B6.1 Numerical Solution of Partial Differential Equations

A brief introduction to the theory of finite difference approximation of partial differential equations

Endre Süli
Mathematical Institute, University of Oxford

April 29, 2021

Contents

Preface. The purpose of these lecture notes is to provide an introduction to computational methods for the approximate solution of partial differential equations (PDEs), by focusing on the construction and the mathematical analysis of the conceptually simplest class of algorithms, finite difference methods for second-order elliptic partial differential equations, initial-boundary-value problems for second-order parabolic equations, and first- and second-order hyperbolic partial differential equations. Only minimal prerequisites in differential and integral calculus, mathematical analysis and linear algebra are assumed.

The notes begin with some basic background from the theory of function spaces that are required in the mathematical analysis of numerical methods for PDEs. The rest of the course focuses on classical techniques for the numerical solution of boundary-value problems for second-order ordinary differential equations and elliptic boundary-value problems, in particular the Poisson equation in two dimensions. Key ideas include: discretization using the finite difference method, stability and convergence analysis, and the use of the discrete maximum principle. The remaining lectures are devoted to the numerical solution of initial-boundary-value problems for second-order parabolic and first- and second-order hyperbolic partial differential equations with topics such as: approximation by finite difference methods, accuracy, stability (including the Courant–Friedrichs–Lewy (CFL) condition) and convergence.

Syllabus and course structure.

Lecture 1: Overview of the lecture course and motivating examples from various applications in the sciences. Basic background from the theory of function spaces.
Lecture 2: Finite difference approximation of two-point boundary-value problems for second-order ODEs. Mesh-dependent inner-products and mesh-dependent norms. Discrete Poincaré inequality.
Lecture 3: Stability, consistency and convergence of finite difference approximations of two-point boundary-value problems.
Lecture 4: Second-order linear elliptic boundary-value problems and their finite difference approximation: uniform meshes on axiparallel domains; nonuniform meshes on nonaxiparallel domains.
Lecture 5: Discrete maximum principle; stability and convergence in the discrete maximum norm.
Lecture 6: Discrete energy estimates; stability and convergence in discrete Sobolev norms.
Lecture 7: Iterative solution of systems of linear equations arising from the discretization of second-order linear elliptic PDEs: linear stationary iterative methods.
Lecture 8: Second-order parabolic initial-value problems and their finite difference approximation: spatial semi-discretization via the method of lines; fully discrete explicit and implicit schemes.
Lecture 9: Discrete Fourier analysis of finite-difference approximations of initial-value problems for second-order linear parabolic PDEs: the Courant–Friedrichs–Lewy (CFL) condition.
Lecture 10: Finite difference approximation of initial-boundary-value problems for second-order parabolic PDEs.
Lecture 11: Discrete maximum principle for finite difference approximations of initial-boundary-value problems for second-order parabolic PDEs; stability and convergence in the discrete maximum norm.
Lecture 12: Discrete energy norm estimates for finite difference approximations of initial-boundary-value problems for second-order parabolic problems: stability, consistency and convergence.
Lecture 13: Finite-difference approximation of second-order linear hyperbolic equations.
Lecture 14: Discrete energy estimates for second-order hyperbolic problems: stability (including the CFL condition), consistency and convergence.
Lecture 15: Finite difference approximation of linear first-order hyperbolic equations: stability (including the CFL condition), consistency and convergence.
Lecture 16: Finite difference approximation of nonlinear first-order hyperbolic conservation laws with convex nonlinearities. The first-order upwind scheme: boundedness of the sequence of approximate solutions in the discrete maximum norm.

Reading List:

[1] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations.* (Cambridge University Press, second edition, 2009). ISBN 978-0-521-73490-5. [Chapters 8–10, 17].

[2] B.S. Jovanović and E. Süli, *Analysis of Finite Difference Schemes for Linear Partial Differential Equations with Generalized Solutions.* (Springer, 2014). ISBN 978-1-447-15461-7. [Sections 2.1, 2.2, 2.3, 3.1, 3.2, 4.1, 4.2].

[3] R. LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations.* (SIAM, 2007). ISBN 978-0-898716-29-0. [Chapter 10].

[4] K.W. Morton and D.F. Mayers, *Numerical Solution of Partial Differential Equations: An Introduction.* (Cambridge University Press, second edition, 2012). ISBN 978-0-521-60793-3. [Chapters 2–7].

Note: These lecture notes will be updated regularly during Michaelmas Term.

Note about the exercises: There will be 4 problem sheets and 4 classes associated with the lectures.

Introduction

Partial differential equations arise in mathematical models of numerous phenomena in science and engineering, and they also frequently occur in problems that originate from economics and finance. In most cases the equations concerned are so complicated that their solution by analytical means (e.g. by Laplace or Fourier transform based techniques or in the form of an infinite series) is either impossible or impracticable, and one has to resort to numerical techniques for their approximate solution. **Lecture 1**

These notes are devoted to the construction and the mathematical analysis of the conceptually simplest class of numerical techniques, finite difference methods, for the approximate solution of elliptic, parabolic and hyperbolic partial differential equations, by considering simple model problems. Preference is given to theoretical results concerning the stability and the accuracy of numerical methods – properties that are of key importance in practical computations.

1 Elements of function spaces

The accuracy of a numerical method for the approximate solution of partial differential equations depend on its capability 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 nonnegative integers. An n-tuple $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$ is called a *multi-index*. The nonnegative integer $|\alpha| := \alpha_1 + \cdots + \alpha_n$ is called the length of the multi-index $\alpha = (\alpha_1, \dots, \alpha_n)$. We denote $(0, \dots, 0)$ by $\mathbf{0}$; clearly $|\mathbf{0}| = 0$.

Let

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

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, we have that

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

We shall frequently write ∂_{x_j} instead of the more cumbersome expression $\frac{\partial}{\partial x_j}$. ⋄

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 $\alpha = (\alpha_1, \dots, \alpha_n)$ with $|\alpha| \le k$. Assuming

that Ω is a *bounded* open set, $C^k(\overline{\Omega})$ 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 $\overline{\Omega}$, the closure of the set Ω, for all $\alpha = (\alpha_1, \ldots, \alpha_n)$ with $|\alpha| \le k$. The linear space $C^k(\overline{\Omega})$ can then be equipped with the norm

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

In particular, when $k = 0$, we shall write $C(\overline{\Omega})$ instead of $C^0(\overline{\Omega})$;

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

Similarly, if $k = 1$,

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

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 all $k \ge 0$. Since $\overline{\Omega} = [0, 1]$, it is clear that u is not continuous on $\overline{\Omega}$; the same is true of its derivatives. Therefore $u \notin C^k(\overline{\Omega})$ for any $k \ge 0$. $\diamond$

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 $|x| := (x_1^2 + \cdots + x_n^2)^{1/2}$ for $x \in \mathbb{R}^n$. Clearly, supp w is the closed unit ball $\{x \in \mathbb{R}^n : |x| \le 1\}$. $\diamond$

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

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

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, $p \ge 1$; 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 an inner product

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

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

We note in passing that a subset of $\mathbb{R}^n$ is said to be a *set of measure zero* if it can be contained in the union of countably many open balls of arbitrarily small total volume. For example, the set of all rational numbers is a set of measure zero in $\mathbb{R}$.

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

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

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

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

The right-hand side is a quadratic polynomial in λ with real coefficients which is nonnegative for all $\lambda \in \mathbb{R}$. Therefore its discriminant is nonpositive, i.e.,

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

and hence the desired inequality. □

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

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

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 $\|u\|_X = (u, u)^{1/2}_X$) is called a Hilbert space if, whenever $\{u_m\}_{m=1}^\infty$ is a Cauchy sequence in X, i.e. a sequence of elements of X such that

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

then there exists a $u \in X$ such that $\lim_{m\to\infty} \|u - u_m\|_X = 0$ (i.e., the sequence $\{u_m\}_{m=1}^\infty$ converges to u in the norm of 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:

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

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 $\overline{\Omega} \subset \Omega$.) Suppose also that there exists a function w_α, locally integrable on Ω, and such that

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

Then we say that w_α is the *weak derivative* of u (of order $|\alpha| = \alpha_1 + \cdots + \alpha_n$) 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, nevertheless, possess a weak derivative. Indeed, for any $v \in C_0^\infty(\Omega)$, we have that

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

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. $w = u' = Du$. ⋄

Now we are ready to give a precise definition of a Sobolev space. Let k be a nonnegative integer. We define (with D^α denoting a weak derivative of order $|\alpha|$)

$$H^k(\Omega) = \{u \in L_2(\Omega) : D^\alpha u \in L_2(\Omega), \quad |\alpha| \le k\}.$$

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

$$\|u\|_{H^k(\Omega)} := \left(\sum_{|\alpha| \le k} \|D^\alpha u\|_{L_2(\Omega)}^2 \right)^{1/2}$$

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 $H^2(\Omega)$.

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

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

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

Similarly,

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

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

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

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 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 $u \in H_0^1(\Omega)$; then, there exists a positive constant $c_\star(\Omega)$, independent of u, such that*

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

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,

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

Thus, by the Cauchy–Schwarz inequality,

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

Analogously,

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

By adding the two inequalities, we obtain

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

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

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 **Lecture 2**

$$\Delta u = 0,$$

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

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left(a_{i,j}(x)\frac{\partial u}{\partial x_i}\right) + \sum_{i=1}^{n} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f(x), \quad x \in \Omega, \tag{2}$$

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

$$\begin{aligned} a_{i,j} &\in C^1(\overline{\Omega}), && i,j = 1,\ldots,n; \\ b_i &\in C(\overline{\Omega}), && i = 1,\ldots,n; \\ c &\in C(\overline{\Omega}), \quad f \in C(\overline{\Omega}), && \text{and} \\ \sum_{i,j=1}^{n} a_{i,j}(x)\xi_i\xi_j &\geq \tilde{c}\sum_{i=1}^{n}\xi_i^2, && \forall\, \xi = (\xi_1,\ldots,\xi_n) \in \mathbb{R}^n, \quad \forall\, x \in \overline{\Omega}; \end{aligned} \tag{3}$$

here $\tilde{c}$ is a positive constant independent of x and ξ. The condition (3) is usually referred to as uniform ellipticity and (2) is called an elliptic equation.

Equation (2) is supplemented with one of the following boundary conditions:

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

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

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

(d) A more general version of the boundary conditions (b) and (c) is

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

where α_j is the angle between the unit outward normal vector ν to $\partial\Omega$ and the Ox_j axis (Oblique derivative boundary condition).

In many physical problems more than one type of boundary condition 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

$$-\sum_{i,j=1}^{n}\frac{\partial}{\partial x_j}\left(a_{i,j}(x)\frac{\partial u}{\partial x_i}\right)+\sum_{i=1}^{n}b_i(x)\frac{\partial u}{\partial x_i}+c(x)u=f(x),\qquad \text{for } x\in\Omega, \tag{4}$$

$$u=0\qquad\qquad \text{on } \partial\Omega, \tag{5}$$

where $a_{i,j}$, b_i, c and f are as in (3).

A function $u\in C^2(\Omega)\cap C(\overline{\Omega})$ satisfying (4) and (5) is called a classical solution of this problem. The theory of partial differential equations tells us that (4), (5) has a unique classical solution, provided $a_{i,j}$, 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 $\mathbb{R}^n$:

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

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

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 (4), (5). Then, for any $v\in C_0^1(\Omega)$,

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

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

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

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,

$$H_0^1(\Omega)=\left\{u\in L^2(\Omega):\frac{\partial u}{\partial x_i}\in L^2(\Omega),\quad i=1,\ldots,n,\quad u=0\text{ on }\partial\Omega\right\}.$$

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

$$\sum_{i,j=1}^{n}\int_\Omega a_{i,j}(x)\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}\,\mathrm{d}x+\sum_{i=1}^{n}\int_\Omega b_i(x)\frac{\partial u}{\partial x_i}v\,\mathrm{d}x+\int_\Omega c(x)uv\,\mathrm{d}x=\int_\Omega f(x)v(x)\,\mathrm{d}x\quad\forall\, v\in C_0^1(\Omega). \tag{6}$$

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

Definition 1 *Let $a_{i,j}\in C(\overline{\Omega})$, $i,j=1,\ldots,n$, $b_i\in C(\overline{\Omega})$, $i=1,\ldots,n$, $c\in C(\overline{\Omega})$, and let $f\in L^2(\Omega)$. A function $u\in H_0^1(\Omega)$ satisfying*

$$\sum_{i,j=1}^{n}\int_\Omega a_{i,j}(x)\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}\,\mathrm{d}x+\sum_{i=1}^{n}\int_\Omega b_i(x)\frac{\partial u}{\partial x_i}v\,\mathrm{d}x+\int_\Omega c(x)uv\,\mathrm{d}x=\int_\Omega f(x)v(x)\,\mathrm{d}x\qquad\forall\, v\in H_0^1(\Omega) \tag{7}$$

is called a weak solution *of (4), (5). All partial derivatives in (7) should be understood as weak derivatives.*

Clearly if u is a classical solution of (4), (5), then it is also a weak solution of (4), (5). However, the converse is not true. If (4), (5) 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 (4), (5).

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

$$a(u,v) := \sum_{i,j=1}^{n} \int_\Omega a_{i,j}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} \,\mathrm{d}x + \sum_{i=1}^{n} \int_\Omega b_i(x) \frac{\partial u}{\partial x_i} v \,\mathrm{d}x + \int_\Omega c(x) uv \,\mathrm{d}x \tag{8}$$

and

$$\ell(v) := \int_\Omega f(x) v(x) \,\mathrm{d}x. \tag{9}$$

With this new notation, problem (7) can be written as follows:

$$\text{find } u \in H_0^1(\Omega) \text{ such that } a(u,v) = \ell(v) \quad \forall\, v \in H_0^1(\Omega). \tag{10}$$

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

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

(a) There exists a $c_0 > 0$ such that $a(v,v) \geq c_0 \|v\|_V^2$ for all $v \in V$;

(b) There exists a $c_1 > 0$ such that $|a(v,w)| \leq c_1 \|v\|_V \|w\|_V$ for all $v, w \in V$;

and let $\ell(\cdot)$ be a linear form on V such that

(c) There exists a $c_2 > 0$ such that $|\ell(v)| \leq c_2 \|v\|_V$ for all $v \in V$.

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

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

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 (4), (5) (or, equivalently, to (10)). Let us recall from Section 1.3 that $H_0^1(\Omega)$ is a Hilbert space with the inner product

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

and the associated norm $\|u\|_{H^1(\Omega)} = (u,u)_{H^1(\Omega)}^{1/2}$. Next we show that $a(\cdot,\cdot)$ and $\ell(\cdot)$, defined by (8) and (9), satisfy the hypotheses (a), (b), (c) of the Lax–Milgram theorem.

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

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

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

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

for all $v \in H_0^1(\Omega)$, where we have used the obvious inequality $\|v\|_{L^2(\Omega)} \le \|v\|_{H^1(\Omega)}$. Letting $c_2 = \|f\|_{L^2(\Omega)}$, we obtain the required bound.

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

$$\begin{aligned}|a(u,v)| &\le \sum_{i,j=1}^n \max_{x\in\overline{\Omega}}|a_{i,j}(x)|\left|\int_\Omega \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}\,\mathrm{d}x\right| + \sum_{i=1}^n \max_{x\in\overline{\Omega}}|b_i(x)|\left|\int_\Omega \frac{\partial u}{\partial x_i}v\,\mathrm{d}x\right| + \max_{x\in\overline{\Omega}}|c(x)|\left|\int_\Omega u(x)v(x)\,\mathrm{d}x\right|\\ &\le c\left\{\sum_{i,j=1}^n\left(\int_\Omega\left|\frac{\partial u}{\partial x_i}\right|^2\mathrm{d}x\right)^{1/2}\left(\int_\Omega\left|\frac{\partial v}{\partial x_j}\right|^2\mathrm{d}x\right)^{1/2} + \sum_{i=1}^n\left(\int_\Omega\left|\frac{\partial u}{\partial x_i}\right|^2\mathrm{d}x\right)^{1/2}\left(\int_\Omega|v|^2\,\mathrm{d}x\right)^{1/2}\right.\\ &\quad\left. + \left(\int_\Omega|u|^2\,\mathrm{d}x\right)^{1/2}\left(\int_\Omega|v|^2\,\mathrm{d}x\right)^{1/2}\right\}\\ &\le c\left\{\left(\int_\Omega|u|^2\,\mathrm{d}x\right)^{1/2} + \sum_{i=1}^n\left(\int_\Omega\left|\frac{\partial u}{\partial x_i}\right|^2\mathrm{d}x\right)^{1/2}\right\}\left\{\left(\int_\Omega|v|^2\,\mathrm{d}x\right)^{1/2} + \sum_{j=1}^n\left(\int_\Omega\left|\frac{\partial v}{\partial x_j}\right|^2\mathrm{d}x\right)^{1/2}\right\},\end{aligned} \tag{11}$$

where

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

By further majorization of the right-hand side in (11),

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

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

It remains to establish (a). Using (3), we deduce that

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

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

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

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

$$c(x) - \frac{1}{2}\sum_{i=1}^n\frac{\partial b_i}{\partial x_i} \ge 0, \quad x \in \overline{\Omega}. \tag{12}$$

Then

$$a(u,u) \geq \tilde{c} \sum_{i=1}^{n} \int_{\Omega} \left| \frac{\partial u}{\partial x_i} \right|^2 \, \mathrm{d}x. \tag{13}$$

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

$$a(u,u) \geq \frac{\tilde{c}}{c_\star} \int_{\Omega} |u|^2 \, \mathrm{d}x. \tag{14}$$

Summing (13) and (14),

$$a(u,u) \geq c_0 \left(\int_{\Omega} |u|^2 \, \mathrm{d}x + \sum_{i=1}^{n} \int_{\Omega} \left| \frac{\partial u}{\partial x_i} \right|^2 \, \mathrm{d}x \right), \tag{15}$$

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 (10); thence problem (4), (5) has a unique weak solution.

We encapsulate this result in the following theorem.

Theorem 2 *Suppose that $a_{i,j} \in C(\overline{\Omega})$, $i,j = 1,\ldots,n$, $b_i \in C^1(\overline{\Omega})$, $i = 1,\ldots,n$, $c \in C(\overline{\Omega})$, $f \in L^2(\Omega)$, and assume that (3) and (12) hold; then the boundary value problem (4), (5) possesses a unique weak solution $u \in H_0^1(\Omega)$. In addition,*

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{c_0} \|f\|_{L^2(\Omega)}. \tag{16}$$

Proof. We only have to prove (16). By (15), (10), the Cauchy–Schwarz inequality and recalling the definition of $\|\cdot\|_{H^1(\Omega)}$,

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

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 $a_{i,j}(x) \equiv 1$, $i = j$, $a_{i,j}(x) \equiv 0$, $i \neq j$, $1 \leq i,j \leq n$, $b_i(x) \equiv 0$, $c(x) \equiv 0$, $f(x) = \operatorname{sgn}(\frac{1}{2} - |x|)$, and $\Omega = (-1,1)^n$, we see that (3) holds with $\tilde{c} = 1$ and (12) is trivially fulfilled. Thus (∗) has a unique weak solution $u \in H_0^1(\Omega)$.

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 implies that the weak formulation of the elliptic boundary value problem (4), (5) is well-posed in the sense of Hadamard; namely, for each $f \in L^2(\Omega)$ there exists a unique (weak) solution $u \in H_0^1(\Omega)$, 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 $H_0^1(\Omega)$ of (4), (5) 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 $H_0^1(\Omega)$ of (4), (5) corresponding to the right-hand side $f_1 - f_2 \in L_2(\Omega)$. Thus, by virtue of (16),

$$\|u_1 - u_2\|_{H^1(\Omega)} \leq \frac{1}{c_0} \|f_1 - f_2|_{L^2(\Omega)}, \tag{17}$$

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 **Lecture 3**

$$\begin{aligned} \mathcal{L}u &= f \qquad \text{in } \Omega, \\ \mathcal{B}u &= g \qquad \text{on } \Gamma := \partial\Omega, \end{aligned} \tag{18}$$

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

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

and

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

or

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

or

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

where α_j is the angle between the unit outward normal vector ν to $\partial\Omega$ and the Ox_j axis, — or some other appropriate boundary condition.

In general, it is impossible to determine the solution of the boundary value problem (18) in closed form. Thus the aim of this chapter is to describe a simple and general numerical technique for the approximate solution of (18), 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 $\overline{\Omega} = \Omega \cup \Gamma$ by a finite set of points

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

where $\Omega_h \subset \Omega$ and $\Gamma_h \subset \Gamma$; $\overline{\Omega}_h$ 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

$$\begin{aligned} \mathcal{L}_h U(x) &= f_h(x), \qquad x \in \Omega_h, \\ l_h U(x) &= g_h(x), \qquad x \in \Gamma_h, \end{aligned} \tag{19}$$

where f_h and g_h are suitable approximations of f and g, respectively. Now (19) 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 $\{U(x) : x \in \overline{\Omega}_h\}$ is an approximation to $\{u(x) : x \in \overline{\Omega}_h\}$, 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 (19) approximates the boundary value problem (18) in some sense, and whether its solution $\{U(x) : x \in \overline{\Omega}_h\}$ approximates $\{u(x) : x \in \overline{\Omega}_h\}$, the values of the exact solution at the mesh-points.

(2) the second problem concerns the effective solution of the discrete problem (19) using techniques from Numerical Linear Algebra.

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

3.1 Finite difference approximation of a two-point boundary-value problem

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:

$$\begin{aligned} -u'' + c(x)u &= f(x), \quad x \in (0,1), \\ u(0) = 0, \;\; u(1) &= 0. \end{aligned} \tag{20}$$

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 $\overline{\Omega}_h = \Omega_h \cup \Gamma_h$. Suppose that u is sufficiently smooth (e.g. $u \in C^4[0,1]$). Then, by Taylor series expansion,

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

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:

$$\begin{aligned} -D_x^+ D_x^- u(x_i) + c(x_i)u(x_i) &\approx \quad f(x_i), \quad i = 1, \ldots, N-1, \\ u(x_0) = 0, \;\; u(x_N) &= 0. \end{aligned} \tag{21}$$

Now (21) indicates that the approximate solution U should be sought as the solution of the system of difference equations:

$$\begin{aligned} -D_x^+ D_x^- U_i + c(x_i)U_i &= \quad f(x_i), \quad i = 1, \ldots, N-1, \\ U_0 = 0, \;\; U_N &= 0. \end{aligned} \tag{22}$$

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 tridiagonal $(N-1)\times(N-1)$ matrix displayed above, and U and F are column vectors of size $N-1$.

3.2 Existence and uniqueness of solutions, stability, consistency, and convergence

We begin the analysis of the finite difference scheme (22) 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 $L_2(0,1)$-inner product

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

Lemma 3 *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*

$$(-D_x^+ D_x^- V, V)_h = \sum_{i=1}^{N} h \left| D_x^- V_i \right|^2. \tag{23}$$

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 $V_0 = V_N = 0$. $\square$

Returning to the finite difference scheme (22), let V be as in the above lemma and note that if $c(x) \geq 0$ then

$$\begin{aligned}(AV, V)_h &= (-D_x^+ D_x^- V + cV, V)_h \\ &= (-D_x^+ D_x^- V, V)_h + (cV, V)_h \\ &\geq \sum_{i=1}^{N} h \left| D_x^- V_i \right|^2 . \end{aligned} \tag{24}$$

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 (22) has a unique solution, $U = A^{-1}F$.

Theorem 3 *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 (22) possesses a unique solution U.*

We note that, by virtue of Theorem 3, the boundary value problem (20) has a unique (weak) solution under the same hypotheses on c and f as in Theorem 3.

Next, we investigate the approximation properties of the difference scheme (22). A key ingredient in our analysis is the fact that the scheme (22) 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 (16). 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*

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

where

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

Using this notation, the inequality (24) can be written

$$(AV, V)_h \geq \|D_x^- V]|_h^2. \tag{25}$$

In fact, employing a discrete version of the Poincaré–Friedrichs inequality (1), stated in Lemma 4 below, we shall prove that

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

where c_0 is a positive constant.

Lemma 4 *(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*

$$\|V\|_h^2 \leq c_\star \|D_x^- V]|_h^2 \tag{26}$$

for all such V.

PROOF. We proceed in the same way as in the proof of (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,

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

We note that the constant $c_\star = 1/2$ in (26). $\square$

Using (26) to bound the right-hand side of (25) from below we obtain

$$(AV, V)_h \geq \frac{1}{c_\star} \|V\|_h^2. \tag{27}$$

Adding (25) to (27) we deduce that

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

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

$$(AV, V)_h \geq c_0 \|V\|_{1,h}^2. \tag{28}$$

Now the stability of the finite difference scheme (22) easily follows.

Theorem 4 *The scheme (22) is stable in the sense that*

$$\|U\|_{1,h} \leq \frac{1}{c_0} \|f\|_h. \tag{29}$$

PROOF. From (28) and (22) we have that

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

and hence (29). $\square$

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

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

Thus,

$$\begin{aligned} Ae_i &= \varphi_i, \qquad i = 1, \ldots, N-1, \\ & \qquad e_0 = 0, \quad e_N = 0, \end{aligned} \tag{30}$$

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

Applying (29) to the finite difference scheme (30), we obtain

$$\|u - U\|_{1,h} = \|e\|_{1,h} \leq \frac{1}{c_0} \|\varphi\|_h. \tag{31}$$

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,

$$\|\varphi\|_h = \left(\sum_{i=1}^{N-1} h \, |\varphi_i|^2 \right)^{1/2} \leq Ch^2. \tag{32}$$

Combining (31) and (32), it follows that

$$\|u - U\|_{1,h} \leq \frac{C}{c_0} h^2. \tag{33}$$

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 (32). 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 (33), it follows that

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

Thus we have proved the following result.

Theorem 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 (20) belongs to $C^4[0,1]$; then*

$$\|u - U\|_{1,h} \leq \frac{1}{8} h^2 \|u^{IV}\|_{C[0,1]}. \tag{34}$$

The analysis of the finite difference scheme (20) 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. (29), for example). A typical stability result for the general finite difference scheme (19) is

$$|||U|||_{\Omega_h} \leq c(\|f_h\|_{\Omega_h} + \|g_h\|_{\Gamma_h}), \tag{35}$$

where $|||\cdot|||_{\Omega_h}$, $\|\cdot\|_{\Omega_h}$ and $\|\cdot\|_{\Gamma_h}$ are mesh-dependent norms involving mesh-points of Ω_h (or $\overline{\Omega}_h$) 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*,

$$\begin{aligned} \varphi_{\Omega_h} &= L_h u - f_h, \qquad \text{in} \quad \Omega_h, \\ \varphi_{\Gamma_h} &= l_h u - g_h, \qquad \text{on} \quad \Gamma_h. \end{aligned}$$

(in the case of the finite difference scheme (20) $\varphi_{\Gamma_h} = 0$, 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 (18), we say that the scheme (19) is *consistent.* If p is the largest positive integer such that

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

(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 (19) is said to provide a *convergent* approximation to (18) in the norm $|||\cdot|||_{\Omega_h}$, if

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

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 6 *Suppose that the finite difference scheme (19) is stable (i.e. (35) holds for all f_h and g_h) and that the scheme is a consistent approximation of (18); then (19) is a convergent approximation of (18), 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

$$L_h e = \phi_{\Omega_h},$$

and similarly,

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

By stability,

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

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 **Lecture 4**

$$\begin{aligned} -\Delta u + c(x)u &= f(x) && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{36}$$

where $\Omega = (0,1) \times (0,1)$, c is a continuous function on $\overline{\Omega}$ 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 $\overline{\Omega}$. 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 (36) 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.

(a) ($f \in C(\overline{\Omega})$) The first step in the construction of the finite difference approximation of (36) 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

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

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 $\Gamma_h = \overline{\Omega}_h \setminus \Omega_h$.

Analogously to (22), the difference scheme is:

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) + c(x_i, y_j)U_{i,j} &= f(x_i, y_j), && \text{for } (x_i, y_j) \in \Omega_h, \\ U &= 0 && \text{on } \Gamma_h. \end{aligned} \tag{37}$$

In an expanded form, this can be written as follows:

$$-\left\{ \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} + \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{h^2} \right\} + c(x_i, y_j)U_{i,j} = f(x_i, y_j), \quad i, j = 1, \ldots, N-1, \tag{38}$$

$$U_{i,j} = 0, \text{ if } i = 0, \ i = N \text{ or if } j = 0, \ j = N. \tag{39}$$

For each i and j, $1 \leq i, j \leq N-1$, the finite difference equation (38) 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 (38), (39) as a system of linear equations

$$AU = F, \tag{40}$$

where

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

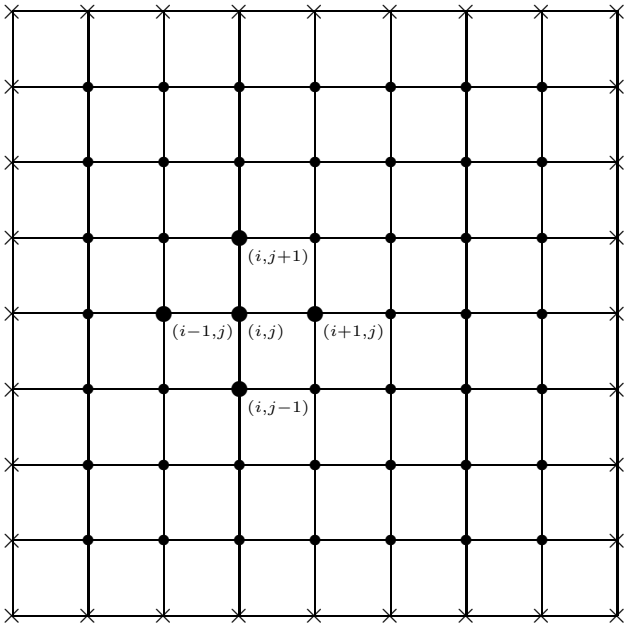


Figure 1: The mesh $\Omega_h(\cdot)$, the boundary mesh $\Gamma_h(\times)$, and a typical 5-point difference stencil.

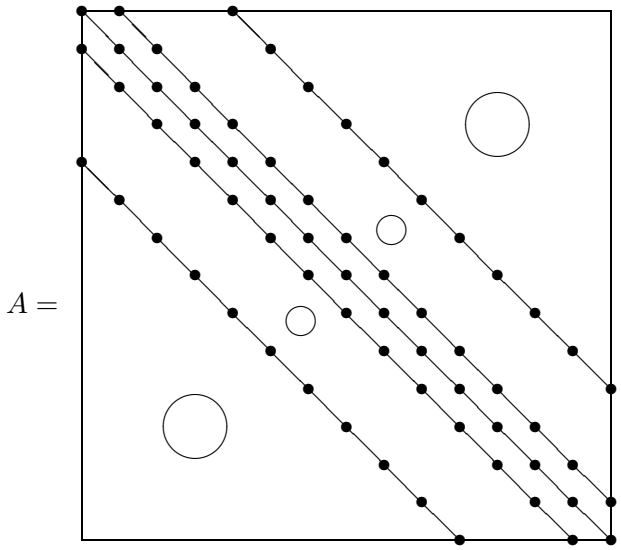


Figure 2: The sparsity structure of the banded matrix A.

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

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.

