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

September 8, 2024

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

Part 1. Overview of the lecture course and motivating examples from various applications in the sciences. Basic background from the theory of function spaces.

Finite difference approximation of two-point boundary-value problems for second-order ODEs. Mesh-dependent inner-products and mesh-dependent norms. Discrete Poincaré–Friedrichs inequality.

Stability, consistency and convergence of finite difference approximations of two-point boundary-value problems.

Part 2. Second-order linear elliptic boundary-value problems and their finite difference approximation: uniform meshes on axiparallel domains; nonuniform meshes on nonaxiparallel domains.

Discrete maximum principle; stability and convergence in the discrete maximum norm.

Discrete energy estimates; stability and convergence in discrete Sobolev norms.

Iterative solution of systems of linear equations arising from the discretization of second-order linear elliptic PDEs: linear stationary iterative methods.

Part 3. 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.

Discrete Fourier analysis of finite-difference approximations of initial-value problems for second-order linear parabolic PDEs: the Courant–Friedrichs–Lewy (CFL) condition.

Finite difference approximation of initial-boundary-value problems for second-order parabolic PDEs.

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.

Discrete energy norm estimates for finite difference approximations of initial-boundary-value problems for second-order parabolic problems: stability, consistency and convergence.

Part 4. Finite-difference approximation of second-order linear hyperbolic equations.

Discrete energy estimates for second-order hyperbolic problems: stability (including the CFL condition), consistency and convergence.

Finite difference approximation of linear first-order hyperbolic equations: stability (including the CFL condition), consistency and convergence.

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].

A note about the problem sheets

There are 4 problem sheets and 4 intercollegiate classes associated with the lectures. Each problem sheet is divided into three sections: A, B and C.

- Section A covers a mixture of background material and examinable material. Section B covers examinable material. Section C contains more challenging questions.
- Sections A and C will not be marked and solutions to questions appearing in these sections are provided to the students. Solutions to questions from Sections A and C will not, normally, be discussed in the classes.
- Section B contains questions on core material; these are of suitable length for the students to attempt in up to 8 hours, for a teaching assistant to mark in 20–30 minutes, and for the class tutor to present in a 90-minute class.

Which lecture videos to watch before attempting the problem sheets?

Problem sheet 1: Watch lecture videos 0, 1, 2, 3;

Problem sheet 2: Watch lecture videos 4, 5, 6, 7;

Problem sheet 3: Watch lecture videos 8, 9, 10, 11;

Problem sheet 4: Watch lecture videos 12, 13, 14, 15, 16.

About these lecture notes

These lecture notes will be updated regularly during Michaelmas Term. If you notice any typographical errors or inaccuracies, please report them to me by email.

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 depends 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 $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$ is called a *multi-index*. The nonnegative integer $|\alpha| := \alpha_1 + \cdots + \alpha_n$ is called the length of the multi-index $\alpha = (\alpha_1, \ldots, \alpha_n)$. We denote $(0, \ldots, 0)$ by $\mathbf{0}$; clearly $|\mathbf{0}| = 0$.

Let

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

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

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

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 $\alpha = (\alpha_1, \ldots, \alpha_n)$ with $|\alpha| \leq k$. 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 $\alpha = (\alpha_1, \dots, \alpha_n)$ with $|\alpha| \le k$. The linear space $C^k(\overline{\Omega})$ can then be equipped with the norm

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

where $x := (x_1, \dots, x_n)$. 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 does not belong to $C^k(\overline{\Omega})$ for any $k \ge 0$. ⋄

The *support*, supp u, of a continuous function u on Ω is defined as the closure in Ω of the set

$$\{x \in \Omega : u(x) \ne 0\}.$$

In other words, supp u is the smallest closed subset of Ω such that $u = 0$ in $\Omega\backslash$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 $\{x \in \mathbb{R}^n : |x| \le 1\}$. ⋄

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 $C_0^\infty(\mathbb{R}^n)$. ⋄

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 an open set $\Omega \subset \mathbb{R}^n$ such that

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

Here, $x := (x_1, \dots, x_n)$ and $\mathrm{d}x := \mathrm{d}x_1 \dots \mathrm{d}x_n$. 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*

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

Proof. By taking $\lambda = 1$ in the proof of the Cauchy–Schwarz inequality above, we deduce that

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

where in the transition to the second line we applied the Cauchy–Schwarz inequality. □

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 linear space X, equipped with an inner product $(\cdot, \cdot)_X$ (and the associated norm $\|u\|_X := (u, u)_X^{1/2}$) is called a Hilbert space if, whenever $\{u_m\}_{m=1}^\infty$ is a 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 that 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:

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

We note here that all integrals on $\partial\Omega$ that arise in the course of partial integration, based on the divergence theorem,[1] have vanished because $v \in C_0^\infty(\Omega)$. However, in the theory of partial differential equations one often has to consider functions u that do not possess the smoothness hypothesized 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

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

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 first weak derivative of the function u defined on Ω by $u(x) = (1 - |x|)_+$. Here, for a real number y, y_+ denotes the nonnegative part of y, defined by $y_+ := \max\{y, 0\}$. 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

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

---

[1] Observe that

$$\int_\Omega \frac{\partial u}{\partial x_i} v\, \mathrm{d}x = \int_\Omega \frac{\partial (uv)}{\partial x_i}\, \mathrm{d}x - \int_\Omega u \frac{\partial v}{\partial x_i}\, \mathrm{d}x = \int_{\partial\Omega} uv\, \nu_i\, \mathrm{d}s(x) - \int_\Omega u \frac{\partial v}{\partial x_i}\, \mathrm{d}x,$$

where ν_i is the i-th component of the unit outward normal vector $\nu = (\nu_1, \ldots, \nu_n)$ to the boundary $\partial\Omega$ of Ω. Here, the first equality follows from the product rule for derivatives, while the second equality follows by applying the divergence theorem to the n-component vector function $(0, \ldots, 0, uv, 0, \ldots, 0)$ whose i-th component is uv while all of the other components are equal to zero, and noting that $\mathrm{div}(0, \ldots, 0, uv, 0, \ldots, 0) = \frac{\partial (uv)}{\partial x_i}$ and $(0, \ldots, 0, uv, 0, \ldots, 0) \cdot \nu = uv\, \nu_i$.

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

Now we are ready to give a precise definition of a Sobolev space. Let k be a 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

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

and the inner product

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

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

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

we can write

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

$|\cdot|_{H^k(\Omega)}$ is called the Sobolev semi-norm (it is only a semi-norm rather than a norm because if $|u|_{H^k(\Omega)} = 0$ for $u \in H^k(\Omega)$ and $k \ge 1$, then 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,

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

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

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

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 $u \in H^1_0(\Omega)$; 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,

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

Thus, by the Cauchy–Schwarz inequality,

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

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 combining the two inequalities (viz. by moving the constants $\frac{1}{2}(b-a)^2$ and $\frac{1}{2}(d-c)^2$ to the left-hand sides of the respective inequalities, and then summing the resulting inequalities), we obtain

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

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

2 Elliptic boundary-value problems

In the first half of this lecture course we shall focus on boundary-value problems for elliptic partial differential equations. Elliptic equations are typified by the Laplace equation[2] **Lecture 2**

$$\Delta u = 0,$$

and its nonhomogeneous 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:

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

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*. In the case of Poisson's equation, for example, $a_{i,j} = \delta_{i,j}$ for $i,j = 1,\dots,n$ (and also $b_i(x) \equiv 0$ for $i = 1,\dots,n$ and $c(x) \equiv 0$), and the ellipticity condition is therefore trivially satisfied, with $\tilde{c} = 1$.

The 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 the boundary $\partial\Omega$ of Ω, and where the derivative in the direction of ν is defined by $\frac{\partial u}{\partial \nu} := \nabla u \cdot \nu$ (*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

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

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 is imposed on $\partial\Omega$ (e.g. $\partial\Omega$ is the union of two disjoint subsets $\partial\Omega_1$ and $\partial\Omega_2$, with a Dirichlet boundary condition is imposed on $\partial\Omega_1$ and a Neumann boundary condition on $\partial\Omega_2$). The study of such mixed boundary-value problems is beyond the scope of these notes.

[2]Recall that in n space dimensions the Laplace operator Δ is defined by $\Delta u := \frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2}$.

We begin by considering the homogeneous Dirichlet boundary-value problem

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

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

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 that $a_{i,j}$, 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 of partial differential equations is inappropriate. Take, for example, Poisson's equation on the cube $\Omega = (-1,1)^n$ in $\mathbb{R}^n$, subject to a zero Dirichlet boundary condition:

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

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 on Ω.

Start of optional material

2.1 Existence and uniqueness of weak solutions

In order to overcome the limitations of the classical theory of partial differential equations and to be able to deal with partial differential equations with "nonsmooth" data such as $(*)$, we generalize the notion of solution by weakening the differentiability requirements on u; this will lead us to the notion of *weak solution*. 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 on $\partial\Omega$, 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-hand side and right-hand side of (6) are both 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*

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

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 does not have a classical solution. Before focusing on this particular boundary-value problem, we consider 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:

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

and

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

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

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

Before proceeding we observe that, for any $f \in L_2(\Omega)$ the mapping $v \in H_0^1(\Omega) \mapsto \ell(v) \in \mathbb{R}$ is a linear functional on $H_0^1(\Omega)$. Similarly, for each fixed $v \in H_0^1(\Omega)$ the mapping $w \in H_0^1(\Omega) \mapsto a(w,v) \in \mathbb{R}$ is a linear functional on $H_0^1(\Omega)$ and for each fixed $w \in H_0^1(\Omega)$ the mapping $v \in H_0^1(\Omega) \mapsto a(w,v) \in \mathbb{R}$ is a linear functional on $H_0^1(\Omega)$; thus $a(\cdot,\cdot)$ is a bilinear functional (or bilinear form) on $H_0^1(\Omega) \times H_0^1(\Omega)$.

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

Theorem 1 *(Lax & Milgram theorem[3]) 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 $|a(w,v)| \leq c_1 \|w\|_V \|v\|_V$ for all $w, v \in V$;

and let $\ell(\cdot)$ be a linear functional 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

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

[3]Lax, P. D.; Milgram, A. N. Parabolic equations. Contributions to the theory of partial differential equations, pp. 167190. Annals of Mathematics Studies, no. 33. Princeton University Press, Princeton, N. J., 1954. For a proof of this result the interested reader is referred to the book of P. Ciarlet: The Finite Element Method for Elliptic Problems, SIAM, Philadelphia, 2002. The digital version of the book is available from https://epubs.siam.org/doi/book/10.1137/1.9780898719208.

We apply the Lax–Milgram theorem with $V = H^1_0(\Omega)$ 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 $H^1_0(\Omega)$ is a Hilbert space with the inner product

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

and the associated norm $\|v\|_{H^1(\Omega)} = (v,v)^{1/2}_{H^1(\Omega)}$. 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 $\alpha, \beta \in \mathbb{R}$,

$$\begin{aligned}\ell(\alpha v_1 + \beta v_2) &= \int_\Omega f(x)\,(\alpha v_1(x) + \beta v_2(x))\,\mathrm{d}x \\ &= \alpha \int_\Omega f(x)\,v_1(x)\,\mathrm{d}x + \beta \int_\Omega f(x)\,v_2(x)\,\mathrm{d}x \\ &= \alpha\ell(v_1) + \beta\ell(v_2), \qquad v_1, v_2 \in H^1_0(\Omega);\end{aligned}$$

hence, $\ell(\cdot)$ is a linear functional on $H^1_0(\Omega)$. 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 $v \in H^1_0(\Omega)$, 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: $|\ell(v)| \le c_2\|v\|_{L_2(\Omega)}$ for all $v \in H^1(\Omega)$.

Next we verify (b). For any fixed $v \in H^1_0(\Omega)$, the mapping $w \in H^1_0(\Omega) \mapsto a(w,v) \in \mathbb{R}$ is linear. Similarly, for any fixed $w \in H^1_0(\Omega)$, the mapping $v \in H^1_0(\Omega) \mapsto a(w,v) \in \mathbb{R}$ is linear. Hence $a(\cdot,\cdot)$ is a bilinear form on $H^1_0(\Omega) \times H^1_0(\Omega)$. By employing the Cauchy–Schwarz inequality, we deduce that

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

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) by applying the inequality $\sqrt{a} + \sqrt{b} \le \sqrt{2}\sqrt{a+b}$, where $a, b \ge 0$, to each of the expressions in the curly brackets, we arrive at the inequality

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

by letting $c_1 := 2nc$, we obtain the desired bound asserted in (b) of the Lax–Milgram theorem.

It remains to verify hypothesis (a) of the Lax–Milgram theorem. Using the uniform ellipticity condition $(3)_4$, we deduce that

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

where we wrote $\frac{\partial v}{\partial x_i} v$ as $\frac{1}{2}\frac{\partial}{\partial x_i}(v^2)$.

In $(3)_1$ we assumed that $a_{i,j} \in C^1(\overline{\Omega})$ for all $i, j = 1, \ldots, n$. As a matter of fact, since the weak formulation of the boundary-value problem stated in Definition 1 does not involve differentiation of the coefficients $a_{i,j}$, and nor has the verification of the conditions of hypotheses (a) and (b) of the Lax–Milgram theorem required that $a_{i,j} \in C^1(\overline{\Omega})$ for all $i, j = 1, \ldots, n$, it will suffice to suppose the weaker requirement that $a_{i,j} \in C(\overline{\Omega})$ for all $i, j = 1, \ldots, n$, which is what was assumed in Definition 1[4]. On the other hand, to proceed with the verification of hypothesis (a) of the Lax–Milgram theorem, we shall have to strengthen our original assumption that $b_i \in C(\overline{\Omega})$, $i = 1, \ldots, n$, and require instead that $b_i \in C^1(\overline{\Omega})$, $i = 1, \ldots, n$.

Integrating by parts in the second term on the right and noting that the boundary integral term arising in the course of partial integration vanishes thanks to the fact that $v|_{\partial\Omega} = 0$, we then obtain

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

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

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

Then,

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

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

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

Summing the inequalities (13) and (14) we deduce that

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

where $c_0 = \tilde{c}/(1+c^\star)$, and hence hypothesis (a) of the Lax–Milgram theorem has also been verified. Having checked all hypotheses of the Lax–Milgram theorem, we deduce the existence of a unique $u \in H^1_0(\Omega)$ satisfying (10); thereby problem (4), (5) has a unique weak solution $u \in H^1_0(\Omega)$.

We record this result in the following theorem.

[4]As a matter of fact, the requirement that $a_{i,j} \in C(\overline{\Omega})$ for $i, j = 1, \ldots, n$ can be further weakened: it suffices to assume that $a_{i,j} \in L_\infty(\Omega)$ for $i, j = 1, \ldots, n$, i.e., that there exists a positive real number M such $|a_{i,j}(x)| \leq M$ for all $i, j = 1, \ldots, n$ and for all $x \in \Omega$, except perhaps for a set of x of measure zero. Thus, for example, the coefficients $a_{i,j}$ may be bounded piecewise continuous functions defined on Ω.

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 the inequality (16). By the inequality (15), the definition (10) of the weak formulation, the Cauchy–Schwarz inequality and by recalling the definition of the norm $\|\cdot\|_{H^1(\Omega)}$, we have that

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

Hence the desired inequality. □

Now we return to our earlier example $(*)$, which has been shown to have no classical solution. However, by applying Theorem 4 with $a_{i,j}(x) \equiv 1$, $i = j$, $a_{i,j}(x) \equiv 0$, $i \neq j$, $1 \leq i, j \leq n$, $b_i(x) \equiv 0$ for $i = 1, \ldots, n$, $c(x) \equiv 0$, $f(x) = \mathrm{sgn}(\frac{1}{2} - |x|)$, and $\Omega = (-1, 1)^n$, we see that inequality (3) holds with $\tilde{c} = 1$ and inequality (12) is trivially satisfied. 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*; that is, 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 the required continuous dependence of the solution of the boundary-value problem on the right-hand side directly follows. ⋄

End of optional material

3 Introduction to the theory of finite difference schemes

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

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

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.

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 computational domain is approximated by a finite set of points, called the finite difference mesh, and second, the derivatives appearing in the differential equation (and, possibly also in the boundary condition(s)) are approximated by divided differences (difference quotients) on the finite difference mesh.

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 of 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 $\max_{1\le i\le n} h_i$ is, the finer the mesh.

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

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

where f_h and g_h are suitable approximations of f and g, respectively. Now (19) is a system of linear algebraic 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 more 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 primarily concerned with the first of these two problems — the question of approximation — although we shall also briefly consider the question of iterative solution of systems of linear algebraic equations by a simple iterative method. More sophisticated iterative methods, so called Krylov subspace iterations, for the solution of large systems of linear algebraic equations, such as those that arise from the approximate solution of partial differential equations, are covered in the fourth-year C6.1 Numerical Linear Algebra course.

