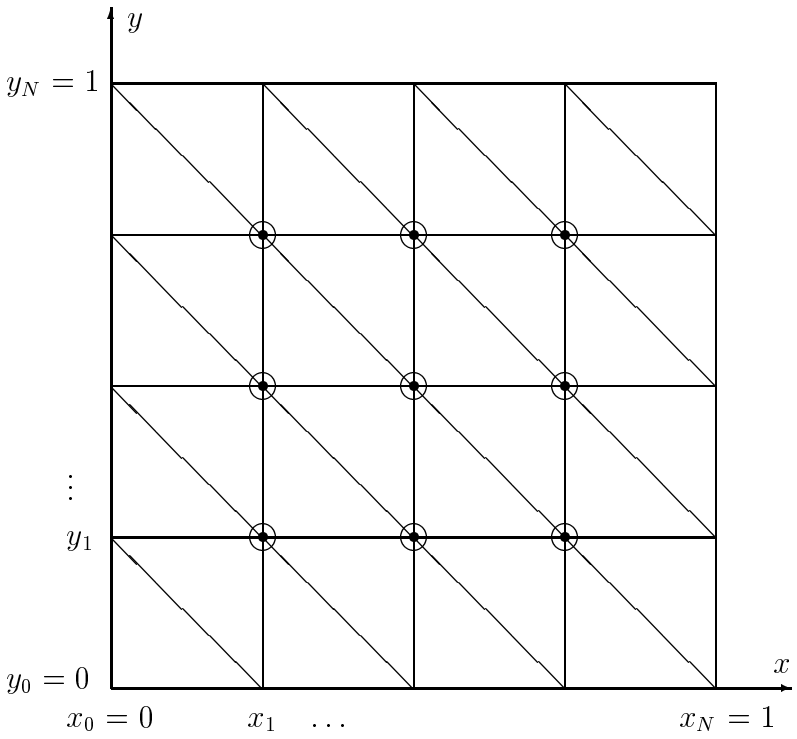


Figure 13: Subdivision (triangulation) of $\bar{\Omega} = [0,1] \times [0,1]$.

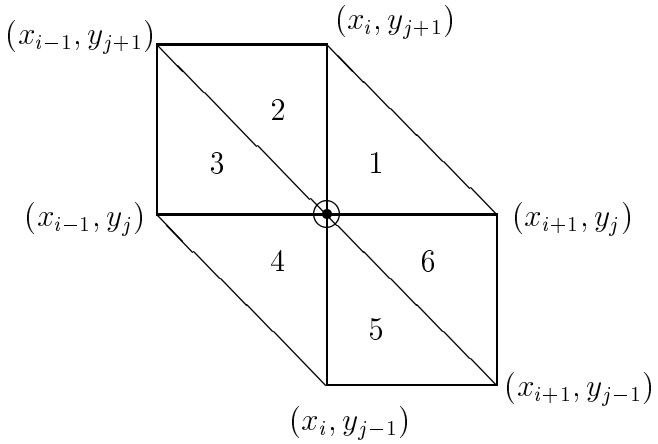


Figure 14: Triangles surrounding the node (x_i, y_j).

(5.13) and the finite element method (5.14) can be written, respectively, as follows:

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

and

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

Let us suppose that $u \in H^2(\Omega) \cap H_0^1(\Omega)$. By the Sobolev embedding theorem $H^2(\Omega) \subset C(\bar{\Omega})$ (cf. also Lemma 4.10 (b)); therefore $u \in C(\bar{\Omega})$. According to Céa's lemma,

$$\|u - u_h\|_a = \min_{v_h \in V_h} \|u - v_h\|_a \le \|u - I_h u\|_a \,, \tag{5.15}$$

where $I_h u$ denotes the continuous piecewise linear interpolant of u on Ω:

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

Clearly $(I_h u)(x_k, y_l) = u(x_k, y_l)$. Since $u \in C(\bar{\Omega})$, $I_h u$ is correctly defined. Let us estimate $\|u - I_h u\|_a$:

$$\begin{aligned} \|u - I_h u\|_a^2 &= \int_\Omega \left| \frac{\partial}{\partial x}(u - I_h u) \right|^2 \mathrm{d}x\,\mathrm{d}y + \int_\Omega \left| \frac{\partial}{\partial y}(u - I_h u) \right|^2 \mathrm{d}x\,\mathrm{d}y \\ &= \sum_{\triangle} \left\{ \int_\triangle \left| \frac{\partial}{\partial x}(u - I_h u) \right|^2 \mathrm{d}x\,\mathrm{d}y + \int_\triangle \left| \frac{\partial}{\partial y}(u - I_h u) \right|^2 \mathrm{d}x\,\mathrm{d}y \right\}, \end{aligned} \tag{5.16}$$

where $\triangle$ is a triangle in the partition of Ω. 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

$$\int_\triangle \left| \frac{\partial}{\partial x}(u - I_h u) \right|^2 \mathrm{d}x\,\mathrm{d}y + \int_\triangle \left| \frac{\partial}{\partial y}(u - I_h u) \right|^2 \mathrm{d}x\,\mathrm{d}y,$$

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

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

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

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

Thus,

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

Therefore,

$$\int_{\triangle} \left| \frac{\partial}{\partial x}(u - I_h u) \right|^2 \mathrm{d}x\,\mathrm{d}y \le 2h^2 \int_{x_i}^{x_{i+1}} \int_{y_j}^{y_{j+1}} \left(\left| \frac{\partial^2 u}{\partial x^2} \right|^2 + \frac{1}{2} \left| \frac{\partial^2 u}{\partial x \partial y} \right|^2 \right) \mathrm{d}x\,\mathrm{d}y. \tag{5.17}$$

Similarly,

$$\int_{\triangle} \left| \frac{\partial}{\partial y}(u - I_h u) \right|^2 \mathrm{d}x\,\mathrm{d}y \le 2h^2 \int_{x_i}^{x_{i+1}} \int_{y_j}^{y_{j+1}} \left(\left| \frac{\partial^2 u}{\partial y^2} \right|^2 + \frac{1}{2} \left| \frac{\partial^2 u}{\partial x \partial y} \right|^2 \right) \mathrm{d}x\,\mathrm{d}y. \tag{5.18}$$

Substituting (5.17) and (5.18) into (5.16),

$$\|u - I_h u\|_a^2 \le 4h^2 \int_\Omega \left(\left| \frac{\partial^2 u}{\partial x^2} \right|^2 + \left| \frac{\partial^2 u}{\partial x \partial y} \right|^2 + \left| \frac{\partial^2 u}{\partial y^2} \right|^2 \right) \mathrm{d}x\,\mathrm{d}y. \tag{5.19}$$

Finally by (5.15) and (5.19),

$$\|u - u_h\|_a \le 2h\,|u|_{H^2(\Omega)}\,. \tag{5.20}$$

Thus we have proved the following result.

Theorem 5.4 *Let u be the weak solution of the boundary value problem (5.12), and let u_h be its piecewise linear finite element approximation defined by (5.14). Suppose that $u \in H^2(\Omega) \cap H_0^1(\Omega)$; then,*

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

Corollary *Under the hypotheses of Theorem 5.4*

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

Proof According to Theorem 5.4,

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

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,

$$\|u - u_h\|_{L_2(\Omega)}^2 \leq \frac{1}{4}\,|u - u_h|_{H^1(\Omega)}^2\,; \tag{5.21}$$

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. □

According to (5.21) and (5.20),

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

This error estimate seems to indicate 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 crude and can be improved to $\mathcal{O}(h^2)$. For this purpose, let us first observe that if $w \in H^2(\Omega) \cap H_0^1(\Omega)$, $\Omega = (0,1) \times (0,1)$, then

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

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,

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

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

$$-\Delta w_g = g \quad \text{in } \Omega, \tag{5.22a}$$
$$w_g = 0 \quad \text{on } \partial\Omega; \tag{5.22b}$$

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

$$|w_g|_{H^2(\Omega)} = \|\Delta w_g\|_{L_2(\Omega)} = \|g\|_{L_2(\Omega)} . \tag{5.23}$$

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,

$$\|u - u_h\|_{L_2(\Omega)} = \sup_{g \in L_2(\Omega)} \frac{(u - u_h, g)}{\|g\|_{L_2(\Omega)}}. \tag{5.24}$$

Given $g \in L_2(\Omega)$, let $w_g \in H_0^1(\Omega)$ denote the weak solution of the problem (5.22), i.e.

$$a(w_g, v) = l_g(v) \quad \forall v \in H_0^1(\Omega), \tag{5.25}$$

where

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

Consider the finite element approximation of (5.25):

$$\text{find } w_{gh} \in V_h \text{ such that } a(w_{gh}, v_h) = l_g(v_h) \quad \forall v_h \in V_h. \tag{5.26}$$

From (5.25), (5.26) and the error bound (5.20), we deduce that

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

and therefore, by (5.23),

$$\|w_g - w_{gh}\|_a \leq 2h \|g\|_{L_2(\Omega)} . \tag{5.27}$$

Now,

$$\begin{aligned}(u - u_h, g) &= (g, u - u_h) = l_g(u - u_h) \\ &= a(w_g, u - u_h) = a(u - u_h, w_g). \end{aligned} \tag{5.28}$$

Because $w_{gh} \in V_h$, (5.11) implies that

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

and therefore, by (5.28),

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

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 (5.20) and (5.27)

$$(u - u_h, g) \leq 4h^2 |u|_{H^2(\Omega)} \cdot \|g\|_{L_2(\Omega)} . \tag{5.29}$$

Substituting (5.29) into the right-hand side of (5.24), 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.