4.1 Existence and uniqueness of solutions, stability, consistency, and convergence

Next we show that (37) 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

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

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

Lemma 5 *Suppose that V is a function defined on $\overline{\Omega}_h$ and that $V = 0$ on Γ_h; then*

$$(-D_x^+ D_x^- V, V)_h + (-D_y^+ D_y^- V, V)_h = \sum_{i=1}^{N} \sum_{j=1}^{N-1} h^2 |D_x^- V_{i,j}|^2 + \sum_{i=1}^{N-1} \sum_{j=1}^{N} h^2 |D_y^- V_{i,j}|^2. \tag{41}$$

Proof. The identity (41) is a direct consequence of (23) and the analogous identity for $-D_y^+ D_y^-$. □

Returning to the analysis of the finite difference scheme (37), we note that, since $c(x, y) \geq 0$ on $\overline{\Omega}$, by (41) we have

$$\begin{aligned} (AV, V)_h &= (-D_x^+ D_x^- V - D_y^+ D_y^- V + cV, V)_h \\ &= (-D_x^+ D_x^- V, V)_h + (-D_y^+ D_y^- V, V)_h + (cV, V)_h \\ &\geq \sum_{i=1}^{N} \sum_{j=1}^{N-1} h^2 |D_x^- V_{i,j}|^2 + \sum_{i=1}^{N-1} \sum_{j=1}^{N} h^2 |D_y^- V_{i,j}|^2, \end{aligned} \tag{42}$$

for any V defined on $\overline{\Omega}_h$ 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 (42) yields:

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

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

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 (37). Thus the solution of the finite difference scheme (37) may be found by solving the system of linear equations (40).

In order to prove the stability of the finite difference scheme (37), we introduce (similarly as in one dimension) the mesh–dependent norms

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

and

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

where

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

and

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

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

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

With this new notation, the inequality (42) takes the following form:

$$(AV, V)_h \geq \|D_x^- V]|_x^2 + \|D_y^- V]|_y^2. \tag{43}$$

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

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

where c_0 is a positive constant.

Lemma 6 *(Discrete Poincaré–Friedrichs inequality.)*
Let V be a function defined on $\overline{\Omega}_h$ and such that $V = 0$ on Γ_h; then there exists a constant c_*, independent of V and h, such that

$$\|V\|_h^2 \leq c_* \left(\|D_x^- V]|_x^2 + \|D_y^- V]|_y^2 \right) \tag{44}$$

for all such V.

Proof. The inequality (44) is a straightforward consequence of its one-dimensional counterpart (26). It follows from (26) that, for each fixed j, $1 \leq j \leq N-1$,

$$\sum_{i=1}^{N-1} h |V_{i,j}|^2 \leq \frac{1}{2} \sum_{i=1}^{N} h |D_x^- V_{i,j}|^2. \tag{45}$$

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

$$\sum_{j=1}^{N-1} h |V_{i,j}|^2 \leq \frac{1}{2} \sum_{j=1}^{N} h |D_y^- V_{i,j}|^2. \tag{46}$$

We multiply (45) by h and sum through j, $1 \leq j \leq N-1$, multiply (46) by h and sum through i, $1 \leq i \leq N-1$, and add these two inequalities to obtain

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

Hence (44) with $c_* = \frac{1}{4}$. $\square$

Now (43) and (44) imply that

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

Finally, combining this with (43) and recalling the definition of the norm $\|\cdot\|_{1,h}$, we obtain

$$(AV, V)_h \geq c_0\|V\|_{1,h}^2, \tag{47}$$

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

Theorem 7 *The scheme (37) is stable in the sense that*

$$\|U\|_{1,h} \leq \frac{1}{c_0}\|f\|_h. \tag{48}$$

PROOF. The proof of this inequality is identical to that of (29) and is therefore omitted. □

4.1.1 Convergence in the class of classical solutions

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

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

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

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

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

Let

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

then

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

By virtue of (48),

$$\|u - U\|_{1,h} = \|e\|_{1,h} \leq \frac{1}{c_0}\|\varphi\|_h. \tag{49}$$

Noting that

$$|\varphi_{i,j}| \leq \frac{h^2}{12}\left(\left\|\frac{\partial^4 u}{\partial x^4}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^4 u}{\partial y^4}\right\|_{C(\overline{\Omega})}\right),$$

we deduce that the truncation error, φ, satisfies

$$\|\varphi\|_h \leq \frac{h^2}{12}\left(\left\|\frac{\partial^4 u}{\partial x^4}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^4 u}{\partial y^4}\right\|_{C(\overline{\Omega})}\right). \tag{50}$$

Finally (49) and (50) yield the following result.

Theorem 8 *Let $f \in C(\overline{\Omega})$, $c \in C(\overline{\Omega})$, with $c(x,y) \geq 0$, $(x,y) \in \overline{\Omega}$, and suppose that the corresponding weak solution of the boundary value problem (36) belongs to $C^4(\overline{\Omega})$; then*

$$\|u - U\|_{1,h} \leq \frac{5h^2}{48} \left(\left\| \frac{\partial^4 u}{\partial x^4} \right\|_{C(\overline{\Omega})} + \left\| \frac{\partial^4 u}{\partial y^4} \right\|_{C(\overline{\Omega})} \right). \tag{51}$$

PROOF. Recall that $c_0 = (1 + c_*)^{-1}, c_* = \frac{1}{4}$, so that $1/c_0 = \frac{5}{4}$, and combine (49) and (50). □

According to this result, the five-point difference scheme (37) for the boundary value problem (36) is second-order convergent, provided u is sufficiently smooth.

In general, however, even if f and c are smooth functions, the corresponding solution, u, of (36) will not be a smooth function because the boundary, Γ, of the domain, Ω, is a non-smooth curve. Thus, the hypothesis $u \in C^4(\overline{\Omega})$ is unrealistic.

Our analysis has another limitation: it has been performed under the assumption that $f \in C(\overline{\Omega})$ 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 (38) to cater for the fact that f is not necessarily continuous on $\overline{\Omega}$. **Lecture 5**

The idea is to replace $f(x_i, y_j)$ in (38) by a cell-average of f,

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

where

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

This, seemingly *ad hoc* approach, has the following justification. Integrating the partial differential equation $-\Delta u + cu = f$ over the cell $K_{i,j}$, noting that $\Delta u = \nabla \cdot (\nabla u) = \mathrm{div}(\nabla u)$, and using the divergence theorem we have that

$$-\int_{\partial K_{i,j}} \frac{\partial u}{\partial \nu}\,\mathrm{d}l + \int_{K_{i,j}} cu\,\mathrm{d}x\,\mathrm{d}y = \int_{K_{i,j}} f\,\mathrm{d}x\,\mathrm{d}y \tag{$**$}$$

where $\partial K_{i,j}$ is the boundary of $K_{i,j}$, and ν the unit outward normal to $\partial K_{i,j}$. The normal vectors to $\partial K_{i,j}$ 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 $\mathrm{meas}(K_{i,j}) = h^2$, we obtain

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

Remark 1 *Finite difference schemes that arise from integral formulations of a differential equation, such as ($**$), are called finite volume methods.* ◇

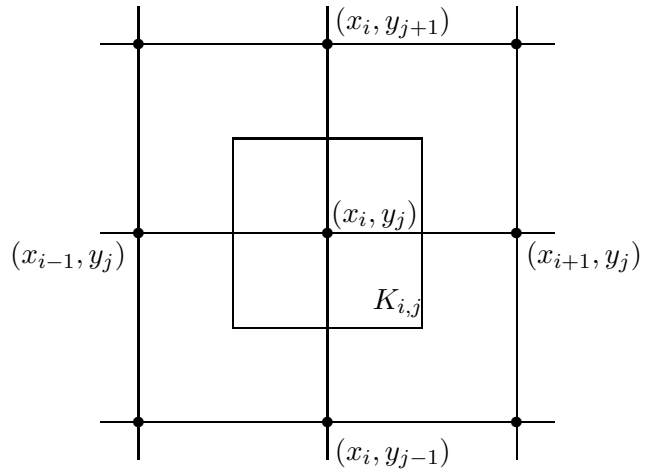


Figure 3: The cell $K_{i,j}$

Clearly, $Tf_{i,j}$ is well defined for f in $L_2(\Omega)$:

$$\begin{aligned} |Tf_{i,j}| &= \frac{1}{h^2}\left|\int_{K_{i,j}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y\right| \\ &\le \frac{1}{h^2}\left(\int_{K_{i,j}} 1^2\,\mathrm{d}x\,\mathrm{d}y\right)^{1/2}\left(\int_{K_{i,j}} |f(x,y)|^2\,\mathrm{d}x\,\mathrm{d}y\right)^{1/2} \\ &= \frac{1}{h}\|f\|_{L^2(K_{i,j})} \le \frac{1}{h}\|f\|_{L^2(\Omega)}. \end{aligned} \tag{52}$$

Thus we define our finite difference (or, more precisely, finite volume) approximation of (36) by

$$\begin{aligned} -(D_x^+D_x^-U_{i,j} + D_y^+D_y^-U_{i,j}) + c(x_i,y_j)U_{i,j} &= Tf_{i,j}, \qquad && \text{for } (x_i,y_j)\in\Omega_h, \\ U &= 0, && \text{on } \Gamma_h. \end{aligned} \tag{53}$$

Since we have not changed the difference operator on the left-hand side, the argument presented on page 28 still applies, and therefore (53) has a unique solution, U.

Theorem 9 *The scheme (53) is stable in the sense that*

$$\|U\|_{1,h} \le \frac{1}{c_0}\|f\|_{L^2(\Omega)}. \tag{54}$$

Proof. According to (47) and (52),

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

and hence (54). □

Having established the stability of the scheme (53), we consider the question of its accuracy. Let us define the global error, e, as before,

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

Clearly,

$$\begin{aligned} Ae_{i,j} &= Au(x_i, y_j) - AU_{i,j} \\ &= Au(x_i, y_j) - Tf_{i,j} \\ &= -(D_x^+ D_x^- u(x_i, y_j) + D_y^+ D_y^- u(x_i, y_j)) + c(x_i, y_j)u(x_i, y_j) \\ &\quad + \left(T\left(\frac{\partial^2 u}{\partial x^2}\right)(x_i, y_j) + T\left(\frac{\partial^2 u}{\partial y^2}\right)(x_i, y_j) - T(cu)(x_i, y_j) \right). \end{aligned} \tag{55}$$

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,

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

(55) 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,

$$\begin{aligned} Ae &= D_x^+ \varphi_1 + D_y^+ \varphi_2 + \psi \qquad && \text{in } \Omega_h, \\ e &= 0 && \text{on } \Gamma_h. \end{aligned} \tag{56}$$

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

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

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

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

we shall proceed in a different way. According to (47),

$$\begin{aligned} c_0 \|e\|_{1,h}^2 &\le (Ae, e)_h \\ &= (D_x^+ \varphi_1, e)_h + (D_y^+ \varphi_2, e)_h + (\psi, e)_h. \end{aligned} \tag{57}$$

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,

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

Thus,

$$(D_x^+\varphi_1, e)_h \leq |\!|\!|\varphi_1]\!|\!|_x |\!|\!|D_x^- e]\!|\!|_x. \tag{58}$$

Similarly,

$$(D_y^+\varphi_2, e)_h \leq |\!|\!|\varphi_2|\!|\!|_y |\!|\!|D_y^- e]\!|\!|_y \tag{59}$$

(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

$$(\psi, e)_h \leq |\!|\!|\psi|\!|\!|_h |\!|\!|e|\!|\!|_h. \tag{60}$$

Upon substituting (58)–(60) into (57) we obtain

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

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

Lemma 7 *The global error, e, of the finite difference scheme (53) satisfies*

$$|\!|\!|e|\!|\!|_{1,h} \leq \frac{1}{c_0} (|\!|\!|\varphi_1]\!|\!|_x^2 + |\!|\!|\varphi_2]\!|\!|_y^2 + |\!|\!|\psi|\!|\!|_h^2)^{1/2}, \tag{61}$$

where φ_1, φ_2, and ψ are defined by

$$\varphi_1(x_i, y_j) = \frac{1}{h} \int_{y_j - h/2}^{y_j + h/2} \frac{\partial u}{\partial x}(x_i - h/2, y)\, \mathrm{d}y - D_x^- u(x_i, y_j), \tag{62}$$

$$\varphi_2(x_i, y_j) = \frac{1}{h} \int_{x_i - h/2}^{x_i + h/2} \frac{\partial u}{\partial y}(x, y_j - h/2)\, \mathrm{d}x - D_y^- u(x_i, y_j), \tag{63}$$

$$\psi(x_i,y_j) = (cu)(x_i,y_j) - \frac{1}{h^2}\int_{x_i-h/2}^{x_i+h/2}\int_{y_j-h/2}^{y_j+h/2}(cu)(x,y)\,\mathrm{d}x\,\mathrm{d}y, \qquad (64)$$
$$i = 1,\ldots,N-1, \quad j = 1,\ldots,N.$$

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

$$|\varphi_1(x_i,y_j)| \le \frac{h^2}{24}\left(\left\|\frac{\partial^3 u}{\partial x\partial y^2}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^3 u}{\partial x^3}\right\|_{C(\overline{\Omega})}\right), \qquad (65)$$

$$|\varphi_2(x_i,y_j)| \le \frac{h^2}{24}\left(\left\|\frac{\partial^3 u}{\partial x^2\partial y}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^3 u}{\partial y^3}\right\|_{C(\overline{\Omega})}\right), \qquad (66)$$

$$|\psi(x_i,y_j)| \le \frac{h^2}{24}\left(\left\|\frac{\partial^2 (cu)}{\partial x^2}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^2 (cu)}{\partial y^2}_{C(\overline{\Omega})}\right\|\right), \qquad (67)$$

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

Theorem 10 *Let $f \in L_2(\Omega)$, $c \in C^2(\overline{\Omega})$ with $c(x,y) \ge 0$, $(x,y) \in \overline{\Omega}$, and suppose that the corresponding weak solution of the boundary value problem (36) belongs to $C^3(\overline{\Omega})$. Then*

$$\|u-U\|_{1,h} \le \frac{5}{96}h^2 M_3, \qquad (68)$$

where

$$M_3 = \left\{\left(\left\|\frac{\partial^3 u}{\partial x\partial y^2}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^3 u}{\partial x^3}\right\|_{C(\overline{\Omega})}\right)^2 + \left(\left\|\frac{\partial^3 u}{\partial x^2 y}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^3 u}{\partial y^3}\right\|_{C(\overline{\Omega})}\right)^2 + \left(\left\|\frac{\partial^2 (cu)}{\partial x^2}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^2 (cu)}{\partial y^2}\right\|_{C(\overline{\Omega})}\right)^2\right\}^{1/2}.$$

Proof. Recalling that $1/c_0 = 5/4$ and substituting (65) - (67) into the right-hand side of (61), (68) immediately follows. □

4.1.2 Convergence in the class of weak solutions

Comparing (68) with (51), we see that while the smoothness requirement on the solution has been relaxed from $u \in C^4(\overline{\Omega})$ to $u \in C^3(\overline{\Omega})$, second-order convergence has been retained.

The hypothesis $u \in C^3(\overline{\Omega})$ can be further relaxed by using integral representations of φ_1, φ_2 and ψ instead of Taylor series expansions. We show how this is done for φ_1; φ_2 and ψ are handled analogously. The key idea is to use the Newton–Leibniz formula

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

Thus, denoting $x_{i\pm1/2} = x_i \pm h/2$ and $y_{j\pm1/2} = y_j \pm h/2$, we have

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

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,

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

Consequently,

$$\|\varphi_1]|_x^2 \le \frac{h^4}{32} \left(\left\| \frac{\partial^3 u}{\partial x^3} \right\|_{L^2(\Omega)}^2 + \left\| \frac{\partial^3 u}{\partial x \partial y^2} \right\|_{L^2(\Omega)}^2 \right). \tag{69}$$

Analogously,

$$\|\varphi_2]|_y^2 \le \frac{h^4}{32} \left(\left\| \frac{\partial^3 u}{\partial y^3} \right\|_{L^2(\Omega)}^2 + \left\| \frac{\partial^3 u}{\partial x^2 \partial y} \right\|_{L^2(\Omega)}^2 \right). \tag{70}$$

In order to estimate ψ, we note that

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

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,

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

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

$$\|\psi\|_h^2 \le \frac{3h^4}{64} \left(\left\| \frac{\partial^2 w}{\partial x^2} \right\|_{L^2(\Omega)}^2 + \left\| \frac{\partial^2 w}{\partial y^2} \right\|_{L^2(\Omega)}^2 + 4 \left\| \frac{\partial^2 w}{\partial x \partial y} \right\|_{L^2(\Omega)}^2 \right). \tag{71}$$

Substituting (69)–(71) into the right-hand side of (61) and recalling that $1/c_0 = 4/5$, we obtain the following result.

Theorem 11 *Let $f \in L^2(\Omega)$, $c \in C^2(\overline{\Omega})$, with $c(x,y) \geq 0$, $(x,y) \in \overline{\Omega}$, and suppose that the corresponding weak solution of the boundary value problem (36) belongs to $H^3(\Omega)$. Then*

$$\|u - U\|_{1,h} \leq Ch^2 \|u\|_{H^3(\Omega)}, \tag{72}$$

where C is a positive constant (computable from (69)–(71)).

It can be shown that the error estimate (72) 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, (72) is an optimal error estimate for the difference scheme (53), but (68) is not.

We have used integral representations of differences to show the bounds (69)–(71). Alternatively one can use the following abstract device.

Start of optional material

Lemma 8 *(The Bramble–Hilbert Lemma) Suppose $\Phi : H^k(\Omega) \to \mathbb{R}$ is a linear functional, 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 (69) for φ_1. Let $K = [-1/2, 1/2] \times [-1/2, 1/2]$, and consider the affine mapping

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

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

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

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,

$$|\Phi(\bar{u})| \leq \int_{-1/2}^{1/2} \left| \frac{\partial \bar{u}}{\partial s}(0,t) \right| \mathrm{d}t + 2 \max_{(s,t) \in K} |\bar{u}(s,t)| \, . \tag{73}$$

Lemma 9 *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

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

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

Equipped with the inequalities (a) and (b), we return to (73). 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

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

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

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

and therefore,

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

Consequently,

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

Therefore,

$$\|\varphi_1]|_x \leq C_1 h^2 \, |u|_{H^3(\Omega)} \, . \tag{74}$$

Similarly,

$$\|\varphi_2]|_y \leq C_2 h^2 \, |u|_{H^3(\Omega)} \tag{75}$$

and

$$\|\psi\|_h \leq C_3 h^2 \, |u|_{H^2(\Omega)} \, . \tag{76}$$

The bounds (74)–(76) derived by using the Bramble–Hilbert lemma are essentially the same as those obtained earlier by integral representations, and stated in (69)–(71). There is, however, an important practical difference: while the constants involved in (69)–(71) are known, those which appear in (74)–(76) (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.

End of optional material

4.2 Nonaxiparallel domains and nonuniform meshes

We have carried out an error analysis of finite difference schemes for the partial differential equation **Lecture 6**

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

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

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

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

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

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. (73) 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

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

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

$$\hbar\!\!\!^{-}_j = \frac{1}{2}(k_{j+1} + k_j).$$

Let

$$\begin{aligned}
D_y^+U_j &= \frac{U_{j+1}-U_j}{\hbar\!\!\!^{-}_j}, \quad D_y^-U_j = \frac{U_j - U_{j-1}}{k_j},\\
D_y^+D_y^-U_j &= \frac{1}{\hbar\!\!\!^{-}_j}\left(\frac{U_{j+1}-U_j}{k_{j+1}} - \frac{U_j-U_{j-1}}{k_j}\right).
\end{aligned}$$

So on a general non-uniform mesh

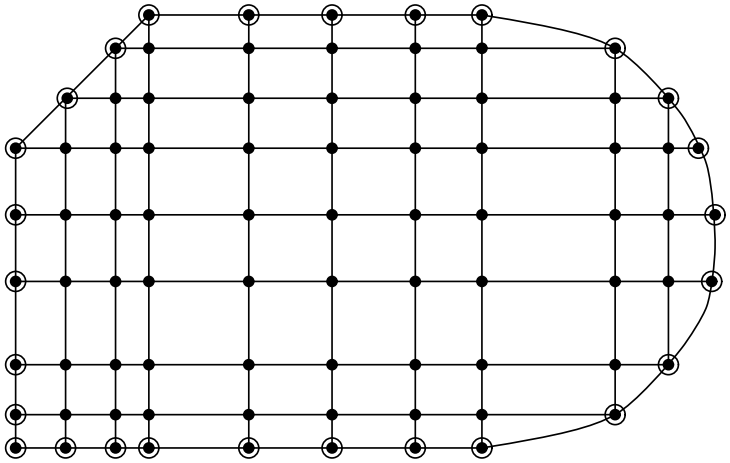
$$\overline{\Omega}_h = \{(x_i,y_j) : x_{i+1}-x_i = h_i,\ y_{j+1}-y_j = k_j\},$$

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

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

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



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

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

The finite difference approximation of this boundary value problem is

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) &= f(x_i, y_j) && \text{in } \Omega_h, \\ U_{i,j} &= 0 && \text{on } \Gamma_h. \end{aligned}$$

Equivalently,

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

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

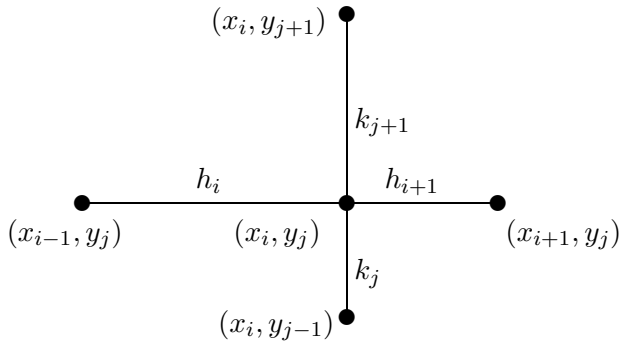


Figure 5: Five-point stencil on a non-uniform mesh.

4.3 The discrete maximum principle

The maximum principle is a key property of elliptic equations. Under a suitable sign-conditions on the source term in the equation and the coefficients of the differential operator, it (roughly speaking) ensures that the maximum value of the solution is attained at the boundary of the domain rather than at an interior point, and if the maximum value of the solution happens to have been attained at an interior point, then the solution must be constant.

To motivate the discussion that will follow, let us begin by considering the two-point boundary-value problem

$$-u''(x) = f(x), \quad x \in (a,b); \qquad u(a) = A, \quad u(b) = B.$$

By integrating twice and imposing the boundary conditions in order to fix the integration constants, one finds that

$$u(x) = \frac{b-x}{b-a}\int_a^x (t-a)f(t)\,\mathrm{d}t + \frac{x-a}{b-a}\int_x^b (b-t)f(t)\,\mathrm{d}t + \left(1 - \frac{x-a}{b-a}\right)A + \frac{x-a}{b-a}B, \qquad a \le x \le b.$$

Hence, if $f(x) \le 0$ for all $x \in [a,b]$, then

$$u(x) \le \left(1 - \frac{x-a}{b-a}\right)A + \frac{x-a}{b-a}B, \qquad a \le x \le b,$$

i.e. the solution curve is below the line connecting the points with coordinates (a,A) and (b,B), and therefore, in particular

$$u(x) \le \max(A,B), \qquad a \le x \le b.$$

Hence the maximum value of u is attained at the boundary, — a property that is usually referred to as *maximum principle*.

Analogously, if $f(x) \ge 0$ for all $x \in [a,b]$, then

$$u(x) \ge \left(1 - \frac{x-a}{b-a}\right)A + \frac{x-a}{b-a}B, \qquad a \le x \le b,$$

i.e. the solution curve is above the line connecting the points with coordinates (a,A) and (b,B), and therefore, in particular

$$u(x) \ge \min(A,B), \qquad a \le x \le b.$$

Hence the minimum value of u is attained at the boundary, — a property that is usually referred to as *minimum principle*.

It would be far too tedious to use a direct calculation to prove a maximum principle for the multidimensional counterpart of the two-point boundary-value problem considered above: i.e., for

$$-\Delta u = f(x), \quad x \in \Omega, \qquad u|_{\partial\Omega} = g,$$

where $\Omega \subset \mathbb{R}^n$ is a bounded open set, $f \in C(\Omega)$ and $g \in C(\partial\Omega)$. We shall therefore show the maximum principle for this problem by an indirect, contradiction-based, argument.

Suppose first that $f(x) < 0$ for all $x \in \Omega$ and that $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is a (classical) solution to the above boundary-value problem, i.e. $-\Delta u(x) = f(x)$ for all $x \in \Omega$ and $u|_{\partial\Omega} = g$. We shall prove that the maximum value of u is then attained on $\partial\Omega$. Suppose otherwise, that u attains its maximum value at $x_0 \in \Omega$. Then,

$$\frac{\partial u}{\partial x_i}(x_0) = 0, \qquad i = 1, \ldots, n$$

and

$$\frac{\partial^2 u}{\partial x_i^2}(x_0) \le 0, \qquad i = 1, \ldots, n.$$

Hence,

$$-\Delta u(x_0) = -\sum \frac{\partial^2 u}{\partial x_i^2}(x_0) \geq 0,$$

which contradicts the assumption that $f(x) < 0$ for all $x \in \Omega$. The maximum value of u must be therefore attained on $\partial\Omega$.

Let us now show that a maximum principle still holds under the weaker assumption $f(x) \leq 0$ for all $x \in \Omega$. To this end, we consider the auxiliary function $v \in C^2(\Omega) \cap C(\overline{\Omega})$ defined by

$$v(x) := u(x) + \frac{\varepsilon}{2n}(x_1^2 + \cdots + x_n^2),$$

where $\varepsilon > 0$. Then, $-\Delta v(x) = -\Delta u(x) - \varepsilon = f(x) - \varepsilon < 0$ for all $x \in \Omega$. Hence, by what we have previously proved, v attains its maximum value on the boundary $\partial\Omega$ of Ω. Consequently,

$$\begin{aligned}
\max_{x\in\partial\Omega} u(x) &= \max_{x\in\partial\Omega} \left[v(x) - \frac{\varepsilon}{2n}(x_1^2 + \cdots + x_n^2) \right] \\
&\geq \max_{x\in\partial\Omega} v(x) - \max_{x\in\partial\Omega} \left[\frac{\varepsilon}{2n}(x_1^2 + \cdots + x_n^2) \right] \\
&= \max_{x\in\overline{\Omega}} v(x) - \max_{x\in\partial\Omega} \left[\frac{\varepsilon}{2n}(x_1^2 + \cdots + x_n^2) \right] \\
&= \max_{x\in\overline{\Omega}} v(x) - \frac{\varepsilon}{2n} \max_{x\in\partial\Omega} |x|^2.
\end{aligned}$$

As $v(x) = u(x) + \frac{\varepsilon}{2n}|x|^2 \geq u(x)$, it follows that

$$\max_{x\in\partial\Omega} u(x) \geq \max_{x\in\overline{\Omega}} u(x) - \frac{\varepsilon}{2n} \max_{x\in\partial\Omega} |x|^2$$

for all $\varepsilon > 0$. Since the expression on the left-hand side of this inequality is independent of ε, as is the first term on the right-hand side, by passing to the limit $\varepsilon \to 0_+$ we deduce that

$$\max_{x\in\partial\Omega} u(x) \geq \max_{x\in\overline{\Omega}} u(x).$$

As $\partial\Omega \subset \overline{\Omega}$, trivially $\max_{x\in\overline{\Omega}} u(x) \geq \max_{x\in\partial\Omega} u(x)$. Therefore,

$$\boxed{\max_{x\in\partial\Omega} u(x) = \max_{x\in\overline{\Omega}} u(x).}$$

Thus we have shown that, if $f(x) \leq 0$ in Ω, then the maximum value of u is attained on the boundary $\partial\Omega$ of the domain Ω, which completes the proof of the *maximum principle*.

Analogously, if $-\Delta u = f$ in Ω, $u|_{\partial\Omega} = g$, and $f(x) \geq 0$ in Ω, then $-u$ is the solution of the partial differential equation $-\Delta(-u) = -f \leq 0$. Therefore $-u$ attains its maximum value on the boundary $\partial\Omega$ of the domain Ω. Equivalently, u attains its minimum value on $\partial\Omega$; hence, u satisfies a *minimum principle* in this case, i.e.,

$$\boxed{\min_{x\in\partial\Omega} u(x) = \min_{x\in\overline{\Omega}} u(x).}$$

Our objective is now to construct a finite difference approximation of the elliptic boundary-value problem $-\Delta u = f$, $u|_{\partial\Omega} = g$, and show that a discrete counterpart of the maximum principle satisfied by the function u holds for its finite difference approximation U. For the sake of ease of exposition we shall confine ourselves to the case of two space dimensions and consider a general nonaxiparallel domain, such as the one depicted in Fig.4, and a general non-uniform mesh

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

The Laplace operator, Δ, is approximated by $D_x^+ D_x^- + D_y^+ D_y^-$, with the difference operators $D_x^+ D_x^-$, $D_y^+ D_y^-$ defined above. The finite difference approximation of the Dirichlet problem

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

is then given by

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) &= f(x_i, y_j) \qquad \text{in } \Omega_h, \\ U_{i,j} &= g(x_i, y_j) \qquad \text{on } \Gamma_h. \end{aligned} \tag{77}$$

Equivalently,

$$\begin{aligned} -\frac{1}{\hbar_i}\left(\frac{U_{i+1,j}-U_{i,j}}{h_{i+1}} - \frac{U_{i,j}-U_{i-1,j}}{h_i}\right) - \frac{1}{\hbar_j}\left(\frac{U_{i,j+1}-U_{i,j}}{k_{j+1}} - \frac{U_{i,j}-U_{i,j-1}}{k_j}\right) &= f(x_i, y_j) \quad \text{in } \Omega_h, \\ U_{i,j} &= g(x_i, y_j) \quad \text{on } \Gamma_h. \end{aligned}$$

Suppose that $f(x_i, y_j) < 0$ for all $(x_i, y_j) \in \Omega_h$ and that the maximum value of U is attained at a point $(x_{i_0}, y_{j_0}) \in \Omega_h$. Clearly,

$$\left(\frac{1}{\hbar_i}\left(\frac{1}{h_{i+1}} + \frac{1}{h_i}\right) + \frac{1}{\hbar_j}\left(\frac{1}{k_{j+1}} + \frac{1}{k_j}\right)\right) U_{i,j} = \frac{U_{i+1,j}}{\hbar_i\, h_{i+1}} + \frac{U_{i-1,j}}{\hbar_i\, h_i} + \frac{U_{i,j+1}}{\hbar_j\, k_{j+1}} + \frac{U_{i,j-1}}{\hbar_j\, k_j} + f(x_i, y_j)$$

for any $(x_i, y_j) \in \Omega_h$. Therefore, because $U_{i_0\pm1,j_0} \le U_{i_0,j_0}$ and $U_{i_0,j_0\pm1} \le U_{i_0,j_0}$, and $f(x_{i_0}, y_{j_0}) < 0$, it follows that

$$\left(\frac{1}{\hbar_{i_0}}\left(\frac{1}{h_{i_0+1}} + \frac{1}{h_{i_0}}\right) + \frac{1}{\hbar_{j_0}}\left(\frac{1}{k_{j_0+1}} + \frac{1}{k_{j_0}}\right)\right) U_{i_0 j_0} < \frac{U_{i_0 j_0}}{\hbar_{i_0}\, h_{i_0+1}} + \frac{U_{i_0 j_0}}{\hbar_{i_0}\, h_{i_0}} + \frac{U_{i_0 j_0}}{\hbar_{j_0}\, k_{j_0+1}} + \frac{U_{i_0 j_0}}{\hbar_{j_0}\, k_{j_0}}.$$

Note, however, that the expressions on the two sides of this equality are equal, which means that we have run into a contradiction. Thus we have shown that if $f(x_i, y_j) < 0$ for all $(x_i, y_j) \in \Omega_h$ then the maximum value of U is attained on the boundary Γ_h of Ω_h, which completes the proof of the *discrete maximum principle* in this case:

$$\boxed{\max_{(x_i,y_j)\in\Gamma_h} U_{i,j} = \max_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j}.}$$

Now suppose that $f(x_i, y_j) \le 0$ for all $(x_i, y_j) \in \Omega_h$. We define the auxiliary mesh function V by

$$V_{i,j} := U_{i,j} + \frac{\varepsilon}{4}(x_i^2 + y_j^2) \qquad \text{for } (x_i, y_j) \in \overline{\Omega}_h.$$

Hence,

$$-(D_x^+ D_x^- V_{i,j} + D_y^+ D_y^- V_{i,j}) = -(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) - \varepsilon = f(x_i, y_j) - \varepsilon < 0 \qquad \text{in } \Omega_h,$$

which then implies that the maximum value of V is attained on Γ_h. Therefore,

$$\begin{aligned} \max_{(x_i,y_j)\in\Gamma_h} U_{i,j} &= \max_{(x_i,y_j)\in\Gamma_h} \left[V_{i,j} + \frac{\varepsilon}{4}(x_i^2 + y_j^2)\right] \\ &\ge \max_{(x,y)\in\Gamma_h} V_{i,j} - \frac{\varepsilon}{4} \max_{(x_i,y_j)\in\Gamma_h} (x_i^2 + y_j^2) \\ &= \max_{(x_i,y_j)\in\overline{\Omega}_h} V_{i,j} - \frac{\varepsilon}{4} \max_{(x_i,y_j)\in\Gamma_h} (x_i^2 + y_j^2). \end{aligned}$$

As, by definition, $V_{i,j} \geq U_{i,j}$ for $(x_i, y_j) \in \overline{\Omega}_h$, it follows that

$$\max_{(x_i,y_j)\in\Gamma_h} U_{i,j} \geq \max_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j} - \frac{\varepsilon}{4} \max_{(x_i,y_j)\in\Gamma_h} (x_i^2 + y_j^2) \qquad \forall\, \varepsilon > 0.$$

By passing to the limit $\varepsilon \to 0_+$ it then follows that

$$\max_{(x_i,y_j)\in\Gamma_h} U_{i,j} \geq \max_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j}.$$

As $\Gamma_h \subset \overline{\Omega}_h$, trivially $\max_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j} \geq \max_{(x_i,y_j)\in\Gamma_h} U_{i,j}$, and therefore we have shown that if $f(x_i, y_j) \leq 0$ for all $(x_i, y_j) \in \Omega_h$, then the *discrete maximum principle* holds:

$$\max_{(x_i,y_j)\in\Gamma_h} U_{i,j} = \max_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j}.$$

Analogously, if $f(x_i, y_j) \geq 0$ for all $(x_i, y_j) \in \Omega_h$, then a *discrete minimum principle* holds:

$$\boxed{\min_{(x_i,y_j)\in\Gamma_h} U_{i,j} = \min_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j}.}$$

Our objective in the next section is to use the discrete maximum and minimum principles we have established to prove the stability of the finite difference scheme (77) with respect to perturbations in the boundary data.