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 **Lecture 3**
consider the following two-point boundary-value problem for a second-order linear (ordinary) differential equation:

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

where f and c are real-valued functions, which are defined and continuous on the interval $[0,1]$ and $c(x) \geq 0$ for all $x \in [0,1]$.

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 $x_i := ih$, $i = 0,\dots,N$. Formally, $\Omega_h := \{x_i : i = 1,\dots,N-1\}$ is the set of interior mesh-points, $\Gamma_h := \{x_0, x_N\}$ the set of boundary mesh-points and $\overline{\Omega}_h := \Omega_h \cup \Gamma_h$ the set of all mesh-points. 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

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

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

and

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

D_x^+ and D_x^- are called the *forward* and *backward first divided difference* operator, respectively, and $D_x^+ D_x^-$ $(= D_x^- D_x^+)$ is called the (symmetric) *second divided difference* operator. The difference operator D_x^0, called the *central first divided difference* operator, is defined by

$$D_x^0 u(x_i) := \frac{1}{2}\left(D_x^+ u(x_i) + D_x^- u(x_i)\right) = \frac{u(x_{i+1}) - u(x_{i-1})}{2h} \quad (= u'(x_i) + \mathcal{O}(h^2)).$$

Thus we replace the second derivative u'' in the differential equation by the second divided difference $D_x^+ D_x^- u(x_i)$; hence,

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

Now (21) indicates that the approximate solution U (not to be confused with the exact solution u) should be sought as the solution of the system of difference equations:

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

This is, in fact, a system of $N-1$ linear algebraic equations for the $N-1$ unknowns, U_i, $i = 1, \dots, N-1$. Using matrix notation, the linear system can be written as follows:

$$\begin{bmatrix} \frac{2}{h^2} + c(x_1) & -\frac{1}{h^2} & & & 0 \\ -\frac{1}{h^2} & \frac{2}{h^2} + c(x_2) & -\frac{1}{h^2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{h^2} & \frac{2}{h^2} + c(x_{N-2}) & -\frac{1}{h^2} \\ 0 & & & -\frac{1}{h^2} & \frac{2}{h^2} + c(x_{N-1}) \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-2} \\ U_{N-1} \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{N-2}) \\ f(x_{N-1}) \end{bmatrix},$$

or, more compactly, $AU = F$, where A is the symmetric tridiagonal $(N-1) \times (N-1)$ matrix displayed above, and U and F are column vectors of size $N-1$, corresponding to the $N-1$ 'interior' mesh-points $x_1, \dots, x_{N-1}$ contained in the open interval $(0, 1)$.

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 nonsingular (i.e., $\det A \neq 0$), and therefore invertible. We shall do so by developing a technique which we shall, in subsequent sections, extend to the finite difference approximation of partial differential equations. The purpose of this section is to introduce the key ideas through the finite difference approximation (21) of the simple two-point boundary-value problem (20).

For this purpose, we introduce, for two functions V and W defined at the interior mesh-points x_i, $i = 1, \dots, N-1$, 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

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

The argument that we shall develop is based on mimicking, at the discrete level, the following procedure based on integration-by-parts, noting that the solution of the boundary-value problem (20) satisfies the homogeneous boundary conditions $u(0) = 0$ and $u(1) = 0$ at the end-points of the interval $[0, 1]$:

$$\int_0^1 (-u''(x) + c(x)u(x))\, u(x)\, \mathrm{d}x = \int_0^1 |u'(x)|^2 + c(x)|u(x)|^2\, \mathrm{d}x \geq \int_0^1 |u'(x)|^2\, \mathrm{d}x, \tag{23}$$

thanks to the assumption that $c(x) \geq 0$ for all $x \in [0,1]$. Thus if, for example, f is identically zero on $[0,1]$, then so is $-u''+c(x)u$, and thanks to the inequality (23) the function u' is then also identically equal to zero on $[0,1]$. Consequently, u is a constant function on $[0,1]$, but because $u(0) = 0$ and $u(1) = 0$, the constant function u must be identically equal to 0. In other words, the only solution to the homogeneous boundary-value problem (i.e., the boundary-value problem with $f(x) \equiv 0$ for $x \in [0,1]$) is the function $u(x) \equiv 0$, $x \in [0,1]$. For the finite difference approximation of the boundary-value problem, if we could show by an analogous argument that the homogeneous system of linear algebraic equations corresponding to $f(x_i) = 0$, $i = 1, \ldots, N-1$, has the trivial solution $U_i = 0$, $i = 0, \ldots, N$, as its unique solution, then the desired invertibility of the matrix A would directly follow.

Our key technical tool to this end is the following summation-by-parts identity, which is the discrete counterpart of the integration-by-parts identity $(-u'', u) = (u', u') = \|u'\|^2_{L_2((0,1))}$ satisfied by the function u, obeying the homogeneous boundary conditions $u(0) = 0$, $u(1) = 0$, used in (23) above.

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,*

$$(-D_x^+ D_x^- V, V)_h = \sum_{i=1}^{N} h \left|D_x^- V_i\right|^2 . \tag{24}$$

Proof. By recalling the definition of the inner product $(\cdot,\cdot)_h$ and the definition of $D_x^+ D_x^- V_i$ we have that

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

where in the transition to the third line we shifted the index in the first summation, and in the transition to the fourth line we made use of the fact that, by hypothesis, $V_0 = V_N = 0$. $\square$

Returning to the finite difference scheme (22), let V be as in the above lemma and note that as, by hypothesis, $c(x) \geq 0$ for all $x \in [0,1]$, we have that

$$\begin{aligned}
(AV, V)_h &= (-D_x^+ D_x^- V + cV, V)_h \\
&= (-D_x^+ D_x^- V, V)_h + (cV, V)_h \\
&\geq \sum_{i=1}^{N} h \left|D_x^- V_i\right|^2 .
\end{aligned} \tag{25}$$

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 $V = 0$. It therefore follows that A is a nonsingular matrix, and thereby (22) has a unique solution, $U = A^{-1}F$. We record this result in the next theorem.

Theorem 3 *Suppose that c and f are continuous real-valued functions defined on the interval $[0,1]$, and $c(x) \geq 0$ for all $x \in [0,1]$; then, the finite difference scheme (22) possesses a unique solution U.*

We note in passing that, thanks to Theorem 3, the boundary-value problem (20) has a unique (weak) solution under the hypotheses on c and f asserted in Theorem 3.

Remark 1 *In the discussion preceding Theorem 3 we used the symbol A to denote the matrix of the system of linear equations that arises from the finite difference approximation as well as the finite difference operator $V \mapsto -D_x^+ D_x^- V + cV$. Similarly, we used the symbol U to denote the vector $(U_1, \ldots, U_{N-1})^{\mathrm{T}}$ of unknowns representing the solution of the system of linear algebraic equations $AU = F$ as well as the mesh function defined on the finite difference mesh $\overline{\Omega}_h$ with the understanding that $U_0 = U_N = 0$. For the sake of notational simplicity we shall continue to use these conventions throughout: i.e., we shall use the same notation for matrices and finite difference operators, and we shall use the same notation for vectors and mesh functions defined over finite difference meshes. It will be clear from the context which of the two interpretations of the same symbol is intended.*

Next, we investigate the approximation properties of the finite 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

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

is the norm induced by the inner product

$$(V,W]_h := \sum_{i=1}^{N} h\, V_i\, W_i.$$

Using this notation, the inequality (25) can be rewritten as follows:

$$(AV,V)_h \geq \|D_x^- V]|_h^2. \tag{26}$$

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

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

where c_0 is a positive constant, independent of h.

Lemma 4 *(Discrete Poincaré–Friedrichs inequality.) Let V be a function defined on the finite difference mesh $\{x_i := ih : i = 0, \ldots, N\}$, where $h := 1/N$ and $N \geq 2$, and such that $V_0 = V_N = 0$; then, there exists a positive constant $c_\star$, independent of V and h, such that*

$$\|V\|_h^2 \leq c_\star \|D_x^- V]|_h^2 \tag{27}$$

for all such V.

Proof. We proceed in the same way as in the proof of inequality (1). We begin by noting that, thanks to the definition of $D_x^- V_i$ and by use of the Cauchy–Schwarz inequality,

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

Therefore, because $\sum_{i=1}^{N-1} i = \frac{1}{2}(N-1)N$ and $Nh = 1$, we have that

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

We note that the constant $c_\star = 1/2$ in the inequality (27). □

Using the inequality (27) to bound the right-hand side of the inequality (26) from below we obtain

$$(AV, V)_h \ge \frac{1}{c_\star} \|V\|_h^2. \tag{28}$$

Adding the inequality (26) to the inequality (28) we arrive at the inequality

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

Letting $c_0 = (1 + c_\star)^{-1}$ it follows that

$$(AV, V)_h \ge c_0 \|V\|_{1,h}^2. \tag{29}$$

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

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

$$\|U\|_{1,h} \le \frac{1}{c_0} \|f\|_h. \tag{30}$$

Proof. From the inequality (29) and the definition (22) of the finite difference scheme we have that

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

and hence the inequality (30). □

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,

$$
\begin{aligned}
Ae_i = \varphi_i, \qquad i &= 1, \ldots, N-1, \\
e_0 = 0, \quad e_N &= 0,
\end{aligned} \tag{31}
$$

where $\varphi_i := Au(x_i) - f(x_i) = u''(x_i) - D_x^+ D_x^- u(x_i)$ is the *consistency error* (sometimes also called the *truncation error*). By applying the inequality (30) to the finite difference scheme (31), we obtain

$$
\|u - U\|_{1,h} = \|e\|_{1,h} \leq \frac{1}{c_0} \|\varphi\|_h. \tag{32}
$$

It remains to estimate $\|\varphi\|_h$. We showed on page 14 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 exists a positive constant C, independent of h, such that

$$
|\varphi_i| \leq C h^2.
$$

Consequently,

$$
\|\varphi\|_h = \left(\sum_{i=1}^{N-1} h \, |\varphi_i|^2 \right)^{1/2} \leq C h^2. \tag{33}
$$

By combining the inequalities (32) and (33) it follows that

$$
\|u - U\|_{1,h} \leq \frac{C}{c_0} h^2. \tag{34}
$$

In fact, a more careful treatment of the remainder term in the Taylor series expansion on p.14 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

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

in inequality (33). 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 inequality (34) 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 $c(x) \geq 0$ for all $x \in [0,1]$, and suppose that the corresponding (weak) solution of the boundary-value problem (20) belongs to $C^4([0,1])$; then*

$$
\|u - U\|_{1,h} \leq \frac{1}{8} h^2 \|u^{IV}\|_{C([0,1])}. \tag{35}
$$

The analysis of the simple finite difference scheme (22) 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. inequality (30), for example). A typical stability result for the general finite difference scheme (19) is

$$|||U|||_{\Omega_h} \leq C_1 \left(\|f_h\|_{\Omega_h} + \|g_h\|_{\Gamma_h}\right), \tag{36}$$

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_1 is a positive constant, independent of h.

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

$$\begin{aligned} \varphi_{\Omega_h} &:= \mathcal{L}_h u - f_h, \qquad && \text{in } \Omega_h, \\ \varphi_{\Gamma_h} &:= \mathcal{B}_h u - g_h, \qquad && \text{on } \Gamma_h. \end{aligned}$$

(in the case of the finite difference scheme (20) $\varphi_{\Gamma_h} = 0$, and therefore φ_{Γ_h} did not appear 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 the boundary-value problem (18), we say that the scheme (19) is *consistent*. If p is the largest positive integer such that

$$\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h} \leq C_2 h^p \quad \text{as} \quad h \to 0,$$

(where C_2 is a positive constant independent of h) for all sufficiently smooth u, the scheme is said to have *order of accuracy* (or *order of consistency*) p.

The finite difference scheme (19) is said to provide a *convergent* approximation to the solution u of the boundary-value problem (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

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

(where C is a positive constant independent of the mesh-size 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), involving linear finite difference operators $\mathcal{L}_h$ and $\mathcal{B}_h$, is stable (i.e., the inequality (36) holds for all f_h and g_h) and that the scheme is a consistent approximation of the boundary-value problem (18); then the finite difference scheme (19) is a convergent approximation of the boundary-value problem (18), and the order of convergence q is not smaller then the order of accuracy (order of consistency) p.*

PROOF. We define the *global error* $e := u - U$. Then, thanks to the assumed linearity of $\mathcal{L}_h$, we have that

$$\mathcal{L}_h e = \mathcal{L}_h(u - U) = \mathcal{L}_h u - \mathcal{L}_h U = \mathcal{L}_h u - f_h.$$

Thus

$$\mathcal{L}_h e = \varphi_{\Omega_h}.$$

Similarly, thanks to the assumed linearity of $\mathcal{B}_h$, we have that

$$\mathcal{B}_h e = \varphi_{\Gamma_h}.$$

By the assumed stability of the scheme it then follows that

$$|||u - U|||_{\Omega_h} = |||e|||_{\Omega_h} \leq C_1(\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h}),$$

and hence the stated result with $q \geq p$ thanks to the assumed consistency of order p of the finite difference scheme. That completes the proof. $\square$

Thus, paraphrasing Theorem 3.6, *stability* and *consistency* imply *convergence*. This abstract result is at the heart of the convergence 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**

$$\begin{aligned} -\Delta u + c(x,y)u &= f(x,y) \qquad && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{37}$$

where $\Omega = (0,1) \times (0,1)$, c is a continuous function on $\overline{\Omega}$ and $c(x,y) \geq 0$. 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 $L_2(\Omega)$. As f need not be continuous on Ω, the boundary-value problem (37) need not have a classical solution – only a weak solution exists. This gives rise to technical difficulties: in particular, we cannot use a Taylor series expansion to estimate the size of the consistency 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 (37) 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

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

Similarly as in Section 3, we consider the set of interior mesh-points

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

and the set of boundary mesh-points $\Gamma_h := \overline{\Omega}_h \setminus \Omega_h$.

Analogously to (22), the finite difference scheme is:

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) + c(x_i, y_j)U_{i,j} &= f(x_i, y_j) \qquad && \text{for } (x_i, y_j) \in \Omega_h, \\ U &= 0 && \text{on } \Gamma_h. \end{aligned} \tag{38}$$

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

$$-\left\{ \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} + \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{h^2} \right\} + c(x_i, y_j)U_{i,j} = f(x_i, y_j),$$
$$i, j = 1, \ldots, N-1, \tag{39}$$

$$U_{i,j} = 0 \quad \text{if } i = 0,\ i = N \text{ or if } j = 0,\ j = N. \tag{40}$$

For each i and j, $1 \leq i, j \leq N-1$, the finite difference equation (39) involves five values of the approximate solution U: $U_{i,j}$, $U_{i-1,j}$, $U_{i+1,j}$, $U_{i,j-1}$, $U_{i,j+1}$, and is therefore frequently referred to as the *five-point difference scheme*. It is again possible to write (39), (40) as a system of linear algebraic equations

$$AU = F, \tag{41}$$

where now

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

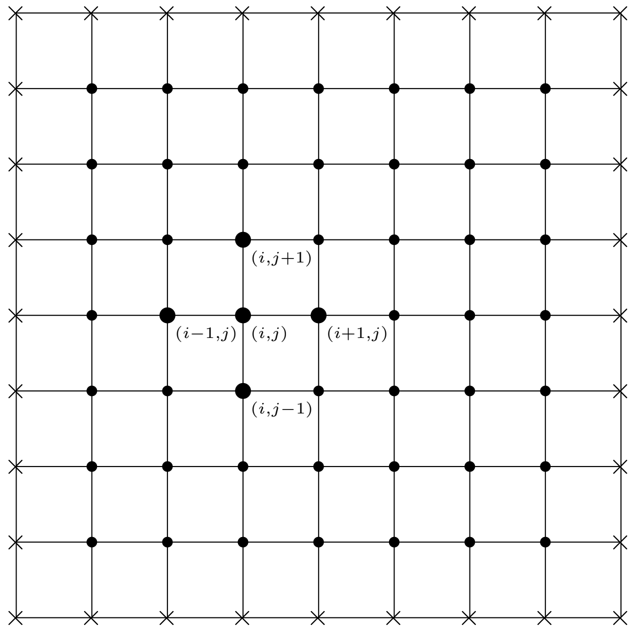


Figure 1: The mesh $\Omega_h(\cdot)$, the boundary mesh $\Gamma_h(\times)$, and a typical five-point difference stencil.