6 Finite difference approximation of evolutionary problems

In Sections 3–5 we considered numerical methods for the approximate solution of elliptic equations. This section is devoted to finite difference methods for time-dependent problems described by parabolic and hyperbolic equations.

6.1 Finite difference methods for parabolic equations

Let Ω 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 $u(x,t)$, $x \in \Omega$, $t \in (0,T]$:

$$\frac{\partial u}{\partial t} - \sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}(a_{ij}(x,t)\frac{\partial u}{\partial x_i}) + \sum_{i=1}^{n} b_i(x,t)\frac{\partial u}{\partial x_i} + c(x,t)u = f(x,t), \quad x \in \Omega, \quad t \in (0,T], \tag{6.1}$$

$$u(x,t) = 0, \qquad x \in \Gamma, \quad t \in [0,T], \tag{6.2}$$

$$u(x,0) = u_0(x), \qquad x \in \bar{\Omega}, \tag{6.3}$$

where, for the sake of consistency between the boundary condition (6.2) and the initial condition (6.3), we shall assume that the initial datum u_0 satisfies: $u_0(x) = 0$, $x \in \Gamma$. Suppose that $u_0 \in L_2(\Omega)$, and that there exists a positive constant $\tilde{c}$ such that

$$\sum_{i,j=1}^{n} a_{ij}(x,t)\xi_i\xi_j \geq \tilde{c}\sum_{i=1}^{n}\xi_i^2, \quad \forall \xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n, \quad \forall x \in \bar{\Omega}, \quad t \in [0,T]. \tag{6.4}$$

We shall also assume that

$$a_{ij} \in C^1(\bar{Q}), \quad b_i \in C^1(\bar{Q}), \quad i,j = 1, \ldots, n,$$
$$c \in C^0(\bar{Q}), \quad f \in L^2(Q),$$

and that

$$c(x,t) - \frac{1}{2}\sum_{i=1}^{n}\frac{\partial b_i}{\partial x_i}(x,t) \geq 0, \quad (x,t) \in \bar{Q}, \tag{6.5}$$

similarly as in the elliptic case.

A partial differential equation of the form (6.1) 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 convection-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 of a parabolic initial boundary value problem is more technical than the proof of the corresponding result for an elliptic boundary value problem and so it is omitted. Instead, we shall assume that (6.1)–(6.3) has a unique solution and we shall investigate its decay in t (t typically signifies time), and 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 (6.1) with u, noting that $u(x, t) = 0$, $x \in \Gamma$, integrating by parts, and employing (6.4) and (6.5),

$$\left(\frac{\partial u}{\partial t}(\cdot, t), u(\cdot, t)\right) + \tilde{c} \sum_{i=1}^{n} \left\| \frac{\partial u}{\partial x_i}(\cdot, t)\right\|^2_{L_2(\Omega)} \leq (f(\cdot, t), u(\cdot, t)).$$

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 (1.1), 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,

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} \|u(\cdot, t)\|^2_{L_2(\Omega)} + K \|u(\cdot, t)\|^2_{L_2(\Omega)} &\leq \|f(\cdot, t)\|_{L_2(\Omega)} \|u(\cdot, t)\|_{L_2(\Omega)}\\ &\leq \frac{1}{2K} \|f(\cdot, t)\|^2_{L_2(\Omega)} + \frac{K}{2} \|u(\cdot, t)\|^2_{L_2(\Omega)}.\end{aligned}$$

Thence,

$$\frac{\mathrm{d}}{\mathrm{d}t} \|u(\cdot, t)\|^2_{L_2(\Omega)} + K \|u(\cdot, t)\|^2_{L_2(\Omega)} \leq \frac{1}{K} \|f(\cdot, t)\|^2_{L_2(\Omega)}.$$

Multiplying both sides by 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

$$\|u(\cdot,t)\|^2_{L_2(\Omega)} \le \mathrm{e}^{-Kt}\,\|u_0\|^2_{L_2(\Omega)} + \frac{1}{K}\int_0^t \mathrm{e}^{-K(t-\tau)}\,\|f(\cdot,\tau)\|^2_{L_2(\Omega)}\;\mathrm{d}\tau. \tag{6.6}$$

Assuming that (6.1)–(6.3) has a solution, (6.6) implies that the solution is unique. Indeed, if u_1 and u_2 are solutions of (6.1)–(6.3), then $u = u_1 - u_2$ satisfies (6.1)–(6.3) with $f \equiv 0$ and $u_0 \equiv 0$; therefore, by (6.6), $u \equiv 0$, i.e. $u_1 \equiv u_2$.

Let us also look at the special case when $f \equiv 0$ in (6.1). This corresponds to considering the evolution of the solution from the initial datum, u_0, in the absence of external forces. In this case (6.6) yields

$$\|u(\cdot,t)\|^2_{L_2(\Omega)} \le \mathrm{e}^{-Kt}\,\|u_0\|^2_{L_2(\Omega)}\,, \quad t \ge 0. \tag{6.7}$$

In other words, the energy, $\frac{1}{2}\,\|u(\cdot,t)\|^2_{L_2(\Omega)}$ decays (dissipates) exponentially fast. Since $K = \tilde{c}/c_\star$, we have

$$\|u(\cdot,t)\|^2_{L_2(\Omega)} \le \mathrm{e}^{-\tilde{c}t/c_\star}\,\|u_0\|^2_{L_2(\Omega)}\,, \quad t \ge 0, \tag{6.8}$$

and we deduce that the rate of dissipation depends on the lower bound, $\tilde{c}$, on the diffusion coefficients (i.e. the smaller $\tilde{c}$, the slower the decay of the energy).

In the next section we consider some simple finite difference schemes for the numerical solution of parabolic initial boundary value problems. Analogous results can be proved when the spatial discretisation is based on the finite difference method. In order to simplify the presentation, we restrict ourselves to the heat equation in one space dimension.

6.1.1 Explicit and implicit schemes

We consider the following simple model problem for the heat equation in one space dimension. Let $Q = \Omega \times (0,T]$, where $\Omega = (0,1)$, $T > 0$;

$$\begin{aligned}
&\text{find } u(x,t) \text{ such that}\\
&\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t), \quad x \in (0,1), \quad t \in (0,T],\\
&u(0,t) = 0, \quad u(1,t) = 0, \quad t \in [0,T],\\
&u(x,0) = u_0(x), \quad x \in [0,1].
\end{aligned} \tag{6.9}$$

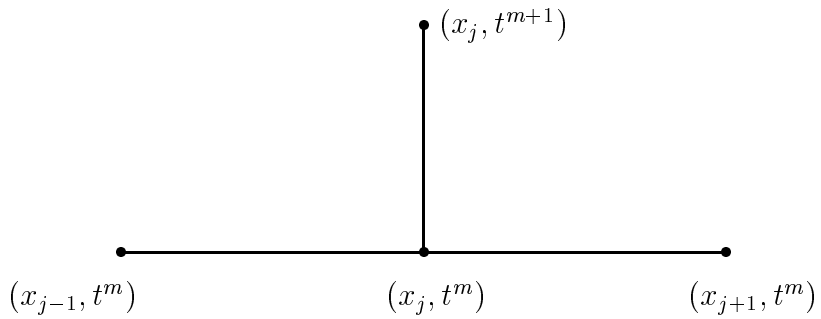


Figure 15: Four-point stencil for the explicit scheme.

We describe two schemes for the numerical solution of (6.9). They both use the same discretisation of $\partial^2 u/\partial x^2$, but while the first scheme (called the explicit scheme) employs a forward difference in t to approximate $\partial u/\partial t$, the second (called the implicit scheme) uses a backward difference in t.

The explicit 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

$$\bar{Q}_h^{\Delta t} = \{(x_j, t^m) : x_j = jh,\ 0 \leq j \leq N;\ t^m = m \cdot \Delta t,\ 0 \leq m \leq M\}.$$

On $\bar{Q}_h^{\Delta t}$ we approximate (6.9) by the following finite difference scheme:

$$\begin{aligned}
&\text{find } U_j^m, \quad 0 \leq j \leq N, \ \ 0 \leq m \leq M, \text{ such that} \\
&\frac{U_j^{m+1} - U_j^m}{\Delta t} = D_x^+ D_x^- U_j^m + f(x_j, t^m), \quad 1 \leq j \leq N-1, \ \ 0 \leq m \leq M-1, \\
&U_0^m = 0, \quad U_N^m = 0, \quad 0 \leq m \leq M, \\
&U_j^0 = u_0(x_j), \quad 0 \leq j \leq N,
\end{aligned} \tag{6.10}$$

where U_j^m represents the approximation of $u(x_j, t^m)$, the value of u at the mesh-point (x_j, t^m).