4.4 Stability in the discrete maximum norm

Consider the finite difference scheme (77). Our first result asserts the existence of a solution to (77) as well as its uniqueness.

Lemma 10 *The finite difference scheme* (77) *has a unique solution.*

PROOF. We begin by noting that (77) is, in fact, a system of linear algebraic equations for the values $U_{i,j}$ such that $(x_i, y_j) \in \Omega_h$, so if the total number of mesh points contained in Ω_h is denoted by M_h, then the system of linear algebraic equations concerned has an $M_h \times M_h$ matrix, and showing the existence of a unique solution to the finite difference scheme (77) is therefore equivalent to showing that this system of linear algebraic equations has a unique solution, which amounts to showing that the matrix of the linear system is invertible. The matrix of the linear system associated with (77) is invertible if, and only if, the corresponding homogeneous system of linear algebraic equation has the zero vector as its only solution, which is, in turn, equivalent to showing that the finite difference scheme (77) with $f(x_i, y_j) = 0$ for all $(x_i, y_j) \in \Omega_h$ and $g(x_i, y_j) = 0$ for all $(x_i, y_j) \in \Gamma_h$ has the trivial solution as its only solution, i.e. that $U_{i,j} = 0$ for all $(x_i, y_j) \in \overline{\Omega}_h$. Let us therefore consider

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) &= 0 \qquad && \text{in } \Omega_h, \\ U_{i,j} &= 0 && \text{on } \Gamma_h. \end{aligned} \tag{78}$$

The existence of a solution to (78) is obvious: the mesh-function U, with $U_{i,j} = 0$ for all $(x_i, y_j) \in \overline{\Omega}_h$ is clearly a solution. According to the discrete maximum principle, for any solution U of the finite difference scheme (78),

$$0 = \max_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j},$$

while according to the discrete minimum principle

$$0 = \min_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j}.$$

Therefore the only solution is the trivial solution. This then implies the existence of a unique solution to (77). $\square$

We are now ready to embark on the analysis of the stability of the scheme (77) with respect to perturbations in the boundary data.

Consider the mesh functions $U^{(1)}$ and $U^{(2)}$, which satisfy, respectively:

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j}^{(1)} + D_y^+ D_y^- U_{i,j}^{(1)}) &= f(x_i, y_j) && \text{in } \Omega_h, \\ U_{i,j}^{(1)} &= g^{(1)}(x_i, y_j) && \text{on } \Gamma_h \end{aligned} \tag{79}$$

and

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j}^{(2)} + D_y^+ D_y^- U_{i,j}^{(2)}) &= f(x_i, y_j) && \text{in } \Omega_h, \\ U_{i,j}^{(2)} &= g^{(2)}(x_i, y_j) && \text{on } \Gamma_h \end{aligned} \tag{80}$$

for given boundary data $g^{(1)}$ and $g^{(2)}$. Let $U := U^{(1)} - U^{(2)}$ and $g := g^{(1)} - g^{(2)}$. Then, by subtracting (80) from (79) we find that U solves

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) &= 0 && \text{in } \Omega_h, \\ U_{i,j} &= g(x_i, y_j) && \text{on } \Gamma_h. \end{aligned} \tag{81}$$

By the discrete maximum principle we have from (81) that

$$\max_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j} = \max_{(x_i,y_j)\in\Gamma_h} U_{i,j} = \max_{(x_i,y_j)\in\Gamma_h} g(x_i, y_j) \leq \max_{(x_i,y_j)\in\Gamma_h} U_{i,j} = \max_{(x_i,y_j)\in\Gamma_h} |g(x_i, y_j)|.$$

In other words, for all $(x_i, y_j) \in \overline{\Omega}_h$,

$$U_{i,j} \leq \max_{(x_i,y_j)\in\Gamma_h} |g(x_i, y_j)|. \tag{82}$$

It follows from (81) that $-U$ solves

$$\begin{aligned} -(D_x^+ D_x^- (-U)_{i,j} + D_y^+ D_y^- (-U)_{i,j}) &= 0 && \text{in } \Omega_h, \\ (-U)_{i,j} &= -g(x_i, y_j) && \text{on } \Gamma_h, \end{aligned} \tag{83}$$

where $(-U)_{i,j} = -U_{i,j}$. Hence, also,

$$-U_{i,j} = (-U)_{i,j} \leq \max_{(x_i,y_j)\in\Gamma_h} |-g(x_i, y_j)| = \max_{(x_i,y_j)\in\Gamma_h} |g(x_i, y_j)| \tag{84}$$

for all $(x_i, y_j) \in \overline{\Omega}_h$. By combining (82) and (84) we have the inequality

$$|U_{i,j}| \leq \max_{(x_i,y_j)\in\Gamma_h} |g(x_i, y_j)|$$

for all $(x_i, y_j) \in \overline{\Omega}_h$. and hence,

$$\max_{(x_i,y_j)\in\overline{\Omega}_h} |U_{i,j}| \leq \max_{(x_i,y_j)\in\Gamma_h} |g(x_i, y_j)|.$$

By recalling the definitions of U and g, we have thereby shown that

$$\max_{(x_i,y_j)\in\overline{\Omega}_h} |U_{i,j}^{(1)} - U_{i,j}^{(2)}| \leq \max_{(x_i,y_j)\in\Gamma_h} |g^{(1)}(x_i, y_j) - g^{(2)}(x_i, y_j)|. \tag{85}$$

The inequality (85) expresses continuous dependence of the solution U to the finite difference scheme with respect to the boundary data g: it ensures that small perturbations in the boundary data result in small perturbations of the associated solution, a property that is referred to as *stability of the solution with respect to perturbations in the boundary data*.

4.5 Iterative solution of linear systems: linear stationary iterative methods

Before embarking on our discussion of the main topic of this section, we require a few technical tools. **Lecture 7**
Let us start by considering the finite difference approximation of the eigenvalue problem:

$$\begin{aligned} -u''(x) + c\,u(x) &= \lambda u(x), \qquad x \in (0,1), \\ u(0) &= 0, \quad u(1) = 0, \end{aligned}$$

where $c \geq 0$. A nontrivial solution $u(x) \not\equiv 0$ is called an *eigenfunction*, and the corresponding $\lambda \in \mathbb{C}$ for which such a nontrivial solution exists is called an *eigenvalue*. A simple calculation reveals that there is an infinite sequence of eigenfunctions u_k and eigenvalues λ_k, $k = 1, 2, \ldots$, where

$$u_k(x) = \sin(k\pi x) \quad \text{and} \quad \lambda_k = c + k^2\pi^2, \qquad k = 1, 2, \ldots.$$

Clearly, $c + \pi^2 \leq \lambda_k$ for all $k = 1, 2, \ldots$, and $\lambda_k \to +\infty$ as $k \to +\infty$.

The finite difference approximation of this eigenvalue problem on the mesh $\{x_i : i = 0, \ldots, N\}$ of uniform spacing $h = 1/N$, with $N \geq 2$, and $x_i = ih$, $i = 0, \ldots, N$, is given by

$$\begin{aligned} -\frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} + c\,U_i &= \Lambda U_i, \quad i = 1, \ldots, N-1, \\ U_0 &= 0, \quad U_N = 0. \end{aligned}$$

Once again, a simple calculation yields the nontrivial solution: $U_i := U_k(x_i)$ where

$$U_k(x) = \sin(k\pi x), \quad x \in \{x_0, x_1, \ldots, x_N\} \quad \text{and} \quad \Lambda_k = c + \frac{4}{h^2}\sin^2\frac{k\pi h}{2}, \qquad k = 1, 2, \ldots, N-1.$$

This can be verified by inserting

$$U_i = \sin(k\pi x_i) \quad \text{and} \quad U_{i\pm1} = \sin(k\pi x_{i\pm1})$$

into the finite difference scheme and noting that

$$\sin(k\pi x_{i\pm1}) = \sin(k\pi(x_i \pm h)) = \sin(k\pi x_i)\cos(k\pi h) \pm \cos(k\pi x_i)\sin(k\pi h) \quad \text{and} \quad 1 - \cos(k\pi h) = 2\sin^2\frac{k\pi h}{2}$$

for $k = 1, 2, \ldots, N-1$ and $i = 1, 2, \ldots, N-1$.

Using matrix notation the finite difference approximation of the eigenvalue problem can be written as

$$\begin{bmatrix} \frac{2}{h^2} + c & -\frac{1}{h^2} & & & \bigcirc \\ -\frac{1}{h^2} & \frac{2}{h^2} + c & -\frac{1}{h^2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{h^2} & \frac{2}{h^2} + c & -\frac{1}{h^2} \\ \bigcirc & & & -\frac{1}{h^2} & \frac{2}{h^2} + c \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-2} \\ U_{N-1} \end{bmatrix} = \Lambda \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-2} \\ U_{N-1} \end{bmatrix},$$

or, more compactly, $AU = \Lambda U$, where A is the symmetric tridiagonal $(N-1) \times (N-1)$ matrix displayed above, and $U = (U_1, \ldots, U_{N-1})^{\mathrm{T}}$ is a column vector of size $N-1$. The calculation performed above implies that the eigenvalues of the matrix A are

$$\Lambda_k = c + \frac{4}{h^2}\sin^2\frac{k\pi h}{2}, \qquad k = 1, 2, \ldots, N-1.$$

Clearly, $c + \pi^2 \leq \Lambda_k \leq c + \frac{4}{h^2}$ for all $k = 1, 2, \ldots, N-1$. The first of these inequalities follows by noting that $\sin x \geq \frac{2}{\pi}x$ for $x \in [0, \frac{\pi}{2}]$; the second inequality is the consequence of $0 \leq \sin x \leq 1$ for all $x \in \mathbb{R}$.

Example 1 *Suppose that $\Omega = (0,1)^2$, the open unit square in $\mathbb{R}^2$, and consider the problem*

$$\begin{aligned} -\Delta u + cu &= \lambda u \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \Gamma := \partial\Omega, \end{aligned}$$

where $c \geq 0$ is a given real number. A simple calculation shows that there is, once again, an infinite sequence of eigenfunctions and associated eigenvalues:

$$u_{k,m}(x,y) = \sin(k\pi x)\sin(m\pi y), \quad \lambda_{k,m} = c + (k^2+m^2)\pi^2, \qquad k,m = 1,2,\ldots.$$

The finite difference approximation of this eigenvalue problem posed on a uniform mesh $\{(x_i,y_j) : i,j = 0,\ldots,N\}$ of spacing $h = 1/N$, $N \geq 2$, in the x and y directions, is the following:

$$\begin{aligned} -\frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} - \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{h^2} + c\,U_{i,j} &= \Lambda U_{i,j}, \qquad i,j = 1,\ldots,N-1, \\ U_{i,j} &= 0 \qquad \text{for } (x_i,y_j) \in \Gamma_h, \end{aligned}$$

where, Γ_h is the set of mesh points on Γ. This can be rewritten as an algebraic eigenvalue problem of the form $AU = \Lambda U$, where now A is a symmetric $(N-1)^2 \times (N-1)^2$ matrix with positive eigenvalues

$$\Lambda_{k,m} = c + \frac{4}{h^2}\left(\sin^2\frac{k\pi}{2} + \sin^2\frac{m\pi}{2}\right),$$

with $c + 2\pi^2 \leq \Lambda_{k,m} \leq c + \frac{8}{h^2}$, and eigenvectors/(discrete) eigenfunctions $U_{i,j} = U_{k,m}(x_i,y_j)$, where

$$U_{k,m}(x,y) = \sin(k\pi x)\sin(m\pi y),$$

for $i,j = 1,\ldots,N-1$ and $k,m = 1,\ldots,N-1$.

Let us consider now the boundary-value problem:

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

where $c \geq 0$ and $f \in C([0,1])$. The finite difference approximation of this boundary-value problem on the mesh $\{x_i : i = 0,\ldots,N\}$ of uniform spacing $h = 1/N$, with $N \geq 2$, and $x_i = ih$, $i = 0,\ldots,N$, is given by

$$\begin{aligned} -\frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} + c\,U_i &= f(x_i), \quad i = 1,\ldots,N-1, \\ U_0 = 0, \quad U_N &= 0. \end{aligned} \tag{86}$$

In terms of matrix notation, this can be rewritten as a system of linear algebraic equations of the form

$$AU = F \tag{87}$$

where A is the same $(N-1)\times(N-1)$ symmetric tridiagonal matrix, with distinct positive eigenvalues Λ_k, $k = 1,\ldots,N-1$, as above, $F = (f(x_1),\ldots,f(x_{N-1}))^{\mathrm{T}}$, and $U = (U_1,\ldots,U_{N-1})^{\mathrm{T}}$ is the associated vector of unknowns.

Similarly, if one considers the elliptic boundary-value problem

$$\begin{aligned} -\Delta u + cu &= f(x,y) \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \Gamma := \partial\Omega, \end{aligned}$$

where $c \geq 0$ is a given real number and $f \in C(\Omega)$, whose finite difference approximation posed on a uniform mesh $\{(x_i, y_j) : i, j = 0, \ldots, N\}$ of spacing $h = 1/N$, $N \geq 2$, in the x and y directions, is

$$\begin{aligned} -\frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} - \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{h^2} + c\,U_{i,j} &= f(x_i, y_j), \qquad i, j = 1, \ldots, N-1, \\ U_{i,j} &= 0 \qquad \text{for } (x_i, y_j) \in \Gamma_h, \end{aligned} \tag{88}$$

where, Γ_h is the set of mesh points on Γ, then this, too, can be rewritten as a system of linear algebraic equations of the form $AU = F$, where now A is a symmetric $(N-1)^2 \times (N-1)^2$ matrix with positive eigenvalues, given in Example 1 above.

Motivated by these examples, we shall be interested in developing a simple iterative method for the solution of systems of linear algebraic equations of the form $AU = F$, where $A \in \mathbb{R}^{M \times M}$ is a symmetric matrix with positive eigenvalues, which are contained in a nonempty closed interval $[\alpha, \beta]$, with $0 < \alpha < \beta$, $U \in \mathbb{R}^M$ is the vector of unknowns and $F \in \mathbb{R}^M$ is a given vector. To this end, we consider the following iteration for the approximate solution of the linear system $AU = F$.

$$U^{j+1} := U^j - \tau(AU^j - F), \qquad j = 0, 1, \ldots, \tag{89}$$

where $U^0 \in \mathbb{R}^M$ is a given initial guess, and $\tau > 0$ is a parameter to be chosen so as to ensure that the sequence of iterates $\{U^j\}_{j=0}^{\infty} \subset \mathbb{R}^M$ converges to $U \in \mathbb{R}^M$ as $j \to \infty$. We begin by observing that $U = U - \tau(AU - F)$. Therefore, upon subtraction of (89) from this equality we find that

$$U - U^{j+1} = U - U^j - \tau A(U - U^j) = (I - \tau A)(U - U^j), \qquad j = 0, 1, \ldots, \tag{90}$$

where $I \in \mathbb{R}^{M \times M}$ is the identity matrix. Consequently,

$$U - U^j = (I - \tau A)^j (U - U^0), \qquad j = 1, 2, \ldots.$$

Recall that if $\|\cdot\|$ is a(ny) norm on $\mathbb{R}^M$, then the *induced matrix norm* is defined, for a matrix $B \in \mathbb{R}^{M \times M}$ by

$$\|B\| := \sup_{V \in \mathbb{R}^M} \frac{\|BV\|}{\|V\|}.$$

Thanks to this definition, $\|BV\| \leq \|B\|\|V\|$ for any $V \in \mathbb{R}^M$, and hence, by induction $\|B^j V\| \leq \|B\|^j \|V\|$ for all $j = 1, 2 \ldots$ and all $V \in \mathbb{R}^M$. Therefore,

$$\|U - U^j\| = \|(I - \tau A)^j (U - U^0)\| \leq \|I - \tau A\|^j \|U - U^0\|. \tag{91}$$

In order to continue, we need to bound $\|I - \tau A\|$, and to this end we need a few tools from linear algebra; we shall therefore make a brief detour. Our first observation is that $\mathbb{R}^M$ is a finite-dimensional linear space, and in a finite-dimensional linear spaces all norms are equivalent.[1] Therefore, if the sequence $\{U^j\}_{j=0}^{\infty}$ converges to U in one particular norm on $\mathbb{R}^M$, it will also converge to U in any other norm on $\mathbb{R}^M$. For the sake of simplicity of the exposition we shall therefore assume that the norm $\|\cdot\|$ on $\mathbb{R}^M$ appearing in the inequality above is the Euclidean norm:

$$\|V\| := \left(\sum_{i=1}^{M} V_i^2 \right)^{1/2}, \qquad V = (V_1, \ldots, V_M)^{\mathrm{T}} \in \mathbb{R}^M.$$

A symmetric matrix $B \in \mathbb{R}^{M \times M}$ has real eigenvalues, and the associated set of orthonormal eigenvectors spans the whole of $\mathbb{R}^M$. Denoting by $\{e_i\}_{i=1}^M$ the (orthonormal) eigenvectors of B and by λ_i, $i = 1, \ldots, M$,

[1] Suppose that $\mathcal{V}$ is a linear space and $\|\cdot\|_1$ and $\|\cdot\|_2$ are two norms on $\mathcal{V}$; then $\|\cdot\|_1$ and $\|\cdot\|_2$ are said to be *equivalent* if there exist positive constants C_1 and C_2 such that $C_1\|V\|_1 \leq \|V\|_2 \leq C_2\|V\|_1$ for all $V \in \mathcal{V}$. For the details of the proof of the assertion that any two norms on a finite-dimensional linear space are equivalent, see, for example, the webpage http://mathonline.wikidot.com/equivalence-of-norms-in-a-finite-dimensional-linear-space#toc0

the corresponding eigenvalues, for any vector $V = \alpha_1 e_1 \cdots + \alpha_M e_M$, expanded in terms of the eigenvectors of B, then, thanks to orthonormality, the Euclidean norms of V and BV are, respectively,

$$\|V\| = \left(\sum_{i=1}^{M} \alpha_i^2 \right)^{1/2} \quad \text{and} \quad \|BV\| = \left(\sum_{i=1}^{M} \alpha_i^2 \lambda_i^2 \right)^{1/2}.$$

Clearly, $\|BV\| \leq \max_{i=1,\ldots,M} |\lambda_i| \, \|V\|$ for all $V \in \mathbb{R}^M$, and the inequality becomes an equality if V happens to be the eigenvector of B associated with the largest in absolute value eigenvalue of B. Therefore, $\|B\| = \max_{i=1,\ldots,M} |\lambda_i|$.

We are now ready to return to (91) to find that $\|I - \tau A\|$ appearing on its right-hand side is equal to the largest in absolute value of the symmetric matrix $I - \tau A$. As the eigenvalues of A are assumed to belong to the interval $[\alpha, \beta]$, where $0 < \alpha < \beta$, and the parameter τ is by assumption positive, the eigenvalues of $I - \tau A$ are contained in the interval $[1 - \tau\beta, 1 - \tau\alpha]$, whereby $\|I - \tau A\| \leq \max\{|1 - \tau\beta|, |1 - \tau\alpha|\}$. As $\tau > 0$ is a free parameter, to be suitably chosen, we would like to select it so that the iterative method (89) converge as fast as possible, and to this end we see from (91) that it is desirable to choose τ so that $\|I - \tau A\|$ is as small as possible, and less than 1. We shall therefore seek $\tau > 0$ so as to ensure that

$$\min_{\tau > 0} \max\{|1 - \tau\beta|, |1 - \tau\alpha|\} < 1.$$

By plotting the piecewise linear functions $\tau \mapsto |1 - \tau\beta|$ and $\tau \mapsto |1 - \tau\alpha|$ for $\tau \in [0, \infty)$, we see that their graphs intersect at $\tau = 0$ and at $\tau = \frac{2}{\alpha+\beta}$, and the continuous piecewise linear function $\tau \mapsto \max\{|1 - \tau\beta|, |1 - \tau\alpha|\}$ attains its minimum at $\tau = \frac{2}{\alpha+\beta}$. Thus,

$$\min_{\tau > 0} \max\{|1 - \tau\beta|, |1 - \tau\alpha|\} = \max\{|1 - \tau\beta|, |1 - \tau\alpha|\}|_{\tau = \frac{2}{\alpha+\beta}} = \frac{\beta - \alpha}{\beta + \alpha} < 1.$$

In summary then, the iterative method proposed for the solution of the linear system $AU = F$ is the one stated in (89), with $\tau := \frac{\beta-\alpha}{\beta+\alpha} < 1$, and $[\alpha, \beta]$ being a closed subinterval of $(0, \infty)$ that contains all eigenvalues of the symmetric matrix $A \in \mathbb{R}^{M \times M}$.

Example 2 *In the case of the finite difference scheme* (86), $\alpha = c + \pi^2$ *and* $\beta = c + \frac{4}{h^2}$, *while in the case of* (88), $\alpha = c + 2\pi^2$ *and* $\beta = c + \frac{8}{h^2}$. *In both cases*

$$\frac{\beta - \alpha}{\beta + \alpha} = 1 - \textit{Const.}\, h^2;$$

thus, while the sequence of iterates $\{U^j\}_{j=0}^{\infty}$ *defined by the iterative method* (89) *is guaranteed to converge to the exact solution* U *of the linear system* $AU = F$, *the speed of convergence will deteriorate as* $h \to 0$.

We note that by multiplying (90) by the matrix A and recalling that $AU = F$, one has that

$$F - AU^{j+1} = (I - \tau A)(F - AU^j),$$

and therefore, proceeding as above

$$\|F - AU^{(j)}\| \leq \|I - \tau A\|^j \|F - AU^0\| \leq \left(\frac{\beta - \alpha}{\beta + \alpha} \right)^j \|F - AU^0\|. \tag{92}$$

As F, A and the initial guess U^0 are available, as are α and β, it is possible to quantify the number of iterations required to ensure that the Euclidean norm of the so-called *residual* $F - AU^j$ of the j-th iterate is smaller than a chosen tolerance $\texttt{TOL} > 0$: a sufficient condition for this is that the right-hand side of (92) is smaller than TOL, which will hold as soon as

$$j > \log \frac{\|F - AU^0\|}{\texttt{TOL}} \left[\log \left(\frac{\beta + \alpha}{\beta - \alpha} \right) \right]^{-1}.$$

In the case of the two examples considered the right-hand side of this inequality is $\sim$ Const. $h^{-2} \log(1/\texttt{TOL})$.

5 Finite difference approximation of parabolic equations

The final section of these lecture notes is concerned with the construction and mathematical analysis of finite difference methods for the numerical solution of parabolic equations. As a simple yet representative model problem we shall focus on the unsteady diffusion equation (heat equation) in one space dimension: **Lecture 8**

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \tag{93}$$

which we shall consider for $x \in (-\infty, \infty)$ and $t \geq 0$, subject to the initial condition

$$u(x,0) = u_0(x), \qquad x \in (-\infty, \infty),$$

where u_0 is a given function.

The solution of this initial-value problem can be expressed explicitly in terms of the initial datum u_0. As the expression for the solution of the initial-value problem provides helpful insight into the behaviour of solutions of parabolic partial differential equations, which we shall try to mimic in the course of their numerical approximation, we shall summarize here briefly the derivation of this expression.

We recall that the Fourier transform of a function v is defined by

$$\hat{v}(\xi) = F[v](\xi) = \int_{-\infty}^{\infty} v(x)\, \mathrm{e}^{-\imath x\xi}\, \mathrm{d}x.$$

We shall assume henceforth that the functions under consideration are sufficiently smooth and that they decay to 0 as $x \to \pm\infty$ sufficiently quickly in order to ensure that our manipulations make sense.

By Fourier-transforming the partial differential equation (93) we obtain

$$\int_{-\infty}^{\infty} \frac{\partial u}{\partial t}(x,t)\, \mathrm{e}^{-\imath x\xi}\, \mathrm{d}x = \int_{-\infty}^{\infty} \frac{\partial^2 u}{\partial x^2}(x,t)\, \mathrm{e}^{-\imath x\xi}\, \mathrm{d}x.$$

After (formal) integration by parts on the right-hand side and ignoring boundary terms at $\pm\infty$, we obtain

$$\frac{\partial}{\partial t}\hat{u}(\xi,t) = (\imath\xi)^2 \hat{u}(\xi,t),$$

whereby

$$\hat{u}(\xi,t) = \mathrm{e}^{-t\xi^2}\hat{u}(\xi,0),$$

and therefore

$$u(x,t) = F^{-1}\left(\mathrm{e}^{-t\xi^2}\hat{u}_0\right).$$

The inverse Fourier transform of a function is defined by

$$v(x) = F^{-1}[\hat{v}](x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{v}(\xi)\mathrm{e}^{\imath x\xi}\, \mathrm{d}\xi.$$

Thus, after some lengthy calculations whose details we omit, we find that

$$u(x,t) = F^{-1}\left(\mathrm{e}^{-t\xi^2}\hat{u}_0(\xi)\right) = \int_{-\infty}^{\infty} w(x-y,t)u_0(y)\, \mathrm{d}y,$$

where the function w, defined by

$$w(x,t) = \frac{1}{\sqrt{4\pi t}}\, \mathrm{e}^{-x^2/(4t)},$$

is called the **heat kernel**. So, finally,

$$u(x,t) = \frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} \mathrm{e}^{-(x-y)^2/(4t)}u_0(y)\, \mathrm{d}y. \tag{94}$$

This formula gives an explicit expression of the solution of the heat equation (93) in terms of the initial datum u_0. Because $w(x,t) > 0$ for all $x \in (-\infty, \infty)$ and all $t > 0$, and

$$\int_{-\infty}^{\infty} w(y,t)\,\mathrm{d}y = 1 \qquad \text{for all } t > 0,$$

we deduce from (94) that if u_0 is a bounded continuous function, then

$$\sup_{x\in(-\infty,+\infty)}|u(x,t)| \le \sup_{x\in(-\infty,\infty)}|u_0(x)|, \qquad t > 0. \tag{95}$$

In other words, the 'largest' and 'smallest' values of $u(\cdot,t)$ at $t > 0$ cannot exceed those of $u_0(\cdot)$. Similar bounds on the 'magnitude' of the solution at future times in terms of the 'magnitude' of the initial datum can be obtained in other norms as well, and we shall focus here on the L_2 norm in particular. We will show, using Parseval's identity, that the L_2 norm of the solution, at any time $t > 0$, is bounded by the L_2 norm of the initial datum. We shall then try to mimic this property when using various numerical approximations of the initial-value problem for the heat equation.

Lemma 11 (Parseval's identity) *Let $L_2(-\infty,\infty)$ denote the set of all complex-valued square-integrable functions defined on the real line. Suppose that $u \in L_2(-\infty,\infty)$. Then, $\hat{u} \in L_2(-\infty,\infty)$, and the following equality holds:*

$$\|u\|_{L_2(-\infty,\infty)} = \frac{1}{\sqrt{2\pi}}\,\|\hat{u}\|_{L_2(-\infty,\infty)},$$

where

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

Proof. We begin by observing that

$$\begin{aligned}
\int_{-\infty}^{\infty} \hat{u}(\xi)\,v(\xi)\,\mathrm{d}\xi &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} u(x)\,\mathrm{e}^{-\imath x\xi}\,\mathrm{d}x\right) v(\xi)\,\mathrm{d}\xi \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} v(\xi)\,\mathrm{e}^{-\imath x\xi}\,\mathrm{d}\xi\right) u(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} u(x)\,\hat{v}(x)\,\mathrm{d}x.
\end{aligned}$$

We then take (where, for a complex-valued function w, we denote by $\overline{w}$ the complex conjugate of w)

$$v(\xi) = \overline{\hat{u}(\xi)} = 2\pi F^{-1}[\overline{u}](\xi), \qquad \xi \in (-\infty,\infty),$$

and substitute this into the identity above to complete the proof. □

Returning to the equation (93), we thus have by Parseval's identity that

$$\|u(\cdot,t)\|_{L_2(-\infty,\infty)} = \frac{1}{\sqrt{2\pi}}\,\|\hat{u}(\cdot,t)\|_{L_2(-\infty,\infty)}, \qquad t > 0,$$

and therefore

$$\begin{aligned}
\|u(\cdot,t)\|_{L_2(-\infty,\infty)} &= \frac{1}{\sqrt{2\pi}}\,\|\mathrm{e}^{-t\xi^2}\,\hat{u}_0(\cdot)\|_{L_2(-\infty,\infty)} \\
&\le \frac{1}{\sqrt{2\pi}}\,\|\hat{u}_0\|_{L_2(-\infty,\infty)} \\
&= \|u_0\|_{L_2(-\infty,\infty)}, \qquad t > 0.
\end{aligned}$$

Thus we have shown that

$$\|u(\cdot,t)\|_{L_2(-\infty,\infty)} \leq \|u_0\|_{L_2(-\infty,\infty)} \qquad \text{for all } t > 0. \tag{96}$$

This is a useful result as it can be used to deduce stability of the solution of the equation (93) with respect to perturbations of the initial datum in a sense which we shall now explain. Suppose that u_0 and $\tilde{u}_0$ are two functions contained in $L_2(-\infty,\infty)$ and denote by u and $\tilde{u}$ the solutions to (93) resulting from the initial functions u_0 and $\tilde{u}_0$, respectively. Then $u - \tilde{u}$ solves the heat equation with initial datum $u_0 - \tilde{u}_0$, and therefore, by (96), we have that

$$\|u(\cdot,t) - \tilde{u}(\cdot,t)\|_{L_2(-\infty,\infty)} \leq \|u_0 - \tilde{u}_0\|_{L_2(-\infty,\infty)} \qquad \text{for all } t > 0. \tag{97}$$

This inequality implies continuous dependence of the solution on the initial function: small perturbations in u_0 in the $L_2(-\infty,\infty)$ norm will result in small perturbations in the associated analytical solution $u(\cdot,t)$ in the $L_2(-\infty,\infty)$ norm for all $t > 0$.