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

and A is an $(N-1)^2 \times (N-1)^2$ sparse matrix of banded structure (i.e. a sparse matrix whose nonzero entries are confined to a diagonal band, comprising the main diagonal and zero or more diagonals on either side). A typical row of the matrix contains five nonzero 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 a solution, stability, consistency, and convergence

Next we show that (38) has a unique solution. We proceed analogously 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 $(v, w) := \int_\Omega v(x, y)\, w(x, y)\, \mathrm{d}x\, \mathrm{d}y$. The next result is a direct extension of Lemma 3 from the univariate case to the case of two space dimensions.

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

$$(-D_x^+ D_x^- V, V)_h + (-D_y^+ D_y^- V, V)_h = \sum_{i=1}^{N} \sum_{j=1}^{N-1} h^2 |D_x^- V_{i,j}|^2 + \sum_{i=1}^{N-1} \sum_{j=1}^{N} h^2 |D_y^- V_{i,j}|^2. \quad (42)$$

Proof. The identity (42) is a direct consequence of (24) and the analogous identity for $-D_y^+ D_y^-$. □

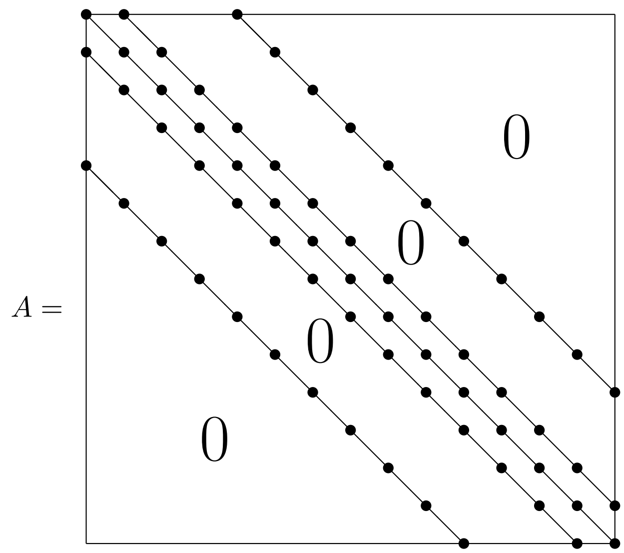


Figure 2: The sparsity structure of the banded matrix A.

Returning to the analysis of the finite difference scheme (38), we shall now proceed in much the same way as in the univariate case considered in the previous section. We note that, since $c(x,y) \geq 0$ on $\overline{\Omega}$, by (42) we have that

$$\begin{aligned} (AV,V)_h &= (-D_x^+D_x^-V - D_y^+D_y^-V + cV, V)_h \\ &= (-D_x^+D_x^-V, V)_h + (-D_y^+D_y^-V, V)_h + (cV,V)_h \\ &\geq \sum_{i=1}^{N}\sum_{j=1}^{N-1} h^2|D_x^-V_{i,j}|^2 + \sum_{i=1}^{N-1}\sum_{j=1}^{N} h^2|D_y^-V_{i,j}|^2, \end{aligned} \tag{43}$$

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 nonsingular matrix. Indeed if $AV = 0$, then (43) yields:

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

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

Since $V = 0$ on Γ_h, these imply that $V \equiv 0$. Thus $AV = 0$ if and only if $V = 0$. Hence A is nonsingular, and $U = A^{-1}F$ is the unique solution of (38). Thus the solution of the finite difference scheme (38) may be found by solving the system of linear algebraic equations (41).

In order to prove the stability of the finite difference scheme (38), we introduce (similarly as in the univariate case) 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 $\|\cdot\|_{H^1(\Omega)}$, defined by

$$\|u\|_{H^1(\Omega)} := \left(\|u\|^2_{L_2(\Omega)} + \left\| \frac{\partial u}{\partial x} \right\|^2_{L_2(\Omega)} + \left\| \frac{\partial u}{\partial y} \right\|^2_{L_2(\Omega)} \right)^{1/2} .$$

With this new notation, the inequality (43) can be rewritten in the following compact form:

$$(AV, V)_h \geq \|D_x^- V]|_x^2 + \|D_y^- V]|_y^2. \tag{44}$$

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.) Suppose that V is 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*

$$\|V\|_h^2 \leq c_* \left(\|D_x^- V]|_x^2 + \|D_y^- V]|_y^2 \right) \tag{45}$$

for all such V.

Proof. The inequality (45) is a straightforward consequence of its univariate counterpart (27). It follows from (27) that, for each fixed j, $1 \leq j \leq N-1$,

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

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

$$\sum_{j=1}^{N-1} h|V_{i,j}|^2 \leq \frac{1}{2} \sum_{j=1}^{N} h|D_y^- V_{i,j}|^2. \tag{47}$$

We first multiply (46) by h and sum through j, $1 \leq j \leq N-1$, then multiply (47) by h and sum through i, $1 \leq i \leq N-1$, and finally 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 we arrive at (45) with $c_* = \frac{1}{4}$. That completes the proof. □

Now the inequalities (44) and (45) imply that

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

Finally, combining this inequality with (44) and recalling the definition of the norm $\|\cdot\|_{1,h}$, we obtain

$$(AV, V)_h \geq c_0\|V\|_{1,h}^2, \tag{48}$$

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

Using the inequality (48) we can now prove the stability of the finite difference scheme (38).

Theorem 7 *The finite difference scheme (38) is stable in the sense that*

$$\|U\|_{1,h} \leq \frac{1}{c_0}\|f\|_h. \tag{49}$$

Proof. The proof of this inequality is identical to that of the stability inequality (30) in the univariate case. From (48) and (38) 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 we arrive at the desired inequality (49). □

4.1.1 Convergence in the class of classical solutions

Having established stability of the finite difference scheme (38), we turn to the question of its 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 with remainder terms in the x and y coordinate directions, respectively, we have that

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

where $\xi_i \in [x_{i-1}, x_{i+1}]$, $\eta_j \in [y_{j-1}, y_{j+1}]$, and $f_{i,j} := f(x_i, y_j)$.

We define the *consistency error* (or *truncation error*) of the finite difference scheme (38) by

$$\varphi_{i,j} := Au(x_i, y_j) - f_{i,j}.$$

Then, by the calculations above,

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

and

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

Thanks to the stability result (49), we therefore have that

$$\|u - U\|_{1,h} = \|e\|_{1,h} \le \frac{1}{c_0}\|\varphi\|_h. \tag{50}$$

To arrive at a bound on the global error $e := u - U$ in the norm $\|\cdot\|_{1,h}$ it therefore remains to bound $\|\varphi\|_h$ and insert the resulting bound in the right-hand side of (50). Indeed, by noting that

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

we deduce that the consistency error, φ, satisfies

$$\|\varphi\|_h \le \frac{h^2}{12}\left(\left\|\frac{\partial^4 u}{\partial x^4}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^4 u}{\partial y^4}\right\|_{C(\overline{\Omega})}\right). \tag{51}$$

Finally (50) and (51) yield the following result.

Theorem 8 *Let $f \in C(\overline{\Omega})$, $c \in C(\overline{\Omega})$, with $c(x,y) \ge 0$, $(x,y) \in \overline{\Omega}$, and suppose that the corresponding weak solution of the boundary-value problem (37) belongs to $C^4(\overline{\Omega})$; then,*

$$\|u - U\|_{1,h} \le \frac{5h^2}{48}\left(\left\|\frac{\partial^4 u}{\partial x^4}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^4 u}{\partial y^4}\right\|_{C(\overline{\Omega})}\right). \tag{52}$$

PROOF. Recall that $c_0 = (1 + c_*)^{-1}$ and $c_* = \frac{1}{4}$, so that $1/c_0 = \frac{5}{4}$, and combine (50) and (51). $\square$

According to this result, the five-point difference scheme (38) for the boundary-value problem (37) is second-order convergent, provided that u is sufficiently smooth. As in the univariate case, we have deduced second-order convergence of the finite difference scheme from its stability and its second-order consistency, under the assumption that the exact solution u is sufficiently smooth, i.e., that $u \in C^4(\overline{\Omega})$, and therefore, because $c \in C(\overline{\Omega})$ by hypothesis, necessarily $f = -\Delta u + cu \in C(\overline{\Omega})$.

In general, however, even if f and c are smooth functions, the corresponding solution, u, of (37) will not be a smooth function because the boundary, Γ, of the domain, $\Omega = (0,1)^2$, is not a smooth curve. Thus, the hypothesis $u \in C^4(\overline{\Omega})$ is unrealistic.[5]

Our analysis has another limitation: it has been performed under the assumption that, at the very least, $f \in C(\overline{\Omega})$, which was required in order to ensure that the values of the function f are correctly defined at the mesh-points. However, in physical applications one often has to consider differential equations where f is not a continuous function on Ω, but 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 22.

[5]We note in passing that the regularity $u \in C^4(\overline{\Omega})$ can be guaranteed by assuming suitable, so called, *compatibility conditions* on the function f, which, for example in the special case when c is identically zero, require that the function f and its first and second partial derivatives vanish at the four corners of the square domain $\Omega = (0,1)^2$. However, we shall not consider such situations involving compatibility conditions here.

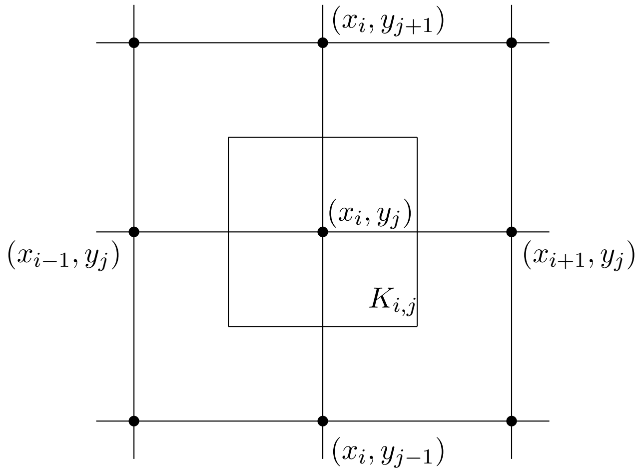


Figure 3: The cell $K_{i,j}$

(b) ($f \in L_2(\Omega)$). We retain the same finite difference mesh as in case (a), but we shall modify the right-hand side in the finite difference scheme (39) to cater for the fact that f is no longer assumed to be a continuous function on $\overline{\Omega}$. **Lecture 5**

The idea is to replace $f(x_i, y_j)$ in (39) 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. By 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

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

where $\partial K_{i,j}$ is the boundary of $K_{i,j}$, and ν is the unit outward normal to $\partial K_{i,j}$. The outward normal vectors to the faces of $\partial K_{i,j}$ 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 (x_i, y_j), $(x_{i\pm1}, y_j)$, $(x_i, y_{j\pm1})$ marked by "$\bullet$" in Fig. 3. Thus, by approximating the second integral on the left by mid-point quadrature, continuing to assume that $c \in C(\overline{\Omega})$, 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 2 *Finite difference schemes that arise from integral formulations of a differential equation, such as (**), are called finite volume methods.* $\diamond$

Clearly, $Tf_{i,j}$ is well defined for f in $L_2(\Omega)$ (in fact, $Tf_{i,j}$ is well defined even if $f \in L_1(\Omega)$ only); indeed,

$$\begin{aligned}
|Tf_{i,j}| &= \frac{1}{h^2}\left|\int_{K_{i,j}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y\right| \\
&\le \frac{1}{h^2}\left(\int_{K_{i,j}} 1^2\,\mathrm{d}x\,\mathrm{d}y\right)^{1/2}\left(\int_{K_{i,j}} |f(x,y)|^2\,\mathrm{d}x\,\mathrm{d}y\right)^{1/2} \\
&= \frac{1}{h}\|f\|_{L_2(K_{i,j})} \le \frac{1}{h}\|f\|_{L_2(\Omega)}.
\end{aligned} \tag{53}$$

Thus we define our finite difference (or, more precisely, finite volume) approximation of (37) by

$$\begin{aligned}
-(D_x^+D_x^-U_{i,j} + D_y^+D_y^-U_{i,j}) + c(x_i,y_j)U_{i,j} = Tf_{i,j}, &\qquad \text{for } (x_i,y_j)\in\Omega_h, \\
U = 0, &\qquad \text{on } \Gamma_h.
\end{aligned} \tag{54}$$

Since we have not changed the difference operator on the left-hand side, the argument presented on page 24 still applies, and therefore (54) has a unique solution, U.

Theorem 9 *The scheme (54) is stable in the sense that*

$$\|U\|_{1,h} \le \frac{1}{c_0}\|Tf\|_h \quad \left(\le \frac{1}{c_0}\|f\|_{L_2(\Omega)}\right). \tag{55}$$

PROOF. According to (48) and (53),

$$\begin{aligned}
c_0\|U\|_{1,h}^2 &\le (AU,U)_h = (Tf,U)_h \\
&\le \|Tf\|_h\|U\|_h \le \|Tf\|_h\|U\|_{1,h} \\
&\le \|f\|_{L_2(\Omega)}\|U\|_{1,h},
\end{aligned}$$

and hence (55). □

Having established the stability of the scheme (54), 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,

$$\begin{aligned}
Ae_{i,j} &= Au(x_i,y_j) - AU_{i,j} \\
&= Au(x_i,y_j) - Tf_{i,j} \\
&= -(D_x^+D_x^-u(x_i,y_j) + D_y^+D_y^-u(x_i,y_j)) + c(x_i,y_j)u(x_i,y_j) \\
&\quad + \left(T\left(\frac{\partial^2 u}{\partial x^2}\right)(x_i,y_j) + T\left(\frac{\partial^2 u}{\partial y^2}\right)(x_i,y_j) - T(cu)(x_i,y_j)\right).
\end{aligned} \tag{56}$$

By 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,

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

the equality (56) 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,

$$\begin{aligned}
Ae &= D_x^+\varphi_1 + D_y^+\varphi_2 + \psi \qquad && \text{in } \Omega_h,\\
e &= 0 && \text{on } \Gamma_h.
\end{aligned} \tag{57}$$

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

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

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 the inequality (48) and because $Ae = \varphi$, we have that

$$\begin{aligned}
c_0\|e\|_{1,h}^2 &\leq (Ae, e)_h\\
&= (D_x^+\varphi_1, e)_h + (D_y^+\varphi_2, e)_h + (\psi, e)_h.
\end{aligned} \tag{58}$$

Let us focus on the first two terms on the right-hand side of (58). Our plan is to use summations by parts to 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, we then have that

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

Thus,

$$(D_x^+\varphi_1, e)_h \le \|\varphi_1]|_x \|D_x^- e]|_x. \tag{59}$$

Similarly,

$$(D_y^+\varphi_2, e)_h \le \|\varphi_2]|_y \|D_y^- e]|_y \tag{60}$$

(see page 25 for the definitions of the mesh-dependent norms $\|\cdot]|_x$ and $\|\cdot]|_y$). By the Cauchy–Schwarz inequality we also have that

$$(\psi, e)_h \le \|\psi\|_h \|e\|_h. \tag{61}$$

By substituting the inequalities (59)–(61) into the inequality (58) we obtain

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

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

Lemma 7 *The global error, e, of the finite difference scheme (54) satisfies the inequality*

$$\|e\|_{1,h} \le \frac{1}{c_0}(\|\varphi_1]|_x^2 + \|\varphi_2]|_y^2 + \|\psi\|_h^2)^{1/2}, \tag{62}$$

where φ_1, φ_2, and ψ are defined by

$$\varphi_1(x_i, y_j) := \frac{1}{h}\int_{y_j-h/2}^{y_j+h/2} \frac{\partial u}{\partial x}(x_i - h/2, y)\,\mathrm{d}y - D_x^- u(x_i, y_j), \tag{63}$$

for $i = 1, \ldots, N$, $j = 1, \ldots, N-1$;

$$\varphi_2(x_i, y_j) := \frac{1}{h}\int_{x_i-h/2}^{x_i+h/2} \frac{\partial u}{\partial y}(x, y_j - h/2)\,\mathrm{d}x - D_y^- u(x_i, y_j), \tag{64}$$

for $i = 1, \ldots, N-1$, $j = 1, \ldots, N$; and

$$\psi(x_i, y_j) := (cu)(x_i, y_j) - \frac{1}{h^2}\int_{x_i-h/2}^{x_i+h/2}\int_{y_j-h/2}^{y_j+h/2} (cu)(x, y)\,\mathrm{d}x\,\mathrm{d}y, \tag{65}$$

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

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

$$|\varphi_1(x_i, y_j)| \le \frac{h^2}{24}\left(\left\|\frac{\partial^3 u}{\partial x \partial y^2}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^3 u}{\partial x^3}\right\|_{C(\overline{\Omega})}\right), \tag{66}$$

$$|\varphi_2(x_i, y_j)| \le \frac{h^2}{24}\left(\left\|\frac{\partial^3 u}{\partial x^2 \partial y}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^3 u}{\partial y^3}\right\|_{C(\overline{\Omega})}\right), \tag{67}$$

$$|\psi(x_i, y_j)| \le \frac{h^2}{24}\left(\left\|\frac{\partial^2 (cu)}{\partial x^2}\right\|_{C(\overline{\Omega})} + \left\|\frac{\partial^2 (cu)}{\partial y^2}\right\|_{C(\overline{\Omega})}\right), \tag{68}$$

and by using these to bound $\|\varphi_1]|_x$, $\|\varphi_2]|_y$ and $\|\psi]|_h$ on the right-hand side of the inequality (62) we arrive at the following theorem.