Clearly, (6.10) is a 4-point difference scheme involving the values of U at the mesh-points

$$(x_{j-1}, t^m),\ (x_j, t^m),\ (x_{j+1}, t^m),\ (x_j, t^{m+1}),$$

shown in Fig. 15. The scheme (6.10) is applied as follows. First we set $m = 0$. Since U_{j-1}^0, U_j^0, U_{j+1}^0 are given by the initial condition $U_j^0 = u_0(x_j)$, $j = 0, \ldots, N$, the values U_j^1, $j = 0, \ldots, N$, can be computed from (6.10):

$$\begin{aligned}
&U_j^1 = U_j^0 + \frac{\Delta t}{h^2}(U_{j+1}^0 - 2U_j^0 + U_{j-1}^0) + \Delta t \cdot f(x_j, t^0), \quad j = 1, \ldots, N-1, \\
&U_0^1 = 0, \quad U_N^1 = 0;
\end{aligned}$$

the values of U on the time-level $t = t^1 = 1 \cdot \Delta t$ can be calculated explicitly from U_j^0, $j = 0, \ldots, N$, and hence the terminology *explicit scheme*.

Suppose we have already calculated U_j^m, $j = 0, \ldots, N$, the values of U on time level $t^m = m \cdot \Delta t$. The values of U on the next time level $t^{m+1} = (m+1) \cdot \Delta t$ can be obtained from (6.10):

$$\begin{aligned} U_j^{m+1} &= U_j^m + \frac{\Delta t}{h^2}(U_{j+1}^m - 2U_j^m + U_{j-1}^m) + \Delta t \cdot f(x_j, t^m), \quad j = 1, \ldots, N-1, \\ U_0^{m+1} &= 0 \quad U_N^{m+1} = 0, \end{aligned}$$

for any m, $0 \le m \le M-1$.

The implicit scheme. Alternatively, one can approximate the time derivative by a backward difference, which gives rise to the following *implicit scheme*:

$$\begin{aligned} &\text{find } U_j^m, \quad 0 \le j \le N, \ \ 0 \le m \le M, \text{ such that} \\ &\frac{U_j^{m+1} - U_j^m}{\Delta t} = D_x^+ D_x^- U_j^{m+1} + f(x_j, t^{m+1}), \quad 1 \le j \le N-1, \ \ 0 \le m \le M-1, \\ &U_0^{m+1} = 0, \quad U_N^{m+1} = 0, \quad 0 \le m \le M-1, \\ &U_j^0 = u_0(x_j), \quad 0 \le j \le N, \end{aligned} \tag{6.11}$$

where U_j^m represents the approximation of $u(x_j, t^m)$, the value of u at the mesh-point (x_j, t^m). Equivalently, (6.11) can be written

$$\begin{aligned} -\frac{\Delta t}{h^2}U_{j+1}^{m+1} + \left(\frac{2\Delta t}{h^2} + 1\right) U_j^{m+1} - \frac{\Delta t}{h^2}U_{j-1}^{m+1} &= U_j^m + \Delta t \cdot f(x_j, t^{m+1}), \\ &\qquad 1 \le j \le N-1, \\ U_0^{m+1} = 0, \quad U_N^{m+1} &= 0, \end{aligned} \tag{6.12}$$

for each m, $0 \le m \le M-1$.

This is, again, a 4-point finite difference scheme, but it involves the values of U at the mesh-points

$$(x_{j-1}, t^{m+1}),\ (x_j, t^{m+1}),\ (x_{j+1}, t^{m+1}),\ (x_j, t^m),$$

shown in Fig. 16. The implicit scheme (6.12) is implemented as follows. First we set $m = 0$; then, (6.12) is a system of linear equations with a tridiagonal matrix, and the right-hand side can be computed from the initial datum $U_j^0 = u_0(x_j)$, and the forcing function $f(x_j, t^1)$. Suppose we have already computed U_j^m, $j = 0, \ldots, N$, the values of U on time level $t^m = m \cdot \Delta t$. The values of U on the next time level $t^{m+1} = (m+1) \cdot \Delta t$ are obtained by solving the system of linear equations (6.12).

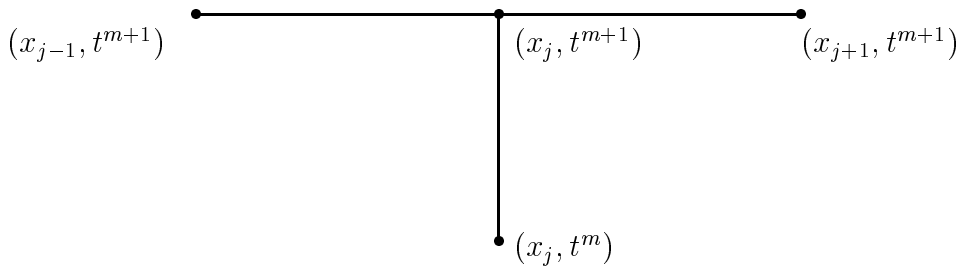


Figure 16: Four-point stencil for the implicit scheme.

6.1.2 Stability of explicit and implicit schemes

We shall study the stability of the schemes (6.10) and (6.11) simultaneously, by embedding them into a one-parameter family of finite difference schemes:

$$\begin{aligned}
&\text{find } U_j^m, \quad 0 \le j \le N, \;\; 0 \le m \le M, \text{ such that} \\
&\frac{U_j^{m+1} - U_j^m}{\Delta t} = D_x^+ D_x^- (\theta U_j^{m+1} + (1-\theta) U_j^m) + f(x_j, t^{m+\theta}), \qquad \begin{array}{l} 1 \le j \le N-1, \\ 0 \le m \le M-1, \end{array} \\
&U_0^m = 0, \quad U_N^m = 0, \quad 0 \le m \le M, \\
&U_j^0 = u_0(x_j), \quad 0 \le j \le N,
\end{aligned} \tag{6.13}$$

where $0 \le \theta \le 1$. Recall that

$$\begin{aligned}
(V, W)_h &= \sum_{j=1}^{N-1} h V_j W_j, \\
\|V\|_h &= (V, V)_h^{1/2}.
\end{aligned}$$

Taking the inner product of (6.13) with

$$U^{m+\theta} := \theta U^{m+1} + (1-\theta) U^m,$$

we get

$$\left(\frac{U^{m+1} - U^m}{\Delta t}, U^{m+\theta} \right)_h - (D_x^+ D_x^- U^{m+\theta}, U^{m+\theta})_h = (f^{m+\theta}, U^{m+\theta})_h,$$

where $f_j^{m+\theta} = f^{m+\theta}(x_j) = f(x_j, t^{m+\theta})$. Let

$$\|V\|_h = \left(\sum_{j=1}^{N} h \, |V_j|^2 \right)^{1/2}.$$

Noting that $U_0^{m+\theta} = 0$, $U_N^{m+\theta} = 0$, it follows from Lemma 3.1 that

$$-(D_x^+ D_x^- U^{m+\theta}, U^{m+\theta})_h = \left\| D_x^- U^{m+\theta} \right]\!\!\right|_h^2 .$$

Thus,

$$\left(\frac{U^{m+1} - U^m}{\Delta t}, U^{m+\theta} \right)_h + \left\| D_x^- U^{m+\theta} \right]\!\!\right|_h^2 = (f^{m+\theta}, U^{m+\theta})_h .$$

Since

$$U^{m+\theta} = \Delta t(\theta - \tfrac{1}{2}) \frac{U^{m+1} - U^m}{\Delta t} + \frac{U^{m+1} + U^m}{2},$$

it follows that

$$\Delta t(\theta - \tfrac{1}{2}) \left\| \frac{U^{m+1} - U^m}{\Delta t} \right\|_h^2 + \frac{\|U^{m+1}\|_h^2 - \|U^m\|_h^2}{2\Delta t} + \left\| D_x^- U^{m+\theta} \right]\!\!\right|_h^2 = (f^{m+\theta}, U^{m+\theta})_h . \quad (6.14)$$

Suppose $\theta \in [1/2, 1]$; then, $\theta - 1/2 \geq 0$, and therefore

$$\begin{aligned} \frac{\|U^{m+1}\|_h^2 - \|U^m\|_h^2}{2\Delta t} + \left\| D_x^- U^{m+\theta} \right]\!\!\right|_h^2 &\leq (f^{m+\theta}, U^{m+\theta})_h \\ &\leq \left\| f^{m+\theta} \right\|_h \left\| U^{m+\theta} \right\|_h . \end{aligned}$$

According to the discrete Poincaré–Friedrichs inequality (3.9),

$$\left\| U^{m+\theta} \right\|_h^2 \leq \frac{1}{2} \left\| D_x^- U^{m+\theta} \right]\!\!\right|_h^2 .$$

Thus

$$\frac{\|U^{m+1}\|_h^2 - \|U^m\|_h^2}{2\Delta t} + 2 \left\| U^{m+\theta} \right\|_h^2 \leq \frac{1}{2} \left\| f^{m+\theta} \right\|_h^2 + \frac{1}{2} \left\| U^{m+\theta} \right\|_h^2 ,$$

so that

$$\left\| U^{m+1} \right\|_h^2 \leq \|U^m\|_h^2 + \Delta t \left\| f^{m+\theta} \right\|_h^2 .$$

Summing through m,

$$\left\| U^k \right\|_h^2 \leq \left\| U^0 \right\|_h^2 + \sum_{m=0}^{k-1} \Delta t \left\| f^{m+\theta} \right\|_h^2 , \quad (6.15)$$

for all k, $1 \leq k \leq M$.