The inequality (96) is therefore a relevant property, which we shall try to mimic with our numerical approximations of the equation (93).

5.1 Finite difference approximation of the heat equation

We take our computational domain to be

$$\{(x,t) \in (-\infty,\infty) \times [0,T]\},$$

where $T > 0$ is a given final time. We then consider a finite difference mesh with spacing $\Delta x > 0$ in the x-direction and spacing $\Delta t = T/M$ in the t-direction, with $M \geq 1$, and we approximate the partial derivatives appearing in the differential equation using divided differences as follows. Let $x_j = j\Delta x$ and $t_m = m\Delta t$, and note that

$$\frac{\partial u}{\partial t}(x_j,t_m) \approx \frac{u(x_j,t_{m+1}) - u(x_j,t_m)}{\Delta t}$$

and

$$\frac{\partial^2 u}{\partial x^2}(x_j,t_m) \approx \frac{u(x_{j+1},t_m) - 2u(x_j,t_m) + u(x_{j-1},t_m)}{(\Delta x)^2}.$$

This then motivates us to approximate the heat equation (93) at the point (x_j,t_m) by the following numerical method, called the **explicit Euler scheme**:

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = \frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2}, \quad j = 0, \pm 1, \pm 2, \ldots$$

$$U_j^0 = u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots$$

Equivalently, we can write this as

$$U_j^{m+1} = U_j^m + \mu(U_{j+1}^m - 2U_j^m + U_{j-1}^m),$$

$$U_j^0 = u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots$$

where $\mu = \frac{\Delta t}{(\Delta x)^2}$. Thus, U_j^{m+1} can be explicitly calculated, for all $j = 0, \pm 1, \pm 2, \ldots$, from the values U_{j+1}^m, U_j^m, and U_{j-1}^m from the previous time level.

Alternatively, if instead of time level m the expression on the right-hand side of the explicit Euler scheme is evaluated on the time level $m+1$, we arrive at the **implicit Euler scheme**:

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = \frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2}, \quad j = 0, \pm 1, \pm 2, \ldots$$

$$U_j^0 = u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots .$$

The explicit and implicit Euler schemes are special cases of a more general one-parameter family of numerical methods for the heat equation, called the **θ-method**, which is a convex combination of the two Euler schemes, with a parameter $\theta \in [0,1]$. The θ-method is defined as follows:

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = (1-\theta)\,\frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2} + \theta\,\frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2},$$

$$U_j^0 = u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots ,$$

where $\theta \in [0,1]$ is a parameter. For $\theta = 0$ it coincides with the explicit Euler scheme, for $\theta = 1$ it is the implicit Euler scheme, and for $\theta = 1/2$ it is the arithmetic average of the two Euler schemes, and is called the **Crank–Nicolson scheme**.

5.1.1 Accuracy of the θ-method

Our aim in this section is to assess the accuracy of the θ-method for the Dirichlet initial-boundary-value problem for the heat equation. The consistency error of the θ-method is defined by

$$T_j^m = \frac{u_j^{m+1} - u_j^m}{\Delta t} - (1-\theta)\,\frac{u_{j+1}^m - 2u_j^m + u_{j-1}^m}{(\Delta x)^2} - \theta\,\frac{u_{j+1}^{m+1} - 2u_j^{m+1} + u_{j-1}^{m+1}}{(\Delta x)^2},$$

where

$$u_j^m \equiv u(x_j, t_m).$$

We shall explore the size of the consistency error by performing a Taylor series expansion about a suitable point. We begin by noting that

$$u_j^{m+1} = \left[u + \frac{1}{2}\Delta t u_t + \frac{1}{2}\left(\frac{1}{2}\Delta t\right)^2 u_{tt} + \frac{1}{6}\left(\frac{1}{2}\Delta t\right)^3 u_{ttt} + \cdots \right]_j^{m+1/2},$$

$$u_j^m = \left[u - \frac{1}{2}\Delta t u_t + \frac{1}{2}\left(\frac{1}{2}\Delta t\right)^2 u_{tt} - \frac{1}{6}\left(\frac{1}{2}\Delta t\right)^3 u_{ttt} + \cdots \right]_j^{m+1/2}.$$

Therefore,

$$\frac{u_j^{m+1} - u_j^m}{\Delta t} = \left[u_t + \frac{1}{24}(\Delta t)^2 u_{ttt} + \cdots \right]_j^{m+1/2}.$$

Similarly,

$$\begin{aligned}
&(1-\theta)\,\frac{u_{j+1}^m - 2u_j^m + u_{j-1}^m}{(\Delta x)^2} + \theta\,\frac{u_{j+1}^{m+1} - 2u_j^{m+1} + u_{j-1}^{m+1}}{(\Delta x)^2} \\
&\quad = \left[u_{xx} + \frac{1}{12}(\Delta x)^2 u_{xxxx} + \frac{2}{6!}(\Delta x)^4 u_{xxxxxx} + \cdots \right]_j^{m+1/2} \\
&\qquad + \left(\theta - \frac{1}{2}\right)\Delta t \left[u_{xxt} + \frac{1}{12}(\Delta x)^2 u_{xxxxt} + \cdots \right]_j^{m+1/2} \\
&\qquad\qquad + \frac{1}{8}(\Delta t)^2 \left[u_{xxtt} + \cdots \right]_j^{m+1/2}.
\end{aligned}$$

Combining these, we deduce that

$$\begin{aligned} T_j^m = & \boxed{[u_t - u_{xx}]_j^{m+1/2}} \\ & + \left[\left(\frac{1}{2} - \theta \right) \Delta t \, u_{xxt} - \frac{1}{12} (\Delta x)^2 \, u_{xxxx} \right]_j^{m+1/2} \\ & + \left[\frac{1}{24} (\Delta t)^2 \, u_{ttt} - \frac{1}{8} (\Delta t)^2 \, u_{xxtt} \right]_j^{m+1/2} \\ & + \left[\frac{1}{12} \left(\frac{1}{2} - \theta \right) \Delta t \, (\Delta x)^2 \, u_{xxxxt} - \frac{2}{6!} (\Delta x)^4 \, u_{xxxxxx} \right]_j^{m+1/2} + \cdots . \end{aligned}$$

Note however that the term contained in the box vanishes, as u is a solution to the heat equation. Hence,

$$T_j^m = \begin{cases} \mathcal{O}\left((\Delta x)^2 + (\Delta t)^2 \right) & \text{for } \theta = 1/2, \\ \mathcal{O}\left((\Delta x)^2 + \Delta t \right) & \text{for } \theta \neq 1/2. \end{cases}$$

Thus, in particular, the explicit and implicit Euler schemes have consistency error

$$T_j^m = \mathcal{O}\left((\Delta x)^2 + \Delta t \right),$$

while the Crank–Nicolson scheme has consistency error

$$T_j^m = \mathcal{O}\left((\Delta x)^2 + (\Delta t)^2 \right).$$

5.2 Stability of finite difference schemes

Lecture 9

In order to be able to replicate the stability property (96) at the discrete level, we require an appropriate notion of stability. We shall say that a finite difference scheme for the unsteady heat equation is **(practically) stable in the ℓ_2 norm**, if

$$\|U^m\|_{\ell_2} \leq \|U^0\|_{\ell_2}, \qquad m = 1, \ldots, M,$$

where

$$\|U^m\|_{\ell_2} = \left(\Delta x \sum_{j=-\infty}^{\infty} |U_j^m|^2 \right)^{1/2}.$$

We shall use the semidiscrete Fourier transform to explore the stability of finite difference schemes.

Definition 2 *The semidiscrete Fourier transform of a function U defined on the infinite mesh $x_j = j\Delta x$, $j = 0, \pm 1, \pm 2, \ldots$, is:*

$$\hat{U}(k) = \Delta x \sum_{j=-\infty}^{\infty} U_j \, \mathrm{e}^{-\imath k x_j}, \qquad k \in [-\pi/\Delta x, \pi/\Delta x].$$

We shall also require the inverse semidiscrete Fourier transform, as well the discrete counterpart of Parseval's identity that connect these transforms, analogously as in the case of the Fourier transform and its inverse considered earlier.

Definition 3 *Let $\hat{U}$ be defined on the interval $[-\pi/\Delta x, \pi/\Delta x]$. The inverse semidiscrete Fourier transform of $\hat{U}$ is defined by*

$$U_j := \frac{1}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} \hat{U}(k) \, \mathrm{e}^{\imath k j \Delta x} \, \mathrm{d}k.$$

We then have the following result.

Lemma 12 (Discrete Parseval's identity) *Let*

$$\|U\|_{\ell_2} = \left(\Delta x \sum_{j=-\infty}^{\infty} |U_j|^2 \right)^{1/2} \quad \text{and} \quad \|\hat{U}\|_{L_2} = \left(\int_{-\pi/\Delta x}^{\pi/\Delta x} |\hat{U}(k)|^2 \, \mathrm{d}k \right)^{1/2}.$$

If $\|U\|_{\ell_2}$ is finite, then also $\|\hat{U}\|_{L_2}$ is finite, and

$$\|U\|_{\ell_2} = \frac{1}{\sqrt{2\pi}} \|\hat{U}\|_{L_2}.$$

The proof of this result is very similar to the proof of Lemma 11, and we shall therefore leave it to the reader as an exercise.

5.2.1 Stability analysis of the explicit Euler scheme

We are now ready to embark on the stability analysis of the explicit Euler scheme. By inserting

$$U_j^m = \frac{1}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} \mathrm{e}^{\imath k j \Delta x} \hat{U}^m(k) \, \mathrm{d}k$$

into the Euler scheme we deduce that

$$\frac{1}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} \mathrm{e}^{\imath k j \Delta x} \frac{\hat{U}^{m+1}(k) - \hat{U}^m(k)}{\Delta t} \, \mathrm{d}k = \frac{1}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} \frac{\mathrm{e}^{\imath k (j+1) \Delta x} - 2\mathrm{e}^{\imath k j \Delta x} + \mathrm{e}^{\imath k (j-1) \Delta x}}{(\Delta x)^2} \hat{U}^m(k) \, \mathrm{d}k.$$

Therefore, we have that

$$\frac{1}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} \mathrm{e}^{\imath k j \Delta x} \frac{\hat{U}^{m+1}(k) - \hat{U}^m(k)}{\Delta t} \, \mathrm{d}k = \frac{1}{2\pi} \int_{-\pi/\Delta x}^{\pi/\Delta x} \mathrm{e}^{\imath k j \Delta x} \frac{\mathrm{e}^{\imath k \Delta x} - 2 + \mathrm{e}^{-\imath k \Delta x}}{(\Delta x)^2} \hat{U}^m(k) \, \mathrm{d}k.$$

By comparing the left-hand side with the right-hand side we deduce that

$$\hat{U}^{m+1}(k) = \hat{U}^m(k) + \mu(\mathrm{e}^{\imath k \Delta x} - 2 + \mathrm{e}^{-\imath k \Delta x}) \hat{U}^m(k)$$

for all **wave numbers** $k \in [-\pi/\Delta x, \, \pi/\Delta x]$, and we thus deduce that

$$\hat{U}^{m+1}(k) = \lambda(k) \hat{U}^m(k),$$

where

$$\lambda(k) = 1 + \mu(\mathrm{e}^{\imath k \Delta x} - 2 + \mathrm{e}^{-\imath k \Delta x})$$

is the **amplification factor** and

$$\mu := \frac{\Delta t}{(\Delta x)^2}$$

is called the CFL number (after Richard Courant, Kurt Friedrichs, and Hans Levy, who first performed an analysis of this kind).[2] By the discrete Parseval identity stated in Lemma 12 we have that

$$\begin{aligned} \|U^{m+1}\|_{\ell_2} &= \frac{1}{\sqrt{2\pi}} \|\hat{U}^{m+1}\|_{L_2} \\ &= \frac{1}{\sqrt{2\pi}} \|\lambda \hat{U}^m\|_{L_2} \\ &\leq \frac{1}{\sqrt{2\pi}} \max_k |\lambda(k)| \, \|\hat{U}^m\|_{L_2} \\ &= \max_k |\lambda(k)| \, \|U^m\|_{\ell_2}. \end{aligned}$$

[2] Richard Courant, Kurt Friedrichs, and Hans Lewy (*Über die partiellen Differenzengleichungen der mathematischen Physik*. Mathematische Annalen, 100:32–74, 1928).

In order to mimic the bound (96) we would like to ensure that

$$\|U^{m+1}\|_{\ell_2} \leq \|U^m\|_{\ell_2}, \qquad m = 0, 1, \ldots, M-1.$$

Thus we demand that

$$\max_k |\lambda(k)| \leq 1,$$

i.e., that

$$\max_k |1 + \mu(\mathrm{e}^{\imath k\Delta x} - 2 + \mathrm{e}^{-\imath k\Delta x})| \leq 1.$$

Using Euler's formula

$$\mathrm{e}^{\imath\varphi} = \cos\varphi + \imath \sin\varphi$$

and the trigonometric identity

$$1 - \cos\varphi = 2\sin^2\frac{\varphi}{2}$$

we can restate this as follows:

$$\max_k \left|1 - 4\mu \sin^2\left(\frac{k\Delta x}{2}\right)\right| \leq 1.$$

Equivalently, we need to ensure that

$$-1 \leq 1 - 4\mu\sin^2\left(\frac{k\Delta x}{2}\right) \leq 1 \quad \forall\, k \in [-\pi/\Delta x, \pi/\Delta x].$$

This holds if, and only if, $\mu = \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2}$. Thus we have shown the following result.

Theorem 12 *Suppose that U_j^m is the solution of the explicit Euler scheme*

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = \frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2}, \quad j = 0, \pm 1, \pm 2, \ldots,$$

$$U_j^0 = u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots,$$

and $\mu = \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2}$. Then,

$$\|U^m\|_{\ell_2} \leq \|U^0\|_{\ell_2}, \qquad m = 1, 2, \ldots, M. \tag{98}$$

In other words the explicit Euler scheme is **conditionally practically stable**, the condition for stability being that $\mu = \Delta t/\Delta x^2 \leq 1/2$. One can also show that if $\mu > 1/2$, then (98) will fail. In other words, once Δx has been chosen, one must choose Δt so that $\Delta t/\Delta x^2 \leq 1/2$ in order to ensure that the bound (98) holds.

5.2.2 Stability analysis of the implicit Euler scheme

We shall now perform a similar analysis for the **implicit Euler scheme** for the heat equation (93), which is defined as follows:

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = \frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2}, \quad j = 0, \pm 1, \pm 2, \ldots$$

$$U_j^0 = u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots.$$

Equivalently,

$$U_j^{m+1} - \mu(U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}) = U_j^m$$

$$U_j^0 = u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots,$$

where, again,

$$\mu = \frac{\Delta t}{(\Delta x)^2}.$$

Using an identical argument as for the explicit Euler scheme, we find that the amplification factor is now

$$\lambda(k) = \frac{1}{1 + 4\mu \sin^2\left(\frac{k\Delta x}{2}\right)}.$$

Clearly,

$$\max_k |\lambda(k)| \leq 1$$

for all values of

$$\mu = \frac{\Delta t}{(\Delta x)^2}.$$

Thus we have the following result.

Theorem 13 *Suppose that U_j^m is the solution of the implicit Euler scheme*

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = \frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2}, \quad j = 0, \pm 1, \pm 2, \ldots,$$

$$U_j^0 = u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots.$$

Then, for all $\Delta t > 0$ and $\Delta x > 0$,

$$\|U^m\|_{\ell_2} \leq \|U^0\|_{\ell_2}, \qquad m = 1, 2, \ldots, M. \tag{99}$$

In other words, the implicit Euler scheme is **unconditionally practically stable**, meaning that the bound (99) holds without any restrictions on Δx and Δt.

5.3 Von Neumann stability

In certain situations, practical stability is too restrictive and we need a less demanding notion of stability. **Lecture 10**
The one below, due to John von Neumann, is called **von Neumann stability**.

Definition 4 *We shall say that a finite difference scheme for the unsteady heat equation on the time interval $[0, T]$ is **von Neumann stable** in the ℓ_2 norm, if there exists a positive constant $C = C(T)$ such that*

$$\|U^m\|_{\ell_2} \leq C\|U^0\|_{\ell_2}, \qquad m = 1, \ldots, M = \frac{T}{\Delta t},$$

where

$$\|U^m\|_{\ell_2} = \left(\Delta x \sum_{j=-\infty}^{\infty} |U_j^m|^2\right)^{1/2}.$$

Clearly, practical stability implies von Neumann stability, with stability constant $C = 1$. As the **stability constant** C in the definition of von Neumann stability may dependent on T, and when it does then, typically, $C(T) \to +\infty$ as $T \to +\infty$, it follows that, unlike practical stability which is meaningful for $m = 1, 2, \ldots$, von Neumann stability makes sense on finite time intervals $[0, T]$ (with $T < \infty$) and for the limited range of $0 \leq m \leq T/\Delta t$, only.

Von Neumann stability of a finite difference scheme can be easily verified by using the following result.

Lemma 13 *Suppose that the semidiscrete Fourier transform of the solution* $\{U_j^m\}_{j=1}^{\infty}$, $m = 0, 1, \ldots, \frac{T}{\Delta t}$, *of a finite difference scheme for the heat equation satisfies*

$$\hat{U}^{m+1}(k) = \lambda(k)\hat{U}^m(k)$$

and

$$|\lambda(k)| \leq 1 + C_0\Delta t \qquad \forall\, k \in [-\pi/\Delta x, \pi/\Delta x].$$

Then the scheme is von Neumann stable. In particular, if $C_0 = 0$ *then the scheme is practically stable.*

PROOF: By Parseval's identity for the semidiscrete Fourier transform we have that

$$\begin{aligned}
\|U^{m+1}\|_{\ell_2} &= \frac{1}{\sqrt{2\pi}}\|\hat{U}^{m+1}\|_{L_2} \\
&= \frac{1}{\sqrt{2\pi}}\,\|\lambda\hat{U}^m\|_{L_2} \\
&\leq \frac{1}{\sqrt{2\pi}}\max_k|\lambda(k)|\;\|\hat{U}^m\|_{L_2} \\
&= \max_k|\lambda(k)|\;\|U^m\|_{\ell_2}.
\end{aligned}$$

Hence,

$$\|U^{m+1}\|_{\ell_2} \leq (1 + C_0\Delta t)\|U^m\|_{\ell_2}, \qquad m = 0, 1, \ldots, M-1.$$

Therefore,

$$\|U^m\|_{\ell_2} \leq (1 + C_0\Delta t)^m\|U^0\|_{\ell_2}, \qquad m = 1, \ldots, M.$$

As $1 + C_0\Delta t \leq \mathrm{e}^{C_0\Delta t}$ and $(1 + C_0\Delta t)^m \leq \mathrm{e}^{C_0 m\Delta t} \leq \mathrm{e}^{C_0 T}$ for all $M = 1, \ldots, M$, it follows that

$$\|U^m\|_{\ell_2} \leq \mathrm{e}^{C_0 T}\|U^0\|_{\ell_2}, \qquad m = 1, 2, \ldots, M,$$

meaning that von Neumann stability holds, with stability constant $C = \mathrm{e}^{C_0 T}$. □

5.4 Stability of the θ-scheme

The explicit and implicit Euler schemes are special cases of a more general one-parameter family of numerical methods for the heat equation, called the **θ-scheme**, which is a convex combination of the two Euler schemes, with a parameter $\theta \in [0, 1]$. The θ-scheme is defined as follows:

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = (1-\theta)\,\frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2} + \theta\,\frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2},$$

$$U_j^0 = u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots,$$

where $\theta \in [0, 1]$ is a parameter. For $\theta = 0$ it coincides with the explicit Euler scheme, for $\theta = 1$ it is the implicit Euler scheme, and for $\theta = 1/2$ it is the arithmetic average of the two Euler schemes, and is called the **Crank–Nicolson scheme**.

To analyse the practical stability of the θ-scheme in the ℓ_2 norm, we shall use Lemma 13 with $C_0 = 0$. Suppose that

$$U_j^m = [\lambda(k)]^m\,\mathrm{e}^{\imath k x_j}.$$

Substitution of this 'Fourier mode' into the θ-scheme gives the equality

$$\lambda(k) - 1 = -4(1-\theta)\,\mu\,\sin^2\left(\frac{k\Delta x}{2}\right) - 4\theta\,\mu\,\lambda(k)\sin^2\left(\frac{k\Delta x}{2}\right).$$

Therefore,

$$\lambda(k) = \frac{1 - 4(1-\theta)\mu \sin^2\left(\frac{k\Delta x}{2}\right)}{1 + 4\theta\mu \sin^2\left(\frac{k\Delta x}{2}\right)}.$$

For practical stability, we demand that

$$|\lambda(k)| \leq 1 \qquad \forall\, k \in [-\pi/\Delta x, \pi/\Delta x],$$

which holds if, and only if,

$$2(1-2\theta)\mu \leq 1.$$

Thus we have shown that:

- For $\theta \in [1/2, 1]$ the θ-scheme is **unconditionally practically stable**;
- For $\theta \in [0, 1/2)$ the θ-scheme is **conditionally practically stable**, the stability condition being that

$$\mu \leq \frac{1}{2(1-2\theta)}.$$

5.5 Boundary-value problems for parabolic problems

When a parabolic partial differential equation is considered on a bounded spatial domain, one needs to impose boundary conditions on the boundary of the domain. Here we shall concentrate on the simplest case, when a Dirichlet boundary is imposed at both endpoints of the spatial domain, which we take to be the nonempty bounded open interval (a, b). We shall therefore consider the following Dirichlet initial–boundary value problem for the heat equation:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad a < x < b, \quad 0 < t \leq T,$$

subject to the initial condition

$$u(x, 0) = u_0(x), \qquad x \in [a, b],$$

and the following Dirichlet boundary conditions at $x = a$ and $x = b$:

$$u(a, t) = A(t), \quad u(b, t) = B(t), \qquad t \in (0, T].$$

Remark 2 *We note in passing that the Neumann initial-boundary-value problem for the heat equation is:*

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad a < x < b, \quad 0 < t \leq T,$$

subject to the initial condition

$$u(x, 0) = u_0(x), \qquad x \in [a, b],$$

and the Neumann boundary conditions

$$\frac{\partial u}{\partial x}(a, t) = A(t), \quad \frac{\partial u}{\partial x}(b, t) = B(t), \qquad t \in (0, T].$$

An example of a mixed Dirichlet–Neumann initial-boundary-value problem for the heat equation is

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad a < x < b, \quad 0 < t \leq T,$$

subject to the initial condition

$$u(x, 0) = u_0(x), \qquad x \in [a, b],$$

and the mixed Dirichlet–Neumann boundary conditions

$$u(a, t) = A(t), \quad \frac{\partial u}{\partial x}(b, t) = B(t), \qquad t \in (0, T].$$

5.5.1 θ-scheme for the Dirichlet initial-boundary-value problem

Our aim in this section is to construct a numerical approximation of the Dirichlet initial-boundary-value problem based on the θ-scheme. Let $\Delta x = (b-a)/J$ and $\Delta t = T/M$, and define

$$x_j := a + j\Delta x, \quad j = 0, \ldots, J, \qquad t_m := m\Delta t, \quad m = 0, \ldots, M.$$

We approximate the Dirichlet initial-boundary-value problem with the following θ-scheme:

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = (1-\theta)\,\frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2} + \theta\,\frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2},$$

for $j = 1, \ldots, J-1$, $m = 0, 1, \ldots, M-1$,

$$U_j^0 = u_0(x_j), \qquad j = 1, \ldots, J-1,$$

$$U_0^{m+1} = A(t_{m+1}), \;\; U_J^{m+1} = B(t_{m+1}), \;\; m = 0, \ldots, M-1.$$

In order to implement this scheme it is helpful to rewrite it as a system of linear algebraic equations to compute the values of the approximate solution on time-level $m+1$ from those on time-level m. We have that

$$\begin{aligned}[1 - \theta\mu\delta^2]U_j^{m+1} &= [1 + (1-\theta)\mu\delta^2]U_j^m,\\ U_j^0 &= u_0(x_j), \qquad 1 \le j \le J-1,\end{aligned}$$

$$U_0^{m+1} = A(t_{m+1}), \qquad U_J^{m+1} = B(t_{m+1}), \;\; 0 \le m \le M-1,$$

where

$$\delta^2 U_j := U_{j+1} - 2U_j + U_{j-1}.$$

The matrix form of this system of linear equations is therefore the following. We consider the symmetric tridiagonal $(J-1)\times(J-1)$ matrix:

$$\mathcal{A} = \begin{pmatrix} -2 & 1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & 0 & \ldots & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 1 & -2 \end{pmatrix}.$$

Let $\mathcal{I}$ be the $(J-1)\times(J-1)$ identity matrix $\mathcal{I} = \mathrm{diag}(1,\, 1,\, 1, \ldots,\, 1,\, 1)$. Then, the θ-scheme can be written as

$$(\mathcal{I} - \theta\mu\mathcal{A})\mathbf{U}^{m+1} = (\mathcal{I} + (1-\theta)\mu\mathcal{A})\mathbf{U}^m + \theta\mu\mathbf{F}^{m+1} + (1-\theta)\mu\mathbf{F}^m$$

for $m = 0, 1, \ldots, M-1$, where

$$\mathbf{U}^m = (U_1^m,\, U_2^m,\, \ldots,\, U_{J-2}^m,\, U_{J-1}^m)^{\mathrm{T}}$$

and

$$\mathbf{F}^m = (A(t_m),\, 0,\, \ldots,\, 0,\, B(t_m))^{\mathrm{T}}.$$

Matlab code for the Crank–Nicolson scheme

```
% cn.m - Crank--Nicolson scheme for the heat equation.
% Save this file as cn.m
% Run this by typing cn at the Matlab command line, and choose the value of N when prompted.
%
N = input('N? ');
dx = 1/N; x = dx:dx:1-dx; N1 = N-1;
dt = dx/2; mu = dt/dx^2;
% u = max([1-2.*abs(0.5-x); 0*x])';
u = (sin(pi*x).*exp(3*x))';
x1 = [0, x, 1];
u1 = [0, u', 0];
hold off; plot(x1,u1,'linewidth',2)
text(0.71,0.75,'t = 0','fontsize',15)
A = (-2.)  * eye(N1);
for i = 1:N1-1
A(i,i+1) = 1; A(i+1,i) = 1;
end
A1 = eye(N1) - (1/2) * mu * A;
A2 = eye(N1) + (1/2) * mu * A;
grid;
hold on;
pause;
for i = 1:50
u = A1\(A2 * u);
u1 = [0, u', 0];
plot(x1,u1,'b','linewidth',2);
text(.41,0.45,'t=20*dt','fontsize',15)
end
```

5.5.2 The discrete maximum principle

We shall now try to prove a bound, analogous to (95), for the θ-scheme **Lecture 11**

Theorem 14 (Discrete maximum principle for the θ-scheme)
The θ-scheme for the Dirichlet initial-boundary-value problem for the heat equation, with $0 \le \theta \le 1$ and $\mu(1-\theta) \le \frac{1}{2}$, yields a sequence of numerical approximations $\{U_j^m\}_{j=0,\ldots,J;\ m=0,\ldots,M}$ satisfying

$$U_{\min} \le U_j^m \le U_{\max}$$

where

$$U_{\min} = \min\left\{\min\{U_0^m\}_{m=0}^M,\ \min\{U_j^0\}_{j=0}^J,\ \min\{U_J^m\}_{m=0}^M\right\}$$

and

$$U_{\max} = \max\left\{\max\{U_0^m\}_{m=0}^M,\ \max\{U_j^0\}_{j=0}^J,\ \max\{U_J^m\}_{m=0}^M\right\}.$$

Proof: We rewrite the θ-scheme as

$$(1+2\theta\mu)\,U_j^{m+1} = \theta\mu\left(U_{j+1}^{m+1} + U_{j-1}^{m+1}\right) + (1-\theta)\mu\left(U_{j+1}^m + U_{j-1}^m\right) + [1-2(1-\theta)\mu]\,U_j^m, \qquad (100)$$

and recall that, by hypothesis,

$$\theta\mu \ge 0 \qquad (1-\theta)\mu \ge 0, \qquad 1-2(1-\theta)\mu \ge 0.$$

Suppose that U attains its maximum value at an internal mesh point U_j^{m+1}, $1 \le j \le J-1,\ 0 \le m \le M-1$. If this is not the case, the proof is complete. We define

$$U^\star = \max\{U_{j+1}^{m+1},\, U_{j-1}^{m+1},\, U_{j+1}^m,\, U_{j-1}^m,\, U_j^m\}.$$

Then,

$$(1+2\theta\mu)\,U_j^{m+1} \;\le\; 2\theta\mu U^\star + 2(1-\theta)\mu U^\star + [1-2(1-\theta)\mu]U^\star = (1+2\theta\mu)\,U^\star, \tag{101}$$

and therefore

$$U_j^{m+1} \le U^\star.$$

However, also,

$$U^\star \le U_j^{m+1},$$

as U_j^{m+1} is assumed to be the overall maximum value. Hence,

$$U_j^{m+1} = U^\star.$$

Thus the maximum value is also attained at the points neighbouring (x_j, t_{m+1}) present in the scheme.[3]

The same argument applies to these neighbouring points, and we can then repeat this process until the boundary at $x = a$ or $x = b$ or at $t = 0$ is reached, and this will happen in a finite number of steps. The maximum is therefore attained at a boundary point. Similarly, the minimum is attained at a boundary point. □

In summary then, for

$$\mu(1-\theta) \le \frac{1}{2}$$

the θ-scheme satisfies the discrete maximum principle. This is clearly more demanding than the ℓ_2-stability condition:

$$\mu(1-2\theta) \le \frac{1}{2} \qquad \text{for} \quad 0 \le \theta \le \tfrac{1}{2}.$$

For example, the Crank–Nicolson scheme is unconditionally stable in the ℓ_2 norm, yet it only satisfies the discrete maximum principle when $\mu := \frac{\Delta t}{(\Delta x)^2} \le 1$.

5.5.3 Convergence analysis of the θ-scheme in the maximum norm

We close our discussion of finite difference schemes for the heat equation (93) in one space-dimension with the convergence analysis of the θ-scheme for the Dirichlet initial-boundary-value problem. We begin by rewriting the scheme as follows:

$$(1+2\theta\mu)\,U_j^{m+1} = \theta\mu\left(U_{j+1}^{m+1} + U_{j-1}^{m+1}\right) + (1-\theta)\mu\left(U_{j+1}^{m} + U_{j-1}^{m}\right) + [1-2(1-\theta)\mu]\,U_j^m.$$

The scheme is considered subject to the initial condition

$$U_j^0 = u_0(x_j), \qquad j = 1, \ldots, J-1,$$

and the boundary conditions

$$U_0^{m+1} = A(t_{m+1}), \;\; U_J^{m+1} = B(t_{m+1}), \;\; m = 0, \ldots, M-1.$$

[3] To see that the maximum value $U_j^{m+1} = U^*$ is attained at *each* of points neighbouring (x_j, t_{m+1}) present in the scheme, first observe that if: (a) $\theta = 0$, then U_{j+1}^{m+1} and U_{j-1}^{m+1} are absent from the right-hand side of (100); (b) if $\theta = 1$ then U_{j+1}^{m} and U_{j-1}^{m} are absent from the right-hand side of (100); (c) if $2(1-\theta)\mu = 1$, then U_j^m is absent from the right-hand side of (100), and (d) if $\theta \notin \{0, 1, 1-\frac{1}{2\mu}\}$, then U_{j+1}^{m+1}, U_{j-1}^{m+1}, U_{j+1}^{m}, U_{j-1}^{m}, and U_j^m are all present on the right-hand side of (100). There are therefore four different cases to be discussed: (a), (b), (c) and (d). Suppose that we are in case (d) (the cases (a), (b) and (c) being dealt with identically); if one of U_{j+1}^{m+1}, U_{j-1}^{m+1}, U_{j+1}^{m}, U_{j-1}^{m}, and U_j^m were strictly smaller than $U_j^{m+1} = U^*$, then, by returning to the transition from (100) to (101), we would deduce (101) from (100), but now with the $\le$ symbol in (101) replaced by $<$, which would then imply that $U_j^{m+1} < U^*$. This would, however, contradict the equality $U_j^{m+1} = U^*$ we have already proved. Thus the value $U^{m+1} = U^*$ is attained at *each* of the five point neighbouring (x_j, t_{m+1}).