Theorem 10 *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, u, of the boundary-value problem (37) belongs to $C^3(\overline{\Omega})$; then,*

$$\|u - U\|_{1,h} \leq \frac{5}{96} h^2 M_3, \tag{69}$$

where

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

Proof. By recalling that $1/c_0 = 5/4$ and substituting the bounds (66)–(68) into the right-hand side of the inequality (62), the inequality (69) immediately follows. □

4.1.2 Convergence in the class of weak solutions that belong to $H^3(\Omega)$

Comparing (69) with (52), 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 (also known as the *fundamental theorem of calculus*):

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

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

$$\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 \bigg|_{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 \bigg|_{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,

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

However, since

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

it follows that

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

Consequently,

$$\|\varphi_1]|_x^2 \le \frac{h^4}{32} \left(\left\| \frac{\partial^3 u}{\partial x^3} \right\|_{L_2(\Omega)}^2 + \left\| \frac{\partial^3 u}{\partial x \partial y^2} \right\|_{L_2(\Omega)}^2 \right). \tag{70}$$

Analogously,

$$\|\varphi_2]|_y^2 \le \frac{h^4}{32} \left(\left\| \frac{\partial^3 u}{\partial y^3} \right\|_{L_2(\Omega)}^2 + \left\| \frac{\partial^3 u}{\partial x^2 \partial y} \right\|_{L_2(\Omega)}^2 \right). \tag{71}$$

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

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

Hence,

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

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

$$\|\psi\|_h^2 \leq \frac{3h^4}{64} \left(\left\| \frac{\partial^2 w}{\partial x^2} \right\|_{L_2(\Omega)}^2 + \left\| \frac{\partial^2 w}{\partial y^2} \right\|_{L_2(\Omega)}^2 + 4 \left\| \frac{\partial^2 w}{\partial x \partial y} \right\|_{L_2(\Omega)}^2 \right). \tag{72}$$

By substituting the bounds (70)–(72) into the right-hand side of the inequality (62), noting that $1/c_0 = 4/5$ and recalling the definition of the Sobolev norm $\|\cdot\|_{H^3(\Omega)}$, 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 (37) belongs to $H^3(\Omega)$; then,*

$$\|u - U\|_{1,h} \leq Ch^2 \|u\|_{H^3(\Omega)}, \tag{73}$$

where C is a positive constant (computable from (70)–(72)).

It can be shown that the error estimate (73) is best possible in the sense that further weakening of the regularity hypothesis on u leads to a loss of second-order convergence. Error estimates where the highest possible order of accuracy has been attained with the minimum hypotheses on the smoothness of the solution are called *optimal error estimates*. Thus, for example, (73) is an optimal error estimate for the finite difference scheme (54), but (69) is not.

We have used integral representations of differences to show the bounds (70)–(72). 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

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

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

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

Here $|\cdot|_{H^k(\Omega)}$ and $\|\cdot\|_{H^k(\Omega)}$ are the Sobolev semi-norm and Sobolev norm, defined in Section 1.3.

Proof. See P. Ciarlet: The Finite Element Method for Elliptic Problems, SIAM, Philadelphia, 2002. The digital version of the book is available from https://epubs.siam.org/doi/book/10.1137/1.9780898719208 for details. $\square$

We shall use the Bramble–Hilbert lemma to re-derive the bound (70) for φ_1. Let $K = [-1/2, 1/2] \times [-1/2, 1/2]$, and consider the affine mapping

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

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

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

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

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

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

$$|\Phi(\bar{u})| \le \int_{-1/2}^{1/2} \left|\frac{\partial \bar{u}}{\partial s}(0,t)\right| \mathrm{d}t + 2 \max_{(s,t)\in K} |\bar{u}(s,t)|\,. \tag{74}$$

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. We begin by proving the inequality stated in (a); we shall then prove the inequality under (b).

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

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

Thus,

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

Integrating both sides on the last inequality with respect to s and t we have that

$$\begin{aligned}
\int_{-1/2}^{1/2} \left|\frac{\partial v}{\partial s}(0,t)\right|\,\mathrm{d}t &\leq \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial v}{\partial s}(s,t)\right|\,\mathrm{d}s\,\mathrm{d}t + \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial^2 v}{\partial s^2}(\sigma,t)\right|\,\mathrm{d}\sigma\,\mathrm{d}t, \\
&\leq \left(\int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial v}{\partial s}(s,t)\right|^2\,\mathrm{d}s\,\mathrm{d}t\right)^{1/2} + \left(\int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial^2 v}{\partial s^2}(\sigma,t)\right|^2\,\mathrm{d}\sigma\,\mathrm{d}t\right)^{1/2} \\
&= \left\|\frac{\partial v}{\partial s}\right\|_{L_2(K)} + \left\|\frac{\partial^2 v}{\partial s^2}\right\|_{L_2(K)}.
\end{aligned}$$

Finally, using the inequality

$$a + b \leq \sqrt{2}(a^2+b^2)^{1/2}, \qquad a, b \geq 0,$$

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

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

Integrating both sides with respect to s and t it follows that

$$\begin{aligned}
|v(x,y)| &\leq \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} |v(s,t)|\,\mathrm{d}s\,\mathrm{d}t + \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial v}{\partial s}(\sigma,t)\right|\,\mathrm{d}\sigma\,\mathrm{d}t \\
&\quad + \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial v}{\partial t}(s,\tau)\right|\,\mathrm{d}s\,\mathrm{d}\tau + \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \left|\frac{\partial^2 v}{\partial s\,\partial t}(\sigma,\tau)\right|\,\mathrm{d}\sigma\,\mathrm{d}\tau \\
&\leq \|v\|_{L_2(K)} + \left\|\frac{\partial v}{\partial s}\right\|_{L_2(K)} + \left\|\frac{\partial v}{\partial t}\right\|_{L_2(K)} + \left\|\frac{\partial^2 v}{\partial s\,\partial t}\right\|_{L_2(K)} \\
&\leq 2\|v\|_{H^2(K)} \quad \forall (x,y) \in K.
\end{aligned}$$

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

Equipped with the inequalities (a) and (b) we now return to the inequality (74). It follows that

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

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

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

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

Hence, there exists a constant C_1 such that

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

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

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

Therefore,

$$\|\varphi_1]|_x \leq C_1 h^2 \, |u|_{H^3(\Omega)} \, . \tag{75}$$

Similarly,

$$\|\varphi_2]|_y \leq C_2 h^2 \, |u|_{H^3(\Omega)} \tag{76}$$

and

$$\|\psi\|_h \leq C_3 h^2 \, |u|_{H^2(\Omega)} \, . \tag{77}$$

The bounds (75)–(77) derived by using the Bramble–Hilbert lemma are essentially the same as those obtained earlier by integral representations, and stated in (70)–(72). There is, however, an important practical difference: while the constants involved in (70)–(72) are known, those which appear in (75)–(77) (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 this partial differential equation by

$$\begin{aligned}
&-\frac{1}{h}\left[a_1(x_{i+1/2}, y_j)\frac{U_{i+1,j} - U_{i,j}}{h} - a_1(x_{i-1/2}, y_j)\frac{U_{i,j} - U_{i-1,j}}{h}\right] \\
&-\frac{1}{h}\left[a_2(x_i, y_{j+1/2})\frac{U_{i,j+1} - U_{i,j}}{h} - a_2(x_i, y_{j-1/2})\frac{U_{i,j} - U_{i,j-1}}{h}\right] \\
&+ b_1(x_i, y_j)\frac{U_{i+1,j} - U_{i-1,j}}{2h} + b_2(x_i, y_j)\frac{U_{i,j+1} - U_{i,j-1}}{2h} \\
&+ c(x_i, y_j)U_{i,j} = \frac{1}{h^2}\int_{x_{i-1/2}}^{x_{i+1/2}}\int_{y_{j-1/2}}^{y_{j+1/2}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y.
\end{aligned}$$

This is still a five point difference scheme that is second order consistent.

When Ω has a curved boundary, a nonuniform 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

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

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

$$\bar{k}_j := \frac{1}{2}(k_{j+1} + k_j).$$

Let

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

Note that, whereas on a uniform mesh $D_x^- U_{i+1} = D_x^+ U_i$ and $D_y^- U_{j+1} = D_y^+ U_j$, on nonuniform meshes this is no longer the case. For the same reason, on a nonuniform mesh $D_x^+ D_x^- U_i \neq D_x^- D_x^+ U_i$ and $D_y^+ D_y^- U_j \neq D_y^- D_y^+ U_j$.

On a general nonuniform mesh

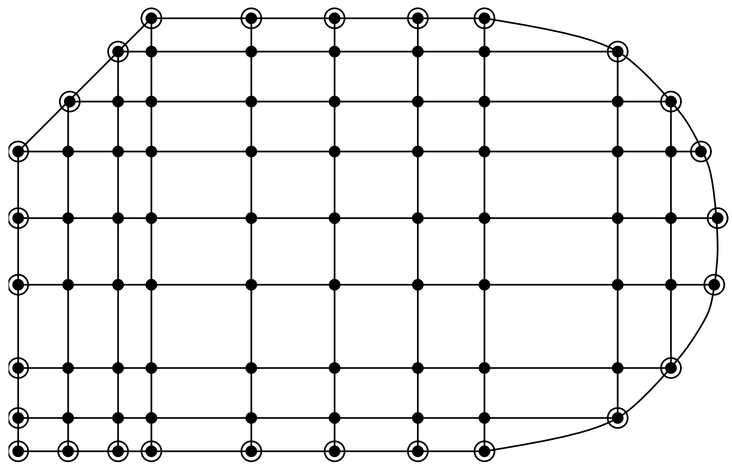
$$\overline{\Omega}_h := \{(x_i, y_j) \in \overline{\Omega} \,:\, x_{i+1} - x_i = h_{i+1},\ y_{j+1} - y_j = k_{j+1}\},$$

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

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

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



$\bullet\ \Omega_h$; $\odot\ \Gamma_h$, $\overline{\Omega}_h = \Omega_h \cup \Gamma_h$.

Figure 4: Nonuniform mesh $\overline{\Omega}_h$.

The finite difference approximation of this boundary-value problem is

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

Equivalently,

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

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

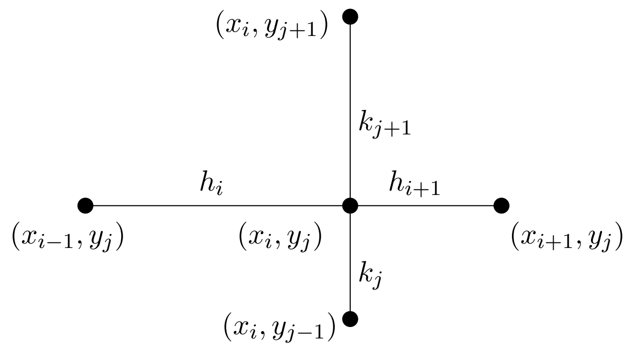


Figure 5: Five-point stencil on a nonuniform mesh.

4.3 The discrete maximum principle

The maximum principle is a key property of elliptic equations. Under suitable sign-conditions imposed 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 is 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

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

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

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

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

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

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

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

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

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

$$u(x) \geq \min(A, B), \qquad a \leq x \leq b.$$

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,

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

and

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

Hence,

$$-\Delta u(x_0) = -\sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2}(x_0) \ge 0,$$

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 $f(x) \le 0$ for all $x \in \Omega$. To this end, we consider the auxiliary function $v \in C^2(\Omega) \cap C(\overline{\Omega})$ defined by

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

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,

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

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

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

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

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

As $\partial\Omega \subset \overline{\Omega}$, trivially $\max_{x \in \overline{\Omega}} u(x) \ge \max_{x \in \partial\Omega} u(x)$. Therefore, these two inequalities yield that

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

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 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 nonuniform mesh

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

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 as in Section 4.2. 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

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

Equivalently,

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

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,

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

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

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

Note, however, that the expressions on the two sides of this inequality 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:

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

Now suppose that $f(x_i, y_j) \leq 0$ 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,

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

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 deduce from these two inequalities that if $f(x_i, y_j) \leq 0$ for all $(x_i, y_j) \in \Omega_h$, then the *discrete maximum principle* holds:

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

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/minimum principle we have established to prove the stability of the finite difference scheme (78) with respect to perturbations in the boundary data.

4.4 Stability in the discrete maximum norm

Consider the finite difference scheme (78) on the nonuniform mesh formulated in Section 4.2. Our first result asserts the existence of a solution to (78) as well as its uniqueness.

Lemma 10 *The finite difference scheme* (78) *has a unique solution.*

Proof. We note that (78) 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 (78) is therefore equivalent to showing that this system of linear algebraic equations has a unique solution, which amounts to proving that the matrix of the linear system is invertible. The matrix of the linear system associated with (78) 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 (78) 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

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) &= 0 \qquad && \text{in } \Omega_h, \\ U_{i,j} &= 0 \qquad && \text{on } \Gamma_h. \end{aligned} \tag{79}$$

The existence of a solution to (79) 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 (79),

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

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 (78). □

We are now ready to embark on the analysis of the stability of the scheme (78) with respect to perturbations in the boundary data.

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

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j}^{(1)} + D_y^+ D_y^- U_{i,j}^{(1)}) &= f(x_i, y_j) \qquad && \text{in } \Omega_h, \\ U_{i,j}^{(1)} &= g^{(1)}(x_i, y_j) \qquad && \text{on } \Gamma_h \end{aligned} \tag{80}$$

and

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j}^{(2)} + D_y^+ D_y^- U_{i,j}^{(2)}) &= f(x_i, y_j) \qquad && \text{in } \Omega_h, \\ U_{i,j}^{(2)} &= g^{(2)}(x_i, y_j) \qquad && \text{on } \Gamma_h \end{aligned} \tag{81}$$

for given boundary data $g^{(1)}$ and $g^{(2)}$. Let $U := U^{(1)} - U^{(2)}$ and $g := g^{(1)} - g^{(2)}$. Then, by subtracting (81) from (80) we find that U solves

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) &= 0 \qquad && \text{in } \Omega_h, \\ U_{i,j} &= g(x_i, y_j) \qquad && \text{on } \Gamma_h. \end{aligned} \tag{82}$$

By the discrete maximum principle we have from (82) that

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

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

$$U_{i,j} \leq \max_{(x_i,y_j)\in\Gamma_h} |g(x_i, y_j)|. \tag{83}$$

It follows from (82) that $-U$ solves

$$\begin{aligned} -(D_x^+ D_x^- (-U)_{i,j} + D_y^+ D_y^- (-U)_{i,j}) &= 0 \qquad && \text{in } \Omega_h, \\ (-U)_{i,j} &= -g(x_i, y_j) \qquad && \text{on } \Gamma_h, \end{aligned} \tag{84}$$

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

$$-U_{i,j} = (-U)_{i,j} \leq \max_{(x_i,y_j)\in\Gamma_h} |-g(x_i,y_j)| = \max_{(x_i,y_j)\in\Gamma_h} |g(x_i,y_j)| \tag{85}$$

for all $(x_i, y_j) \in \overline{\Omega}_h$. By combining (83) and (85) 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

$$\max_{(x_i,y_j)\in\overline{\Omega}_h} |U^{(1)}_{i,j} - U^{(2)}_{i,j}| \leq \max_{(x_i,y_j)\in\Gamma_h} |g^{(1)}(x_i,y_j) - g^{(2)}(x_i,y_j)|. \tag{86}$$

The inequality (86) 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* (in the discrete maximum norm, in this case).

4.5 Iterative solution of linear systems: linear stationary iterative methods

Lecture 7

Before embarking on our discussion of the main topic of this section, we require a few technical tools. 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 $c \geq 0$ is a real number. A nontrivial solution $u(x) \not\equiv 0$ of this boundary-value problem 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 $\{x_i := ih : i = 0, \ldots, N\}$ of uniform spacing $h := 1/N$, with $N \geq 2$, is given by

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

Again, we seek nontrivial solutions, and a simple calculation yields that $U_i := U^k(x_i)$, 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

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

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 becomes

$$\begin{bmatrix} \frac{2}{h^2}+c & -\frac{1}{h^2} & & & 0 \\ -\frac{1}{h^2} & \frac{2}{h^2}+c & -\frac{1}{h^2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{h^2} & \frac{2}{h^2}+c & -\frac{1}{h^2} \\ 0 & & & -\frac{1}{h^2} & \frac{2}{h^2}+c \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-2} \\ U_{N-1} \end{bmatrix} = \Lambda \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-2} \\ U_{N-1} \end{bmatrix},$$

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

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

and the corresponding eigenvectors are, respectively, $(U^k(x_1), \ldots, U^k(x_{N-1}))^{\mathrm{T}}$, $k = 1, \ldots, N-1$.