The inequality (6.15) can be thought of as the discrete version of (6.6). If follows from (6.15) that

$$\max_{1 \leq k \leq M} \left\| U^k \right\|_h^2 \leq \left\| U^0 \right\|_h^2 + \sum_{m=0}^{M-1} \Delta t \left\| f^{m+\theta} \right\|_h^2 ,$$

i.e.

$$\max_{1\le k\le M} \left\|U^k\right\|_h \le \left[\left\|U^0\right\|_h^2 + \sum_{m=0}^{M-1} \Delta t \left\|f^{m+\theta}\right\|_h^2\right]^{1/2}, \tag{6.16}$$

which expresses the continuous dependence of the solution of the finite difference scheme (6.13) 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 (6.13) is stable without any limitations on the time step in terms of h. In other words, the scheme (6.13) is *unconditionally stable* for $\theta \in [1/2, 1]$.

Now let us consider the case $\theta \in [0, 1/2)$. First suppose that $f \equiv 0$. Then, according to (6.14),

$$\frac{\|U^{m+1}\|_h^2 - \|U^m\|_h^2}{2\Delta t} + \left\|D_x^- U^{m+\theta}\right]\Big|_h^2 = \Delta t(\tfrac{1}{2} - \theta) \left\|\frac{U^{m+1} - U^m}{\Delta t}\right\|_h^2. \tag{6.17}$$

Recalling (6.13) and the fact that $f \equiv 0$, it follows that

$$\frac{U^{m+1} - U^m}{\Delta t} = D_x^+ D_x^- U^{m+\theta}.$$

Moreover, a simple calculation based on the inequality $(a-b)^2 \le 2a^2 + 2b^2$ shows that

$$\left\|D_x^+ D_x^- U^{m+\theta}\right\|_h^2 \le \frac{4}{h^2} \left\|D_x^- U^{m+\theta}\right]\Big|_h^2. \tag{6.18}$$

Thus, (6.17) implies that

$$\frac{\|U^{m+1}\|_h^2 - \|U^m\|_h^2}{2\Delta t} + \left\|D_x^- U^{m+\theta}\right]\Big|_h^2 \le \frac{4\Delta t}{h^2}(\tfrac{1}{2} - \theta) \left\|D_x^- U^{m+\theta}\right]\Big|_h^2,$$

i.e.

$$\frac{\|U^{m+1}\|_h^2 - \|U^m\|_h^2}{2\Delta t} + \left(1 - \frac{2\Delta t(1-2\theta)}{h^2}\right) \left\|D_x^- U^{m+\theta}\right]\Big|_h^2 \le 0.$$

Let us assume that

$$\Delta t \le \frac{h^2}{2(1-2\theta)}, \quad \theta \in [0, 1/2); \tag{6.19}$$

then,

$$\left\|U^{m+1}\right\|_h^2 \le \|U^m\|_h^2, \quad m = 0, \ldots, M-1,$$

and hence,

$$\max_{1\le k\le M} \left\|U^k\right\|_h \le \left\|U^0\right\|_h.$$

Thus, again, the scheme is stable, but only if (6.19) holds. In other words, for $\theta \in [0, 1/2)$ the scheme (6.13) is *conditionally stable*, the condition being (6.19) (when $f \equiv 0$).

Let us suppose that $\theta \in [0, 1/2)$, as before, but consider the general situation when f is not identically zero. We shall prove that (6.13) is still only conditionally stable, and, in particular, that the explicit scheme, corresponding to $\theta = 0$, is conditionally stable.

Recalling (6.14),

$$\frac{\|U^{m+1}\|_h^2 - \|U^m\|_h^2}{2\Delta t} + \left\|D_x^- U^{m+\theta}\right]\!\right|_h^2 \leq \left\|f^{m+\theta}\right\|_h \left\|U^{m+\theta}\right\|_h + \Delta t(\tfrac{1}{2} - \theta) \left\|\frac{U^{m+1} - U^m}{\Delta t}\right\|_h^2. \quad (6.20)$$

By (6.13), for any $\epsilon \in (0, 1)$,

$$\begin{aligned}
\left\|\frac{U^{m+1} - U^m}{\Delta t}\right\|_h^2 &= \left\|D_x^+ D_x^- U^{m+\theta} + f^{m+\theta}\right\|_h^2 \\
&\leq \left(\left\|D_x^+ D_x^- U^{m+\theta}\right\|_h + \left\|f^{m+\theta}\right\|_h\right)^2 \\
&\leq (1+\epsilon) \left\|D_x^+ D_x^- U^{m+\theta}\right\|_h^2 + (1+\epsilon^{-1}) \left\|f^{m+\theta}\right\|_h^2 \\
&\leq (1+\epsilon)\frac{4}{h^2} \left\|D_x^- U^{m+\theta}\right]\!\right|_h^2 + (1+\epsilon^{-1}) \left\|f^{m+\theta}\right\|_h^2,
\end{aligned}$$

where (6.18) has been applied in the last line. Substituting into (6.20),

$$\begin{aligned}
&\frac{\|U^{m+1}\|_h^2 - \|U^m\|_h^2}{2\Delta t} + \left(1 - \Delta t(\tfrac{1}{2} - \theta) \cdot \frac{4(1+\epsilon)}{h^2}\right) \left\|D_x^- U^{m+\theta}\right]\!\right|_h^2 \\
&\qquad \leq \left\|f^{m+\theta}\right\|_h \left\|U^{m+\theta}\right\|_h + \Delta t(\tfrac{1}{2} - \theta)(1+\epsilon^{-1}) \left\|f^{m+\theta}\right\|_h^2. \qquad (6.21)
\end{aligned}$$

According to the discrete Poincaré–Friedrichs inequality (3.9),

$$\left\|U^{m+\theta}\right\|_h^2 \leq \tfrac{1}{2} \left\|D_x^- U^{m+\theta}\right]\!\right|_h^2,$$

and therefore,

$$\begin{aligned}
\left\|f^{m+\theta}\right\|_h \left\|U^{m+\theta}\right\|_h &\leq \frac{1}{8\epsilon^2} \left\|f^{m+\theta}\right\|_h^2 + 2\epsilon^2 \left\|U^{m+\theta}\right\|_h^2 \\
&\leq \frac{1}{8\epsilon^2} \left\|f^{m+\theta}\right\|_h^2 + \epsilon^2 \left\|D_x^- U^{m+\theta}\right]\!\right|_h^2. \qquad (6.22)
\end{aligned}$$

Substituting (6.22) into (6.21),

$$\begin{aligned}
&\frac{\|U^{m+1}\|_h^2 - \|U^m\|_h^2}{2\Delta t} + \left(1 - \Delta t \frac{2(1-2\theta)(1+\epsilon)}{h^2} - \epsilon^2\right) \left\|D_x^- U^{m+\theta}\right]\!\right|_h^2 \\
&\qquad \leq \frac{1}{8\epsilon^2} \left\|f^{m+\theta}\right\|_h^2 + \Delta t(\tfrac{1}{2} - \theta)(1+\epsilon^{-1}) \left\|f^{m+\theta}\right\|_h^2.
\end{aligned}$$

Let us suppose that

$$\Delta t \leq \frac{h^2}{2(1-2\theta)}(1-\epsilon), \quad \theta \in [0, 1/2),$$

where ϵ is a fixed real number, $\epsilon \in (0,1)$. Then

$$1 - \Delta t \, \frac{2(1-2\theta)(1+\epsilon)}{h^2} - \epsilon^2 \geq 0,$$

so that

$$\left\|U^{m+1}\right\|_h^2 \leq \|U^m\|_h^2 + \frac{\Delta t}{4\epsilon^2}\left\|f^{m+\theta}\right\|_h^2 + \Delta t^2(1-2\theta)(1+\epsilon^{-1})\left\|f^{m+\theta}\right\|_h^2.$$

Letting $c_\epsilon = 1/(4\epsilon^2) + \Delta t(1-2\theta)(1+\epsilon^{-1})$, upon summation through all m this implies that

$$\max_{1\leq k\leq M}\left\|U^k\right\|_h^2 \leq \left\|U^0\right\|_h^2 + c_\epsilon \sum_{m=0}^{M-1} \Delta t \left\|f^{m+\theta}\right\|_h^2.$$

Taking the square root of both sides, we deduce that for $\theta \in [0, 1/2)$ the scheme (6.13) is conditionally stable in the sense that

$$\max_{1\leq k\leq M}\left\|U^k\right\|_h \leq \left[\left\|U^0\right\|_h^2 + c_\epsilon \sum_{m=0}^{M-1} \Delta t \left\|f^{m+\theta}\right\|_h^2\right]^{1/2}, \tag{6.23}$$

provided

$$\Delta t \leq \frac{h^2}{2(1-2\theta)}(1-\epsilon), \quad 0 < \epsilon < 1. \tag{6.24}$$

To summarise: when $\theta \in [1/2, 1]$, the difference scheme (6.13) is unconditionally stable. In the particular the implicit scheme, corresponding to $\theta = 1$, and the Crank–Nicolson scheme, corresponding to $\theta = 1/2$, are both unconditionally stable, and (6.16) holds. When $\theta \in [0, 1/2)$, the scheme (6.13) is conditionally stable, subject to the time step limitation (6.24). In particular the explicit scheme, corresponding to $\theta = 0$, is only conditionally stable.