The **consistency error** for the θ-scheme is defined by

$$T_j^m = \frac{u_j^{m+1} - u_j^m}{\Delta t} - (1-\theta)\,\frac{u_{j+1}^m - 2u_j^m + u_{j-1}^m}{(\Delta x)^2} - \theta\,\frac{u_{j+1}^{m+1} - 2u_j^{m+1} + u_{j-1}^{m+1}}{(\Delta x)^2},$$

where $u_j^m \equiv u(x_j, t_m)$, and therefore

$$(1+2\theta\mu)\,u_j^{m+1} = \theta\mu\left(u_{j+1}^{m+1} + u_{j-1}^{m+1}\right) + (1-\theta)\mu\left(u_{j+1}^m + u_{j-1}^m\right) + [1-2(1-\theta)\mu]\,u_j^m + \Delta t T_j^m.$$

Let us define the **global error**, that is the discrepancy at a mesh-point between the exact solution and its numerical approximation, by

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

It then follows that

$$e_0^{m+1} = 0,\ e_J^{m+1} = 0,\ e_j^0 = 0, \quad j = 0, \ldots, J,$$

and

$$(1+2\theta\mu)\,e_j^{m+1} = \theta\mu\left(e_{j+1}^{m+1} + e_{j-1}^{m+1}\right) + (1-\theta)\mu\left(e_{j+1}^m + e_{j-1}^m\right) + [1-2(1-\theta)\mu]\,e_j^m + \Delta t T_j^m.$$

We define,

$$E^m = \max_{0\le j\le J} |e_j^m| \quad \text{and} \quad T^m = \max_{0\le j\le J} |T_j^m|.$$

As, by hypothesis,

$$\theta\mu \ge 0, \qquad (1-\theta)\mu \ge 0, \qquad 1-2(1-\theta)\mu \ge 0,$$

we have that

$$(1+2\theta\mu)E^{m+1} \le 2\theta\mu E^{m+1} + E^m + \Delta t T^m.$$

Hence,

$$E^{m+1} \le E^m + \Delta t\, T^m.$$

As $E^0 = 0$, upon summation,

$$\begin{aligned} E^m &\le \Delta t \sum_{n=0}^{m-1} T^n \\ &\le m\Delta t \max_{0\le n\le m-1} T^n \\ &\le T \max_{0\le m\le M} \max_{1\le j\le J-1} |T_j^m|, \end{aligned}$$

which then implies that

$$\max_{0\le j\le J} \max_{0\le m\le M} |u(x_j, t_m) - U_j^m| \le T \max_{1\le j\le J-1} \max_{0\le m\le M} |T_j^m|.$$

Recall that the consistency error of the θ-scheme is

$$T_j^m = \begin{cases} \mathcal{O}\left((\Delta x)^2 + (\Delta t)^2\right) & \text{for } \theta = 1/2, \\ \mathcal{O}\left((\Delta x)^2 + \Delta t\right) & \text{for } \theta \ne 1/2. \end{cases}$$

It therefore follows that for the explicit and implicit Euler schemes, which have consistency error

$$T_j^m = \mathcal{O}\left((\Delta x)^2 + \Delta t\right),$$

one has the following bound on the global error:

$$\max_{0\le j\le J}\ \max_{0\le m\le M}\ |u(x_j,t_m)-U_j^m|\le \text{Const.}\left((\Delta x)^2+\Delta t\right),$$

while for the Crank–Nicolson scheme, which has consistency error

$$T_j^m=\mathcal{O}\left((\Delta x)^2+(\Delta t)^2\right),$$

one has

$$\max_{0\le j\le J}\ \max_{0\le m\le M}\ |u(x_j,t_m)-U_j^m|\le \text{Const.}\left((\Delta x)^2+(\Delta t)^2\right).$$

The results developed in this section can be easily extended to multidimensional axiparallel domains, such as rectangular or L-shaped domains in two space-dimensions whose edges are parallel with the x and y, axes, or cuboid-shaped domains in three space-dimensions whose faces are parallel with the co-ordinate planes. For more complicated computational domains, such as those with nonaxiparallel or curved faces, finite difference meshes with uneven spacing need to be used for points inside the computational domain that are closest to the boundary of the domain, or if a mesh with even spacing is used, then 'ghost-points', which lie outside the computational domains, need to be introduced. For further details, we refer, for example, to R. LeVeque, Finite Difference Methods for Ordinary and Partial Differential Equations. SIAM, 2007. ISBN: 978-0-898716-29-0; or to K.W. Morton and D.F. Mayers, Numerical Solution of Partial Differential Equations: An Introduction, 2nd Edition, CUP, 2005. ISBN: 978-0-521607-93-3.

In the next section we shall confine ourselves to discussing the construction of finite difference schemes for the unsteady heat-equation in two space-dimensions on a rectangular spatial domain.

5.6 Finite difference approximation of parabolic equations in two space-dimensions

Start of optional material

Consider the heat equation

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2},\qquad (x,y)\in\Omega:=(a,b)\times(c,d),\ \ t\in(0,T],$$

subject to the initial condition

$$u(x,y,0)=u_0(x,y),\qquad (x,y)\in[a,b]\times[c,d],$$

and the Dirichlet boundary condition

$$u|_{\partial\Omega}=B(x,y,t),\qquad (x,y)\in\partial\Omega,\ \ t\in(0,T],$$

where $\partial\Omega$ is the boundary of Ω. We begin by considering the explicit Euler finite difference approximation of this problem.

5.6.1 The explicit Euler scheme

Let

$$\delta_x^2 U_{i,j}:=U_{i+1,j}-2U_{i,j}+U_{i-1,j},$$

and

$$\delta_y^2 U_{i,j}:=U_{i,j+1}-2U_{i,j}+U_{i,j-1}.$$

Let, further, $\Delta x:=(b-a)/J_x$, $\Delta y:=(d-c)/J_y$, $\Delta t:=T/M$, and define

$$\begin{aligned} x_i&=a+i\Delta x, && i=0,\dots,J_x,\\ y_j&=c+j\Delta y, && j=0,\dots,J_y,\\ t_m&=m\Delta t, && m=0,\dots,M. \end{aligned}$$

The explicit Euler finite difference approximation of the unsteady heat equation on the space-time domain $\overline{\Omega} \times [0, T]$ is then the following:

$$\frac{U_{i,j}^{m+1} - U_{i,j}^{m}}{\Delta t} = \frac{\delta_x^2 U_{i,j}^{m}}{(\Delta x)^2} + \frac{\delta_y^2 U_{i,j}^{m}}{(\Delta y)^2},$$

for $i = 1, \ldots, J_x - 1$, $j = 1, \ldots, J_y - 1$, $m = 0, 1, \ldots, M - 1$, subject to the initial condition

$$U_{i,j}^{0} = u_0(x_i, y_j), \quad i = 0, \ldots, J_x, \;\; j = 0, \ldots, J_y,$$

and the boundary condition

$$U_{i,j}^{m} = B(x_i, y_j, t_m), \;\; \text{at the boundary mesh points, for } m = 1, \ldots, M.$$

5.6.2 The implicit Euler scheme

The implicit Euler scheme is defined analogously. Let $\Delta x := (b - a)/J_x$, $\Delta y := (d - c)/J_y$, $\Delta t := T/M$, and define

$$\begin{aligned} x_i &= a + i\Delta x, & i &= 0, \ldots, J_x, \\ y_j &= b + j\Delta y, & j &= 0, \ldots, J_y, \\ t_m &= m\Delta t, & m &= 0, \ldots, M. \end{aligned}$$

The implicit Euler finite difference scheme for the problem under consideration is then

$$\frac{U_{i,j}^{m+1} - U_{i,j}^{m}}{\Delta t} = \frac{\delta_x^2 U_{i,j}^{m+1}}{(\Delta x)^2} + \frac{\delta_y^2 U_{i,j}^{m+1}}{(\Delta y)^2},$$

for $i = 1, \ldots, J_x - 1$, $j = 1, \ldots, J_y - 1$, $m = 0, 1, \ldots, M - 1$, subject to the initial condition

$$U_{i,j}^{0} = u_0(x_i, y_j), \quad i = 0, \ldots, J_x, \;\; j = 0, \ldots, J_y,$$

and the boundary condition

$$U_{i,j}^{m+1} = B(x_i, y_j, t_{m+1}), \;\; \text{at the boundary mesh points, for } m = 0, \ldots, M - 1.$$

5.6.3 The θ-scheme

By taking the convex combination of the explicit and implicit Euler schemes, with a parameter $\theta \in [0, 1]$, with $\theta = 0$ corresponding to the explicit Euler scheme and $\theta = 1$ to the implicit Euler scheme, we obtain a one-parameter family of schemes, called the θ-scheme. It is defined as follows.

Let $\Delta x := (b - a)/J_x$, $\Delta y := (d - c)/J_y$, $\Delta t := T/M$, and, for $\theta \in [0, 1]$, consider the finite difference scheme

$$\frac{U_{i,j}^{m+1} - U_{i,j}^{m}}{\Delta t} = (1 - \theta) \left(\frac{\delta_x^2 U_{i,j}^{m}}{(\Delta x)^2} + \frac{\delta_y^2 U_{i,j}^{m}}{(\Delta y)^2} \right) + \theta \left(\frac{\delta_x^2 U_{i,j}^{m+1}}{(\Delta x)^2} + \frac{\delta_y^2 U_{i,j}^{m+1}}{(\Delta y)^2} \right),$$

for $i = 1, \ldots, J_x - 1$, $j = 1, \ldots, J_y - 1$, $m = 0, 1, \ldots, M - 1$, subject to the initial condition

$$U_{i,j}^{0} = u_0(x_i, y_j), \quad i = 0, \ldots, J_x, \;\; j = 0, \ldots, J_y,$$

and the boundary condition

$$U_{i,j}^{m+1} = B(x_i, y_j, t_{m+1}), \;\; \text{at the boundary mesh points, for } m = 0, \ldots, M - 1.$$

The practical stability of the θ-scheme (in the absence of boundary conditions now, i.e. for the pure initial-value problem rather than the initial-boundary-value problem) in the ℓ^2 norm is easily assessed by inserting the Fourier mode

$$U_{i,j}^m = [\lambda(k_x, k_y)]^m \, \mathrm{e}^{\imath(k_x x_i + k_y y_j)}$$

into the scheme. This gives

$$\lambda - 1 = -4(1-\theta)\left[\mu_x \sin^2\left(\frac{k_x \Delta x}{2}\right) + \mu_y \sin^2\left(\frac{k_y \Delta y}{2}\right)\right] - 4\theta\lambda\left[\mu_x \sin^2\left(\frac{k_x \Delta x}{2}\right) + \mu_y \sin^2\left(\frac{k_y \Delta y}{2}\right)\right],$$

where

$$\mu_x = \frac{\Delta t}{(\Delta x)^2}, \qquad \mu_y = \frac{\Delta t}{(\Delta y)^2}.$$

Hence,

$$\lambda = \frac{1 - 4(1-\theta)\left[\mu_x \sin^2\left(\frac{k_x \Delta x}{2}\right) + \mu_y \sin^2\left(\frac{k_y \Delta y}{2}\right)\right]}{1 + 4\theta\left[\mu_x \sin^2\left(\frac{k_x \Delta x}{2}\right) + \mu_y \sin^2\left(\frac{k_y \Delta y}{2}\right)\right]}.$$

For practical stability in the ℓ_2 norm, we require that

$$|\lambda(k_x, k_y)| \le 1 \quad \forall\, (k_x, k_y) \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right] \times \left[-\frac{\pi}{\Delta y}, \frac{\pi}{\Delta y}\right].$$

Thus, we demand that

$$-1 \le \frac{1 - 4(1-\theta)\left[\mu_x + \mu_y\right]}{1 + 4\theta\left[\mu_x + \mu_y\right]} \le 1,$$

which can be restated in the following equivalent form:

$$2(1-2\theta)(\mu_x + \mu_y) \le 1.$$

For example, the implicit Euler scheme ($\theta = 1$) and the Crank–Nicolson scheme ($\theta = 1/2$) are unconditionally stable, while the explicit Euler scheme ($\theta = 0$) is only conditionally stable, the stability condition being that Δx, Δy, and Δt satisfy the following inequality:

$$\mu_x + \mu_y \equiv \Delta t\left(\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2}\right) \le \frac{1}{2}.$$

Under a suitable condition the θ-scheme for the initial-boundary-value problem also satisfies a discrete maximum principle. To see this, we rewrite the θ-scheme as

$$\begin{aligned}(1 + 2\theta(\mu_x + \mu_y))U_{i,j}^{m+1} = {} & (1 - 2(1-\theta)(\mu_x + \mu_y))U_{i,j}^m \\ & + (1-\theta)\mu_x(U_{i+1,j}^m + U_{i-1,j}^m) \\ & + (1-\theta)\mu_y(U_{i,j+1}^m + U_{i,j-1}^m) \\ & + \theta\mu_x(U_{i+1,j}^{m+1} + U_{i-1,j}^{m+1}) \\ & + \theta\mu_y(U_{i,j+1}^{m+1} + U_{i,j-1}^{m+1}),\end{aligned}$$

for $i = 1, \ldots, J_x - 1$, $j = 1, \ldots, J_y - 1$, $m = 0, 1, \ldots, M-1$, subject to the initial condition

$$U_{i,j}^0 = u_0(x_i, y_j), \quad i = 0, \ldots, J_x, \;\; j = 0, \ldots, J_y,$$

and the boundary condition

$$U_{i,j}^m = B(x_i, y_j, t_m), \;\; \text{at the boundary mesh points, for } m = 1, \ldots, M.$$

Theorem 15 *Suppose that*

$$(\mu_x + \mu_y)(1-\theta) \le \frac{1}{2}, \qquad \theta \in [0,1].$$

Then, the θ-scheme satisfies the following discrete maximum principle:

$$U_{\min} \le U_{i,j}^m \le U_{\max},$$

where

$$U_{\min} = \min\left\{\min\{U_{i,j}^0\}_{i,j=0}^{J_x,J_y},\ \min\{U_{i,j}^m\}_{m=0}^{M}|_{(x_i,y_j)\in\partial\Omega}\right\}$$

and

$$U_{\max} = \max\left\{\max\{U_{i,j}^0\}_{i,j=0}^{J_x,J_y},\ \max\{U_{i,j}^m\}_{m=0}^{M}|_{(x_i,y_j)\in\partial\Omega}\right\}.$$

PROOF: The proof proceeds by an obvious modification of the proof of the discrete maximum principle for the θ-scheme in one space-dimension. □

In summary, then, for

$$(\mu_x + \mu_y)(1-\theta) \le \frac{1}{2}$$

the θ-scheme satisfies the discrete maximum principle. This condition is more demanding than the one for the ℓ_2-stability of the scheme, which requires that

$$(\mu_x + \mu_y)(1-2\theta) \le \frac{1}{2} \qquad \text{for} \quad 0 \le \theta \le \tfrac{1}{2}.$$

For example, the Crank–Nicolson scheme is unconditionally stable in the ℓ_2 norm, but for the discrete maximum principle to hold we had to assume that

$$\mu_x + \mu_y = \frac{\Delta t}{(\Delta x)^2} + \frac{\Delta t}{(\Delta y)^2} \le 1.$$

We close our discussion of the θ-scheme with its error analysis. The starting point is to rewrite the scheme as follows:

$$\begin{aligned}(1+2\theta(\mu_x+\mu_y))U_{i,j}^{m+1} = (1-2(1-\theta)(\mu_x+\mu_y))U_{i,j}^m \\ + (1-\theta)\mu_x(U_{i+1,j}^m + U_{i-1,j}^m) \\ + (1-\theta)\mu_y(U_{i,j+1}^m + U_{i,j-1}^m) \\ + \theta\mu_x(U_{i+1,j}^{m+1} + U_{i-1,j}^{m+1}) \\ + \theta\mu_y(U_{i,j+1}^{m+1} + U_{i,j-1}^{m+1}),\end{aligned}$$

for $i = 1,\ldots,J_x-1$, $j = 1,\ldots,J_y-1$, $m = 0,1,\ldots,M-1$, subject to the initial condition

$$U_{i,j}^0 = u_0(x_i,y_j), \quad i = 0,\ldots,J_x, \ \ j = 0,\ldots,J_y,$$

and the boundary condition

$$U_{i,j}^m = B(x_i,y_j,t_m), \ \text{ at the boundary mesh points, for } m = 1,\ldots,M.$$

Suppose further that

$$(\mu_x + \mu_y)(1-\theta) \le \frac{1}{2}, \qquad \theta \in [0,1].$$

The consistency error of the θ-scheme is defined as follows:

$$T_{i,j}^m := \frac{u_{i,j}^{m+1} - u_{i,j}^m}{\Delta t} - (1-\theta)\left(\frac{\delta_x^2 u_{i,j}^m}{(\Delta x)^2} + \frac{\delta_y^2 u_{i,j}^m}{(\Delta y)^2}\right) - \theta\left(\frac{\delta_x^2 u_{i,j}^{m+1}}{(\Delta x)^2} + \frac{\delta_y^2 u_{i,j}^{m+1}}{(\Delta y)^2}\right),$$

where

$$u_{i,j}^m \equiv u(x_i, y_j, t_m).$$

By performing some elementary but tedious Taylor series expansions, one can deduce that

$$T_{i,j}^m = \begin{cases} \mathcal{O}\left((\Delta x)^2 + (\Delta y)^2 + (\Delta t)^2\right) & \theta = 1/2, \\ \mathcal{O}\left((\Delta x)^2 + (\Delta y)^2 + \Delta t\right) & \theta \neq 1/2. \end{cases}$$

It follows from the definition of the consistency error $T_{i,j}^m$ for the θ-scheme that

$$\begin{aligned} (1 + 2\theta(\mu_x + \mu_y))u_{i,j}^{m+1} = (1 - 2(1-\theta)(\mu_x + \mu_y))u_{i,j}^m \\ + (1-\theta)\mu_x(u_{i+1,j}^m + u_{i-1,j}^m) \\ + (1-\theta)\mu_y(u_{i,j+1}^m + u_{i,j-1}^m) \\ + \theta\mu_x(u_{i+1,j}^{m+1} + u_{i-1,j}^{m+1}) \\ + \theta\mu_y(u_{i,j+1}^{m+1} + u_{i,j-1}^{m+1}) \\ + \Delta t\, T_{i,j}^m, \end{aligned}$$

for $i = 1, \ldots, J_x - 1$, $j = 1, \ldots, J_y - 1$, $m = 0, 1, \ldots, M-1$. We define the **global error** as

$$e_{i,j}^m := u(x_i, y_j, t_m) - U_{i,j}^m.$$

Then, $e_{i,j}^0 = 0$ and $e_{i,j}^m = 0$ for $(x_i, y_j) \in \partial\Omega$, and

$$\begin{aligned} (1 + 2\theta(\mu_x + \mu_y))e_{i,j}^{m+1} = (1 - 2(1-\theta)(\mu_x + \mu_y))e_{i,j}^m \\ + (1-\theta)\mu_x(e_{i+1,j}^m + e_{i-1,j}^m) \\ + (1-\theta)\mu_y(e_{i,j+1}^m + e_{i,j-1}^m) \\ + \theta\mu_x(e_{i+1,j}^{m+1} + e_{i-1,j}^{m+1}) \\ + \theta\mu_y(e_{i,j+1}^{m+1} + e_{i,j-1}^{m+1}) \\ + \Delta t\, T_{i,j}^m. \end{aligned}$$

We further define,

$$E^m := \max_{i,j} |e_{i,j}^m| \quad \text{and} \quad T^m := \max_{i,j} |T_{i,j}^m|.$$

As, by hypothesis,

$$1 - 2(1-\theta)(\mu_x + \mu_y) \geq 0,$$

we have

$$(1 + 2\theta(\mu_x + \mu_y))E^{m+1} \leq 2\theta(\mu_x + \mu_y)E^{m+1} + E^m + \Delta t T^m.$$

Hence,

$$E^{m+1} \leq E^m + \Delta t\, T^m, \qquad m = 0, 1, \ldots, M-1.$$

As $E^0 = 0$, upon summation we deduce that

$$\begin{aligned} E^m &\leq \Delta t \sum_{n=0}^{m-1} T^n \\ &\leq m\Delta t \max_{0 \leq n \leq m-1} T^n \\ &\leq T \max_{0 \leq m \leq M} \max_{1 \leq j \leq J-1} |T_{i,j}^m|, \end{aligned}$$

and we have that

$$\max_{i,j} \max_{0\le m\le M} |u(x_i, y_j, t_m) - U^m_{i,j}| \le T \max_{i,j} \max_{0\le m\le M} |T^m_{i,j}|.$$

The explicit and implicit Euler schemes therefore satisfy:

$$\max_{i,j} \max_{0\le m\le M} |u(x_i, y_j, t_m) - U^m_{i,j}| \le \text{Const.} \left((\Delta x)^2 + (\Delta y)^2 + \Delta t\right),$$

where in the case of the explicit Euler scheme we are assuming that $\mu_x + \mu_y \le \frac{1}{2}$, while for the Crank–Nicolson scheme we have that

$$\max_{i,j} \max_{0\le m\le M} |u(x_i, y_j, t_m) - U^m_{i,j}| \le \text{Const.} \left((\Delta x)^2 + (\Delta y)^2 + (\Delta t)^2\right),$$

assuming that $\mu_x + \mu_y \le 1$.

5.6.4 The alternating direction (ADI) method

Except for $\theta = 0$ corresponding to the explicit Euler scheme, for all other values of $\theta \in (0, 1]$ the θ-scheme is an implicit scheme, and its implementation therefore involves the solution of large systems of linear algebraic equations. This is true, in particular, for the Crank–Nicolson scheme corresponding to $\theta = \frac{1}{2}$. Our objective here is to propose a more economical scheme, which replaces the tedious task of solving such large systems of algebraic equations with the successive solution of smaller linear systems in the x and y co-ordinate directions respectively, alternating between solves in the x and y co-ordinate directions. The resulting finite difference scheme is called the alternating direction (or ADI) scheme. We describe its construction starting from the Crank–Nicolson scheme, which has the form:

$$\left(1 - \frac{1}{2}\mu_x\delta_x^2 - \mu_y\frac{1}{2}\delta_y^2\right) U^{m+1}_{i,j} = \left(1 + \frac{1}{2}\mu_x\delta_x^2 + \mu_y\frac{1}{2}\delta_y^2\right) U^m_{i,j},$$

for $i = 1, \ldots, J_x - 1$, $j = 1, \ldots, J_y - 1$, $m = 0, 1, \ldots, M - 1$, subject to the initial condition

$$U^0_{i,j} = u_0(x_i, y_j), \quad i = 0, \ldots, J_x, \;\; j = 0, \ldots, J_y,$$

and the boundary condition

$$U^m_{i,j} = B(x_i, y_j, t_m), \;\; \text{at the boundary mesh points, for } m = 1, \ldots, M.$$

Let us modify this scheme (subject to the same initial and boundary conditions) to:

$$\left(1 - \frac{1}{2}\mu_x\delta_x^2\right)\left(1 - \mu_y\frac{1}{2}\delta_y^2\right) U^{m+1}_{i,j} = \left(1 + \frac{1}{2}\mu_x\delta_x^2\right)\left(1 + \mu_y\frac{1}{2}\delta_y^2\right) U^m_{i,j}.$$

By introducing the intermediate level $U^{m+1/2}$, we can rewrite the last equality in the following equivalent form:

$$\left(1 - \frac{1}{2}\mu_x\delta_x^2\right) U^{m+1/2}_{i,j} = \left(1 + \frac{1}{2}\mu_y\delta_y^2\right) U^m_{i,j}, \qquad (1)$$

$$\left(1 - \frac{1}{2}\mu_y\delta_y^2\right) U^{m+1}_{i,j} = \left(1 + \frac{1}{2}\mu_x\delta_x^2\right) U^{m+1/2}_{i,j}. \qquad (2)$$

The equivalence is seen by applying

$$\left(1 + \frac{1}{2}\mu_x\delta_x^2\right) \text{ to eq. (1)} \quad \text{and} \quad \left(1 - \frac{1}{2}\mu_x\delta_x^2\right) \text{ to eq. (2)}.$$

The stability in the ℓ^2 norm of the ADI scheme (for the pure initial-value problem now, i.e. with no boundary conditions assumed) is easily seen by substituting the Fourier mode

$$U_{i,j}^m = [\lambda(k_x, k_y)]^m \mathrm{e}^{\imath(k_x x_i + k_y y_j)}$$

into the scheme. Hence,

$$\lambda(k_x, k_y) = \frac{\left(1 - 2\mu_x \sin^2 \frac{1}{2} k_x \Delta x\right)\left(1 - 2\mu_y \sin^2 \frac{1}{2} k_x \Delta y\right)}{\left(1 + 2\mu_x \sin^2 \frac{1}{2} k_x \Delta x\right)\left(1 + 2\mu_y \sin^2 \frac{1}{2} k_x \Delta y\right)}.$$

Clearly,

$$|\lambda(k_x, k_y)| \leq 1 \quad \forall\, (k_x, k_y) \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right] \times \left[-\frac{\pi}{\Delta y}, \frac{\pi}{\Delta y}\right].$$

Consequently, the ADI scheme is unconditionally stable in the ℓ_2 norm. The consistency error of the ADI scheme can be shown (again, after tedious Taylor series expansions) to be

$$T_{i,j}^m = \mathcal{O}\left((\Delta x)^2 + (\Delta y)^2 + (\Delta t)^2\right).$$

The ADI scheme satisfies a discrete maximum principle for $\mu_x \leq 1$ and $\mu_y \leq 1$. The proof of this is similar to the case of the θ-scheme in one space-dimension (cf. the textbook by K.W. Morton and D.F. Mayers, Numerical Solution of Partial Differential Equations: An Introduction, 2nd Edition, CUP, 2005. ISBN: 978-0-521607-93-3. pp. 64–65).

End of optional material

6 Finite difference approximation of hyperbolic equations

In this section we shall be concerned with the finite difference approximation of the second-order linear wave equation **Lecture 12**

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = f(x,t),$$

where $c > 0$ is the wave speed and f is a given source term.

In the simplest case when f is identically zero and the equation is considered on the whole real line, $x \in \mathbb{R}$, by supplying two initial conditions

$$\begin{aligned} u(x,0) &= u_0(x) \quad \text{for } x \in \mathbb{R}, \\ \frac{\partial u}{\partial t}(x,0) &= u_1(x) \quad \text{for } x \in \mathbb{R}, \end{aligned}$$

where u_0 and u_1 are continuous functions defined on $\mathbb{R}$, the solution is given by d'Alembert's formula

$$u(x,t) = \frac{1}{2}\left[u_0(x-ct) + u_0(x+ct)\right] + \frac{1}{2c}\int_{x-ct}^{x+ct} u_1(\xi)\,\mathrm{d}\xi.$$

More generally, if f is a continuous function on $\mathbb{R}\times[0,\infty)$, there is still an explicit formula for the solution

$$u(x,t) = \frac{1}{2}\left[u_0(x-ct) + u_0(x+ct)\right] + \frac{1}{2c}\int_{x-ct}^{x+ct} u_1(\xi)\,\mathrm{d}\xi + \frac{1}{2c}\int_0^t \int_{x-c(t-\tau)}^{x+c(t-\tau)} f(s,\tau)\,\mathrm{d}s\,\mathrm{d}\tau.$$

In this section, we shall be interested in a problem of the above form, but in the physically more realistic setting of a nonempty bounded closed spatial interval $[a,b]$ of the real line, where $a < b$, and on a finite time-interval $[0,T]$, where $T > 0$. In this case, in addition to the two initial conditions stated above, boundary conditions need to be prescribed at $x = a$ and $x = b$, and the problem under consideration thus becomes an initial-boundary-value problem.

6.1 Second-order hyperbolic equations: initial-boundary-value problem and energy estimate

Consider the closed interval $[a,b]$ of the real line, with $a < b$, and let $T > 0$. Suppose, in addition, that $T > 0$. We shall be concerned with the finite difference approximation of the initial-boundary-value problem

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} &= f(x,t) \qquad \text{for } (x,t) \in (a,b)\times(0,T], \\ u(x,0) &= u_0(x) \qquad \text{for } x \in [a,b], \\ \frac{\partial u}{\partial t}(x,0) &= u_1(x) \qquad \text{for } x \in [a,b], \\ u(a,t) = 0 \quad \text{and} \quad u(b,t) &= 0 \qquad \text{for } t \in [0,T]. \end{aligned} \tag{102}$$

Here, f is assumed to be a continuous real-valued function defined on $(a,b)\times[0,T]$, u_0 and u_1 are supposed to be continuous real-valued functions defined on $[a,b]$, and we shall assume compatibility of the initial data with the boundary conditions, in the sense that u_0 and u_1 will be required to vanish at both $x = a$ and $x = b$. As before, $c > 0$ is the wave speed.

Before embarking on the construction and the analysis of the finite difference approximations of (102), it is worth emphasizing that our key analytical tools will be 'discrete energy inequalities', which will imply the stability of the finite difference schemes under consideration, and which will also play a key role in their convergence analysis. We shall consider two finite difference schemes — an implicit scheme and an

explicit scheme — and the derivations of the corresponding discrete energy inequalities for these will be guided by the derivation of an energy inequality for the initial-boundary-value problem (102). We shall therefore begin by describing the derivation of the energy inequality (or energy estimate) satisfied by the solution of the initial-boundary-value problem (102). As the proof of existence of a solution to the initial-boundary-value problem (102) is beyond the scope of these lecture notes, we shall simply suppose here that a solution u to (102) exists and that u is sufficiently smooth, so that the calculations to be performed below are meaningful.

We begin by multiplying the partial differential equation $(102)_1$ by the time derivative of u, and we then integrate the resulting expression over the interval $[a, b]$; thus,

$$\int_a^b \frac{\partial^2 u}{\partial t^2}(x,t)\,\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x - c^2\int_a^b \frac{\partial^2 u}{\partial x^2}(x,t)\,\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x = \int_0^t f(x,t)\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x. \tag{103}$$

As $u(a,t) = 0$ and $u(b,t) = 0$ for all $t \in [0,T]$, it follows that

$$\frac{\partial u}{\partial t}(a,t) = 0 \quad \text{and} \quad \frac{\partial u}{\partial t}(b,t) = 0 \quad \text{for all } t \in [0,T].$$

Thus, by performing partial integration with respect to x in the second term on the left-hand side of (103), we arrive at the following equality:

$$\int_a^b \frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}(x,t)\right)\,\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x + c^2\int_a^b \frac{\partial u}{\partial x}(x,t)\,\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial x}(x,t)\right)\,\mathrm{d}x = \int_0^t f(x,t)\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x. \tag{104}$$

Clearly,

$$\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right)\,\frac{\partial u}{\partial t} = \frac{1}{2}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right)^2 \quad \text{and} \quad \frac{\partial u}{\partial x}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial x}\right) = \frac{1}{2}\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial x}\right)^2,$$

and therefore

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,t)\,\mathrm{d}x + \frac{c^2}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b \left(\frac{\partial u}{\partial x}\right)^2(x,t)\,\mathrm{d}x = \int_a^b f(x,t)\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x. \tag{105}$$

In the special case when f is identically zero, the right-hand side of (105) vanishes, and after integrating the resulting expression from 0 to t, for any $t \in (0,T]$, we deduce that

$$\frac{1}{2}\int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,t)\,\mathrm{d}x + \frac{c^2}{2}\int_a^b \left(\frac{\partial u}{\partial x}\right)^2(x,t)\,\mathrm{d}x = \frac{1}{2}\int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,0)\,\mathrm{d}x + \frac{c^2}{2}\int_a^b \left(\frac{\partial u}{\partial x}\right)^2(x,0)\,\mathrm{d}x. \tag{106}$$

If we view the expression on the left-hand side of the equality (106) as the 'total energy' at time t and the right-hand side as the 'initial total energy', then the equality (106) can be understood to be expressing conservation of the total energy during the course of the evolution of the solution from time 0 to time $t \in (0,T]$, in the absence of a source term.