Clearly, $c+8 \le \Lambda_k \le c + \frac{4}{h^2}$ for all $k = 1, 2, \ldots, N-1$. The first of these inequalities follows by noting that $\Lambda_k \ge \Lambda_1$ for $k = 1, \ldots, N-1$ and $\sin x \ge \frac{2\sqrt{2}}{\pi}x$ for $x \in [0, \frac{\pi}{4}]$ (recall that $h \in [0, \frac{1}{2}]$ because $N \ge 2$, whereby $0 < \frac{\pi h}{2} \le \frac{\pi}{4}$); the second inequality is the consequence of $0 \le \sin^2 x \le 1$ 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*

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

where $c \ge 0$ 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 finite difference mesh $\{(x_i, y_j) := (ih, jh) : i, j = 0, \ldots, N\}$ of spacing $h = 1/N$, $N \ge 2$, in the x and y directions, is

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

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

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

with $c + 16 \le \Lambda_{k,m} \le c + \frac{8}{h^2}$, 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:

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

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

$$\begin{aligned} -\frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} + c\,U_i = f(x_i), \quad i = 1,\dots,N-1, \\ U_0 = 0, \quad U_N = 0. \end{aligned} \tag{87}$$

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

$$AU = F \tag{88}$$

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

Similarly, if one considers the elliptic boundary-value problem

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

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

$$\begin{aligned} -\frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} - \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{h^2} + c\,U_{i,j} = f(x_i,y_j), \qquad & i,j = 1,\dots,N-1, \\ U_{i,j} = 0 \qquad & \text{for } (x_i,y_j) \in \Gamma_h, \end{aligned} \tag{89}$$

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, $\Lambda_{k,m}$, $k,m = 1,\dots,N-1$, given in Example 1 above.

Motivated by these examples, we shall be interested in developing a simple iterative method for the approximate 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 $AU = F$.

$$U^{(j+1)} := U^{(j)} - \tau(AU^{(j)} - F), \qquad j = 0,1,\dots, \tag{90}$$

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 (90) from this equality we find that

$$U - U^{(j+1)} = U - U^{(j)} - \tau A(U - U^{(j)}) = (I - \tau A)(U - U^{(j)}), \qquad j = 0,1,\dots, \tag{91}$$

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

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

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

$$\|B\| := \sup_{V\in\mathbb{R}^M\setminus\{0\}} \frac{\|BV\|}{\|V\|}.$$

Thanks to this definition, $\|BV\| \leq \|B\|\|V\|$ for all $V \in \mathbb{R}^M$, and hence, by induction $\|B^jV\| \leq \|B\|^j\|V\|$ for all $j = 1, 2\dots$ and all $V \in \mathbb{R}^M$. Therefore, with $B := I - \tau A$ and $V := U - U^{(0)}$, we have that

$$\|U - U^{(j)}\| = \|(I - \tau A)^j(U - U^{(0)})\| \leq \|I - \tau A\|^j\|U - U^{(0)}\|. \tag{92}$$

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.[6] 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:

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

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, $i = 1, \dots, M$, the corresponding eigenvalues, for any vector $V = \alpha_1 e_1 + \cdots + \alpha_M e_M$, expanded in terms of the eigenvectors of B, thanks to orthonormality the Euclidean norms of V and BV can be expressed, respectively, as follows:

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

Clearly, $\|BV\| \leq \max_{i=1,\dots,M} |\lambda_i|\,\|V\|$ 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, $\|B\| = \max_{i=1,\dots,M} |\lambda_i|$, where now $\|\cdot\|$ is the matrix norm induced by the Euclidean norm.

We are now ready to return to (92) to find that $\|I - \tau A\|$ appearing on the right-hand side of (92), where again $\|\cdot\|$ denotes the matrix norm induced by the Euclidean norm, is equal to the largest in absolute value eigenvalue 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 (90) converges as fast as possible, and to this end we see from (92) 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 nonnegative piecewise linear functions $\tau \mapsto |1 - \tau\beta|$ and $\tau \mapsto |1 - \tau\alpha|$ for $\tau \in [0, \infty)$, we see that they vanish at $\tau = 1/\beta$ and $\tau = 1/\alpha$, respectively; their graphs intersect at $\tau = 0$ and at $\tau = \frac{2}{\alpha+\beta}$. As $0 < \alpha < \beta$, clearly $0 < 1/\beta < 1/\alpha$. Next, by plotting the continuous piecewise linear

[6] 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

function $\tau \mapsto \max\{|1-\tau\beta|, |1-\tau\alpha|\}$ for $\tau \in [0,\infty)$, we observe that it attains its minimum at $\tau = \frac{2}{\alpha+\beta}$ where $1-\tau\beta = \tau\alpha - 1$. 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 approximate solution of the linear system $AU = F$ is the one stated in (90), with $\tau := \frac{2}{\beta+\alpha}$, 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* (87), $\alpha = c + 8$ *and* $\beta = c + \frac{4}{h^2}$, *while in the case of* (89), $\alpha = c + 16$ *and* $\beta = c + \frac{8}{h^2}$. *In both cases*

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

thus, while the sequence of iterates $\{U^{(j)}\}_{j=0}^{\infty}$ *defined by the iterative method* (90) *is guaranteed to converge to the exact solution* U *of the linear system* $AU = F$, *the right-hand side in the inequality*

$$\|U - U^{(j)}\| \le \left(\frac{\beta-\alpha}{\beta+\alpha}\right)^j \|U - U^{(0)}\| \tag{93}$$

will gradually deteriorate as $h \to 0$. *By 'deteriorate' we mean that the smaller the value of the mesh-size* h, *the closer the fraction* $(\beta-\alpha)/(\beta+\alpha) \in (0,1)$ *will be to* 1, *and therefore the slower the convergence of the sequence appearing on the right-hand side of the inequality* (93) *will be to* 0 *as* $j \to \infty$. *The deceleration of the convergence to* 0 *of the right-hand side of the inequality* (93) *as* $j \to \infty$ *with decreasing* h *does not, of course, automatically imply a corresponding deceleration of the convergence of* $\|U - U^{(j)}\|$ *to* 0 *as* $j \to \infty$ *with decreasing* h, *as the right-hand side of the inequality* (93) *is merely an upper bound on the left-hand side. That this is however the case can be verified by numerical experiments, which indicate that the smaller the mesh-size* h *the slower the convergence of* $\|U - U^{(j)}\|$ *to* 0 *will be as* $j \to \infty$.

We note that by multiplying (91) by the matrix A and recalling that $AU = F$, one has that

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

and therefore, by proceeding as above,

$$\|F - AU^{(j)}\| \le \|I - \tau A\|^j \|F - AU^{(0)}\| \le \left(\frac{\beta-\alpha}{\beta+\alpha}\right)^j \|F - AU^{(0)}\|. \tag{94}$$

As α and β are available (in the case of the simple boundary-value problems considered here, at least) as are F, A and the initial guess $U^{(0)}$, 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 becomes smaller than a chosen tolerance $\texttt{TOL} > 0$: a sufficient condition for this is that the right-hand side of (94) is smaller than $\texttt{TOL}$, which will hold as soon as

$$j > \log \frac{\|F - AU^{(0)}\|}{\texttt{TOL}} \left[\log\left(\frac{\beta+\alpha}{\beta-\alpha}\right)\right]^{-1}. \tag{95}$$

In the case of the two boundary-value problems considered above,

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

and therefore (because $\log(1 - \text{Const.}h^2) \sim -\text{Const.}h^2$ as $h \to 0$) the right-hand side of the inequality (95) is $\sim \text{Const.}\, h^{-2} \log(1/\texttt{TOL})$. We see in particular that the smaller the value of the mesh-size h the larger the number of iterations j will need to be to ensure that $\|F - AU^{(j)}\| < \texttt{TOL}$.

5 Finite difference approximation of parabolic equations

This penultimate section of the 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**

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \tag{96}$$

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 explicit expression for the analytical solution of the initial-value problem (96).

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 formal manipulations make sense.

By Fourier-transforming the partial differential equation (96) 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,

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} \mathrm{e}^{-(x-y)^2/(4t)} u_0(y)\,\mathrm{d}y. \tag{97}$$

This formula gives an explicit expression for the solution of the heat equation (96) 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 (97) that if u_0 is a bounded continuous function, then

$$\sup_{x\in(-\infty,+\infty)} |u(x,t)| \le \sup_{x\in(-\infty,\infty)} |u_0(x)|, \qquad t > 0. \tag{98}$$

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 stated below, 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. $\square$

Returning to the equation (96), 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

$$\|u(\cdot,t)\|_{L_2((-\infty,\infty))} \le \|u_0\|_{L_2((-\infty,\infty))} \qquad \text{for all } t>0. \tag{99}$$

This is a useful result as it can be used to deduce stability of the solution of the equation (96) 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 (96) 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 (99), we have that

$$\|u(\cdot,t)-\tilde{u}(\cdot,t)\|_{L_2((-\infty,\infty))} \le \|u_0-\tilde{u}_0\|_{L_2((-\infty,\infty))} \qquad \text{for all } t>0. \tag{100}$$

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

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 $M\ge 1$, 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 (96) at the point (x_j,t_m) by the following numerical method, called the *explicit Euler scheme*:

$$\frac{U_j^{m+1}-U_j^m}{\Delta t} = \frac{U_{j+1}^m-2U_j^m+U_{j-1}^m}{(\Delta x)^2}, \quad j=0,\pm1,\pm2,\dots, \quad m=0,\dots,M-1,$$

$$U_j^0 := u_0(x_j), \qquad j=0,\pm1,\pm2,\dots.$$

Equivalently, we can write this as

$$U_j^{m+1} = U_j^m + \mu(U_{j+1}^m-2U_j^m+U_{j-1}^m), \quad j=0,\pm1,\pm2,\dots, \quad m=0,\dots,M-1,$$

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

where $\mu = \frac{\Delta t}{(\Delta x)^2}$. Thus, U_j^{m+1} can be explicitly calculated, for all $j = 0, \pm 1, \pm 2, \dots$, 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*:

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

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

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:

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = (1-\theta)\,\frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2} + \theta\,\frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2}, \quad \begin{cases} j = 0, \pm 1, \pm 2, \dots, \\ m = 0, \dots, M-1, \end{cases}$$

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

where $\theta \in [0, 1]$ is a parameter. For $\theta = 0$ the θ-scheme 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*.

Numerical methods of this kind are called *fully-discrete approximations*. An alternative approach is to approximate the spatial partial derivative only in the heat equation, resulting in the following initial-value problem for a system of ordinary differential equations:

$$\frac{\mathrm{d}U_j(t)}{\mathrm{d}t} = \frac{U_{j+1}(t) - 2U_j(t) + U_{j-1}(t)}{(\Delta x)^2}, \qquad j = 0, \pm 1, \pm 2, \dots,$$

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

This is called a spatially semi-discrete approximation, because no discretization with respect to the temporal variable t has taken place. Because an initial-value problem for the heat equation is considered for $x \in (-\infty, \infty)$, the spatially semidiscrete approximation consists of an infinite system of ordinary differential equations. Had the range of x been limited to a bounded interval (a, b) of the real line instead, and had, in conjunction with the initial condition, boundary conditions been supplied at $x = a$ and $x = b$, spatial semi-discretization of such an initial-boundary-value problem for the heat equation would have resulted in a system consisting of a finite number of ordinary differential equations, coupled to algebraic equation that stem from the spatial discretization of the boundary conditions. Such a system of differential-algebraic equations (DAEs) could then have been solved approximately by any standard method for the numerical solution of DAEs (such as, for example, the Matlab solvers `ode15s` and `ode23t`). Because no discretization in time was performed in the first place, this approach is usually referred to as the *method of lines*.

5.1.1 Accuracy of the θ-method

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

$$T_j^m := \frac{u_j^{m+1} - u_j^m}{\Delta t} - (1-\theta)\,\frac{u_{j+1}^m - 2u_j^m + u_{j-1}^m}{(\Delta x)^2} - \theta\,\frac{u_{j+1}^{m+1} - 2u_j^{m+1} + u_{j-1}^{m+1}}{(\Delta x)^2}, \quad \begin{cases} j = 0, \pm 1, \pm 2, \dots, \\ m = 0, \dots, M-1, \end{cases}$$

where

$$u_j^m := u(x_j, t_m).$$

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

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

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,

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

Combining these, we deduce that

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

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

In order to be able to replicate the stability property (99) 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 **Lecture 9**

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

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, defined below, to explore the stability of the finite difference schemes under consideration. In order to avoid complicating the discussion with the inclusion of technical details that concern the convergence of various infinite sums, we shall simply assume throughout that all infinite sums considered converge.

Definition 2 *The semidiscrete Fourier transform of a function U defined on the infinite mesh with mesh-points $x_j = j\Delta x$, $j = 0, \pm 1, \pm 2, \ldots$, 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 as 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 so is $\|\hat{U}\|_{L_2}$, 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 the proof to the reader as an exercise.

With all technical prerequisites in place, we are now ready to discuss the stability of the various finite difference schemes under consideration. We begin by exploring the practical stability of the explicit and implicit Euler schemes. We shall prove in particular that the explicit Euler scheme is conditionally practically stable (the condition required for stability being that $\mu := \Delta t/(\Delta x)^2 \leq 1$), while the implicit Euler scheme will be shown to be unconditionally practically stable.

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 for the heat equation (96). By inserting

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

into the explicit Euler scheme we deduce that

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

Therefore, we have that

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

By comparing the left-hand side with the right-hand side we deduce that the two integrands are identically equal,[7] and therefore

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

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 Lewy, who first performed an analysis of this kind).[8] By the discrete Parseval identity stated in Lemma 12 we have that

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

In order to mimic the bound (99) we would like to ensure that

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

Thus we demand that

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

[7] This is a consequence of the fact that the semidiscrete Fourier transform and its inverse are injective (one-to-one) mappings.

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

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*

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

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

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

$$\|U^m\|_{\ell_2} \leq \|U^0\|_{\ell_2}, \qquad m = 1, 2, \dots, M. \tag{101}$$

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 (101) 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 (101) 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 (96), which is defined as follows:

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

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

Equivalently,

$$U_j^{m+1} - \mu(U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}) = U_j^m, \quad j = 0, \pm 1, \pm 2, \dots, \quad m = 0, \dots, M-1,$$

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

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,

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

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*

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

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

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

$$\|U^m\|_{\ell_2} \le \|U^0\|_{\ell_2}, \qquad m = 1, 2, \ldots, M. \tag{102}$$

In other words, the implicit Euler scheme is *unconditionally practically stable*, meaning that the bound (102) 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. The one below, due to John von Neumann, is called *von Neumann stability.* **Lecture 10**

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*

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

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 $0 \le m \le T/\Delta t$, 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 $\{U_j^m\}_{j=-\infty}^{\infty}$, $m = 0, 1, \ldots, \frac{T}{\Delta t}$, of a finite difference scheme for the heat equation satisfies*

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

and there exists a nonnegative constant C_0 such that

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

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

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

Hence,

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

Therefore,

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

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

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

meaning that von Neumann stability holds, with stability constant $C = \mathrm{e}^{C_0 T}$. In particular if $C_0 = 0$, then $C = 1$, and practical stability follows. ☐

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:

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} = (1-\theta)\,\frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2} + \theta\,\frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2}, \qquad \begin{cases} j = 0, \pm1, \pm2, \dots, \\ m = 0, \dots, M-1, \end{cases}$$

$$U_j^0 := u_0(x_j), \qquad j = 0, \pm1, \pm2, \dots,$$

where $\theta \in [0, 1]$ is a parameter. As we have noted in Section 5.1, for $\theta = 0$ the θ-scheme 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 then called the *Crank–Nicolson scheme*.

To analyze 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 the amplification factor of the scheme is

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

For practical stability, we demand that

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

which holds if, and only if,

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

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 at 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) of $\mathbb{R}$. 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 $x = b$:

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

It will be assumed throughout this section that the boundary conditions are compatible with the initial condition in the sense that $A(0) = u_0(a)$ and $B(0) = u_0(b)$.