6.1.3 Error analysis of difference schemes for the heat equation

In this section we investigate the accuracy of the finite difference scheme (6.13) for the numerical solution of the initial boundary value problem (6.9).

We define the truncation error of the scheme (6.13) by

$$\varphi_j^{m+\theta} = \frac{u(x_j, t^{m+1}) - u(x_j, t^m)}{\Delta t} - D_x^+ D_x^- [\theta u(x_j, t^{m+1}) + (1-\theta)u(x_j, t^m)] - f(x_j, t^{m+\theta}), \qquad \begin{matrix} 1 \leq j \leq N-1, \\ 0 \leq m \leq M-1, \end{matrix}$$

and the global error by

$$e_j^m = u(x_j, t^m) - U_j^m.$$

It is easily seen that e_j^m satisfies the following finite difference scheme:

$$\begin{aligned}
&\frac{e_j^{m+1} - e_j^m}{\Delta t} - D_x^+ D_x^- [\theta e_j^{m+1} + (1-\theta) e_j^m] = \varphi_j^{m+\theta}, \qquad \begin{array}{l} 1 \le j \le N-1, \\ 0 \le m \le M-1. \end{array} \\
&e_0^m = 0, \quad e_N^m = 0, \quad 0 \le m \le M, \\
&e_j^0 = 0, \quad 0 \le j \le N.
\end{aligned}$$

According to the stability results proved in Section 6.1.2,

$$\max_{1 \le m \le M} \|u^m - U^m\|_h \le \left[\sum_{k=0}^{M-1} \Delta t \left\| \varphi^{k+\theta} \right\|_h^2 \right]^{1/2}, \quad \theta \in [1/2, 1], \tag{6.25}$$

by (6.16), and

$$\max_{1 \le m \le M} \|u^m - U^m\|_h \le \left[c_\epsilon \sum_{k=0}^{M-1} \Delta t \left\| \varphi^{k+\theta} \right\|_h^2 \right]^{1/2}, \quad \theta \in [0, 1/2), \tag{6.26}$$

provided

$$\Delta t \le \frac{h^2}{2(1-2\theta)}(1-\epsilon), \quad 0 < \epsilon < 1, \quad \theta \in [0, 1/2).$$

In either case we have to estimate $\left\| \varphi^{m+\theta} \right\|_h$. Using the differential equation, $\varphi_j^{m+\theta}$ can be written as

$$\begin{aligned}
\varphi_j^{m+\theta} = &\left[\frac{u(x_j, t^{m+1}) - u(x_j, t^m)}{\Delta t} - \frac{\partial u}{\partial t}(x_j, t^{m+\theta}) \right] \\
&+ \left[\frac{\partial^2 u}{\partial x^2}(x_j, t^{m+\theta}) - D_x^+ D_x^- (\theta u(x_j, t^{m+1}) + (1-\theta) u(x_j, t^m)) \right].
\end{aligned} \tag{6.27}$$

In order to estimate the size of the truncation error, $\varphi_j^{m+\theta}$, we expand it into a Taylor series about the point $(x_j, t^{m+1/2})$.

$$\begin{aligned}
u_j^{m+1} &= \left[u + \frac{\Delta t}{2}\frac{\partial u}{\partial t} + \frac{1}{2}\left(\frac{\Delta t}{2}\right)^2 \frac{\partial^2 u}{\partial t^2} + \frac{1}{6}\left(\frac{\Delta t}{2}\right)^3 \frac{\partial^3 u}{\partial t^3} + \dots \right]_j^{m+1/2} \\
u_j^m &= \left[u - \frac{\Delta t}{2}\frac{\partial u}{\partial t} + \frac{1}{2}\left(\frac{\Delta t}{2}\right)^2 \frac{\partial^2 u}{\partial t^2} - \frac{1}{6}\left(\frac{\Delta t}{2}\right)^3 \frac{\partial^3 u}{\partial t^3} + \dots \right]_j^{m+1/2}.
\end{aligned}$$

If we subtract the second of these expansions from the first, all the even-numbered terms will cancel, and we obtain

$$\frac{u(x_j,t^{m+1})-u(x_j,t^m)}{\Delta t} = \left[\frac{\partial u}{\partial t}+\frac{1}{24}(\Delta t)^2\frac{\partial^3 u}{\partial t^3}+\ldots\right]_j^{m+1/2}. \tag{6.28}$$

Also, since

$$D_x^+D_x^-u(x_j,t^{m+1}) = \left[\frac{\partial^2 u}{\partial x^2}+\frac{1}{12}h^2\frac{\partial^4 u}{\partial x^4}+\frac{2}{6!}h^4\frac{\partial^6 u}{\partial x^6}+\ldots\right]_j^{m+1},$$

expanding the right-hand side about the point $(x_j,t^{m+1/2})$,

$$\begin{aligned} D_x^+D_x^-u(x_j,t^{m+1}) = &\left[\frac{\partial^2 u}{\partial x^2}+\frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}+\frac{2h^4}{6!}\frac{\partial^6 u}{\partial x^6}+\ldots\right]_j^{m+1/2}\\ &+\frac{\Delta t}{2}\left[\frac{\partial^3 u}{\partial x^2\partial t}+\frac{h^2}{12}\frac{\partial^5 u}{\partial x^4\partial t}+\ldots\right]_j^{m+1/2}\\ &+\frac{1}{2}\left(\frac{\Delta t}{2}\right)^2\left[\frac{\partial^4 u}{\partial x^2\partial t^2}+\ldots\right]_j^{m+1/2}. \end{aligned}$$

There is a similar expansion for $D_x^+D_x^-u(x_j,t^m)$; combining these we obtain:

$$\begin{aligned} D_x^+D_x^-[\theta u(x_j,t^{m+1})+(1-\theta)u(x_j,t^m)] = &\left[\frac{\partial^2 u}{\partial x^2}+\frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}+\frac{2h^4}{6!}\frac{\partial^6 u}{\partial x^6}+\ldots\right]_j^{m+1/2}\\ &+(\theta-\tfrac{1}{2})\Delta t\left[\frac{\partial^3 u}{\partial x^2\partial t}+\frac{h^2}{12}\frac{\partial^5 u}{\partial x^4\partial t}+\ldots\right]_j^{m+1/2}\\ &+\frac{1}{8}(\Delta t)^2\left[\frac{\partial^4 u}{\partial x^2\partial t^2}+\ldots\right]_j^{m+1/2}. \end{aligned} \tag{6.29}$$

Substituting (6.28) and (6.29) into (6.27):

$$\begin{aligned} \varphi_j^{m+\theta} = &\left[(\tfrac{1}{2}-\theta)\Delta t\frac{\partial^3 u}{\partial x^2\partial t}-\frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}\right]_j^{m+1/2}\\ &+(\Delta t)^2\left[\frac{1}{24}\frac{\partial^3 u}{\partial t^3}-\frac{1}{8}\frac{\partial^4 u}{\partial x^2\partial t^2}\right]_j^{m+1/2}\\ &+h^2\left[\frac{1}{12}(\tfrac{1}{2}-\theta)\Delta t\frac{\partial^5 u}{\partial x^4\partial t}-\frac{2}{6!}h^2\frac{\partial^6 u}{\partial x^6}+\ldots\right]_j^{m+1/2}\\ &+f(x_j,t^{m+1/2})-f(x_j,t^{m+\theta}). \end{aligned}$$

Thence

$$\left|\varphi_j^{m+\theta}\right| \le \frac{h^2}{12}M_{4x}+\frac{\Delta t^2}{24}(M_{3t}+3M_{2x2t})+H.O.T., \quad \theta=\tfrac{1}{2}, \tag{6.30}$$

$$\left|\varphi_j^{m+\theta}\right| \le \left|\tfrac{1}{2}-\theta\right| \Delta t(M_{2t}+2M_{2x1t}) + \frac{h^2}{12}M_{4x} + H.O.T., \quad \theta \neq \tfrac{1}{2}, \tag{6.31}$$

where

$$M_{kxlt} = \max_{(x,t)\in\bar{Q}} \left| \frac{\partial^{k+l}}{\partial x^k \partial t^l} u(x,t) \right|.$$

Substituting (6.30) into (6.25) and (6.31) into (6.25) or (6.26) we obtain the following error bounds:

$$\max_{1\le m\le M} \|u^m - U^m\|_h \le C_1(h^2+\Delta t^2), \quad \theta = \tfrac{1}{2}, \tag{6.32}$$

where C_1 is a positive constant, independent of h and Δt;

$$\max_{1\le m\le M} \|u^m - U^m\|_h \le C_2(h^2+\Delta t), \quad \theta \in (1/2, 1], \tag{6.33}$$

where C_2 is a positive constant, independent of h and Δt. Moreover,

$$\max_{1\le m\le M} \|u^m - U^m\|_h \le C_3(h^2+\Delta t), \quad \theta \in [0, 1/2), \tag{6.34}$$

where $C_3 = (c_\epsilon)^{1/2} \cdot C_2$, provided that

$$\Delta t \le \frac{h^2}{2(1-2\theta)}(1-\epsilon), \quad \epsilon \in (0,1), \quad \theta \in [0, 1/2).$$

Thus we deduce that the Crank–Nicolson scheme ($\theta = 1/2$) converges in the norm $\|\cdot\|_h$ 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 difference 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 this theory is beyond the scope of these notes.