After multiplying (105) by 2 and defining

$$\mathcal{L}^2(u(\cdot,t)) := \int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,t)\,\mathrm{d}x + c^2\int_a^b \left(\frac{\partial u}{\partial x}\right)^2(x,t)\,\mathrm{d}x$$

for $t \in [0,T]$, the equality (105) can be rewritten as

$$\mathcal{L}^2(u(\cdot,t)) = \mathcal{L}^2(u(\cdot,0)) \qquad \text{for all } t \in [0,T].$$

It is this 'energy equality' that we shall try to mimic in our stability analysis of the finite difference approximations of the initial-boundary-value problem (105) when f is identically 0. We note in passing

that the mapping $u \mapsto \max_{t\in[0,T]}[\mathcal{L}^2(u(\cdot,t))]^{1/2}$ is a norm on the linear space of continuous functions u defined on $[a,b]\times[0,T]$ such that $u(a,t) = u(b,t) = 0$ for all $t \in [0,T]$, and whose first partial derivatives with respect to x and t are continuous functions defined on $[a,b]\times[0,T]$.

More generally, if f is not identically zero, then (105) implies that

$$\mathcal{L}^2(u(\cdot,t)) = \mathcal{L}^2(u(\cdot,0)) + 2\int_0^t \int_a^b f(x,t)\frac{\partial u}{\partial t}(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau.$$

As

$$2\alpha\beta \le \alpha^2 + \beta^2, \qquad \text{for all } \alpha,\beta \in \mathbb{R},$$

it follows that

$$\begin{aligned}\mathcal{L}^2(u(\cdot,t)) &\le \mathcal{L}^2(u(\cdot,0)) + \int_0^t\int_a^b f^2(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau + \int_0^t\int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau \\ &\le \mathcal{L}^2(u(\cdot,0)) + \int_0^t\int_a^b f^2(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau + \int_0^t \mathcal{L}^2(u(\cdot,\tau))\,\mathrm{d}\tau.\end{aligned} \tag{107}$$

To proceed, we require the following result, called *Gronwall's Lemma*.

Lemma 14 *(Gronwall's Lemma) Suppose that A and B are continuous real-valued nonnegative functions defined on $[0,T]$, and B is a nondecreasing function of its argument. Suppose further that*

$$A(t) \le B(t) + \int_0^t A(s)\,\mathrm{d}s$$

for all $t \in [0,T]$; then

$$A(t) \le \mathrm{e}^t\, B(t)$$

for all $t \in [0,T]$.

Proof: Clearly,

$$\mathrm{e}^{-t}\,A(t) - \mathrm{e}^{-t}\int_0^t A(s)\,\mathrm{d}s \le \mathrm{e}^{-t}\,B(t),$$

and therefore, equivalently,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\mathrm{e}^{-t}\int_0^t A(s)\,\mathrm{d}s\right] \le \mathrm{e}^{-t}\,B(t).$$

Hence, by integrating and observing that the expression in the square brackets on the left-hand side of the last inequality vanishes at $t = 0$, we find that

$$\mathrm{e}^{-t}\int_0^t A(s)\,\mathrm{d}s \le \int_0^t \mathrm{e}^{-s}\,B(s)\,\mathrm{d}s.$$

Multiplying this inequality by e^t, and because B is by hypothesis a nondecreasing function, whereby $B(s) \le B(t)$ for all $s \in [0,t]$, we have that

$$\int_0^t A(s)\,\mathrm{d}s \le \mathrm{e}^t\,B(t)\int_0^t \mathrm{e}^{-s}\,\mathrm{d}s = \mathrm{e}^t\,B(t)\,(1-\mathrm{e}^{-t}) = \mathrm{e}^t\,B(t) - B(t).$$

By substituting this into the right-hand side of the assumed inequality, it follows that $A(t) \le B(t) + \mathrm{e}^t\,B(t) - B(t) = \mathrm{e}^t\,B(t)$, as has been asserted. That completes the proof. □

We now return to (107) and set

$$A(t) := \mathcal{L}^2(u(\cdot,t)) \quad \text{and} \quad B(t) := \mathcal{L}^2(u(\cdot,0)) + \int_0^t\int_a^b f^2(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau$$

It then follows from Gronwall's inequality that $A(t) \leq \mathrm{e}^t B(t)$, that is

$$\mathcal{L}^2(u(\cdot,t)) \leq \mathrm{e}^t \left(\mathcal{L}^2(u(\cdot,0)) + \int_0^t \int_a^b f^2(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau \right),$$

with

$$\mathcal{L}^2(u(\cdot,t)) := \int_a^b \left(\frac{\partial u}{\partial t}\right)^2 (x,t)\,\mathrm{d}x + c^2 \int_a^b \left(\frac{\partial u}{\partial x}\right)^2 (x,t)\,\mathrm{d}x$$

and

$$\mathcal{L}^2(u(\cdot,0)) := \int_a^b \left(\frac{\partial u}{\partial t}\right)^2 (x,0)\,\mathrm{d}x + c^2 \int_a^b \left(\frac{\partial u}{\partial t}\right)^2 (x,0)\,\mathrm{d}x = \|u_1\|^2_{L_2(a,b)} + c^2\,|u_0|^2_{H^1(a,b)},$$

which is the desired energy inequality satisfied by the solution. It provides a bound on the (square of the) norm of the solution in terms of the (square of the) norm of the initial data and the (square of the) L^2 norm of the source term f. We shall mimic the derivation of this energy inequality in the stability analysis of the implicit and explicit finite difference approximations of the initial-boundary-value problem (102) in the general case when f is not identically zero.

6.2 The implicit scheme: stability, consistency and convergence

Lecture 13

For $M \geq 2$, we define $\Delta t := T/M$, and for $J \geq 2$ the spatial step is taken to be $\Delta x := (b-a)/J$. We let $x_j := a + j\Delta x$ for $j = 0, 1, \ldots, J$ and $t_m := m\Delta t$ for $m = 0, 1, \ldots, M$. On the space-time mesh $\{(x_j, t_m) : 0 \leq j \leq J,\ 0 \leq m \leq M\}$ we consider the finite difference scheme

$$\begin{aligned}
\frac{U_j^{m+1} - 2U_j^m + U_j^{m-1}}{\Delta t^2} - c^2 \frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{\Delta x^2} &= f(x_j, t_{m+1}) && \text{for } \begin{cases} j = 1, \ldots, J-1, \\ m = 1, \ldots, M-1, \end{cases} \\
U_j^0 &= u_0(x_j) && \text{for } j = 0, 1, \ldots, J, \\
U_j^1 &= U_j^0 + \Delta t\, u_1(x_j) && \text{for } j = 1, 2, \ldots, J-1, \\
U_0^m = 0 \text{ and } \quad U_J^m &= 0 && \text{for } m = 1, \ldots, M.
\end{aligned} \tag{108}$$

The second numerical initial condition, featuring in equation $(108)_3$, stems from the observation that, if $\frac{\partial^2 u}{\partial t^2} \in C([a,b] \times [0,T])$, then

$$\frac{u(x_j, \Delta t) - U_j^0}{\Delta t} = \frac{u(x_j, \Delta t) - u(x_j, 0)}{\Delta t} = \frac{\partial u}{\partial t}(x_j, 0) + \mathcal{O}(\Delta t) = u_1(x_j) + \mathcal{O}(\Delta t),$$

upon ignoring the $\mathcal{O}(\Delta t)$ term, at the cost of replacing $u(x_j, \Delta t)$ by its numerical approximation U_j^1.

Once the values of U_j^{m-1} and U_j^m, for $j = 0, \ldots, J$, have been computed (or have been specified by the initial data, in the case of $m = 1$), the subsequent values U_j^{m+1}, $j = 0, \ldots, J$, need to be computed by solving a system of $J-1$ linear algebraic equations for the $J-1$ unknowns U_j^{m+1}, $j = 0, \ldots, J-1$. The finite difference scheme (108) is therefore usually referred to as the *implicit scheme* for the initial-boundary-value problem (102).

Stability of the implicit scheme. We shall consider the inner products

$$(U, V) := \sum_{j=1}^{J-1} \Delta x\, U_j\, V_j,$$

$$(U, V] := \sum_{j=1}^{J} \Delta x\, U_j\, V_j,$$

and the associated norms, respectively, $\|\cdot\|$ and $\|\cdot]|$, defined by $\|U\| := (U,U)^{\frac{1}{2}}$ and $\|U]| := (U,U]^{\frac{1}{2}}$.

Note that for two mesh functions defined on the computational mesh $\{x_j : j = 1,\dots,J-1\}$ one has that

$$(A-B,A) = \frac{1}{2}(\|A\|^2 - \|B\|^2) + \frac{1}{2}\|A-B\|^2.$$

Thus, by taking $A = U^{m+1} - U^m$ and $B = U^m - U^{m-1}$, we have that

$$(U^{m+1} - 2U^m + U^{m-1}, U^{m+1} - U^m) = \frac{1}{2}(\|U^{m+1} - U^m\|^2 - \|U^m - U^{m-1}\|^2) + \frac{1}{2}\|U^{m+1} - 2U^m + U^{m-1}\|^2.$$

We note further that, similarly as above,

$$(A-B,A] = \frac{1}{2}(|A]|^2 - |B]|^2) + \frac{1}{2}|A-B]|^2.$$

Hence, by performing a summation by parts and then taking $A = D_x^+ U^{m+1}$ and $B = D_x^+ U^m$ we have

$$\begin{aligned}\left(-D_x^+ D_x^- U^{m+1}, U^{m+1} - U^m\right) &= (D_x^- U^{m+1}, D_x^-(U^{m+1} - U^m)] \\ &= (D_x^- U^{m+1} - D_x^- U^m, D_x^- U^{m+1}] \\ &= \frac{1}{2}(|D_x^- U^{m+1}]|^2 - |D_x^- U^m]|^2) + \frac{1}{2}|D_x^-(U^{m+1} - U^m)]|^2.\end{aligned}$$

By taking the $(\cdot,\cdot)$ inner product of $(115)_1$ with $U^{m+1} - U^m$ and using the identities stated above we therefore obtain:

$$\begin{aligned}&\frac{1}{2}\left(\left\|\frac{U^{m+1} - U^m}{\Delta t}\right\|^2 - \left\|\frac{U^m - U^{m-1}}{\Delta t}\right\|^2\right) + \frac{1}{2}\,\Delta t^2 \left\|\frac{U^{m+1} - 2U^m + U^{m-1}}{\Delta t^2}\right\|^2 \\ &\quad + \frac{1}{2}(|D_x^- U^{m+1}]|^2 - |D_x^- U^m]|^2) + \frac{1}{2}\,\Delta t^2 \left\|D_x^-\left(\frac{U^{m+1} - U^m}{\Delta t}\right)\right]\right\|^2 = (f(\cdot,t_{m+1}), U^{m+1} - U^m).\end{aligned} \tag{109}$$

In the special case when f is identically zero (109) implies that

$$\left\|\frac{U^{m+1} - U^m}{\Delta t}\right\|^2 + |D_x^- U^{m+1}]|^2 \le \left\|\frac{U^m - U^{m-1}}{\Delta t}\right\|^2 + |D_x^- U^m]|^2. \tag{110}$$

Let us define the nonnegative expression

$$\mathcal{M}^2(U^m) := \left\|\frac{U^{m+1} - U^m}{\Delta t}\right\|^2 + |D_x^- U^{m+1}]|^2.$$

With this notation (110) becomes

$$\mathcal{M}^2(U^m) \le \mathcal{M}^2(U^{m-1}), \qquad \text{for all } m = 1,\dots,M-1,$$

and therefore

$$\mathcal{M}^2(U^m) \le \mathcal{M}^2(U^0), \qquad \text{for all } m = 1,\dots,M-1.$$

One can verify that the mapping $U \mapsto \max_{m\in\{0,\dots,M-1\}}[\mathcal{M}^2(U^m)]^{1/2}$ is a norm on the linear space of mesh functions U defined on the space-time mesh $\{(x_j,t_m) : j = 0,1,\dots,J,\ m = 0,1,\dots,M\}$ such that $U_0^m = U_J^m = 0$ for all $m = 0,1,\dots,M$. Thus we have shown that when f is identically zero the implicit scheme (102) is (unconditionally) stable in this norm.

We now return to the general case when f is not identically zero. Our starting point is the equality (109) and we focus our attention on the term on its right-hand side. By the Cauchy–Schwarz inequality,

$$\begin{aligned}(f(\cdot,t_{m+1}),U^{m+1}-U^m) &\le \|f(\cdot,t_{m+1})\|\,\|U^{m+1}-U^m\| \\ &= \sqrt{\Delta t\,T}\,\|f(\cdot,t_{m+1})\|\,\sqrt{\frac{\Delta t}{T}}\,\left\|\frac{U^{m+1}-U^m}{\Delta t}\right\| \\ &\le \frac{\Delta t\,T}{2}\,\|f(\cdot,t_{m+1})\|^2+\frac{\Delta t}{2T}\,\left\|\frac{U^{m+1}-U^m}{\Delta t}\right\|^2,\end{aligned} \tag{111}$$

where in the transition to the last line we have made use of the elementary inequality

$$\alpha\beta \le \frac{1}{2}\alpha^2+\frac{1}{2}\beta^2, \qquad \text{for } \alpha,\beta\in\mathbb{R}.$$

By substituting (111) into (109) we deduce that

$$\left(1-\frac{\Delta t}{T}\right)\left(\left\|\frac{U^{m+1}-U^m}{\Delta t}\right\|^2+\|D_x^-U^{m+1}]|^2\right) \le \left\|\frac{U^m-U^{m-1}}{\Delta t}\right\|^2+\|D_x^-U^m]|^2+\Delta t\,T\,\|f(\cdot,t_{m+1})\|^2. \tag{112}$$

By recalling the definition of $\mathcal{M}^2(U^m)$ we can rewrite (112) in the following compact form:

$$\left(1-\frac{\Delta t}{T}\right)\mathcal{M}^2(U^m) \le \mathcal{M}^2(U^{m-1})+\Delta t\,T\,\|f(\cdot,t_{m+1})\|^2.$$

As, by assumption, $M\ge 2$, it follows that $\Delta t := T/M \le T/2$, whereby $\Delta t/T \le 1/2$. By noting that

$$1-x \ge \frac{1}{1+2x} \qquad \forall\, x\in\left[0,\frac{1}{2}\right],$$

it follows with $x=\Delta t/T$ that

$$\begin{aligned}\mathcal{M}^2(U^m) &\le \left(1+\frac{2\,\Delta t}{T}\right)\mathcal{M}^2(U^{m-1})+\Delta t\,T\left(1+\frac{2\,\Delta t}{T}\right)\|f(\cdot,t_{m+1})\|^2 \\ &\le \left(1+\frac{2\,\Delta t}{T}\right)\mathcal{M}^2(U^{m-1})+2\,\Delta t\,T\,\|f(\cdot,t_{m+1})\|^2.\end{aligned}$$

To proceed, we require the following result, which is easily proved by induction.

Lemma 15 *Suppose that $M\ge 2$ is an integer, $\{a_m\}_{m=0}^{M-1}$ and $\{b_m\}_{m=1}^{M-1}$ are nonnegative real numbers, $\alpha>0$, and*

$$a_m \le \alpha\,a_{m-1}+b_m \quad \textit{for } m=1,2,\ldots,M-1.$$

Then,

$$a_m \le \alpha^m a_0+\sum_{k=1}^{m}\alpha^{m-k}b_k.$$

We shall apply Lemma 15 with

$$a_m=\mathcal{M}^2(U^m), \quad b_m=2\,\Delta t\,T\,\|f(\cdot,t_{m+1})\|^2, \quad \alpha=1+\frac{2\,\Delta t}{T}$$

to deduce that

$$\mathcal{M}^2(U^m) \le \left(1+\frac{2\,\Delta t}{T}\right)^m\mathcal{M}(U^0)+2\,\Delta t\,T\sum_{k=1}^{m}\left(1+\frac{2\,\Delta t}{T}\right)^{m-k}\|f(\cdot,t^{k+1})\|^2 \quad \text{for } m=1,2,\ldots,M-1.$$

We note that

$$\left(1+\frac{2\,\Delta t}{T}\right)^m \le \left(1+\frac{2\,\Delta t}{T}\right)^M = \left(1+\frac{2\,\Delta t}{T}\right)^{\frac{T}{\Delta t}} \le \mathrm{e}^2,$$

where the last inequality follows from the inequality

$$(1+2x)^{\frac{1}{x}} \le e^2 \quad \forall\, x \in \left(0, \frac{1}{2}\right],$$

which, in turn, follows by noting that $\log(1+x) \le x$ for all $x \ge 0$. Thus we deduce the following stability result for the implicit scheme (102).

Theorem 16 *The implicit finite difference approximation* (102) *of the initial-boundary-value problem* (102)*, on a finite difference mesh of spacing* $\Delta x = (b-a)/J$ *with* $J \ge 2$ *in the* x*-direction and* $\Delta t = T/M$ *with* $M \ge 2$ *in the* t*-direction, is (unconditionally) stable in the sense that*

$$\mathcal{M}^2(U^m) \le \mathrm{e}^2\,\mathcal{M}^2(U^0) + 2\,\mathrm{e}^2\,T\sum_{k=1}^{m}\Delta t\,\|f(\cdot,t_{k+1})\|^2, \quad \textit{for } m = 1,\ldots,M-1,$$

independently of the choice of Δx *and* Δt*.*

Consistency of the implicit scheme. We define the consistency error of the scheme by

$$T_j^{m+1} := \frac{u_j^{m+1}-2u_j^m+u_j^{m-1}}{\Delta t^2} - c^2\frac{u_{j+1}^{m+1}-2u_j^{m+1}+2u_{j-1}^{m+1}}{\Delta x^2} - f(x_j,t_{m+1}), \qquad \begin{cases} j=1,\ldots,J-1,\\ m=1,\ldots,M-1, \end{cases}$$

and

$$T_j^1 := \frac{u_j^1-u_j^0}{\Delta t} - u_1(x_j), \qquad j=1,\ldots,J-1,$$

where $u_j^m := u(x_j,t_m)$. As

$$f(x_j,t_{m+1}) = \frac{\partial^2 u}{\partial t^2}(x_j,t_{m+1}) - c^2\,\frac{\partial^2 u}{\partial x^2}(x_j,t_{m+1}) \quad \text{and} \quad u_1(x_j) = \frac{\partial u}{\partial t}(x_j,0),$$

it follows that

$$T_j^{m+1} := \left(\frac{u_j^{m+1}-2u_j^m+u_j^{m-1}}{\Delta t^2} - \frac{\partial^2 u}{\partial t^2}(x_j,t_{m+1})\right) - c^2\left(\frac{u_{j+1}^{m+1}-2u_j^{m+1}+2u_{j-1}^{m+1}}{\Delta x^2} - \frac{\partial^2 u}{\partial x^2}(x_j,t_{m+1})\right)$$

for $j=1,\ldots,J-1$ and $m=1,\ldots,M-1$ and

$$T_j^1 = \frac{u_j^1-u_j^0}{\Delta t} - \frac{\partial u}{\partial t}(x_j,0)$$

for $j=1,\ldots,J-1$. It follows by Taylor series expansion of u about the point (x_j,t_{m+1}) that

$$\frac{u_j^{m+1}-2u_j^m+u_j^{m-1}}{\Delta t^2} - \frac{\partial^2 u}{\partial t^2}(x_j,t_{m+1}) = \frac{1}{3}\Delta t\left(\frac{\partial^3 u}{\partial t^3}(x_j,\eta_m) - 4\frac{\partial^3 u}{\partial t^3}(x_j,\eta_{m-1})\right),$$

where $\eta_{m-1} \in [t_{m-1},t_{m+1}]$ and $\eta_m \in [t_m,t_{m+1}]$, provided that the third partial derivative of u with respect to t is a continuous function on $[a,b]\times[0,T]$; and

$$\frac{u_{j+1}^{m+1}-2u_j^{m+1}+u_{j-1}^{m+1}}{\Delta x^2} - \frac{\partial^2 u}{\partial x^2}(x_j,t_{m+1}) = \frac{1}{12}\Delta x^2\,\frac{\partial^4 u}{\partial x^4}(\xi_j,t_{m+1})$$

where $\xi_j \in [x_{j-1}, x_{j+1}]$, provided that the fourth partial derivative of u with respect to x is a continuous function on $[a,b] \times [0,T]$. Hence,

$$|T_j^{m+1}| \le \frac{1}{12} c^2 \Delta x^2 M_{4x} + \frac{5}{3} \Delta t M_{3t}, \qquad \begin{cases} j = 1, \ldots, J-1, \\ m = 1, \ldots, M-1, \end{cases} \tag{113}$$

where

$$M_{4x} := \max_{(x,t) \in [a,b] \times [0,T]} \left| \frac{\partial^4 u}{\partial x^4}(x,t) \right| \quad \text{and} \quad M_{3t} := \max_{(x,t) \in [a,b] \times [0,T]} \left| \frac{\partial^3 u}{\partial t^3}(x,t) \right|.$$

Similarly, by Taylor series expansion with an integral remainder term,

$$T_j^1 = \frac{1}{\Delta t} \int_0^{\Delta t} (\Delta t - t) \, \frac{\partial^2 u}{\partial t^2}(x_j, t) \, \mathrm{d}t, \tag{114}$$

and therefore

$$|T_j^1| \le \frac{1}{2} \Delta t \, M_{2t}, \quad j = 1, \ldots, J-1,$$

where

$$M_{2t} := \max_{(x,t) \in [a,b] \times [0,T]} \left| \frac{\partial^2 u}{\partial t^2}(x,t) \right|.$$

Convergence of the implicit scheme. In the rest of the section we shall explore the convergence of the finite difference scheme (102). To this end, we define the *global error*

$$e_j^m := u(x_j, t_m) - U_j^m, \qquad \begin{cases} j = 0, \ldots, J, \\ m = 0, \ldots, M. \end{cases}$$

It follows from the definitions of T_j^{m+1} and T_j^1 that

$$\frac{e_j^{m+1} - 2e_j^m + e_j^{m-1}}{\Delta t^2} - c^2 \frac{e_{j+1}^{m+1} - 2e_j^{m+1} + 2e_{j-1}^{m+1}}{\Delta x^2} = T_j^{m+1}, \qquad \begin{cases} j = 1, \ldots, J-1, \\ m = 1, \ldots, M-1, \end{cases}$$

and

$$e_j^1 = e_j^0 + \Delta t \, T_j^1, \qquad j = 1, \ldots, J-1.$$

Furthermore, $e_j^0 = 0$ for $j = 0, 1, \ldots, J$, and $e_0^m = e_J^m = 0$ for $m = 1, \ldots, M$. Hence, the global error e satisfies an identical finite difference scheme as U, but with $f(x_j, t_{m+1})$ replaced by T_j^{m+1} and $u_1(x_j)$ replaced by T_j^1. It therefore follows from Theorem 16 with U^m replaced by e^m, U^0 replaced by e^0 and $f(x_j, t_{k+1})$ replaced by T_j^{k+1} for $j = 1, \ldots, J-1$ and $k = 1, \ldots, M-1$, that

$$\mathcal{M}^2(e^m) \le \mathrm{e}^2 \, \mathcal{M}^2(e^0) + 2 \, \mathrm{e}^2 \, T \sum_{k=1}^{m} \Delta t \, \left\| T^{k+1} \right\|^2, \quad \text{for } m = 1, \ldots, M-1.$$

Now, because $(J-1)\Delta x \le (b-a)$, it follows from (113) that

$$\max_{1 \le k \le m} \left\| T^{k+1} \right\|^2 = \max_{1 \le k \le m} \sum_{j=1}^{J-1} \Delta x \, |T_j^{k+1}|^2 \le (b-a) \left[\frac{1}{12} c^2 \Delta x^2 M_{4x} + \frac{5}{3} \Delta t M_{3t} \right]^2.$$

On the other hand,

$$\mathcal{M}^2(e^0) = \left\| \frac{e^1 - e^0}{\Delta t} \right\|^2 + \|D_x^- e^1]|^2 = \|T^1\|^2 + \|D_x^- e^1]|^2 \le (b-a) \left[\frac{1}{2} \Delta t \, M_{2t} \right]^2 + \|D_x^- e^1]|^2.$$

As, by recalling (114),

$$D_x^- e_j^1 = D_x^- e_j^0 + \Delta t\, D_x^- T_j^1 = \Delta t\, D_x^- T_j^1 = \int_0^{\Delta t} (\Delta t - t)\, D_x^- \frac{\partial^2 u}{\partial t^2}(x_j, t)\, \mathrm{d}t$$
$$= \frac{1}{\Delta x} \int_0^{\Delta t} (\Delta t - t) \int_{x_{j-1}}^{x_j} \frac{\partial^3 u}{\partial x\, \partial t^2}(x, t)\, \mathrm{d}x\, \mathrm{d}t,$$

we have that

$$|D_x^- e_j^1| \leq \frac{1}{2}\, \Delta t^2\, M_{1x2t}, \qquad \text{where} \quad M_{1x2t} := \max_{(x,t)\in[a,b]\times[0,T]} \left| \frac{\partial^3 u}{\partial x \partial t^2} \right|,$$

whereby

$$\|D_x^- e^1]\|^2 \leq (b-a) \left[\frac{1}{2} \Delta t^2 M_{1x2t} \right]^2.$$

Therefore,

$$\mathcal{M}^2(e^0) \leq (b-a) \left[\frac{1}{2} \Delta t\, M_{2t} \right]^2 + (b-a) \left[\frac{1}{2}\, \Delta t^2 M_{1x2t} \right]^2.$$

Hence, finally,

$$\mathcal{M}^2(e^m) \leq \mathrm{e}^2 (b-a) \left[\frac{1}{2} \Delta t\, M_{2t} \right]^2 + \mathrm{e}^2 (b-a) \left[\frac{1}{2}\, \Delta t^2 M_{1x2t} \right]^2 + \mathrm{e}^2 T^2 (b-a) \left[\frac{1}{12} c^2 \Delta x^2 M_{4x} + \frac{5}{3} \Delta t M_{3t} \right]^2$$

for $m = 1, \ldots, M-1$. Thus, provided that M_{2t}, M_{1x2t}, M_{4x} and M_{3t} are all finite, we have that

$$\max_{m\in\{1,\ldots,M-1\}} [\mathcal{M}^2(u^m - U^m)]^{\frac{1}{2}} = \mathcal{O}(\Delta x^2 + \Delta t),$$

meaning that the implicit scheme exhibits second order convergence with respect to the spatial discretization step Δx and first-order convergence with respect to the temporal discretization step Δt in the norm $\max_{m\in\{1,\ldots,M-1\}} [\mathcal{M}^2(\cdot)]^{\frac{1}{2}}$.

6.3 The explicit scheme: stability, consistency and convergence

Lecture 14

For $M \geq 2$, we define $\Delta t := T/M$, and for $J \geq 2$ the spatial step is taken to be $\Delta x := (b-a)/J$. We let $x_j := a + j\Delta x$ for $j = 0, 1, \ldots, J$ and $t_m := m\Delta t$ for $m = 0, 1, \ldots, M$. On the space-time mesh $\{(x_j, t_m) : 0 \leq j \leq J,\ 0 \leq m \leq M\}$ we consider the finite difference scheme

$$\begin{aligned}
\frac{U_j^{m+1} - 2U_j^m + U_j^{m-1}}{\Delta t^2} - c^2 \frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{\Delta x^2} &= f(x_j, t_m) && \text{for } \begin{cases} j = 1, \ldots, J-1, \\ m = 1, \ldots, M-1, \end{cases} \\
U_j^0 &= u_0(x_j) && \text{for } j = 0, 1, \ldots, J, \\
U_j^1 &= U_j^0 + \Delta t\, u_1(x_j) && \text{for } j = 1, 2, \ldots, J-1, \\
U_0^m = 0 \text{ and } \quad U_J^m &= 0 && \text{for } m = 1, \ldots, M.
\end{aligned} \tag{115}$$

As in the case of the implicit scheme (108), the second numerical initial condition, appearing in $(115)_3$, stems from the observation that

$$\frac{u(x_j, \Delta t) - U_j^0}{\Delta t} = \frac{u(x_j, \Delta t) - u(x_j, 0)}{\Delta t} = \frac{\partial u}{\partial t}(x_j, 0) + \mathcal{O}(\Delta t) = u_1(x_j) + \mathcal{O}(\Delta t),$$

upon ignoring the $\mathcal{O}(\Delta t)$ term, at the cost of replacing $u(x_j, \Delta t)$ by its numerical approximation U_j^1.

A more accurate second numerical initial condition can be obtained by observing that, if f is assumed to be a continuous real-valued function defined on $[a,b]\times[0,T]$, $\frac{\partial^3 u}{\partial t^3}\in C([a,b]\times[0,T])$ and $u_1\in C^4([a,b])$, then

$$\begin{aligned}\frac{u(x_j,\Delta t)-U_j^0}{\Delta t}&=\frac{u(x_j,\Delta t)-u(x_j,0)}{\Delta t}=\frac{\partial u}{\partial t}(x_j,0)+\frac{1}{2}\Delta t\,\frac{\partial^2 u}{\partial t^2}(x_j,0)+\mathcal{O}(\Delta t^2)\\&=u_1(x_j)+\frac{1}{2}\Delta t\left(c^2\frac{\partial^2 u}{\partial x^2}(x_j,0)+f(x_j,0)\right)+\mathcal{O}(\Delta t^2)\\&=u_1(x_j)+\frac{1}{2}\Delta t\left(c^2D_x^+D_x^-u_1(x_j)+f(x_j,0)\right)+\mathcal{O}(\Delta t\,\Delta x^2+\Delta t^2).\end{aligned}$$

One could therefore, instead of $(115)_3$, use the following more accurate second initial condition:

$$U_j^1=U_j^0+\Delta t\,u_1(x_j)+\frac{1}{2}\Delta t^2\left(c^2D_x^+D_x^-u_1(x_j)+f(x_j,0)\right). \tag{116}$$

Once the values of U_j^{m-1} and U_j^m, for $j=0,\ldots,J$, have been computed (or have been specified by the initial data, in the case of $m=1$), the subsequent values U_j^{m+1}, $j=0,\ldots,J$, for $m=1,\ldots,M-1$, can be computed explicitly from (115), without having to solve systems of linear algebraic equations; hence the terminology *explicit scheme*.

Stability of the explicit scheme. We begin our exploration of the properties of the finite difference scheme (115) by investigating its stability. It will transpire from the analysis that will follow that the explicit scheme is, unlike the implicit scheme, which was shown to be unconditionally stable, now only conditionally stable: we shall prove its stability in a certain 'energy norm', whose precise definition which will emerge during the course of our analysis, — the stability condition for the explicit scheme being that $c\Delta t/\Delta x\le c_0$, for some positive constant $c_0\in(0,1)$.