Remark 3 *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

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

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 $j = 1,\dots,J-1$, $m = 0,\dots,M-1$,

$$U_j^0 := u_0(x_j), \qquad j = 0,\dots,J,$$

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

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

$$\begin{aligned}
[1-\theta\mu\delta^2]U_j^{m+1} &= [1+(1-\theta)\mu\delta^2]U_j^m, \quad j = 1,\dots,J-1, \quad m = 0,\dots,M-1,\\
U_j^0 &:= u_0(x_j), \qquad j = 0,\dots,J,\\
U_0^{m+1} &:= A(t_{m+1}), \quad U_J^{m+1} := B(t_{m+1}), \qquad m = 0,\dots,M-1,
\end{aligned}$$

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:

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

Let $\mathcal{I}$ be the $(J-1)\times(J-1)$ identity matrix $\mathcal{I} := \mathrm{diag}(1,\,1,\,1,\dots,\,1,\,1)$. 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 $m = 0,1,\dots,M-1$, where

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

and

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

Thus, for each $m = 0,\dots,M-1$, we are required to solve a system of linear algebraic equations with (the same) tridiagonal matrix $\mathcal{I} - \theta\mu\mathcal{A}$ in order to compute $\mathbf{U}^{m+1}$ from $\mathbf{U}^m$.