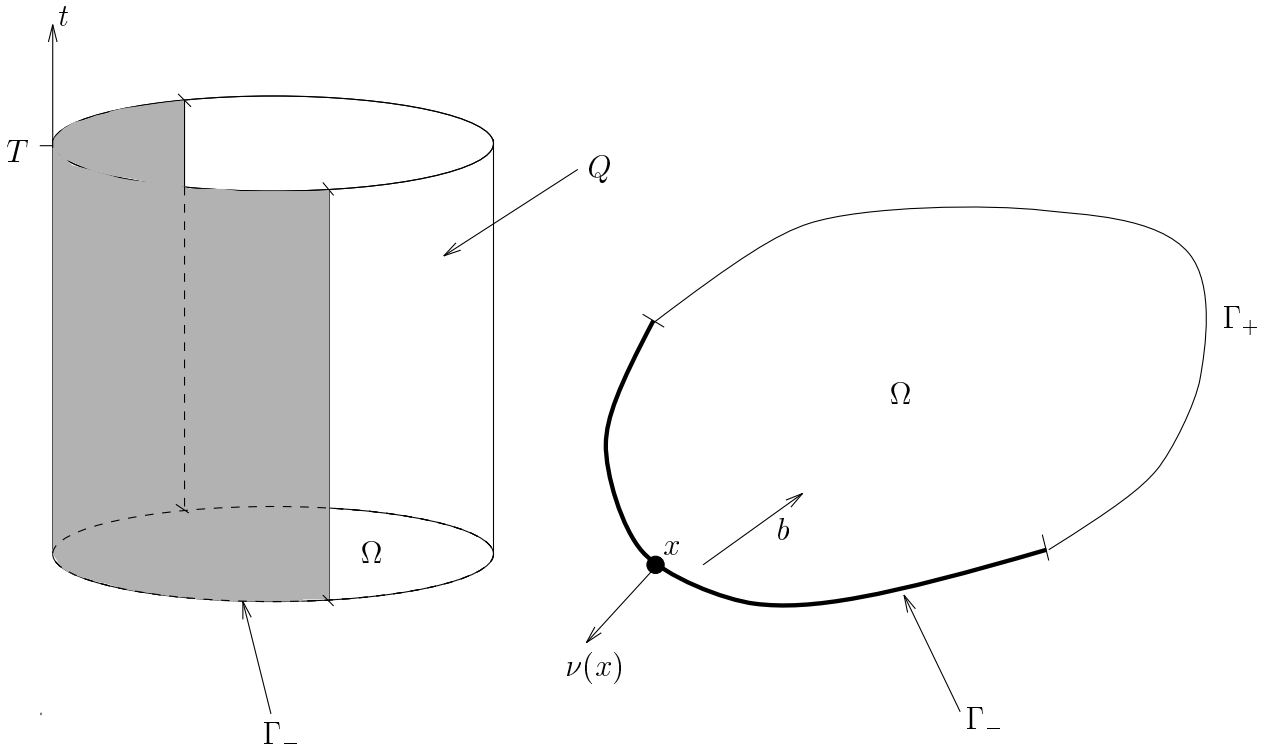
6.2 Finite difference methods for hyperbolic equations

Let Ω be a bounded open set in $\mathbb{R}^n$, $n \ge 1$, with boundary $\Gamma = \partial\Omega$, and let $T > 0$. In $Q = \Omega \times (0, T]$, we consider the initial boundary value problem

$$\frac{\partial u}{\partial t} + \sum_{i=1}^{n} b_i(x) \cdot \frac{\partial u}{\partial x_i} + c(x,t)u = f(x,t), \quad x \in \Omega, \quad t \in (0, T], \tag{6.35}$$

$$u(x,t) = 0, \quad x \in \Gamma_-, \quad t \in [0, T], \tag{6.36}$$

$$u(x,0) = u_0(x) \quad x \in \bar{\Omega}, \tag{6.37}$$



where

$$\Gamma_- = \{x \in \Gamma : b(x) \cdot \nu(x) < 0\},$$

$b = (b_1, \ldots, b_n)$ and $\nu(x)$ denotes the unit outward normal to Γ at $x \in \Gamma$.

Γ_- will be called the inflow boundary. Its complement, $\Gamma_+ = \Gamma \backslash \Gamma_-$, will be referred to as the outflow boundary. It is important to note that, unlike parabolic equations where a boundary condition is specified on the whole of $\Gamma \times [0, T]$, in a hyperbolic initial boundary value problem the boundary condition is only imposed on part of the boundary, namely on $\Gamma_- \times [0, T]$, or else the problem may have no solution.

We shall assume that

$$b_i \in C^1(\bar{\Omega}), \quad i = 1, \ldots, n, \tag{6.38a}$$
$$c \in C(\bar{Q}), \quad f \in L^2(Q), \tag{6.38b}$$
$$u_0 \in L_2(\Omega). \tag{6.38c}$$

In order to ensure consistency between the initial and the boundary condition, we shall suppose that $u_0(x) = 0$, $x \in \Gamma_-$.

The existence of a unique solution (at least for c, $f \in C^1(\bar{Q})$, $u_0 \in C^1(\bar{\Omega})$) can be shown using the method of characteristics. More generally, for b_i, c, f, u_0, obeying the smoothness

requirements of (6.38), a unique solution still exists, but the proof of this result is beyond the scope of these notes. Let us, instead, consider the behaviour of the solution of (6.35)–(6.37) in time.

We make the additional hypothesis:

$$c(x,t) - \frac{1}{2}\sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i}(x) \geq 0, \quad x \in \bar{\Omega}, \;\; t \in [0,T]. \tag{6.39}$$

Taking the inner product of (6.35) with u in $L_2(\Omega)$, we obtain:

$$\begin{aligned}\left(\frac{\partial u}{\partial t}, u\right) + \left(c(\cdot,t) - \frac{1}{2}\sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i}(\cdot), u^2\right) \\ + \frac{1}{2}\int_{\Gamma_+} \left[\sum_{i=1}^{n} b_i(x)\nu_i(x)\right] u^2(x,t)\,\mathrm{d}s(x) = (f,u),\end{aligned} \tag{6.40}$$

where $\nu(x) = (\nu_1(x), \ldots, \nu_n(x))$ is the unit outward normal vector to Γ at $x \in \Gamma$. By virtue of (6.39) and noting that

$$\begin{aligned}\left(\frac{\partial u}{\partial t}, u\right) &= \int_\Omega \frac{\partial u}{\partial t}(x,t) \cdot u(x,t)\,\mathrm{d}x \\ &= \int_\Omega \frac{1}{2}\frac{\partial}{\partial t} u^2(x,t)\,\mathrm{d}x = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u^2(x,t)\,\mathrm{d}x \\ &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|u(\cdot,t)\right\|^2,\end{aligned}$$

it follows from (6.40) that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|u(\cdot,t)\right\|^2 + \frac{1}{2}\int_{\Gamma_+} \left[\sum_{i=1}^{n} b_i(x)\nu_i(x)\right] u^2(x,t)\,\mathrm{d}s(x) \leq (f,u).$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned}(f,u) &\leq \left\|f(\cdot,t)\right\| \cdot \left\|u(\cdot,t)\right\| \\ &\leq \tfrac{1}{2}\left\|f(\cdot,t)\right\|^2 + \tfrac{1}{2}\left\|u(\cdot,t)\right\|^2,\end{aligned}$$

and therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|u(\cdot,t)\right\|^2 + \int_{\Gamma_+} \left[\sum_{i=1}^{n} b_i(x)\nu_i(x)\right] u^2(x,t)\,\mathrm{d}s(x) - \left\|u(\cdot,t)\right\|^2 \leq \left\|f(\cdot,t)\right\|^2, \quad t \in [0,T].$$

Multiplying both sides by e^{-t}, this can be rewritten as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{-t}\left\|u(\cdot,t)\right\|^2 + \mathrm{e}^{-t}\int_{\Gamma_+} \left[\sum_{i=1}^{n} b_i(x)\nu_i(x)\right] u^2(x,t)\,\mathrm{d}s \leq \mathrm{e}^{-t}\left\|f(\cdot,t)\right\|^2, \quad t \in [0,T].$$

Integrating this inequality with respect to t yields

$$\mathrm{e}^{-t}\,\|u(\cdot,t)\|^2 \;+\; \int_0^t \mathrm{e}^{-\tau} \int_{\Gamma_+} \left[\sum_{i=1}^n b_i(x)\nu_i(x)\right] u^2(x,\tau)\,\mathrm{d}s(x)\,\mathrm{d}\tau$$
$$\leq \|u_0\|^2 \;+\; \int_0^t \mathrm{e}^{-\tau}\,\|f(\cdot,\tau)\|^2\;\mathrm{d}\tau, \quad t\in[0,T].$$

Hence

$$\|u(\cdot,t)\|^2 \;+\; \int_0^t \mathrm{e}^{t-\tau} \int_{\Gamma_+} \left[\sum_{i=1}^n b_i(x)\nu_i(x)\right] u^2(x,\tau)\,\mathrm{d}s(x)\,\mathrm{d}\tau$$
$$\leq \mathrm{e}^t\,\|u_0\|^2 \;+\; \int_0^t \mathrm{e}^{t-\tau}\,\|f(\cdot,\tau)\|^2\;\mathrm{d}\tau, \quad t\in[0,T]. \tag{6.41}$$

This, so called, energy inequality expresses the continuous dependence of the solution to (6.35)–(6.37) on the data. In particular it can be used to prove the uniqueness of the solution. Indeed, if u_1 and u_2 are solutions of (6.35)–(6.37), then $u := u_1 - u_2$ also solves (6.35)–(6.37), with $f \equiv 0$ and $u_0 \equiv 0$. Thus, by (6.41), $\|u(\cdot,t)\| = 0$, $t\in[0,T]$ and therefore $u \equiv 0$, i.e. $u_1 \equiv u_2$.