First note that, for any $j\in\{0,\ldots,J\}$ and $m\in\{1,\ldots,M-1\}$,

$$\begin{aligned}U_j^{m+1}-U_j^{m-1}=(U_j^{m+1}-U_j^m)+(U_j^m-U_j^{m-1})&=(U_j^{m+1}+U_j^m)-(U_j^m+U_j^{m-1}),\\U_j^{m+1}-2U_j^m+U_j^{m-1}&=(U_j^{m+1}-U_j^m)-(U_j^m-U_j^{m-1}),\\U_j^{m+1}+2U_j^m+U_j^{m-1}&=(U_j^{m+1}+U_j^m)+(U_j^m+U_j^{m-1}).\end{aligned} \tag{117}$$

The left-hand side of equality $(115)_1$ can be rewritten as

$$\begin{aligned}&\frac{U_j^{m+1}-2U_j^m+U_j^{m-1}}{\Delta t^2}-c^2D_x^+D_x^-U_j^m\\&=\frac{U_j^{m+1}-2U_j^m+U_j^{m-1}}{\Delta t^2}+\frac{c^2\Delta t^2}{4}D_x^+D_x^-\frac{U_j^{m+1}-2U_j^m+U_j^{m-1}}{\Delta t^2}-c^2D_x^+D_x^-\frac{U_j^{m+1}+2U_j^m+U_j^{m-1}}{4}\end{aligned}$$

for $j=1,\ldots,J-1$, where I signifies the identity operator, which maps any mesh function defined on the spatial mesh $\{x_j:j=1,\ldots,J-1\}$ into itself. Insertion of this into $(115)_1$ then yields

$$\left(I+\frac{1}{4}c^2\Delta t^2D_x^+D_x^-\right)\frac{U_j^{m+1}-2U_j^m+U_j^{m-1}}{\Delta t^2}=c^2D_x^+D_x^-\frac{U_j^{m+1}+2U_j^m+U_j^{m-1}}{4}+f(x_j,t_m) \tag{118}$$

for $j=1,\ldots,J-1$, $m=1,\ldots,M-1$. We shall consider the inner products

$$(U,V):=\sum_{j=1}^{J-1}\Delta x\,U_j\,V_j,$$

$$(U,V]:=\sum_{j=1}^{J}\Delta x\,U_j\,V_j,$$

and the associated norms, respectively, $\|\cdot\|$ and $\|\cdot]|$, defined by $\|U\| := (U,U)^{\frac{1}{2}}$ and $\|U]| := (U,U]^{\frac{1}{2}}$, take the $(\cdot,\cdot)$ inner product of $(115)_1$ with $U^{m+1}-U^{m-1}$, making use of $(117)_2$ and the first equality in $(117)_1$ on the left-hand side, and $(117)_2$ and the second equality in $(117)_1$ on the right-hand side. Thus,

$$\begin{aligned}
&\left(\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+D_x^-\right)\frac{U^{m+1}-U^m}{\Delta t},\frac{U^{m+1}-U^m}{\Delta t}\right)-\left(\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+D_x^-\right)\frac{U^m-U^{m-1}}{\Delta t},\frac{U^m-U^{m-1}}{\Delta t}\right)\\
&\quad=-c^2\left(-D_x^+D_x^-\frac{U^{m+1}+U^m}{2},\frac{U^{m+1}+U^m}{2}\right)+c^2\left(-D_x^+D_x^-\frac{U^m+U^{m-1}}{2},\frac{U^m+U^{m-1}}{2}\right)\\
&\qquad+(f(\cdot,t_m),U^{m+1}-U^{m-1}).
\end{aligned}$$

Next, we shall perform summations by parts in the first two terms on the right-hand side, using that, for any mesh-function V defined on $\{x_j : j=0,\dots,J\}$ and such that $V_0=V_J=0$, one has

$$(-D_x^+D_x^-V,V)=(D_x^-V,D_x^-V]=\|D_x^-V]|^2.$$

Using these equalities with $V=\frac{1}{2}(U^{m+1}+U^m)$ and $V=\frac{1}{2}(U^m+U^{m-1})$, we deduce that

$$\begin{aligned}
&\left(\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+D_x^-\right)\frac{U^{m+1}-U^m}{\Delta t},\frac{U^{m+1}-U^m}{\Delta t}\right)-\left(\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+D_x^-\right)\frac{U^m-U^{m-1}}{\Delta t},\frac{U^m-U^{m-1}}{\Delta t}\right)\\
&\quad=-c^2\left(D_x^-\frac{U^{m+1}+U^m}{2},D_x^-\frac{U^{m+1}+U^m}{2}\right]+c^2\left(D_x^-\frac{U^m+U^{m-1}}{2},D_-^x\frac{U^m+U^{m-1}}{2}\right]\\
&\qquad+(f(\cdot,t_m),U^{m+1}-U^{m-1})\\
&\quad=-c^2\left\|D_x^-\frac{U^{m+1}+U^m}{2}\right]\right\|^2+c^2\left\|D_x^-\frac{U^m+U^{m-1}}{2}\right]\right\|^2+(f(\cdot,t_m),U^{m+1}-U^{m-1}).
\end{aligned}$$

This implies, after a rearrangement of terms, that

$$\begin{aligned}
&\left(\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+D_x^-\right)\frac{U^{m+1}-U^m}{\Delta t},\frac{U^{m+1}-U^m}{\Delta t}\right)+c^2\left\|D_x^-\frac{U^{m+1}+U^m}{2}\right]\right\|^2\\
&\quad=\left(\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+D_x^-\right)\frac{U^m-U^{m-1}}{\Delta t},\frac{U^m-U^{m-1}}{\Delta t}\right)+c^2\left\|D_x^-\frac{U^m+U^{m-1}}{2}\right]\right\|^2\\
&\qquad+(f(\cdot,t_m),U^{m+1}-U^{m-1}).
\end{aligned}\tag{119}$$

The second term on the left-hand side of (119) is nonnegative, as is the second term on the right-hand side. We would therefore like to ensure that first term on the left-hand side of (119) and the first term on the right-hand side are also nonnegative. We shall therefore make a short diversion to investigate this question. Letting

$$V_j^m := \frac{U_j^{m+1}-U_j^m}{\Delta t},\qquad j=0,\dots,J,$$

and noting that $V_0^m=V_J^m=0$, it follows that

$$\begin{aligned}
\left(\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+D_x^-\right)V^m,V^m\right)&=\|V^m\|^2+\frac{1}{4}c^2\Delta t^2(D_x^+D_x^-V^m,V^m)\\
&=\|V^m\|^2-\frac{1}{4}c^2\Delta t^2(D_x^-V^m,D_x^-V^m]\\
&=\|V^m\|^2-\frac{1}{4}c^2\Delta t^2\|D_x^-V^m]|^2.
\end{aligned}$$

Now, noting that for any nonnegative real numbers α and β one has $(\alpha-\beta)^2 \le 2\alpha^2+2\beta^2$, it follows that

$$\begin{aligned}\|D_x^- V^m]|^2 &= \sum_{j=1}^{J} \Delta x\, |D_x^- V_j^m|^2 = (\Delta x)^{-1} \sum_{j=1}^{J} (V_j^m - V_{j-1}^m)^2 \\ &\le 2\,(\Delta x)^{-1} \sum_{j=1}^{J} (V_j^m)^2 + (V_{j-1}^m)^2 = 4\,(\Delta x)^{-1} \sum_{j=1}^{J-1} (V_j^m)^2 \\ &= 4\,(\Delta x)^{-2} \sum_{j=1}^{J-1} \Delta x\, (V_j^m)^2 = \left(\frac{2}{\Delta x}\right)^2 \|V\|^2.\end{aligned}$$

Thus we deduce that

$$\left(\left(I + \frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)V^m, V^m\right) \ge \left(1 - \left(\frac{c\,\Delta t}{\Delta x}\right)^2\right)\|V^m\|^2. \tag{120}$$

We shall therefore suppose that the following condition holds, referred to as a Courant–Friedrichs–Lewy (or CFL) condition: there exists a positive constant c_0 such that

$$\frac{c\,\Delta t}{\Delta x} \le c_0 < 1. \tag{121}$$

Assuming that (121) holds, we then have from (120) that

$$\left(\left(I + \frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)\frac{U^{m+1}-U^m}{\Delta t}, \frac{U^{m+1}-U^m}{\Delta t}\right) \ge (1-c_0^2)\left\|\frac{U^{m+1}-U^m}{\Delta t}\right\|^2. \tag{122}$$

We shall therefore proceed by assuming that (121) holds, and define the nonnegative expression

$$\mathcal{N}^2(U^m) := \left(\left(I + \frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)\frac{U^{m+1}-U^m}{\Delta t}, \frac{U^{m+1}-U^m}{\Delta t}\right) + c^2\left\|D_x^-\frac{U^{m+1}-U^m}{2}\right]\right\|^2.$$

With this notation (119) becomes

$$\mathcal{N}^2(U^m) = \mathcal{N}^2(U^{m-1}) + (f(\cdot,t_m), U^{m+1}-U^{m-1}). \tag{123}$$

In the special case when f is identically zero (123) guarantees stability of the explicit scheme under the CFL condition (121); indeed, (123) implies that

$$\mathcal{N}^2(U^m) = \mathcal{N}^2(U^0), \qquad \text{for all } m = 1,\dots,M-1.$$

One can check that the mapping $U \mapsto \max_{m\in\{0,\dots,M-1\}}\left[\mathcal{N}^2(U^m)\right]^{1/2}$ is a norm on the linear space of all mesh functions U defined on the space-time mesh $\{(x_j,t_m) : j = 0,1,\dots,J,\ \ m = 0,1,\dots,M\}$ such that $U_0^m = U_J^m = 0$ for all $m = 0,1,\dots,M$. Thus we have show that, provided the CFL condition (121) holds and f is identically zero, the explicit scheme (115) is (conditionally) stable in this norm.

We now return to the general case when f is not identically zero, under the assumption (121). Our starting point is the equality (123) and we focus our attention on the second term on its right-hand side. For $m \in \{1,\dots,M\}$, let Z^m be the solution of the problem

$$\begin{aligned}\left(I + \frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)Z_j^m &= f(x_j,t_m), \qquad j = 1,\dots,J-1. \\ Z_0^m = Z_J^m &= 0.\end{aligned} \tag{124}$$

To show the existence of a unique solution Z^m to this problem we note that (124) is in fact a system of $J-1$ linear algebraic equations for the $J-1$ unknowns $Z_1^m, \ldots, Z_{J-1}^m$. Therefore (124) will possess a unique solution if, and only if, the corresponding homogeneous problem (i.e. the problem with $f(x_j, t_m)$ replaced by 0 for all $j \in \{1, \ldots, J-1\}$) has $Z_j^m = 0$, $j = 0, \ldots, J$, as its unique solution. Clearly, the homogeneous counterpart of (124) does indeed have Z_j^m, $j = 0, \ldots, J$, as a solution. The fact that this is the *unique solution* to the homogeneous counterpart of (124) follows by noting that, thanks to the inequality (120) with $V^m = Z^m$ and the assumed CFL condition (121), $\|Z^m\|^2 = 0$. Therefore $Z_j^m = 0$ for all $j = 0, \ldots, J$.

Having shown the existence of unique solution to (124), it makes sense to write

$$Z^m = \left(I + \frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)^{-1} f(\cdot, t_m). \tag{125}$$

With this, we return to the second term on the right-hand side of (123) and decompose it as follows:

$$\begin{aligned}(f(\cdot,t^m), U^{m+1} - U^{m-1}) &= (f(\cdot,t^m), U^{m+1} - U^m) + (f(\cdot,t^m), U^m - U^{m-1}) \\ &= \left(\left(I + \frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right) Z^m, U^{m+1} - U^m\right) + \left(\left(I + \frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right) Z^m, U^m - U^{m-1}\right).\end{aligned}$$

We would now like to transfer, in each of the two inner products appearing in the last line, the finite difference operator featuring there from the first entry of the inner product to the second entry in the inner product. To this end, note that for any two mesh functions V, W, defined on the mesh $\{x_j : j = 0, \ldots, J\}$ and such that $V_0 = V_J = 0$ and $W_0 = W_J = 0$, one has, by summation by parts,

$$(D_x^+ D_x^- V, W) = (V, D_x^+ D_x^- W). \tag{126}$$

As $Z_0^m = Z_J^m = 0$, $U_0^{m+1} - U_0^m = U_J^{m+1} - U_J^m = 0$, and $U_0^m - U_0^{m-1} = U_J^m - U_J^{m-1} = 0$, it follows that

$$\begin{aligned}(f(\cdot,t^m), U^{m+1} - U^{m-1}) &= (f(\cdot,t^m), U^{m+1} - U^m) + (f(\cdot,t^m), U^m - U^{m-1}) \\ &= \left(Z^m, \left(I + \frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)\left(U^{m+1} - U^m\right)\right) + \left(Z^m, \left(I + \frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)\left(U^m - U^{m-1}\right)\right) \\ &= \left(\left(I + \frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)\left(U^{m+1} - U^m\right), Z^m\right) + \left(\left(I + \frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)\left(U^m - U^{m-1}\right), Z^m\right).\end{aligned}$$

To simplify our notation, for mesh functions V, W defined on the mesh $\{x_j : j = 0, \ldots, J\}$ and such that $V_0 = V_J = 0$ and $W_0 = W_J = 0$, we shall write

$$[V, W] := \left(\left(I + \frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)V, W\right).$$

We leave it as an exercise to the reader to verify that $[\cdot,\cdot]$ is an inner product: note, to this end, that $[\cdot,\cdot]$ is linear in both of its entries; thanks to (126), $[V, W] = [W, V]$; and, by virtue of (120) and (121), if $[V, V] = 0$ then $V = 0$.

Let $|[\cdot]|$ denote the norm induced by this inner product, i.e., let $|[V]| := [V, V]^{\frac{1}{2}}$. One then has the following Cauchy–Schwarz inequality:

$$[V, W] \leq |[V]|\,|[W]|.$$

With these preparations in place, we are now ready to continue. In terms of this newly introduced notation we therefore have

$$\begin{aligned}(f(\cdot,t^m), U^{m+1} - U^{m-1}) &= (f(\cdot,t^m), U^{m+1} - U^m) + (f(\cdot,t^m), U^m - U^{m-1}) \\ &= [U^{m+1} - U^m, Z^m] + [U^m - U^{m-1}, Z^m] \\ &\leq |[U^{m+1} - U^m]|\,|[Z^m]| + |[U^m - U^{m-1}]|\,|[Z^m]|.\end{aligned}$$

We substitute this into the right-hand side of (123) and, after dividing and multiplying by Δt, we find that

$$\mathcal{N}^2(U^m) \leq \mathcal{N}^2(U^{m-1}) + \Delta t \left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right| |[Z^m]| + \Delta t \left|\left[\frac{U^m-U^{m-1}}{\Delta t}\right]\right| |[Z^m]|. \tag{127}$$

By recalling the definition of $\mathcal{N}^2(U^m)$ and the definition of the norm $|[\cdot]|$ the last inequality can be rewritten as follows

$$\begin{aligned}\left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right|^2 + c^2 \left\| D_x^- \frac{U^{m+1}-U^m}{2}\right]\right|^2 \leq \left|\left[\frac{U^m-U^{m-1}}{\Delta t}\right]\right|^2 + c^2 \left\| D_x^- \frac{U^m-U^{m-1}}{2}\right]\right|^2 \\ + \Delta t \left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right| |[Z^m]| + \Delta t \left|\left[\frac{U^m-U^{m-1}}{\Delta t}\right]\right| |[Z^m]|.\end{aligned} \tag{128}$$

Next we shall make use of the elementary inequality

$$\alpha\beta \leq \alpha^2 + \frac{1}{4}\beta^2, \qquad \text{for } \alpha, \beta \in \mathbb{R}.$$

in the last two terms on the right-hand side of (128):

$$\begin{aligned}\Delta t \left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right| |[Z^m]| &= \sqrt{\frac{\Delta t}{T}} \left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right| \sqrt{\Delta t\, T}\, |[Z^m]| \\ &\leq \frac{\Delta t}{T} \left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right|^2 + \frac{\Delta t\, T}{4} |[Z^m]|^2;\end{aligned}$$

analogously,

$$\Delta t \left|\left[\frac{U^m-U^{m-1}}{\Delta t}\right]\right| |[Z^m]| \leq \frac{\Delta t}{T} \left|\left[\frac{U^m-U^{m-1}}{\Delta t}\right]\right|^2 + \frac{\Delta t\, T}{4} |[Z^m]|^2.$$

We then substitute these inequalities into the right-hand side of (128) and, after a rearrangement of terms and by noting that $1 - \frac{\Delta t}{T} \leq 1 \leq 1 + \frac{\Delta t}{T}$, we arrive at the following inequality:

$$\begin{aligned}&\left(1 - \frac{\Delta t}{T}\right)\left(\left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right|^2 + c^2 \left\| D_x^- \frac{U^{m+1}-U^m}{2}\right]\right|^2\right) \\ &\quad \leq \left(1 + \frac{\Delta t}{T}\right)\left(\left|\left[\frac{U^m-U^{m-1}}{\Delta t}\right]\right|^2 + c^2 \left\| D_x^- \frac{U^m-U^{m-1}}{2}\right]\right|^2\right) + \frac{\Delta t\, T}{2} |[Z^m]|^2.\end{aligned} \tag{129}$$

By recalling the definition of $\mathcal{N}^2(U^m)$ we can rewrite (129) in the following compact form:

$$\mathcal{N}^2(U^m) \leq \frac{T+\Delta t}{T-\Delta t}\mathcal{N}^2(U^{m-1}) + \frac{T^2}{2(T-\Delta t)}\,\Delta t\, |[Z^m]|^2, \quad m = 1, \ldots, M-1.$$

As, by assumption, $M \geq 2$, it follows that $\Delta t := T/M \leq T/2$, whereby $T - \Delta t \geq T/2$; using this the second term on the right-hand side of the last inequality can be simplified, resulting in

$$\mathcal{N}^2(U^m) \leq \frac{T+\Delta t}{T-\Delta t}\mathcal{N}^2(U^{m-1}) + T\,\Delta t\, |[Z^m]|^2, \quad m = 1, \ldots, M-1.$$

To proceed, we shall appeal to Lemma 15, with

$$a_m = \mathcal{N}^2(U^m), \quad b_m = T\,\Delta t\, |[Z^m]|^2, \quad \alpha = \frac{T+\Delta t}{T-\Delta t}$$

to deduce that

$$\mathcal{N}^2(U^m) \leq \left(\frac{T+\Delta t}{T-\Delta t}\right)^m \mathcal{N}^2(U^0) + T\,\Delta t \sum_{k=1}^{m} \left(\frac{T+\Delta t}{T-\Delta t}\right)^{m-k} |[Z^k]|^2, \quad m = 1, \ldots, M-1.$$

Note that

$$\left(\frac{T+\Delta t}{T-\Delta t}\right)^m \leq \left(\frac{T+\Delta t}{T-\Delta t}\right)^M = \left(\frac{1+\frac{\Delta t}{T}}{1-\frac{\Delta t}{T}}\right)^{\frac{T}{\Delta t}}$$

for all $m \in \{0, 1, \ldots, M\}$, with $\Delta t \leq \frac{T}{2}$, and one has

$$\frac{1+x}{1-x} \leq 1 + 4x \quad \forall\, x \in \left[0, \frac{1}{2}\right]$$

and

$$(1+4x)^{\frac{1}{x}} \leq \mathrm{e}^4 \quad \forall\, x \in \left(0, \frac{1}{2}\right],$$

where the second inequality follows by noting that $\log(1+x) \leq x$ for all $x \geq 0$. Hence,

$$\mathcal{N}^2(U^m) \leq \mathrm{e}^4\, \mathcal{N}^2(U^0) + \mathrm{e}^4\, T \sum_{k=1}^{m} \Delta t\, |[Z^k]|^2 \quad \text{for } m = 1, \ldots, M-1.$$

Finally, by recalling the (125), we deduce the following stability result for the explicit finite difference scheme under consideration.

Theorem 17 *Suppose that the CFL condition* (121) *is satisfied. Then, the explicit finite difference approximation* (115) *of the initial-boundary-value problem* (102)*, on a finite difference mesh of spacing* $\Delta x = (b-a)/J$ *with* $J \geq 2$ *in the* x*-direction and* $\Delta t = T/M$ *with* $M \geq 2$ *in the* t*-direction, is (conditionally) stable in the sense that*

$$\mathcal{N}^2(U^m) \leq \mathrm{e}^4\, \mathcal{N}^2(U^0) + \mathrm{e}^4\, T \sum_{k=1}^{m} \Delta t \left\| \left[\left(I + \frac{1}{4} c^2 \Delta t^2 D_x^+ D_x^- \right)^{-1} f(\cdot, t_k) \right] \right\|^2, \quad \textit{for } m = 1, \ldots, M-1.$$

Consistency of the explicit scheme. We define the consistency error of the explicit scheme by

$$T_j^m := \frac{u_j^{m+1} - 2u_j^m + u_j^{m-1}}{\Delta t^2} - c^2 \frac{u_{j+1}^m - 2u_j^m + u_{j-1}^m}{\Delta x^2} - f(x_j, t_m) \quad \text{for} \quad \begin{cases} m = 1, \ldots, M-1, \\ j = 1, \ldots, J-1, \end{cases}$$

and

$$T_j^0 := \frac{u_j^1 - u_j^0}{\Delta t} - u_1(x_j), \quad \text{for } j = 1, \ldots, J-1,$$

where $u_j^m := u(x_j, t_m)$, $j = 0, \ldots, J$, $m = 0, \ldots, M$. Hence, similarly as in the case of the implicit scheme,

$$T_j^m = \left(\frac{u_j^{m+1} - 2u_j^m + u_j^{m-1}}{\Delta t^2} - \frac{\partial^2 u}{\partial t^2}(x_j, t_m) \right) - c^2 \left(\frac{u_{j+1}^m - 2u_j^m + u_{j-1}^m}{\Delta x^2} - \frac{\partial^2 u}{\partial x^2}(x_j, t_m) \right)$$

for $m = 1, \ldots, M-1$ and $j = 1, \ldots, J-1$, and

$$T_j^0 = \frac{u_j^1 - u_j^0}{\Delta t} - \frac{\partial u}{\partial t}(x_j, 0), \quad \text{for } j = 1, \ldots, J-1.$$

By performing Taylor series expansions with respect to t about the point mesh point (x_j, t_m) and then with respect to x about the same mesh point we deduce that

$$T_j^m = \frac{1}{12}\Delta t^2 \frac{\partial^4 u}{\partial t^4}(x_j, \tau_m) + \frac{1}{12}c^2\Delta x^2 \frac{\partial^4 u}{\partial t^4}(\xi_j, t_m), \tag{130}$$

where $\tau_m \in [t_{m-1}, t_{m+1}]$ and $\xi_j \in [x_{j-1}, x_{j+1}]$, provided that the fourth partial derivative of u with respect to t is a continuous function on $[a,b]\times[0,T]$ and the fourth partial derivative of u with respect to x is a continuous function on $[a,b]\times[0,T]$. Also,

$$T_j^0 = \frac{1}{2}\Delta t \frac{\partial u}{\partial t}(x_j, \tau_0),$$

where $\tau_0 \in [0, \Delta t]$, provided that the second partial derivative of u with respect to t is a continuous function on $[a,b]\times[0,T]$. Hence,

$$|T_j^m| \le \frac{1}{12}c^2\Delta x^2\, M_{4x} + \frac{1}{12}\Delta t^2\, M_{4t}, \qquad \begin{cases} j = 1, \dots, J-1, \\ m = 1, \dots, M-1, \end{cases}$$

where

$$M_{4x} := \max_{(x,t)\in[a,b]\times[0,T]} \left|\frac{\partial^4 u}{\partial x^4}(x,t)\right| \quad \text{and} \quad M_{4t} := \max_{(x,t)\in[a,b]\times[0,T]} \left|\frac{\partial^4 u}{\partial t^4}(x,t)\right|,$$

and

$$|T_j^0| \le \frac{1}{2}\Delta t\, M_{2t},$$

where

$$M_{2t} := \max_{(x,t)\in[a,b]\times[0,T]} \left|\frac{\partial^2 u}{\partial t^2}(x,t)\right|.$$

If the more accurate second initial condition (116) is used instead of $(115)_3$, then

$$T_j^0 := \frac{u_j^1 - u_j^0}{\Delta t} - u_1(x_j) - \frac{1}{2}\Delta t \left(c^2 D_x^+ D_x^- u_1(x_j) + f(x_j, 0)\right).$$

In this case, again by Taylor series expansion,

$$|T_j^0| \le \frac{1}{6}\Delta t^2 M_{3t} + \frac{1}{24}c^2\Delta t\, \Delta x^2 M_{4x}.$$

Convergence of the explicit scheme. The *global error* of the finite difference scheme (115) is defined by

$$e_j^m := u(x_j, t_m) - U_j^m, \qquad \begin{cases} j = 0, \dots, J, \\ m = 1, \dots, M-1. \end{cases}$$

Thus, thanks to the definition of the consistency error, we have that

$$\frac{e_j^{m+1} - 2e_j^m + e_j^{m-1}}{\Delta t^2} - c^2 \frac{e_{j+1}^{m+1} - 2e_j^{m+1} + 2e_{j-1}^{m+1}}{\Delta x^2} = T_j^m, \qquad \begin{cases} j = 1, \dots, J-1, \\ m = 1, \dots, M-1, \end{cases}$$

and

$$e_j^1 = e_j^0 + \Delta t\, T_j^0, \qquad j = 1, \dots, J-1.$$

Furthermore, $e_j^0 = 0$ for $j = 0, 1, \dots, J$, and $e_0^m = e_J^m = 0$ for $m = 1, \dots, M$. Hence, the global error e satisfies an identical finite difference scheme as U, but with $f(x_j, t_m)$ replaced by T_j^m and $u_1(x_j)$ replaced

by T_j^0. It therefore follows from Theorem 17 with U^m replaced by e^m, U^0 replaced by e^0 and $f(x_j,t_k)$ replaced by T_j^k for $j=1,\ldots,J-1$ and $k=1,\ldots,M-1$, that

$$\mathcal{N}^2(e^m) \leq \mathrm{e}^4\,\mathcal{N}^2(e^0) + \mathrm{e}^4\,T^2 \max_{1\leq k\leq m}\left|\left[\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)^{-1} T^k\right]\right|^2, \quad \text{for } m=1,\ldots,M-1. \tag{131}$$

It remains to bound the two terms on the right-hand side of this inequality. We note that

$$\begin{aligned}\mathcal{N}^2(e^0) &:= \left(\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)\frac{e^1-e^0}{\Delta t}, \frac{e^1-e^0}{\Delta t}\right) + c^2\left\|D_x^-\frac{e^1-e^0}{2}\right\|^2 \\ &= \left(\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)T^0, T^0\right) + c^2\left\|D_x^- T^0\right\|^2.\end{aligned}$$

By expanding the first term on the right-hand side and then performing summation by parts in the middle term among the three resulting terms, we have

$$\begin{aligned}\left(\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)T^0, T^0\right) + c^2\left\|D_x^- T^0\right\|^2 &= \|T^0\|^2 + \frac{1}{4}c^2\Delta t^2 (D_x^+ D_x^- T^0, T^0) + c^2\|D_x^- T^0]|^2 \\ &= \|T^0\|^2 - \frac{1}{4}c^2\Delta t^2\|D_x^- T^0]|^2 + c^2\|D_x^- T^0]|^2 \\ &\leq \|T^0\|^2 + c^2\|D_x^- T^0]|^2.\end{aligned}$$

As, by Taylor series expansion with a remainder term,

$$T_j^0 = \frac{u_j^1-u_j^0}{\Delta t} - \frac{\partial u}{\partial t}(x_j,0) = \frac{1}{2}\Delta t\frac{\partial^2 u}{\partial t^2}(x_j,\tau),$$

where $\tau\in[0,\Delta t]$, it follows that

$$|T_j^0| \leq \frac{1}{2}\Delta t\, M_{2t},$$

with

$$M_{2t} := \max_{(x,t)\in[a,b]\times[0,T]}\left|\frac{\partial^2 u}{\partial t^2}(x,t)\right|.$$

Hence,

$$\|T^0\|^2 = \Delta t\sum_{j=1}^{J-1}|T_j^0|^2 \leq \frac{1}{4}\Delta t^2 M_{2t}^2.$$

To bound $\|D_x^- T^0]|^2$, we note that, thanks to a Taylor series expansion with an integral remainder term, we have that

$$T_j^0 = \frac{1}{\Delta t}\int_0^{\Delta t}(\Delta t-\tau)\frac{\partial^2 u}{\partial t^2}(x_j,\tau)\,\mathrm{d}\tau.$$

Hence,

$$D_x^- T_j^0 = \frac{1}{\Delta t\,\Delta x}\int_0^{\Delta t}\int_{x_{j-1}}^{x_j}(\Delta t-\tau)\frac{\partial^3 u}{\partial x\,\partial t^2}(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau,$$

whereby

$$|D_x^- T_j^0| \leq \frac{1}{\Delta t\,\Delta x}\,\frac{\Delta t^2}{2}\,\Delta x\, M_{1x2t} = \frac{1}{2}\,\Delta t\, M_{1x2t},$$

with

$$M_{1x2t} = \max_{(x,t)\in[a,b]\times[0,T]}\left|\frac{\partial^3 u}{\partial x\,\partial t^2}(x,t)\right|.$$

Thus we deduce that

$$\|D_x^- T^0]|^2 \leq \frac{1}{4}\,\Delta t^2 M_{1x2t}^2.$$

By collecting the above bounds, we deduce that

$$\begin{aligned}\left(\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)T^0, T^0\right) + c^2\|D_x^- T^0]|^2 &= \|T_0\|^2 + \frac{1}{4}c^2\Delta t^2 (D_x^+ D_x^- T^0, T^0) + c^2\|D_x^- T^0]|^2 \\ &\leq \frac{1}{4}\Delta t^2\, M_{2t}^2 + +c^2\left(\frac{1}{4}\Delta t^2 M_{1x2t}^2\right) \\ &= \mathcal{O}(\Delta t^2).\end{aligned}$$

Thus we have shown that

$$\mathcal{N}^2(e^0) = \mathcal{O}(\Delta t^2).$$

Having bounded the first term on the right-hand side of the inequality (131), we proceed to bound the second term on the right-hand side of (131):

$$\mathrm{e}^4\, T^2 \max_{1\leq k\leq m}\left|\left[\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)^{-1} T^k\right]\right|^2.$$

Letting

$$V^k := \left(I+\frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)^{-1} T^k,$$

it follows that

$$\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)V_j^k = T_j^k \qquad \text{for } j = 1,\ldots,J-1,$$

and $V_0^k = V_J^k = 0$. Taking the $(\cdot,\cdot)$ inner product of both sides of the inequality, using the Cauchy–Schwarz inequality on the right-hand side and the inequality (120) in conjunction with the CFL condition (121), we deduce that

$$(1-c_0^2)\|V^k\|^2 \leq \left(\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)V^k, V^k\right) = |[V^k]|^2 = (T^k, V^k) \leq \|T^k\|\,\|V^k\|,$$

and therefore

$$\|V^k\| \leq (1-c_0^2)^{-1}\|T^k\|.$$

This implies that

$$|[V^k]|^2 \leq (1-c_0^2)^{-1}\|T^k\|^2.$$

Hence, thanks to the definition of V^k above and (130), we have that

$$\begin{aligned}\mathrm{e}^4\, T^2 \max_{1\leq k\leq m}\left|\left[\left(I+\frac{1}{4}c^2\Delta t^2 D_x^+ D_x^-\right)^{-1} T^k\right]\right|^2 &= \mathrm{e}^4\, T^2\, |[V^k]|^2 \leq \mathrm{e}^4\, T^2\,(1-c_0^2)^{-1} \max_{1\leq k\leq m}\|T^k\|^2 \\ &= \mathcal{O}((\Delta x^2+\Delta t^2)^2).\end{aligned}$$

Thus we have also bounded the second term on the right-hand side of the inequality (131); consequently,

$$\mathcal{N}^2(e^m) = \mathcal{O}(\Delta t^2) + \mathcal{O}((\Delta x^2+\Delta t^2)^2).$$

It is worth emphasizing here that the first term on the right-hand side comes from the approximation of the second initial condition, stated in $(115)_3$. If instead of $(115)_3$ one uses the more accurate initial condition (116), then

$$\mathcal{N}^2(e^0) = \mathcal{O}((\Delta t^2 + \Delta t\,\Delta x^2)^2),$$

and therefore in that case

$$\mathcal{N}^2(e^m) = \mathcal{O}((\Delta t^2 + \Delta t\,\Delta x^2)^2) + \mathcal{O}((\Delta x^2 + \Delta t^2)^2) = \mathcal{O}((\Delta x^2 + \Delta t^2)^2).$$

In the first case,

$$\max_{1\le m\le M-1}[\mathcal{N}^2(u^m - U^m)]^{1/2} = \mathcal{O}(\Delta x^2 + \Delta t),$$

while in the second case, when the more accurate approximation (116) of the second initial condition $(102)_3$ is used, then

$$\max_{1\le m\le M-1}[\mathcal{N}^2(u^m - U^m)]^{1/2} = \mathcal{O}(\Delta x^2 + \Delta t^2).$$

This completes the convergence analysis of the explicit scheme (115). We have thus shown that the explicit scheme exhibits second order convergence with respect to the spatial discretization step Δx and first-order convergence with respect to the temporal discretization step Δt in the norm $\max_{m\in\{1,\ldots,M-1\}}[\mathcal{N}^2(\cdot)]^{\frac{1}{2}}$ if the second initial condition $(102)_3$ is approximated by $(115)_3$, but if one uses the more accurate approximation (116) of the second initial condition, then the explicit scheme exhibits second-order convergence with respect to both Δx and Δt in the norm $\max_{m\in\{1,\ldots,M-1\}}[\mathcal{N}^2(\cdot)]^{\frac{1}{2}}$.