Matlab code for the Crank–Nicolson scheme

```
% cn.m - Crank--Nicolson scheme for the heat equation.
% Save this file as cn.m
% Run the program 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 for the θ-scheme in the discrete maximum norm, analogous to (98) satisfied by the solution of the heat equation. Recall that the CFL number is defined by $\mu := \Delta t/(\Delta x)^2$. **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 $0 \leq \theta \leq 1$ and $\mu(1-\theta) \leq \frac{1}{2}$, yields a sequence of numerical approximations $\{U_j^m\}_{j=0,\dots,J;\ m=0,\dots,M}$ satisfying

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

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

$$(1+2\theta\mu)\,U_j^{m+1} = \theta\mu\left(U_{j+1}^{m+1} + U_{j-1}^{m+1}\right) + (1-\theta)\mu\left(U_{j+1}^m + U_{j-1}^m\right) + [1-2(1-\theta)\mu]\,U_j^m, \qquad (103)$$

and recall that, by hypothesis,

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

Suppose that U attains its maximum value at an interior 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,

$$(1+2\theta\mu)\,U_j^{m+1} \le 2\theta\mu U^\star + 2(1-\theta)\mu U^\star + [1-2(1-\theta)\mu]U^\star = (1+2\theta\mu)\,U^\star, \tag{104}$$

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.[9]

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

$$\mu(1-\theta) \le \frac{1}{2} \quad \text{with } \theta \in [0,1]$$

the θ-scheme satisfies the discrete maximum principle. Clearly, this condition is 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 ($\theta = 1/2$) is unconditionally stable in the ℓ_2 norm, yet it only satisfies the discrete maximum principle when $\mu = \frac{\Delta t}{(\Delta x)^2} \le 1$. More generally, for $\theta \in [\frac{1}{2}, 1]$ the θ-scheme is unconditionally stable in the ℓ_2 norm, but it will only satisfy the discrete maximum principle unconditionally when $\theta = 1$ (implicit Euler scheme); for $\theta \in [\frac{1}{2}, 1)$ the validity of the discrete maximum principle is only guaranteed when $\mu(1-\theta) \le \frac{1}{2}$. Concerning the values of $\theta \in [0, \frac{1}{2}]$, except for $\theta = 0$ when the conditions for the validity of the discrete maximum principle and discrete ℓ_2-stability coincide (both require that $\mu \le \frac{1}{2}$), for $\theta \in (0, \frac{1}{2}]$ the inequality $\mu(1-\theta) \le \frac{1}{2}$ is more restrictive than $\mu(1-2\theta) \le \frac{1}{2}$ because, for such θ, $1-\theta > 1-2\theta$.

[9] 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 (103); (b) if $\theta = 1$ then U_{j+1}^m and U_{j-1}^m are absent from the right-hand side of (103); (c) if $2(1-\theta)\mu = 1$, then U_j^m is absent from the right-hand side of (103), 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 (103). 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 (103) to (104), we would deduce (104) from (103), but now with the $\le$ symbol in (104) 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}).

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 (96) 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:

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

for $j = 1, \dots, J-1$ and $m = 0, \dots, M-1$. The scheme is considered subject to the initial condition

$$U_j^0 := u_0(x_j), \qquad j = 0, \dots, J,$$

and the boundary conditions

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

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 $u_j^m := u(x_j, t_m)$, and therefore

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

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

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

and

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

We define,

$$E^m := \max_{0 \le j \le J} |e_j^m| \quad \text{and} \quad T^m := \max_{1 \le j \le J-1} |T_j^m|.$$

As, by hypothesis,

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

we have that

$$(1+2\theta\mu) E^{m+1} \le 2\theta\mu E^{m+1} + E^m + \Delta t T^m, \qquad m = 0, \dots, M-1.$$

Hence,

$$E^{m+1} \le E^m + \Delta t\, T^m, \qquad m = 0, \dots, M-1.$$

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

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

which then implies that

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

Recall that, assuming that u is sufficiently smooth, the consistency error of the θ-scheme is

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

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:

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

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

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

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 coordinate 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 domain, 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

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

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

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

$$U_{i,j}^0 := u_0(x_i, y_j), \quad i = 0, \dots, J_x, \;\; j = 0, \dots, J_y,$$

and the boundary condition

$$U_{i,j}^{m+1} := B(x_i, y_j, t_{m+1}), \;\text{ at the boundary mesh-points, for } m = 0, \dots, M-1.$$

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 the following expression for the amplification factor $\lambda = \lambda(k_x, k_y)$:

$$\lambda - 1 = -4(1-\theta)\left[\mu_x \sin^2\left(\frac{k_x \Delta x}{2}\right) + \mu_y \sin^2\left(\frac{k_y \Delta y}{2}\right)\right] - 4\theta\lambda\left[\mu_x \sin^2\left(\frac{k_x \Delta x}{2}\right) + \mu_y \sin^2\left(\frac{k_y \Delta y}{2}\right)\right],$$

where

$$\mu_x := \frac{\Delta t}{(\Delta x)^2}, \qquad \mu_y := \frac{\Delta t}{(\Delta y)^2}.$$

Hence,

$$\lambda = \lambda(k_x, k_y) = \frac{1 - 4(1-\theta)\left[\mu_x \sin^2\left(\frac{k_x \Delta x}{2}\right) + \mu_y \sin^2\left(\frac{k_y \Delta y}{2}\right)\right]}{1 + 4\theta\left[\mu_x \sin^2\left(\frac{k_x \Delta x}{2}\right) + \mu_y \sin^2\left(\frac{k_y \Delta y}{2}\right)\right]}.$$

For practical stability in the ℓ_2 norm, we require that

$$|\lambda(k_x, k_y)| \le 1 \qquad \forall\, (k_x, k_y) \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right] \times \left[-\frac{\pi}{\Delta y}, \frac{\pi}{\Delta y}\right].$$

Thus, we demand that

$$-1 \le \frac{1 - 4(1-\theta)\,[\mu_x + \mu_y]}{1 + 4\theta\,[\mu_x + \mu_y]} \le 1,$$

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 practically stable, while the explicit Euler scheme ($\theta = 0$) is only conditionally practically 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 $i=1,\dots,J_x-1$, $j=1,\dots,J_y-1$, $m=0,1,\dots,M-1$, subject to the initial condition

$$U_{i,j}^{0} := u_0(x_i,y_j), \quad i=0,\dots,J_x, \;\; j=0,\dots,J_y,$$

and the boundary condition

$$U_{i,j}^{m} := B(x_i,y_j,t_m), \;\text{ at the boundary mesh-points, for } m=1,\dots,M.$$

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 practical stability of the scheme in the ℓ_2 norm, 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 practically 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 by returning to the θ-scheme for the initial-boundary-value problem, and discussing 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 $i = 1, \dots, J_x - 1$, $j = 1, \dots, J_y - 1$, $m = 0, 1, \dots, M - 1$, subject to the initial condition

$$U_{i,j}^0 := u_0(x_i, y_j), \quad i = 0, \dots, J_x, \;\; j = 0, \dots, J_y,$$

and the boundary condition

$$U_{i,j}^m := B(x_i, y_j, t_m), \;\; \text{at the boundary mesh-points, for } m = 1, \dots, M.$$

Suppose further that

$$(\mu_x + \mu_y)(1 - \theta) \leq \frac{1}{2}, \qquad \theta \in [0, 1].$$

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

$$u_{i,j}^m := u(x_i, y_j, t_m).$$

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

$$\begin{aligned}(1 + 2\theta(\mu_x + \mu_y))u_{i,j}^{m+1} &= (1 - 2(1 - \theta)(\mu_x + \mu_y))u_{i,j}^m \\ &\quad + (1 - \theta)\mu_x(u_{i+1,j}^m + u_{i-1,j}^m) + (1 - \theta)\mu_y(u_{i,j+1}^m + u_{i,j-1}^m) \\ &\quad + \theta\mu_x(u_{i+1,j}^{m+1} + u_{i-1,j}^{m+1}) + \theta\mu_y(u_{i,j+1}^{m+1} + u_{i,j-1}^{m+1}) \\ &\quad + \Delta t\, T_{i,j}^m,\end{aligned}$$

for $i = 1, \dots, J_x - 1$, $j = 1, \dots, J_y - 1$, $m = 0, 1, \dots, M - 1$. 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

$$\begin{aligned}(1 + 2\theta(\mu_x + \mu_y))e_{i,j}^{m+1} &= (1 - 2(1 - \theta)(\mu_x + \mu_y))e_{i,j}^m \\ &\quad + (1 - \theta)\mu_x(e_{i+1,j}^m + e_{i-1,j}^m) + (1 - \theta)\mu_y(e_{i,j+1}^m + e_{i,j-1}^m) \\ &\quad + \theta\mu_x(e_{i+1,j}^{m+1} + e_{i-1,j}^{m+1}) + \theta\mu_y(e_{i,j+1}^{m+1} + e_{i,j-1}^{m+1}) \\ &\quad + \Delta t\, T_{i,j}^m \quad \text{for } i = 1, \dots, J_x - 1 \text{ and } j = 1, \dots, J_y - 1.\end{aligned}$$

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 that

$$(1 + 2\theta(\mu_x + \mu_y))E^{m+1} \leq 2\theta(\mu_x + \mu_y)E^{m+1} + E^m + \Delta t T^m, \quad m = 0, \dots, M - 1.$$

Hence,

$$E^{m+1} \leq E^m + \Delta t\, T^m, \qquad m = 0, 1, \dots, M - 1.$$

As $E^0 = 0$, upon summation we deduce that

$$\begin{aligned}E^m &\le \Delta t \sum_{n=0}^{m-1} T^n \\ &\le m\Delta t \max_{0\le n\le m-1} T^n \\ &\le T \max_{0\le m\le M-1} \max_{1\le i\le J_x-1,\, 1\le j\le J_y-1} |T^m_{i,j}|, \qquad m = 1,\dots,M,\end{aligned}$$

and we have that

$$\max_{0\le i\le J_x,\, 0\le j\le J_y} \max_{0\le m\le M} |u(x_i,y_j,t_m) - U^m_{i,j}| \le T \max_{1\le i\le J_x-1,\, 1\le j\le J_y-1} \max_{0\le m\le M-1} |T^m_{i,j}|.$$

The explicit and implicit Euler schemes therefore satisfy the error bound

$$\max_{0\le i\le J_x,\, 0\le j\le J_y} \max_{0\le m\le M} |u(x_i,y_j,t_m) - U^m_{i,j}| \le \text{Const.}\left((\Delta x)^2 + (\Delta y)^2 + \Delta t\right),$$

where in the case of the explicit Euler scheme we are assuming that $\mu_x + \mu_y \le \frac{1}{2}$.

For the Crank–Nicolson scheme, on the other hand, we have that

$$\max_{0\le i\le J_x,\, 0\le j\le J_y} \max_{0\le m\le M} |u(x_i,y_j,t_m) - U^m_{i,j}| \le \text{Const.}\left((\Delta x)^2 + (\Delta y)^2 + (\Delta t)^2\right),$$

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 a large system of linear algebraic equations at each time level. This is true, in particular, in the case of 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 coordinate directions respectively, alternating between solves in the x and y coordinate 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: **Start of optional material**

$$\left(1 - \frac{1}{2}\mu_x\delta_x^2 - \mu_y\frac{1}{2}\delta_y^2\right) U^{m+1}_{i,j} = \left(1 + \frac{1}{2}\mu_x\delta_x^2 + \mu_y\frac{1}{2}\delta_y^2\right) U^m_{i,j}, \tag{105}$$

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

$$U^0_{i,j} := u_0(x_i, y_j), \quad i = 0,\dots,J_x, \;\; j = 0,\dots,J_y,$$

and the boundary condition

$$U^m_{i,j} := B(x_i, y_j, t_m), \;\; \text{at the boundary mesh-points, for } m = 1,\dots,M.$$

Let us modify this scheme (subject to the same initial and boundary conditions) to:

$$\left(1 - \frac{1}{2}\mu_x\delta_x^2\right)\left(1 - \mu_y\frac{1}{2}\delta_y^2\right) U^{m+1}_{i,j} = \left(1 + \frac{1}{2}\mu_x\delta_x^2\right)\left(1 + \mu_y\frac{1}{2}\delta_y^2\right) U^m_{i,j}. \tag{106}$$

As the expressions on the left-hand side and the right-hand side of (106) differ from those in the Crank–Nicolson scheme (105) above, the numerical solution computed from (106) will also differ from the one obtained from the Crank–Nicolson scheme (105). It can be shown however that the consistency error of (106) is still $\mathcal{O}((\Delta x)^2+(\Delta y)^2+(\Delta t)^2)$ as in the case of the Crank–Nicolson scheme; there is therefore no significant loss of accuracy resulting from the replacement of (105) with (106). The benefits of replacing (105) with (106) will be made clear below.

By introducing the intermediate level $U^{m+1/2}$, we can rewrite the last equality in the following equivalent form

$$\left(1-\frac{1}{2}\mu_x\delta_x^2\right)U_{i,j}^{m+1/2}=\left(1+\frac{1}{2}\mu_y\delta_y^2\right)U_{i,j}^{m}, \tag{107}$$

$$\left(1-\frac{1}{2}\mu_y\delta_y^2\right)U_{i,j}^{m+1}=\left(1+\frac{1}{2}\mu_x\delta_x^2\right)U_{i,j}^{m+1/2}. \tag{108}$$

The equivalence of the system (107), (108) to the scheme (106) is seen by applying the finite difference operator

$$\left(1+\frac{1}{2}\mu_x\delta_x^2\right) \text{ to eq. (107)} \qquad \text{and the finite difference operator} \qquad \left(1-\frac{1}{2}\mu_x\delta_x^2\right) \text{ to eq. (108),}$$

and noting that these two finite difference operators commute. Given $U_{i,j}^m$, the equation (107) amounts to solving, for each j, a one-dimensional problem in the x-direction to compute $U_{i,j}^{m+1/2}$; and then, by using the computed values $U_{i,j}^{m+1/2}$ one solves, for each i, a one-dimensional problem in the y-direction using (108) to determine $U_{i,j}^{m+1}$. Thus, by starting from the information at time level $m=0$, where $U_{i,j}^0$ is specified by the initial datum, one proceeds from time level m to time level $m+1$, alternating successively between the x and y directions and advancing from time level m to time level $m+1$ for $m=0,\dots,M-1$.

The practical 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,

$$|\lambda(k_x,k_y)|\le 1 \quad \forall\,(k_x,k_y)\in\left[-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}\right]\times\left[-\frac{\pi}{\Delta y},\frac{\pi}{\Delta y}\right].$$

Consequently, the ADI scheme is unconditionally practically stable in the ℓ_2 norm. The consistency error of the ADI scheme can be shown (again, by 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 $\mu_x\le 1$ and $\mu_y\le 1$. 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 simplest example **Lecture 12** of a second-order linear hyperbolic equation, the linear wave equation

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = f(x,t),$$

where $c > 0$ is the wave speed and f is a given source term.

In the special case when f is identically zero and the equation is considered on the whole real line, $-\infty < x < \infty$, 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 defined on $\mathbb{R}$, u_0 is twice continuously differentiable and u_1 is once continuously differentiable 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 $\mathbb{R} \times [0,\infty)$ such that $\frac{\partial f}{\partial x}$ is a continuous function on $\mathbb{R} \times [0,\infty)$, then 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 where x is confined to a nonempty bounded closed spatial interval $[a,b]$ of the real line, with $a < b$, and where $t \in [0,T]$, with $T > 0$. 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 $T > 0$. We shall be concerned with the finite difference approximation of the initial-boundary-value problem

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} &= f(x,t) && \text{for } (x,t) \in (a,b) \times (0,T], \\ u(x,0) &= u_0(x) && \text{for } x \in [a,b], \\ \frac{\partial u}{\partial t}(x,0) &= u_1(x) && \text{for } x \in [a,b], \\ u(a,t) = 0 \quad \text{and} \quad u(b,t) &= 0 && \text{for } t \in [0,T]. \end{aligned} \tag{109}$$

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 approximation of (109), 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 (109). 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 (109). As the proof of the existence of a solution to the initial-boundary-value problem (109) is not within the scope of these lecture notes, we shall simply suppose here that a solution u to (109) 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 $(109)_1$ by the time derivative of u, and we then integrate the resulting expression over the interval $[a, b]$; thus,

$$\int_a^b \frac{\partial^2 u}{\partial t^2}(x,t)\,\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x - c^2 \int_a^b \frac{\partial^2 u}{\partial x^2}(x,t)\,\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x = \int_a^b f(x,t)\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x. \tag{110}$$

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 (110), we arrive at the following equality:

$$\int_a^b \frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}(x,t)\right)\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x + c^2 \int_a^b \frac{\partial u}{\partial x}(x,t)\,\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial x}(x,t)\right)\mathrm{d}x = \int_a^b f(x,t)\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x. \tag{111}$$

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, by interchanging integration over the spatial interval (a,b) and differentiation with respect to t, we have that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,t)\,\mathrm{d}x + \frac{c^2}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b \left(\frac{\partial u}{\partial x}\right)^2(x,t)\,\mathrm{d}x = \int_a^b f(x,t)\frac{\partial u}{\partial t}(x,t)\,\mathrm{d}x. \tag{112}$$

In the special case when f is identically zero, the right-hand side of (112) vanishes, and after integrating the resulting expression from 0 to t, for any $t \in (0,T]$, we deduce that

$$\frac{1}{2}\int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,t)\,\mathrm{d}x + \frac{c^2}{2}\int_a^b \left(\frac{\partial u}{\partial x}\right)^2(x,t)\,\mathrm{d}x = \frac{1}{2}\int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,0)\,\mathrm{d}x + \frac{c^2}{2}\int_a^b \left(\frac{\partial u}{\partial x}\right)^2(x,0)\,\mathrm{d}x. \tag{113}$$

If we view the expression on the left-hand side of the equality (113) as the 'total energy' at time t and the right-hand side as the 'initial total energy', then the equality (113) 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 (112) 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 (113) 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 argument that we shall try to mimic in our stability analysis of the finite difference approximations of the initial-boundary-value problem (112) 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 (112) implies that

$$\mathcal{L}^2(u(\cdot,t)) = \mathcal{L}^2(u(\cdot,0)) + 2\int_0^t \int_a^b f(x,\tau)\frac{\partial u}{\partial t}(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau.$$

As

$$2\alpha\beta \le \alpha^2+\beta^2, \qquad \text{for all } \alpha,\beta\in\mathbb{R},$$

it follows that

$$\begin{aligned}\mathcal{L}^2(u(\cdot,t)) &\le \mathcal{L}^2(u(\cdot,0)) + \int_0^t\int_a^b f^2(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau + \int_0^t\int_a^b \left(\frac{\partial u}{\partial t}\right)^2(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau \\ &\le \mathcal{L}^2(u(\cdot,0)) + \int_0^t\int_a^b f^2(x,\tau)\,\mathrm{d}x\,\mathrm{d}\tau + \int_0^t \mathcal{L}^2(u(\cdot,\tau))\,\mathrm{d}\tau.\end{aligned} \tag{114}$$

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 nonnegative 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 inequality assumed in the statement of the lemma, 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. $\square$

We now return to (114) 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 lemma 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

$$\mathcal{L}^2(u(\cdot,0)) := \int_a^b \left(\frac{\partial u}{\partial t}\right)^2 (x,0)\,\mathrm{d}x + c^2 \int_a^b \left(\frac{\partial u}{\partial x}\right)^2 (x,0)\,\mathrm{d}x = \|u_1\|^2_{L_2((a,b))} + c^2\,|u_0|^2_{H^1((a,b))},$$

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 (109) in the general case when f is not identically zero.

6.2 The implicit scheme: stability, consistency and convergence

For $M \geq 2$, we define $\Delta t := T/M$, and for $J \geq 2$ the spatial step is taken to be $\Delta x := (b-a)/J$. **Lecture 13**
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

$$\begin{aligned}
\frac{U_j^{m+1} - 2U_j^m + U_j^{m-1}}{(\Delta t)^2} - c^2 \frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2} &= f(x_j, t_{m+1}) && \text{for} \begin{cases} j = 1, \ldots, J-1, \\ m = 1, \ldots, M-1, \end{cases} \\
U_j^0 &:= u_0(x_j) && \text{for } j = 0, 1, \ldots, J, \\
U_j^1 &:= U_j^0 + \Delta t\, u_1(x_j) && \text{for } j = 1, 2, \ldots, J-1, \\
U_0^m := 0 \quad \text{and} \quad U_J^m &:= 0 && \text{for } m = 1, \ldots, M.
\end{aligned} \tag{115}$$

The second numerical initial condition, featuring in equation $(115)_3$, stems from the observation that if $\frac{\partial^2 u}{\partial t^2} \in C([a,b] \times [0,T])$ then

$$\frac{u(x_j, \Delta t) - U_j^0}{\Delta t} = \frac{u(x_j, \Delta t) - u(x_j, 0)}{\Delta t} = \frac{\partial u}{\partial t}(x_j, 0) + \mathcal{O}(\Delta t) = u_1(x_j) + \mathcal{O}(\Delta t);$$