Let us consider a particularly important case when

$$c \equiv 0, \quad f \equiv 0, \quad \text{and} \quad \operatorname{div} b = \sum_{i=1}^n \frac{\partial b_i}{\partial x_i} \equiv 0,$$

where $b(x) = (b_1(x),\ldots,b_n(x))$. Then, by virtue of (6.40),

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\,\|u(\cdot,t)\|^2 + \frac{1}{2}\int_{\Gamma_+} [b(x)\cdot\nu(x)]\,u^2(x,t)\,\mathrm{d}s(x) = 0,$$

and therefore,

$$\|u(\cdot,t)\|^2 + \int_0^t \int_{\Gamma_+} [b(x)\cdot\nu(x)]\,u^2(x,\tau)\,\mathrm{d}s(x)\,\mathrm{d}\tau = \|u_0\|^2\,, \tag{6.42}$$

which expresses the conservation of energy in the physical system modelled by (6.35)–(6.37).

6.2.1 Explicit finite difference scheme

In this section we describe a simple explicit finite difference scheme for the numerical solution of the constant-coefficient hyperbolic equation in one space dimension:

$$\frac{\partial u}{\partial t} + b\frac{\partial u}{\partial x} = f(x,t), \quad x\in(0,1), \quad t\in(0,T], \tag{6.43}$$

subject to the boundary and initial conditions

$$u(x,t) = 0, \qquad x \in \Gamma_-, \quad t \in [0,T], \tag{6.44a}$$
$$u(x,0) = u_0(x), \quad x \in [0,1]. \tag{6.44b}$$

If $b > 0$ then $\Gamma_- = \{0\}$, and if $b < 0$ then $\Gamma_- = \{1\}$. Let us assume, for example, that $b > 0$. Then the appropriate boundary condition is

$$u(0,t) = 0, \quad t \in [0,T]. \tag{6.45}$$

To construct a finite difference approximation of (6.43)–(6.45) let $h = 1/N$ be the mesh-size in the x-direction and $\Delta t = T/M$ the mesh-size in the time-direction, t. Let us also define

$$x_j = jh, \quad j = 0, \ldots, N, \quad t^m = m \cdot \Delta t, \quad m = 0, \ldots, M.$$

At the mesh-point (x_j, t^m), (6.43) is approximated by the explicit finite difference scheme

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} + b \cdot D_x^- U_j^m = f(x_j, t^m), \quad \begin{array}{l} j = 1, \ldots, N, \\ m = 0, \ldots, M-1, \end{array} \tag{6.46}$$

$$U_0^m = 0, \qquad m = 0, \ldots, M, \tag{6.47}$$
$$U_j^0 = u_0(x_j), \quad j = 0, \ldots, N. \tag{6.48}$$

Equivalently,

$$U_j^{m+1} = (1-\mu)U_j^m + \mu U_{j-1}^m + \Delta t f(x_j, t^m), \quad \begin{array}{l} j = 1, \ldots, N, \\ m = 0, \ldots, M-1, \end{array}$$

$$\begin{aligned} U_0^m &= 0, \qquad m = 0, \ldots, M, \\ U_j^0 &= u_0(x_j), \quad j = 0, \ldots, N, \end{aligned}$$

where

$$\mu = \frac{b\Delta t}{h};$$

μ is called the Courant number.

Suppose that $0 \le \mu \le 1$; then,

$$\begin{aligned} \left|U_j^{m+1}\right| &\le (1-\mu)\left|U_j^m\right| + \mu\left|U_{j-1}^m\right| + \Delta t\left|f(x_j, t^m)\right| \\ &\le (1-\mu) \max_{0 \le j \le N} \left|U_j^m\right| + \mu \max_{1 \le j \le N+1} \left|U_{j-1}^m\right| + \Delta t \max_{0 \le j \le N} \left|f(x_j, t^m)\right| \\ &= \max_{0 \le j \le N} \left|U_j^m\right| + \Delta t \max_{0 \le j \le N} \left|f(x_j, t^m)\right|. \end{aligned}$$

Hence

$$\max_{0\le j\le N}\left|U_j^{m+1}\right| \le \max_{0\le j\le N}\left|U_j^m\right| + \Delta t \max_{0\le j\le N}\left|f(x_j,t^m)\right|.$$

Let us define the mesh-dependent norm

$$\|U\|_\infty = \max_{0\le j\le N}|U_j|;$$

then,

$$\left\|U^{m+1}\right\|_\infty \le \|U^m\|_\infty + \Delta t\,\|f(\cdot,t^m)\|_\infty, \qquad m=0,\dots,M-1.$$

Summing through m, we get

$$\max_{1\le k\le M}\left\|U^k\right\|_\infty \le \left\|U^0\right\|_\infty + \sum_{m=0}^{M-1}\Delta t\,\|f(\cdot,t^m)\|_\infty,$$

which expresses the stability of the finite difference scheme (6.46)–(6.48) under the condition

$$0\le\mu=\frac{b\Delta t}{h}\le 1.$$

Thus we have proved that (6.46)–(6.48) is conditionally stable in the $\|\cdot\|_\infty$ norm, the condition being that the Courant number, μ, is in the interval $[0,1]$.

It is possible to show that the scheme (6.46)–(6.48) is also stable in the mesh-dependent L^2-norm, $\|\cdot\|_h$. Recall that

$$\|V\|_h^2 = \sum_{i=1}^N hV_i^2.$$

The associated inner product is

$$(V,W]_h = \sum_{i=1}^N hV_iW_i.$$

Since

$$U_j^m = \frac{U_j^m+U_{j-1}^m}{2} + \frac{U_j^m-U_{j-1}^m}{2},$$

and $U_0^m=0$, it follows that

$$\begin{aligned}(U^m, D_x^- U^m]_h &= \sum_{j=1}^N hU_j^m\frac{U_j^m-U_{j-1}^m}{h}\\ &= \frac{1}{2}\sum_{j=1}^N\{(U_j^m)^2-(U_{j-1}^m)^2\} + \frac{h}{2}\sum_{j=1}^N h\left(\frac{U_j^m-U_{j-1}^m}{h}\right)^2 \qquad (6.49)\\ &= \frac{1}{2}(U_N^m)^2 + \frac{h}{2}\left\|D_x^-U^m\right]\!\!\right\|_h^2.\end{aligned}$$

In addition, since

$$U_j^m = \frac{U_j^{m+1} + U_j^m}{2} - \frac{U_j^{m+1} - U_j^m}{2}, \quad m = 0, \ldots, M-1,$$

we have that

$$\left(\frac{U^{m+1} - U^m}{\Delta t}, U^m \right]_h = \frac{1}{2\Delta t} \left(\left\| U^{m+1} \right]_h^2 - \left\| U^m \right\|_h^2 \right) \tag{6.50}$$

$$- \frac{\Delta t}{2} \left\| \frac{U^{m+1} - U^m}{\Delta t} \right]_h^2, \quad m = 0, \ldots, M-1. \tag{6.51}$$

Thus, taking the $(\cdot, \cdot]_h$-inner product of (6.46) with U^m and using (6.49) and (6.51),

$$\left\| U^{m+1} \right]_h^2 + \Delta t \cdot b (U_N^m)^2 + bh\Delta t \left\| D_x^- U^m \right]_h^2 - \left\| U^m \right\|_h^2$$
$$- \Delta t^2 \left\| \frac{U^{m+1} - U^m}{\Delta t} \right]_h^2 = 2\Delta t (f^m, U^m]_h, \quad m = 0, \ldots, M-1. \tag{6.52}$$

First suppose that $f \equiv 0$; then,

$$\frac{U^{m+1} - U^m}{\Delta t} = -b \cdot D_x^- U^m,$$

so that

$$\left\| U^{m+1} \right]_h^2 + \Delta t \cdot b \left| U_N^m \right|^2 + bh\Delta t (1-\mu) \left\| D_x^- U^m \right]_h^2 = \left\| U^m \right\|_h^2, \quad m = 0, \ldots, M-1.$$

Summing through m,

$$\left\| U^k \right]_h^2 + \sum_{m=0}^{k-1} \Delta t \cdot b \left| U_N^m \right|^2 + bh(1-\mu) \sum_{m=0}^{k-1} \Delta t \left\| D_x^- U^m \right]_h^2 = \left\| U^0 \right]_h^2, \quad k = 1, \ldots, M, \tag{6.53}$$

which proves the stability of the scheme in the case when $f \equiv 0$ under the assumption that

$$0 \leq \mu = \frac{b\Delta t}{h} \leq 1.$$

In particular, if $\mu = 1$, we have that

$$\left\| U^k \right]_h^2 + \sum_{m=0}^{k-1} \Delta t \cdot b \left| U_N^m \right|^2 = \left\| U^0 \right]_h^2, \quad k = 1, \ldots, M,$$

which is the discrete version of the identity (6.41), and expresses conservation of energy in the discrete sense. This is equality is also trivially valid when $\mu = 0$ (i.e. when $b = 0$).