6.4 First-order hyperbolic equations: initial-boundary-value problem and energy estimate

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 **Lecture 15**

$$\frac{\partial u}{\partial t} + \sum_{i=1}^{n} b_i(x) \frac{\partial u}{\partial x_i} + c(x,t)u = f(x,t), \quad x \in \Omega, \;\; t \in (0,T], \tag{132}$$

$$u(x,t) = 0, \quad x \in \Gamma_-, \;\; t \in [0,T], \tag{133}$$

$$u(x,0) = u_0(x) \quad x \in \bar{\Omega}, \tag{134}$$

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 the part of the boundary, namely on $\Gamma_- \times [0,T]$, or else the problem may have no solution.

We shall assume that

$$b_i \in C^1(\bar{\Omega}), \;\; i = 1, \ldots, n, \tag{135}$$

$$c \in C(\bar{Q}), \;\; f \in L_2(Q), \tag{136}$$

$$u_0 \in L^2(\Omega). \tag{137}$$

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 (see A1 Differential Equations). More generally, for b_i, c, f, u_0, obeying the smoothness requirements of (135), 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 (132)–(134) in time.

We make the additional hypothesis:

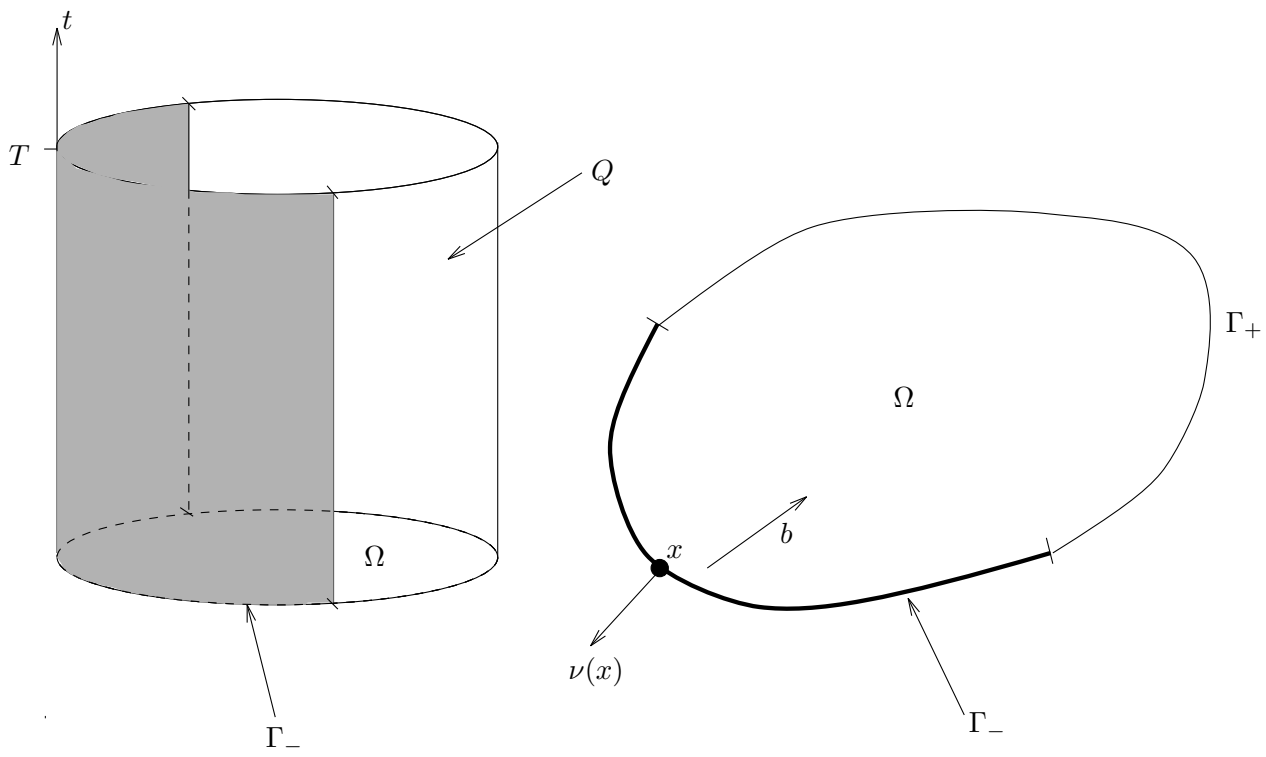
$$c(x,t) - \frac{1}{2} \sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i}(x) \geq 0, \quad x \in \bar{\Omega}, \;\; t \in [0,T]. \tag{138}$$

Taking the inner product of (132) with u in $L^2(\Omega)$, we obtain:

$$\left(\frac{\partial u}{\partial t}, u\right) + \left(c(\cdot,t) - \frac{1}{2} \sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i}(\cdot), u^2\right) + \frac{1}{2} \int_{\Gamma_+} \left[\sum_{i=1}^{n} b_i(x) \nu_i(x)\right] u^2(x,t) \, \mathrm{d}s(x) = (f,u), \tag{139}$$

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 (138) and noting that

$$\begin{aligned}\left(\frac{\partial u}{\partial t}, u\right) &= \int_\Omega \frac{\partial u}{\partial t}(x,t) \, u(x,t) \, \mathrm{d}x \\ &= \int_\Omega \frac{1}{2} \frac{\partial}{\partial t} u^2(x,t) \, \mathrm{d}x = \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega u^2(x,t) \, \mathrm{d}x \\ &= \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|u(\cdot,t)\|^2,\end{aligned}$$



it follows from (139) that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2 + \frac{1}{2}\int_{\Gamma_+}\left[\sum_{i=1}^{n} b_i(x)\nu_i(x)\right] u^2(x,t)\,\mathrm{d}s(x) \le (f,u).$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned}(f,u) &\le \|f(\cdot,t)\|\,\|u(\cdot,t)\| \\ &\le \frac{1}{2}\|f(\cdot,t)\|^2 + \frac{1}{2}\|u(\cdot,t)\|^2,\end{aligned}$$

and therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2 + \int_{\Gamma_+}\left[\sum_{i=1}^{n} b_i(x)\nu_i(x)\right] u^2(x,t)\,\mathrm{d}s(x) - \|u(\cdot,t)\|^2 \le \|f(\cdot,t)\|^2, \quad t\in[0,T].$$

Multiplying both sides by e^{-t}, this can be rewritten as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{-t}\|u(\cdot,t)\|^2 + \mathrm{e}^{-t}\int_{\Gamma_+}\left[\sum_{i=1}^{n} b_i(x)\nu_i(x)\right] u^2(x,t)\,\mathrm{d}s \le \mathrm{e}^{-t}\|f(\cdot,t)\|^2, \quad t\in[0,T].$$

Integrating this inequality with respect to t yields

$$\begin{aligned}&\mathrm{e}^{-t}\|u(\cdot,t)\|^2 + \int_0^t \mathrm{e}^{-\tau}\int_{\Gamma_+}\left[\sum_{i=1}^{n} b_i(x)\nu_i(x)\right] u^2(x,\tau)\,\mathrm{d}s(x)\,\mathrm{d}\tau \\ &\quad\le \|u_0\|^2 + \int_0^t \mathrm{e}^{-\tau}\|f(\cdot,\tau)\|^2\,\mathrm{d}\tau, \quad t\in[0,T].\end{aligned}$$

Hence

$$\|u(\cdot,t)\|^2 + \int_0^t \mathrm{e}^{t-\tau} \int_{\Gamma_+} \left[\sum_{i=1}^n b_i(x)\nu_i(x)\right] u^2(x,\tau)\,\mathrm{d}s(x)\,\mathrm{d}\tau$$
$$\leq \mathrm{e}^t\|u_0\|^2 + \int_0^t \mathrm{e}^{t-\tau}\|f(\cdot,\tau)\|^2\,\mathrm{d}\tau, \quad t\in[0,T]. \tag{140}$$

This, so called, energy inequality expresses the continuous dependence of the solution to (132)–(134) 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 (132)–(134), then $u := u_1 - u_2$ also solves (132)–(134), with $f \equiv 0$ and $u_0 \equiv 0$. Thus, by (140), $\|u(\cdot,t)\| = 0$, $t\in[0,T]$ and therefore $u \equiv 0$, i.e. $u_1 \equiv u_2$.

Let us consider a particularly important case when

$$c \equiv 0, \quad f \equiv 0, \quad \text{and} \quad \mathrm{div}\; b = \sum_{i=1}^n \frac{\partial b_i}{\partial x_i} \equiv 0,$$

where $b(x) = (b_1(x),\ldots,b_n(x))$. Then, by virtue of (139),

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2 + \frac{1}{2}\int_{\Gamma_+} [b(x)\cdot\nu(x)]\,u^2(x,t)\,\mathrm{d}s(x) = 0,$$

and therefore,

$$\|u(\cdot,t)\|^2 + \int_0^t \int_{\Gamma_+} [b(x)\cdot\nu(x)]\,u^2(x,\tau)\,\mathrm{d}s(x)\,\mathrm{d}\tau = \|u_0\|^2,$$

which expresses the conservation of energy in the physical system modelled by (132)–(134).

6.5 Explicit finite difference approximation

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:

$$\frac{\partial u}{\partial t} + b\frac{\partial u}{\partial x} = f(x,t), \quad x\in(0,1), \quad t\in(0,T], \tag{141}$$

subject to the boundary and initial conditions

$$u(x,t) = 0, \qquad x\in\Gamma_-, \quad t\in[0,T], \tag{142}$$
$$u(x,0) = u_0(x), \quad x\in[0,1]. \tag{143}$$

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

$$u(0,t) = 0, \quad t\in[0,T]. \tag{144}$$

To construct a finite difference approximation of (141)–(144) let $\Delta x = 1/J$ 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

$$x_j = j\Delta x, \quad j=0,\ldots,J, \quad t_m = m\,\Delta t, \quad m=0,\ldots,M.$$

At the mesh-point (x_j,t_m), (141) is approximated by the explicit finite difference scheme

$$\frac{U_j^{m+1}-U_j^m}{\Delta t} + b\,D_x^- U_j^m = f(x_j,t_m), \qquad j=1,\ldots,J, \tag{145}$$
$$m=0,\ldots,M-1,$$

subject to the boundary and initial condition, respectively:

$$U_0^m = 0, \qquad m = 0, \ldots, M, \tag{146}$$

$$U_j^0 = u_0(x_j), \qquad j = 0, \ldots, J. \tag{147}$$

Equivalently,

$$U_j^{m+1} = (1-\mu)U_j^m + \mu U_{j-1}^m + \Delta t f(x_j, t_m), \qquad \begin{array}{r} j = 1, \ldots, J, \\ m = 0, \ldots, M-1, \end{array}$$

in conjunction with

$$U_0^m = 0, \qquad m = 0, \ldots, M,$$

$$U_j^0 = u_0(x_j), \qquad j = 0, \ldots, J,$$

where

$$\mu = \frac{b\Delta t}{\Delta x};$$

μ is called the CFL (or Courant–Friedrichs–Lewy) number. The explicit finite difference scheme (145) is frequently called the *first-order upwind scheme.*

Suppose that $0 \le \mu \le 1$; then

$$\begin{aligned} \left|U_j^{m+1}\right| &\le (1-\mu)\left|U_j^m\right| + \mu\left|U_{j-1}^m\right| + \Delta t\left|f(x_j, t_m)\right| \\ &\le (1-\mu)\max_{0\le j\le J}\left|U_j^m\right| + \mu\max_{1\le j\le J+1}\left|U_{j-1}^m\right| + \Delta t\max_{0\le j\le J}\left|f(x_j, t_m)\right| \\ &= \max_{0\le j\le J}\left|U_j^m\right| + \Delta t\max_{0\le j\le J}\left|f(x_j, t_m)\right|. \end{aligned}$$

Hence

$$\max_{0\le j\le J}\left|U_j^{m+1}\right| \le \max_{0\le j\le J}\left|U_j^m\right| + \Delta t\max_{0\le j\le J}\left|f(x_j, t_m)\right|.$$

Let us define the mesh-dependent norm

$$\|U\|_\infty := \max_{0\le j\le J}\left|U_j\right|;$$

then

$$\|U^{m+1}\|_\infty \le \|U^m\|_\infty + \Delta t\|f(\cdot, t_m)\|_\infty, \quad m = 0, \ldots, M-1.$$

Summing through m, we get

$$\max_{1\le k\le M}\|U^k\|_\infty \le \|U^0\|_\infty + \sum_{m=0}^{M-1}\Delta t\|f(\cdot, t_m)\|_\infty, \tag{148}$$

which expresses the stability of the finite difference scheme (145)–(147) under the condition

$$0 \le \mu = \frac{b\Delta t}{\Delta x} \le 1. \tag{149}$$

Thus we have proved that (145)–(147) is conditionally stable, the condition being that the Courant number, μ, is in the interval $[0, 1]$.

It is possible to show that the scheme (145)–(147) is also stable in the mesh-dependent L_2-norm, $\|\cdot]|$ defined by

$$\|V]|^2 = \sum_{i=1}^{J} \Delta x\, V_i^2.$$

The associated inner product is

$$(V, W] := \sum_{i=1}^{J} \Delta x\, V_i W_i.$$

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

$$\begin{aligned}(U^m, D_x^- U^m] &= \sum_{j=1}^{J} \Delta x\, U_j^m \frac{U_j^m - U_{j-1}^m}{\Delta x} \\ &= \frac{1}{2}\sum_{j=1}^{J}\{(U_j^m)^2 - (U_{j-1}^m)^2\} + \frac{\Delta x}{2}\sum_{j=1}^{J}\Delta x \left(\frac{U_j^m - U_{j-1}^m}{\Delta x}\right)^2 \\ &= \frac{1}{2}(U_J^m)^2 + \frac{\Delta x}{2}\|D_x^- U^m]|^2.\end{aligned} \tag{150}$$

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

$$\left(\frac{U^{m+1} - U^m}{\Delta t}, U^m\right] = \frac{1}{2\Delta t}\left(\|U^{m+1}]|^2 - \|U^m]|^2\right) - \frac{\Delta t}{2}\left\|\frac{U^{m+1} - U^m}{\Delta t}\right]\right\|^2, \qquad m = 0, \ldots, M-1. \tag{151}$$

Thus, taking the $(\cdot, \cdot]$-inner product of (145) with U^m and using (150) and (151),

$$\begin{aligned}&\|U^{m+1}]|^2 + \Delta t\, b (U_J^m)^2 + b\,\Delta x\,\Delta t\,\|D_x^- U^m]|^2 - \|U^m]|^2 \\ &\qquad - \Delta t^2 \left\|\frac{U^{m+1} - U^m}{\Delta t}\right]\right\|^2 = 2\Delta t\,(f^m, U^m], \qquad m = 0, \ldots, M-1.\end{aligned} \tag{152}$$

First suppose that $f \equiv 0$ then

$$\frac{U^{m+1} - U^m}{\Delta t} = -b\, D_x^- U^m,$$

so that

$$\|U^{m+1}]|^2 + \Delta t\, b\, |U_J^m|^2 + b\,\Delta x\,\Delta t\,(1-\mu)\|D_x^- U^m]|^2 = \|U^m]|^2, \quad m = 0, \ldots, M-1.$$

Summing through m,

$$\|U^k]|^2 + \sum_{m=0}^{k-1} \Delta t\, b\, |U_J^m|^2 + b\,\Delta x\,(1-\mu)\sum_{m=0}^{k-1}\Delta t\,\|D_x^- U^m]|^2 = \|U^0]|^2, \quad k = 1, \ldots, M, \tag{153}$$

which proves the stability of the scheme in the case when $f \equiv 0$ under the assumption that

$$0 \le \mu = \frac{b\,\Delta t}{\Delta x} \le 1.$$

In particular, if $\mu = 1$, we have that

$$\|U^k]|^2 + \sum_{m=0}^{k-1} \Delta t\, b\, |U_J^m|^2 = \|U^0]|^2, \quad k = 1, \ldots, M,$$

which is the discrete version of the identity (140), and expresses conservation of energy in the discrete sense. More generally, for $0 \le \mu \le 1$, (153) implies

$$\|U^k]|^2 + \sum_{m=0}^{k-1} \Delta t\, b\, |U_J^m|^2 \le \|U^0]|^2, \quad k = 1, \ldots, M.$$

Now let us consider the question of stability in the $\|\cdot]|$-norm in the general case of $f \not\equiv 0$. Since

$$\begin{aligned}\left\| \frac{U^{m+1} - U^m}{\Delta t} \right]\Big|^2 = \|f^m - bD_x^- U^m]|^2 &\le \{\|f^m]| + b\|D_x^- U^m]|\}^2 \\ &\le \left(1 + \frac{1}{\epsilon'}\right) \|f^m]|^2 + (1+\epsilon')b^2 \|D_x^- U^m]|^2, \quad \epsilon' > 0,\end{aligned}$$

and

$$(f^m, U^m] \le \|f^m]|\, \|U^m]| \le \tfrac{1}{2}\|f^m]|^2 + \tfrac{1}{2}\|U^m]|^2,$$

it follows from (152) that

$$\begin{aligned}\|U^{m+1}]|^2 + \Delta t\, b\, |U_n^m|^2 + b\,\Delta x\,\Delta t \left[1 - (1+\epsilon')\frac{b\Delta t}{\Delta x}\right] \|D_x^- U^m]|^2 \\ \le \Delta t \left[\left(1 + \frac{1}{\epsilon'}\right)\Delta t + 1\right] \|f^m]|^2 + (1 + \Delta t)\|U^m]|^2.\end{aligned}$$

Letting $\epsilon = 1 - 1/(1+\epsilon') \in (0,1)$, and assuming

$$0 \le \mu = \frac{b\,\Delta t}{\Delta x} \le 1 - \epsilon,$$

we have, for $m = 0, \ldots, M-1$,

$$\|U^{m+1}]|^2 + \Delta t\, b\, |U_J^m|^2 \le \|U^m]|^2 + \Delta t \left(1 + \frac{\Delta t}{\epsilon}\right) \|f^m]|^2 + \Delta t \|U^m]|^2.$$

Upon summation,

$$\|U^k]|^2 + \left(\sum_{m=0}^{k-1} \Delta t\, b\, |U_J^m|^2\right) \le \|U^0]|^2 + \left(1 + \frac{\Delta t}{\epsilon}\right) \sum_{m=0}^{k-1} \Delta t\, \|f^m]|^2 + \sum_{m=0}^{k-1} \Delta t\, \|U^m]|^2. \tag{154}$$

for $k = 1, \ldots, M$. To complete the proof of stability of the finite difference scheme we require the next lemma, which is easily proved by induction.

Lemma 16 *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*

$$a_k + b_k \le c_k + \sum_{m=0}^{k-1} d_m a_m, \quad k \ge 1; \;\; a_0 + b_0 \le c_0.$$

Then

$$a_k + b_k \le c_k \exp\left(\sum_{m=0}^{k-1} d_m\right), \quad k \ge 1.$$

By applying this lemma to (154) with

$$\begin{aligned}
a_k &= \|U^k]|^2, \quad k \ge 0, \\
b_k &= \sum_{m=0}^{k-1} \Delta t\, b\, |U_J^m|^2, \quad k \ge 1; \;\; b_0 = 0, \\
c_k &= \|U^0]|^2 + \left(1 + \frac{\Delta t}{\epsilon}\right) \sum_{m=0}^{k-1} \Delta t\, \|f^m]|^2, \quad k \ge 1; \;\; c_0 = \|U^0]|^2, \\
d_k &= \Delta t, \quad k = 1, 2, \ldots, M,
\end{aligned}$$

we obtain,

$$\|U^k]|^2 + \sum_{m=0}^{k-1} \Delta t\, b\, |U_J^m|^2 \le \mathrm{e}^{t^k} \left(\|U^0]|^2 + \left(1 + \frac{\Delta t}{\epsilon}\right) \sum_{m=0}^{k-1} \Delta t \|f^m]|^2\right), \quad k = 1, \ldots, M,$$

where $t^k = k\Delta t$. Hence we deduce stability of the scheme, in the sense that

$$\max_{1 \le k \le M} \left(\|U^k]|^2 + \sum_{m=0}^{k-1} \Delta t\, b\, |U_J^m|^2\right) \le \mathrm{e}^T \left(\|U^0]|^2 + \left(1 + \frac{\Delta t}{\epsilon}\right) \sum_{m=0}^{M-1} \Delta t \|f^m]|^2\right). \tag{155}$$

An error bound for the difference scheme (145)–(147) is easily derived from its 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

$$\begin{aligned}
\frac{e_j^{m+1} - e_j^m}{\Delta t} + bD_x^- e_j^m = \varphi_j^m, \qquad & j = 1, \ldots, J, \quad m = 0, \ldots, M-1, \\
e_0^m = 0, \qquad & m = 0, \ldots, M, \\
e_j^0 = 0, \qquad & j = 0, \ldots, J.
\end{aligned}$$

By virtue of the stability inequality established in the first part of this section,

$$\max_{1 \le m \le M} \|e^m\|_\infty \le \sum_{k=0}^{M-1} \Delta t \|\varphi^m\|_\infty. \tag{156}$$

By Taylor series expansion of φ_j^m about the point (x_j, t_m),

$$\varphi_j^m = \frac{1}{2}\Delta t \frac{\partial^2 u}{\partial t^2}(x_j, \tau^m) + \frac{1}{2} b\,\Delta x\, \frac{\partial^2 u}{\partial x^2}(\xi_j, t_m), \quad \tau^m \in (t_m, t_{m+1}), \quad \xi_j \in (x_{j-1}, x_j),$$

so that

$$\left|\varphi_j^m\right| \le \frac{1}{2}(\Delta t M_{2t} + b\,\Delta x\, M_{2x}),$$

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

$$\left|\varphi_j^m\right| \le \frac{1}{2} M(\Delta t + b\,\Delta x) \quad (= \mathcal{O}(\Delta x + \Delta t)). \tag{157}$$

Thus, by (156),

$$\max_{1\le m\le M} \|u^m - U^m\|_\infty \le \frac{1}{2} T M (\Delta t + b\,\Delta x);$$

so the scheme (145)–(147) is first-order convergent.

Analogously, using the stability result (154) in the discrete L_2-norm $\|\cdot\|_h$, (157) implies that

$$\max_{1\le m\le M} \|u^m - U^m\|_\infty \le c_\epsilon^\star \cdot (\Delta t + b\,\Delta x),$$

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.

6.6 Finite difference approximation of scalar nonlinear hyperbolic conservation laws

Lecture 16

Nonlinear hyperbolic conservation laws and systems of nonlinear hyperbolic conservation laws arise in numerous areas of application, fluid dynamics being one such field. Here, we shall confine ourselves to the simplest possible case of an initial value problem for

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0 \quad \text{for } (x,t) \in \mathbb{R} \times (0,\infty), \tag{158}$$

subject to the initial condition $u(x,0) = u_0(x)$, where $u_0 \in C^1(\mathbb{R})$ and has compact support, i.e. u_0 is identically zero outside a bounded closed interval of $\mathbb{R}$. The real-valued function f will be assumed to be twice continuously differentiable on $\mathbb{R}$ and we shall suppose $f(0) = f'(0) = 0$, and $f''(s) \ge 0$ for all $s \in \mathbb{R}$. Under these hypotheses f' is a nondecreasing function, whereby $f'(s) \ge 0$ for all $s \ge 0$. We shall assume further that $|f'(s)| \le f'(|s|)$ for all $s \in \mathbb{R}$. For example $f(s) = \frac{1}{2}s^2$ and $f(s) = \frac{1}{4}s^4 + \frac{1}{2}s^2$ satisfy these hypotheses.

Assuming that there is a $T > 0$ such that a solution $u \in C^1(\mathbb{R} \times [0,T])$ exists, thanks to the chain rule the equation (158) can be rewritten as

$$\frac{\partial u}{\partial t} + f'(u)\frac{\partial u}{\partial x} = 0 \quad \text{for } (x,t) \in \mathbb{R} \times (0,T]. \tag{159}$$

Motivated by the construction of the first-order upwind scheme in the previous section, we decompose

$$f'(u) = [f'(u)]_+ + [f'(u)]_-,$$

where we have used the notation:

$$[x]_+ := \frac{1}{2}(x + |x|) \quad \text{and} \quad [x]_- := \frac{1}{2}(x - |x|).$$

Clearly,

$$x = [x]_+ + [x]_-, \quad |x| = [x]_+ - [x]_-, \quad [x]_+ \geq 0 \quad \text{and} \quad [x]_- \leq 0 \qquad \text{for all } x \in \mathbb{R}.$$

With this notation, we can rewrite (159) as follows:

$$\frac{\partial u}{\partial t} + [f'(u)]_+ \frac{\partial u}{\partial x} + [f'(u)]_- \frac{\partial u}{\partial x} = 0 \quad \text{for } (x,t) \in \mathbb{R} \times (0,T]. \tag{160}$$

We approximation (160) by the following finite difference scheme

$$\begin{aligned} \frac{U_j^{m+1} - U_j^m}{\Delta t} + [f'(U_j^m)]_+ \, D_x^- U_j^m + [f'(U_j^m)]_- \, D_x^+ U_j^m = 0, \qquad j \in \mathbb{Z}, \quad m = 0, \ldots, M-1, \\ U_j^0(x) = u_0(x_j), \quad j \in \mathbb{Z}, \end{aligned} \tag{161}$$

where $\Delta t = T/M$, $M \geq 1$.

We will show that, under a certain CFL condition which we shall state below, the sequence of finite difference approximations $\{U_j^m\}_{j \in \mathbb{Z},\, 0 \leq m \leq M}$ is bounded, similarly as in the case of (148) (but now in terms of the norm of the initial datum only, as there is no source term on the right-hand of the equation (159) under consideration), in the sense that

$$\max_{1 \leq k \leq M} \|U^k\|_\infty \leq \|U^0\|_\infty, \tag{162}$$

where now $\|V\|_\infty := \max_{j \in \mathbb{Z}} |V_j|$.

To this end, we rewrite $(161)_1$ as follows:

$$\begin{aligned} U_j^{m+1} &= U_j^m - \frac{[f'(U_j^m)]_+ \, \Delta t}{\Delta x}(U_j^m - U_{j-1}^m) - \frac{[f'(U_j^m)]_- \, \Delta t}{\Delta x}(U_{j+1}^m - U_j^m) \\ &= \left(1 - \frac{\Delta t}{\Delta x}\left([f'(U_j^m)]_+ - [f'(U_j^m)]_-\right)\right) U_j^m + \frac{[f'(U_j^m)]_+ \, \Delta t}{\Delta x} U_{j-1}^m + \frac{-[f'(U_j^m)]_- \, \Delta t}{\Delta x} U_{j+1}^m \\ &= \left(1 - \frac{|f'(U_j^m)| \, \Delta t}{\Delta x}\right) U_j^m + \frac{[f'(U_j^m)]_+ \, \Delta t}{\Delta x} U_{j-1}^m + \frac{-[f'(U_j^m)]_- \, \Delta t}{\Delta x} U_{j+1}^m \end{aligned} \tag{163}$$

for all $j \in \mathbb{Z}$ and all $m = 0, \ldots, M-1$. Suppose that the following CFL condition holds:

$$\frac{f'(\|U^0\|_\infty) \, \Delta t}{\Delta x} \leq 1. \tag{164}$$

Suppose further, as an inductive hypothesis, that, for some $m \geq 0$,

$$\frac{f'(\|U^k\|_\infty) \, \Delta t}{\Delta x} \leq 1 \qquad \text{for all } k = 0, \ldots, m. \tag{165}$$

Thanks to (164) this inductive hypothesis is satisfied for $m = 0$. Suppose, for the inductive step, that (165) has already been shown to hold for some $m \geq 0$. Because of the assumptions imposed on the

function f, we have that $|f'(U_j^m)| \le f'(|U_j^m|) \le f'(\|U^m\|_\infty)$ for all $j \in \mathbb{Z}$. It then follows from (165) with $k = m$ that

$$\frac{|f'(U_j^m)|\,\Delta t}{\Delta x} \le 1 \qquad \text{for all } j \in \mathbb{Z},$$

and then (163) implies that

$$\begin{aligned}
|U_j^{m+1}| &\le \left(1 - \frac{|f'(U_j^m)|\,\Delta t}{\Delta x}\right)|U_j^m| + \frac{[f'(U_j^m)]_+\,\Delta t}{\Delta x}|U_{j-1}^m| + \frac{-[f'(U_j^m)]_-\,\Delta t}{\Delta x}|U_{j+1}^m| \\
&\le \left(1 - \frac{|f'(U_j^m)|\,\Delta t}{\Delta x}\right)\|U_j^m\|_\infty + \frac{[f'(U_j^m)]_+\,\Delta t}{\Delta x}\|U^m\|_\infty + \frac{-[f'(U_j^m)]_-\,\Delta t}{\Delta x}\|U^m\|_\infty \qquad (166) \\
&= \left(1 - \frac{|f'(U_j^m)|\,\Delta t}{\Delta x}\right)\|U_j^m\|_\infty + \frac{|f'(U_j^m)|\,\Delta t}{\Delta x}\|U^m\|_\infty = \|U^m\|_\infty
\end{aligned}$$

for all $j \in \mathbb{Z}$. Therefore,

$$\|U^{m+1}\|_\infty \le \|U^m\|_\infty.$$

To complete the inductive step it remains to show that (165) holds with m replaced by $m+1$. Let $j_0 \in \mathbb{Z}$ be such that $|U_{j_0}^{m+1}| = \|U^{m+1}\|_\infty$. Then, (166) and the fact that f' is nondecreasing imply that

$$\frac{f'(\|U^{m+1}\|_\infty)\,\Delta t}{\Delta x} = \frac{f'(|U_{j_0}^{m+1}|)\,\Delta t}{\Delta x} \le \frac{f'(\|U^m\|_\infty)\,\Delta t}{\Delta x} \le 1 \qquad \text{for all } k = 0, \ldots, m. \qquad (167)$$

The inequality (167) shows that (165) holds with m replaced by $m+1$, which then completes the inductive step. Thus we have shown that, under the CFL condition (164),

$$\|U^{m+1}\|_\infty \le \|U^m\|_\infty \le \cdots \le \|U^0\|_\infty$$

for all $m = 0, 1, \ldots, M-1$; which completes the proof of the assertion that the sequence of finite difference approximations $\{U_j^m\}_{j\in\mathbb{Z},\, 0\le m\le M}$ is bounded; in particular (162) has been shown to hold.