More generally, for $0 \leq \mu \leq 1$, (6.53) implies

$$\left\| U^k \right]_h^2 + \sum_{m=0}^{k-1} \Delta t \cdot b \left| U_N^m \right|^2 \leq \left\| U^0 \right]_h^2, \quad k = 1, \ldots, M,$$

with strict inequality when $0 < \mu < 1$. Therefore, when $0 < \mu < 1$ the discrete energy dissipates even through, as we have shown in (6.42), the continuous counterpart of the discrete energy is conserved. This feature of the first-order upwind scheme is also quite evident in numerical experiments: as time evolves, the numerical solution will be seen to be smeared in comparison with the analytical solution.

Now let us consider the question of stability in the $\|\cdot\|_h$-norm in the general case of $f \not\equiv 0$. Since

$$\left\| \frac{U^{m+1} - U^m}{\Delta t} \right\|_h^2 = \left\| f^m - bD_x^- U^m \right\|_h^2 \leq \{ \|f^m\|_h + b \left\| D_x^- U^m \right\|_h \}^2$$
$$\leq \left(1 + \frac{1}{\epsilon'} \right) \|f^m\|_h^2 + (1+\epsilon') b^2 \left\| D_x^- U^m \right\|_h^2, \quad \epsilon' > 0,$$

and

$$(f^m, U^m]_h \leq \|f^m\|_h \|U^m\|_h \leq \tfrac{1}{2} \|f^m\|_h^2 + \tfrac{1}{2} \|U^m\|_h^2,$$

it follows from (6.52) that

$$\left\| U^{m+1} \right\|_h^2 + \Delta t \cdot b \left| U_n^m \right|^2 + bh\Delta t \left[1 - (1+\epsilon') \frac{b\Delta t}{h} \right] \left\| D_x^- U^m \right\|_h^2$$
$$\leq \Delta t \left[\left(1 + \frac{1}{\epsilon'} \right) \Delta t + 1 \right] \|f^m\|_h^2 + (1 + \Delta t) \|U^m\|_h^2 .$$

Letting $\epsilon = 1 - 1/(1+\epsilon') \in (0,1)$, and assuming

$$0 \leq \mu = \frac{b\Delta t}{h} \leq 1 - \epsilon,$$

we have, for $m = 0, \ldots, M-1$,

$$\left\| U^{m+1} \right\|_h^2 + \Delta t \cdot b \left| U_N^m \right|^2 \leq \|U^m\|_h^2 + \Delta t \left(1 + \frac{\Delta t}{\epsilon} \right) \|f^m\|_h^2 + \Delta t \|U^m\|_h^2 .$$

Upon summation,

$$\left\| U^k \right\|_h^2 + \left(\sum_{m=0}^{k-1} \Delta t \cdot b \left| U_N^m \right|^2 \right) \leq \left\| U^0 \right\|_h^2 + \left(1 + \frac{\Delta t}{\epsilon} \right) \sum_{m=0}^{k-1} \Delta t \|f^m\|_h^2 + \sum_{m=0}^{k-1} \Delta t \|U^m\|_h^2 . \tag{6.54}$$

for $k = 1, \ldots, M$. The next lemma is easily proved by induction.

Lemma 6.1 *Let (a_k), (b_k), (c_k) and (d_k) be four sequences of non-negative numbers such that the sequence (c_k) is non-decreasing and*

$$a_k + b_k \leq c_k + \sum_{m=0}^{k-1} d_m a_m, \quad k \geq 1; \;\; a_0 + b_0 \leq c_0 .$$

Then

$$a_k + b_k \le c_k \exp\left(\sum_{m=0}^{k-1} d_m\right), \quad k \ge 1.$$

Applying this lemma to (6.54) with

$$\begin{aligned}
a_k &= \left\|U^k\right]\!\right|_h^2, \quad k \ge 0, \\
b_k &= \sum_{m=0}^{k-1} \Delta t \cdot b \left|U_N^m\right|^2, \quad k \ge 1; \;\; b_0 = 0, \\
c_k &= \left\|U^0\right]\!\right|_h^2 + \left(1 + \frac{\Delta t}{\epsilon}\right) \sum_{m=0}^{k-1} \Delta t \left\|f^m\right]\!\right|_h^2, \quad k \ge 1; \;\; c_0 = \left\|U^0\right]\!\right|_h^2, \\
d_k &= \Delta t, \quad k = 1, 2, \ldots, M,
\end{aligned}$$

we obtain,

$$\left|\left|U^k\right]\!\right|_h^2 + \sum_{m=0}^{k-1} \Delta t \cdot b \cdot |U_N^m|^2 \le \mathrm{e}^{t^k} \left(\left\|U^0\right]\!\right|_h^2 + \left(1 + \frac{\Delta t}{\epsilon}\right) \sum_{m=0}^{k-1} \Delta t \left\|f^m\right]\!\right|_h^2\right), \quad k = 1, \ldots, M,$$

and hence stability:

$$\max_{1 \le k \le M} \left(\left\|U^k\right]\!\right|_h^2 + \sum_{m=0}^{k-1} \Delta t \cdot b \cdot |U_N^m|^2\right) \le \mathrm{e}^T \left(\left\|U^0\right]\!\right|_h^2 + \left(1 + \frac{\Delta t}{\epsilon}\right) \sum_{m=0}^{M-1} \Delta t \left\|f^m\right]\!\right|_h^2\right). \quad (6.55)$$

An error estimate for the difference scheme (6.46)–(6.48) is easily derived from stability.

We define the global error, e, and the truncation error, φ, by

$$\begin{aligned}
e_j^m &= u(x_j, t^m) - U_j^m, \\
\varphi_j^m &= \frac{u(x_j, t^{m+1}) - u(x_j, t^m)}{\Delta t} - bD_x^- u(x_j, t^m) - f(x_j, t^m).
\end{aligned}$$

It is easily seen that

$$\begin{aligned}
\frac{e_j^{m+1} - e_j^m}{\Delta t} + bD_x^- e_j^m &= \varphi_j^m, \qquad j = 1, \ldots, N, \quad m = 0, \ldots, M-1, \\
e_0^m &= 0, \qquad m = 0, \ldots, M, \\
e_j^0 &= 0, \qquad j = 0, \ldots, N.
\end{aligned}$$

By virtue of the stability inequality established in the first part of this section,

$$\max_{1 \le m \le M} \|e^m\|_\infty \le \sum_{k=0}^{M-1} \Delta t \, \|\varphi^m\|_\infty . \quad (6.56)$$

By Taylor series expansion of φ_j^m about the point (x_j, t^m),

$$\varphi_j^m = \frac{1}{2}\Delta t \frac{\partial^2 u}{\partial t^2}(x_j, \tau^m) + \frac{1}{2} bh \frac{\partial^2 u}{\partial x^2}(\xi_j, t^m), \quad \tau^m \in (t^m, t^{m+1}), \quad \xi_j \in (x_{j-1}, x_j),$$

so that

$$\left|\varphi_j^m\right| \leq \frac{1}{2}(\Delta t M_{2t} + bh M_{2x}),$$

where

$$M_{kxlt} = \max_{(x,t)\in\bar{Q}} \left| \frac{\partial^{k+l}}{\partial x^k \partial t^l}(x,t) \right|.$$

Defining $M = \max(M_{2t}, M_{2x})$, we have

$$\left|\varphi_j^m\right| \leq \frac{1}{2} M(\Delta t + bh) \quad (= \mathcal{O}(h + \Delta t)). \tag{6.57}$$

Thus, by (6.56),

$$\max_{1\leq m\leq M} \|u^m - U^m\|_\infty \leq \frac{1}{2} T M(\Delta t + bh);$$

so the scheme (6.46)–(6.48) is first-order convergent.

Analogously, using the stability result (6.54) in the discrete L^2-norm $\|\cdot\|_h$, (6.57) implies that

$$\max_{1\leq m\leq M} \|u^m - U^m\|_h \leq c_\epsilon^\star \cdot (\Delta t + bh),$$

where $c_\epsilon^\star = \frac{1}{2} \mathrm{e}^{T/2}(1 + T/\epsilon)^{1/2} T^{1/2} M$.

The analysis presented here can be extended to linear first-order hyperbolic equations with variable coefficients and to hyperbolic problems in more than one space-dimension, as well as to difference schemes on non-uniform meshes.