thus, by ignoring the $\mathcal{O}(\Delta t)$ term and replacing $u(x_j, \Delta t)$ by its numerical approximation U_j^1 we arrive at the numerical initial condition $(115)_3$.

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}, $j = 0, \ldots, J-1$, for each $m = 0, \ldots, M-1$. The finite difference scheme (115) is therefore usually referred to as the *implicit scheme* for the initial-boundary-value problem (109).

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 A and B 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, for two mesh functions A and B defined on the computational mesh $\{x_j \,:\, j=1,\dots,J\}$ we have that

$$(A-B,A] = \frac{1}{2}(\|A]|^2 - \|B]|^2) + \frac{1}{2}\|A-B]|^2.$$

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:

$$\begin{aligned}&\frac{1}{2}\left(\left\|\frac{U^{m+1} - U^m}{\Delta t}\right\|^2 - \left\|\frac{U^m - U^{m-1}}{\Delta t}\right\|^2\right) + \frac{1}{2}(\Delta t)^2 \left\|\frac{U^{m+1} - 2U^m + U^{m-1}}{(\Delta t)^2}\right\|^2 \\ &\quad + \frac{1}{2}c^2(\|D_x^- U^{m+1}]|^2 - \|D_x^- U^m]|^2) + \frac{1}{2}c^2(\Delta t)^2 \left\|D_x^-\left(\frac{U^{m+1} - U^m}{\Delta t}\right)\right]\right|^2 = (f(\cdot, t_{m+1}), U^{m+1} - U^m).\end{aligned} \tag{116}$$

In the special case when f is identically zero the equality (116) implies that

$$\left\|\frac{U^{m+1} - U^m}{\Delta t}\right\|^2 + c^2\|D_x^- U^{m+1}]|^2 \le \left\|\frac{U^m - U^{m-1}}{\Delta t}\right\|^2 + c^2\|D_x^- U^m]|^2. \tag{117}$$

Let us define the nonnegative expression

$$\mathcal{M}^2(U^m) := \left\|\frac{U^{m+1} - U^m}{\Delta t}\right\|^2 + c^2\|D_x^- U^{m+1}]|^2.$$

With this notation (117) 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 (115) 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 (116) and we focus our attention on the term on its right-hand side. By the Cauchy–Schwarz inequality,

$$\begin{aligned}(f(\cdot, t_{m+1}), U^{m+1} - U^m) &\le \|f(\cdot, t_{m+1})\|\, \|U^{m+1} - U^m\| \\ &= \sqrt{\Delta t\, T}\, \|f(\cdot, t_{m+1})\| \sqrt{\frac{\Delta t}{T}} \left\| \frac{U^{m+1} - U^m}{\Delta t} \right\| \\ &\le \frac{\Delta t\, T}{2} \|f(\cdot, t_{m+1})\|^2 + \frac{\Delta t}{2T} \left\| \frac{U^{m+1} - U^m}{\Delta t} \right\|^2,\end{aligned} \tag{118}$$

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 (118) into (116) we deduce that

$$\left(1 - \frac{\Delta t}{T}\right)\left(\left\|\frac{U^{m+1} - U^m}{\Delta t}\right\|^2 + c^2\|D_x^- U^{m+1}]|^2\right) \le \left\|\frac{U^m - U^{m-1}}{\Delta t}\right\|^2 + c^2\|D_x^- U^m]|^2 + \Delta t\, T\, \|f(\cdot, t_{m+1})\|^2. \tag{119}$$

By recalling the definition of $\mathcal{M}^2(U^m)$ we can rewrite (119) 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

$$1 - x \ge \frac{1}{1 + 2x} \qquad \forall x \in \left[0, \tfrac{1}{2}\right],$$

it follows with $x = \Delta t/T$ that, for $m = 1, 2, \dots, M-1$,

$$\begin{aligned}\mathcal{M}^2(U^m) &\le \left(1 + \frac{2\,\Delta t}{T}\right) \mathcal{M}^2(U^{m-1}) + \Delta t\, T \left(1 + \frac{2\,\Delta t}{T}\right) \|f(\cdot, t_{m+1})\|^2 \\ &\le \left(1 + \frac{2\,\Delta t}{T}\right) \mathcal{M}^2(U^{m-1}) + 2\,\Delta t\, T\, \|f(\cdot, t_{m+1})\|^2.\end{aligned}$$

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*

$$a_m \le \alpha\, a_{m-1} + b_m \qquad \text{for } m = 1, 2, \dots, M-1.$$

Then,

$$a_m \le \alpha^m a_0 + \sum_{k=1}^{m} \alpha^{m-k} b_k \qquad \text{for } m = 1, 2, \dots, M-1.$$

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

$$\mathcal{M}^2(U^m) \le \left(1 + \frac{2\,\Delta t}{T}\right)^m \mathcal{M}^2(U^0) + 2\,\Delta t\, T \sum_{k=1}^{m} \left(1 + \frac{2\,\Delta t}{T}\right)^{m-k} \|f(\cdot, t_{k+1})\|^2 \quad \text{for } m = 1, 2, \ldots, M-1.$$

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

$$(1 + 2x)^{\frac{1}{x}} \le \mathrm{e}^2 \quad \forall\, x \in \left(0, \tfrac{1}{2}\right],$$

with $x = \Delta t / T$, which, in turn, follows by noting that $1 + 2x \le \mathrm{e}^{2x}$ for all $x \ge 0$. Thus we deduce the following stability result for the implicit scheme (115).

Theorem 16 *The implicit finite difference approximation* (115) *of the initial-boundary-value problem* (109)*, 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*

$$\mathcal{M}^2(U^m) \le \mathrm{e}^2\, \mathcal{M}^2(U^0) + 2\,\mathrm{e}^2\, T \sum_{k=1}^{m} \Delta t\, \|f(\cdot, t_{k+1})\|^2\,, \quad \text{for } m = 1, \ldots, M-1,$$

independently of the choice of Δx and Δt.

Consistency of the implicit scheme. We define the consistency error of the scheme by

$$T_j^{m+1} := \frac{u_j^{m+1} - 2u_j^m + u_j^{m-1}}{(\Delta t)^2} - c^2 \frac{u_{j+1}^{m+1} - 2u_j^{m+1} + u_{j-1}^{m+1}}{(\Delta x)^2} - f(x_j, t_{m+1}), \qquad \left\{ \begin{array}{l} j = 1, \ldots, J-1, \\ m = 1, \ldots, M-1, \end{array} \right.$$

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

$$T_j^{m+1} := \left(\frac{u_j^{m+1} - 2u_j^m + u_j^{m-1}}{(\Delta t)^2} - \frac{\partial^2 u}{\partial t^2}(x_j, t_{m+1}) \right) - c^2 \left(\frac{u_{j+1}^{m+1} - 2u_j^{m+1} + u_{j-1}^{m+1}}{(\Delta x)^2} - \frac{\partial^2 u}{\partial x^2}(x_j, t_{m+1}) \right)$$

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 $j = 1, \ldots, J-1$. Hence, by Taylor series expansion of $u_j^m = u(x_j, t_m)$ and $u_j^{m-1} = u(x_j, t_{m-1})$ about the point (x_j, t_{m+1}) we have that

$$\frac{u_j^{m+1} - 2u_j^m + u_j^{m-1}}{(\Delta t)^2} - \frac{\partial^2 u}{\partial t^2}(x_j, t_{m+1}) = \frac{1}{3}\Delta t \left(\frac{\partial^3 u}{\partial t^3}(x_j, \eta_m) - 4\frac{\partial^3 u}{\partial t^3}(x_j, \zeta_m) \right),$$

where $\eta_m \in [t_m, t_{m+1}]$ and $\zeta_m \in [t_{m-1}, t_{m+1}]$ and, provided that the third partial derivative of u with respect to t is a continuous function on $[a, b] \times [0, T]$. Similarly, by Taylor series expansion of $u_{j+1}^{m+1} = u(x_{j+1}, t_{m+1})$ and $u_{j-1}^{m+1} = u(x_{j-1}, t_{m+1})$ about the point (x_j, t_{m+1}) we find that

$$\frac{u_{j+1}^{m+1} - 2u_j^{m+1} + u_{j-1}^{m+1}}{(\Delta x)^2} - \frac{\partial^2 u}{\partial x^2}(x_j, t_{m+1}) = \frac{1}{12}(\Delta x)^2 \frac{\partial^4 u}{\partial x^4}(\xi_j, t_{m+1}),$$

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,

$$|T_j^{m+1}| \leq \frac{1}{12}c^2(\Delta x)^2 M_{4x} + \frac{5}{3}\Delta t M_{3t}, \qquad \begin{cases} j = 1, \ldots, J-1, \\ m = 1, \ldots, M-1, \end{cases} \tag{120}$$

where

$$M_{4x} := \max_{(x,t)\in[a,b]\times[0,T]} \left| \frac{\partial^4 u}{\partial x^4}(x, t) \right| \quad \text{and} \quad M_{3t} := \max_{(x,t)\in[a,b]\times[0,T]} \left| \frac{\partial^3 u}{\partial t^3}(x, t) \right|.$$

It remains to bound T_j^1. This time, by performing a Taylor series expansion, but now with an integral remainder term, we get that

$$T_j^1 = \frac{1}{\Delta t}\int_0^{\Delta t} (\Delta t - t)\, \frac{\partial^2 u}{\partial t^2}(x_j, t)\, \mathrm{d}t, \tag{121}$$

and therefore

$$|T_j^1| \leq \frac{1}{2}\Delta t\, M_{2t}, \quad j = 1, \ldots, J-1,$$

where

$$M_{2t} := \max_{(x,t)\in[a,b]\times[0,T]} \left| \frac{\partial^2 u}{\partial t^2}(x, t) \right|.$$

Having bounded the consistency error we are now ready to investigate the convergence of the implicit scheme.

Convergence of the implicit scheme. In the rest of the section we shall explore the convergence of the finite difference scheme (115). 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

$$\frac{e_j^{m+1} - 2e_j^m + e_j^{m-1}}{(\Delta t)^2} - c^2\frac{e_{j+1}^{m+1} - 2e_j^{m+1} + e_{j-1}^{m+1}}{(\Delta x)^2} = T_j^{m+1}, \qquad \begin{cases} j = 1, \ldots, J-1, \\ m = 1, \ldots, M-1, \end{cases}$$

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}, $U_j^0 = u_0(x_j)$

replaced by $e_j^0 = 0$, 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 $j = 1, \dots, J-1$ and $k = 1, \dots, M-1$, that

$$\mathcal{M}^2(e^m) \leq \mathrm{e}^2\, \mathcal{M}^2(e^0) + 2\,\mathrm{e}^2\, T \sum_{k=1}^{m} \Delta t \, \left\| T^{k+1} \right\|^2, \quad \text{for } m = 1, \dots, M-1.$$

Now, because $(J-1)\Delta x \leq (b-a)$, it follows from (120) that

$$\max_{1\leq k\leq m} \left\| T^{k+1} \right\|^2 = \max_{1\leq k\leq m} \sum_{j=1}^{J-1} \Delta x \, |T_j^{k+1}|^2 \leq (b-a) \left[\frac{1}{12} c^2 (\Delta x)^2 M_{4x} + \frac{5}{3} \Delta t M_{3t} \right]^2.$$

On the other hand,

$$\mathcal{M}^2(e^0) = \left\| \frac{e^1 - e^0}{\Delta t} \right\|^2 + c^2 \|D_x^- e^1]|^2 = \|T^1\|^2 + c^2 \|D_x^- e^1]|^2 \leq (b-a) \left[\frac{1}{2} \Delta t \, M_{2t} \right]^2 + c^2 \|D_x^- e^1]|^2.$$

As, by recalling (121),

$$\begin{aligned} D_x^- e_j^1 = D_x^- e_j^0 + \Delta t \, D_x^- T_j^1 &= \Delta t \, D_x^- T_j^1 = \int_0^{\Delta t} (\Delta t - t) \, D_x^- \frac{\partial^2 u}{\partial t^2}(x_j, t) \, \mathrm{d}t \\ &= \frac{1}{\Delta x} \int_0^{\Delta t} (\Delta t - t) \int_{x_{j-1}}^{x_j} \frac{\partial^3 u}{\partial x \, \partial t^2}(x, t) \, \mathrm{d}x \, \mathrm{d}t, \end{aligned}$$

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,

$$\mathcal{M}^2(e^0) \leq (b-a) \left[\frac{1}{2} \Delta t \, M_{2t} \right]^2 + c^2 (b-a) \left[\frac{1}{2} \, (\Delta t)^2 M_{1x2t} \right]^2.$$

Hence, finally,

$$\mathcal{M}^2(e^m) \leq \mathrm{e}^2 (b-a) \left[\frac{1}{2} \Delta t \, M_{2t} \right]^2 + c^2 \mathrm{e}^2 (b-a) \left[\frac{1}{2} \, (\Delta t)^2 M_{1x2t} \right]^2 + 2\mathrm{e}^2 \, T^2 (b-a) \left[\frac{1}{12} c^2 (\Delta x)^2 M_{4x} + \frac{5}{3} \Delta t M_{3t} \right]^2$$

for $m = 1, \dots, M-1$. Thus, provided that M_{2t}, M_{1x2t}, M_{4x} and M_{3t} are all finite, we have that

$$\max_{m\in\{1,\dots,M-1\}} [\mathcal{M}^2(u^m - U^m)]^{\frac{1}{2}} = \mathcal{O}((\Delta x)^2 + \Delta t),$$

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 $\max_{m\in\{1,\dots,M-1\}} [\mathcal{M}^2(\cdot)]^{\frac{1}{2}}$. Thanks to the unconditional stability of the implicit scheme, its convergence is also *unconditional* in the sense that there is no limitation on the size of the time step Δt in terms of the spatial mesh-size Δx for convergence of the sequence of numerical approximations to the solution of the wave equation to occur as Δx and Δt tend to 0.

Next we shall investigate the explicit finite difference approximation of the wave equation. It will be shown that, in contrast with the implicit scheme, the explicit scheme is only *conditionally stable*, and its convergence will therefore also shown to be conditional; specifically, we shall require that

$$\frac{c\,\Delta t}{\Delta x} \leq 1,$$

where $c > 0$ is the wave speed, appearing as the coefficient of $\frac{\partial^2 u}{\partial x^2}$ in the wave equation.

6.3 The explicit scheme: stability, consistency and convergence

For $M \geq 2$, we define $\Delta t := T/M$, and for $J \geq 2$ the spatial step is taken to be $\Delta x := (b-a)/J$. **Lecture 14** 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

$$\begin{aligned}
\frac{U_j^{m+1} - 2U_j^m + U_j^{m-1}}{(\Delta t)^2} - c^2 \frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2} &= f(x_j, t_m) && \text{for } \begin{cases} j = 1, \ldots, J-1, \\ m = 1, \ldots, M-1, \end{cases} \\
U_j^0 &:= u_0(x_j) && \text{for } j = 0, 1, \ldots, J, \\
U_j^1 &:= U_j^0 + \Delta t\, u_1(x_j) && \text{for } j = 1, 2, \ldots, J-1, \\
U_0^m := 0 \text{ and } \quad U_J^m &:= 0 && \text{for } m = 1, \ldots, M.
\end{aligned} \tag{122}$$

As in the case of the implicit scheme (115), the second numerical initial condition, appearing in $(122)_3$, 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 replacing $u(x_j, \Delta t)$ by its numerical approximation U_j^1 at the cost of ignoring the $\mathcal{O}(\Delta t)$ term.

Instead of $(122)_3$, 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

$$\begin{aligned}
\frac{u(x_j, \Delta t) - U_j^0}{\Delta t} &= \frac{u(x_j, \Delta t) - u(x_j, 0)}{\Delta t} = \frac{\partial u}{\partial t}(x_j, 0) + \frac{1}{2}\Delta t\, \frac{\partial^2 u}{\partial t^2}(x_j, 0) + \mathcal{O}((\Delta t)^2) \\
&= u_1(x_j) + \frac{1}{2}\Delta t \left(c^2 \frac{\partial^2 u}{\partial x^2}(x_j, 0) + f(x_j, 0) \right) + \mathcal{O}((\Delta t)^2) \\
&= u_1(x_j) + \frac{1}{2}\Delta t \left(c^2 D_x^+ D_x^- u_1(x_j) + f(x_j, 0) \right) + \mathcal{O}(\Delta t\, (\Delta x)^2 + (\Delta t)^2).
\end{aligned}$$

One could therefore, instead of $(122)_3$, use the following more accurate second initial condition:

$$U_j^1 := U_j^0 + \Delta t\, u_1(x_j) + \frac{1}{2}(\Delta t)^2 \left(c^2 D_x^+ D_x^- u_1(x_j) + f(x_j, 0) \right). \tag{123}$$

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 (122), 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 (122) 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 will emerge during the course of our analysis, — the stability condition for the explicit scheme being that $c\Delta t/\Delta x \leq 10$.

We begin by noting that, for any $j \in \{0, \ldots, J\}$ and $m \in \{1, \ldots, M-1\}$, the following identities hold:

$$\begin{aligned}
U_j^{m+1} - U_j^{m-1} = (U_j^{m+1} - U_j^m) + (U_j^m - U_j^{m-1}) &= (U_j^{m+1} + U_j^m) - (U_j^m + U_j^{m-1}), \\
U_j^{m+1} - 2U_j^m + U_j^{m-1} &= (U_j^{m+1} - U_j^m) - (U_j^m - U_j^{m-1}), \\
U_j^{m+1} + 2U_j^m + U_j^{m-1} &= (U_j^{m+1} + U_j^m) + (U_j^m + U_j^{m-1}).
\end{aligned} \tag{124}$$

The left-hand side of equality $(122)_1$ 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 $j=1,\dots,J-1$. Insertion of this into $(122)_1$ then yields

$$\left(I+\frac{1}{4}c^2(\Delta t)^2D_x^+D_x^-\right)\frac{U_j^{m+1}-2U_j^m+U_j^{m-1}}{(\Delta t)^2}=c^2D_x^+D_x^-\frac{U_j^{m+1}+2U_j^m+U_j^{m-1}}{4}+f(x_j,t_m) \tag{125}$$

for $j=1,\dots,J-1$, $m=1,\dots,M-1$, where I signifies the identity operator, which maps any mesh function defined on the spatial mesh $\{x_j\,:\,j=1,\dots,J-1\}$ into itself. 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 $\|U]|:=(U,U]^{\frac{1}{2}}$. We then take the $(\cdot,\cdot)$ inner product of (125) with $U^{m+1}-U^{m-1}$, making use of $(124)_3$ and $(124)_1$ on the left-hand side, and $(124)_4$ and $(124)_2$ on the right-hand side, together with the equalities

$$(\mathcal{D}(A-B),A+B)=(\mathcal{D}A,A)-(\mathcal{D}B,B),$$

$$(\mathcal{D}(A+B),A-B)=(\mathcal{D}A,A)-(\mathcal{D}B,B),$$

on the left-hand side and the right-hand side, respectively, where the finite difference operator $\mathcal{D}$ satisfies the symmetry property $(\mathcal{D}A,B)=(\mathcal{D}B,A)$; in our case,

$$\mathcal{D}=I+\frac{1}{4}c^2(\Delta t)^2D_x^+D_x^-\qquad\text{and}\qquad\mathcal{D}=c^2D_x^+D_x^-$$

on the left-hand side and right-hand side of (125), respectively, satisfy this symmetry property, which can be verified using summation by parts thanks to the fact that the mesh functions $A=U^{m+1}\pm U^m$ and $B=U^m\pm U^{m-1}$ vanish at x_0 and x_J because of the homogeneous Dirichlet boundary condition $(122)_4$. Thus we obtain the following equality:

$$\begin{aligned}&\left(\left(I+\frac{1}{4}c^2(\Delta t)^2D_x^+D_x^-\right)\frac{U^{m+1}-U^m}{\Delta t},\frac{U^{m+1}-U^m}{\Delta t}\right)-\left(\left(I+\frac{1}{4}c^2(\Delta t)^2D_x^+D_x^-\right)\frac{U^m-U^{m-1}}{\Delta t},\frac{U^m-U^{m-1}}{\Delta t}\right)\\&=-c^2\left(-D_x^+D_x^-\frac{U^{m+1}+U^m}{2},\frac{U^{m+1}+U^m}{2}\right)+c^2\left(-D_x^+D_x^-\frac{U^m+U^{m-1}}{2},\frac{U^m+U^{m-1}}{2}\right)\\&\qquad+(f(\cdot,t_m),U^{m+1}-U^{m-1}).\end{aligned}$$

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

$$\left(\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)\frac{U^{m+1}-U^m}{\Delta t}, \frac{U^{m+1}-U^m}{\Delta t}\right) - \left(\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)\frac{U^{m}-U^{m-1}}{\Delta t}, \frac{U^{m}-U^{m-1}}{\Delta t}\right)$$
$$= -c^2\left(D_x^-\frac{U^{m+1}+U^m}{2}, D_x^-\frac{U^{m+1}+U^m}{2}\right] + c^2\left(D_x^-\frac{U^{m}+U^{m-1}}{2}, D_x^-\frac{U^{m}+U^{m-1}}{2}\right]$$
$$+ (f(\cdot,t_m), U^{m+1}-U^{m-1})$$
$$= -c^2\left\|D_x^-\frac{U^{m+1}+U^m}{2}\right]\Big|^2 + c^2\left\|D_x^-\frac{U^{m}+U^{m-1}}{2}\right]\Big|^2 + (f(\cdot,t_m), U^{m+1}-U^{m-1}).$$

This implies, following a minor rearrangement of terms, that

$$\left(\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)\frac{U^{m+1}-U^m}{\Delta t}, \frac{U^{m+1}-U^m}{\Delta t}\right) + c^2\left\|D_x^-\frac{U^{m+1}+U^m}{2}\right]\Big|^2$$
$$= \left(\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)\frac{U^{m}-U^{m-1}}{\Delta t}, \frac{U^{m}-U^{m-1}}{\Delta t}\right) + c^2\left\|D_x^-\frac{U^{m}+U^{m-1}}{2}\right]\Big|^2 \qquad (126)$$
$$+ (f(\cdot,t_m), U^{m+1}-U^{m-1}).$$

The second term on the left-hand side of (126) 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 (126) and the first term on the right-hand side are also nonnegative. In order to do so we shall make a small diversion to investigate this question. Letting

$$V_j^m := \frac{U_j^{m+1}-U_j^m}{\Delta t}, \qquad j = 0,\ldots,J,$$

and noting that $V_0^m = V_J^m = 0$, it follows that

$$\left(\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)V^m, V^m\right) = \|V^m\|^2 + \frac{1}{4}c^2(\Delta t)^2(D_x^+D_x^-V^m, V^m)$$
$$= \|V^m\|^2 - \frac{1}{4}c^2(\Delta t)^2(D_x^-V^m, D_x^-V^m]$$
$$= \|V^m\|^2 - \frac{1}{4}c^2(\Delta t)^2\|D_x^-V^m]|^2.$$

The left-most expression in this chain of equalities will be nonnegative if and only if

$$\|V^m\|^2 - \frac{1}{4}c^2(\Delta t)^2\|D_x^-V^m]|^2 \geq 0.$$

Our objective is to show that this can be guaranteed by requiring that $c\Delta t/\Delta x \leq 1$. Noting that for any nonnegative real numbers α and β one has $(\alpha-\beta)^2 \leq 2\alpha^2 + 2\beta^2$, it follows that

$$\|D_x^-V^m]|^2 = \sum_{j=1}^{J}\Delta x\,|D_x^-V_j^m|^2 = (\Delta x)^{-1}\sum_{j=1}^{J}(V_j^m - V_{j-1}^m)^2$$
$$\leq 2\,(\Delta x)^{-1}\sum_{j=1}^{J}(V_j^m)^2 + (V_{j-1}^m)^2 = 4\,(\Delta x)^{-1}\sum_{j=1}^{J-1}(V_j^m)^2$$
$$= 4\,(\Delta x)^{-2}\sum_{j=1}^{J-1}\Delta x\,(V_j^m)^2 = \left(\frac{2}{\Delta x}\right)^2\|V\|^2.$$

Thus we deduce that

$$\left(\left(I+\frac{1}{4}c^2(\Delta t)^2D_x^+D_x^-\right)V^m,V^m\right)\geq\left(1-\left(\frac{c\,\Delta t}{\Delta x}\right)^2\right)\|V^m\|^2. \tag{127}$$

We shall therefore suppose that the following condition holds, referred to as a Courant–Friedrichs–Lewy (or CFL) condition:

$$\frac{c\,\Delta t}{\Delta x}\leq 1. \tag{128}$$

Assuming that (128) holds, we then have from (127) that

$$\left(\left(I+\frac{1}{4}c^2(\Delta t)^2D_x^+D_x^-\right)\frac{U^{m+1}-U^m}{\Delta t},\frac{U^{m+1}-U^m}{\Delta t}\right)\geq 0. \tag{129}$$

We shall therefore proceed by assuming that (128) holds, and define the nonnegative expression

$$\mathcal{N}^2(U^m):=\left(\left(I+\frac{1}{4}c^2(\Delta t)^2D_x^+D_x^-\right)\frac{U^{m+1}-U^m}{\Delta t},\frac{U^{m+1}-U^m}{\Delta t}\right)+c^2\left\|D_x^-\frac{U^{m+1}+U^m}{2}\right]\Big|^2.$$

With this notation (126) becomes

$$\mathcal{N}^2(U^m)=\mathcal{N}^2(U^{m-1})+(f(\cdot,t_m),U^{m+1}-U^{m-1}). \tag{130}$$

In the special case when f is identically zero (130) guarantees the stability of the explicit scheme under the CFL condition (128); indeed, (130) 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 $m=0,1,\dots,M$.[10] Thus we have shown that, provided that the CFL condition (128) holds and f is identically zero, the explicit scheme (122) is (conditionally) stable in this norm.

---

[10]If is straightforward to check that the mapping $U\mapsto\max_{m\in\{0,\dots,M-1\}}\left[\mathcal{N}^2(U^m)\right]^{1/2}=:|||U|||$ is a seminorm on the linear space $\mathcal{V}$ 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 $m=0,1,\dots,M$. Indeed, $|||U|||\geq 0$ for all $U\in\mathcal{V}$, $|||0|||=0$, $|||\lambda U|||=|\lambda|\,|||U|||$ for all $\lambda\in\mathbb{R}$ and all $U\in\mathcal{V}$, and $|||U+V|||\leq|||U|||+|||V|||$ for all $U,V\in\mathcal{V}$. To show that $|||\cdot|||$ is in fact a norm, one needs to check in addition that if $|||U|||=0$ for some $U\in\mathcal{V}$ then $U=0$. When $c\Delta t/\Delta x<1$ this is an immediate consequence of (127), which implies that $\|U^{m+1}-U^m\|=0$ for all $m=0,1,\dots,M-1$, and therefore $U_j^{m+1}-U_j^m=0$ for all $j=0,1,\dots,J$, $m=0,1,\dots,M-1$; and this, together with $0=\|D_x^-(U^{m+1}+U^m)]|^2\geq 2\,\|U^{m+1}+U^m\|^2$ (by the discrete Poincaré–Friedrichs inequality), which yields $U_j^{m+1}+U_j^m=0$ for all $j=0,1,\dots,J$, $m=0,1,\dots,M-1$, then implies that $U_j^m=0$ for all $j=0,\dots,J$ and all $m=0,1,\dots,M$, as required. Thus, when $c\Delta t/\Delta x<1$, $|||\cdot|||$ is indeed a norm on $\mathcal{V}$. On the other hand, when $c\Delta t/\Delta x=1$ the inequality (127) can no longer be used to repeat this argument. It can be shown, however, that (127) can be improved:

$$\begin{aligned}\left(\left(I+\frac{1}{4}c^2(\Delta t)^2D_x^+D_x^-\right)V^m,V^m\right)&=\left(\left(I+\frac{1}{4}(\Delta x)^2D_x^+D_x^-\right)V^m,V^m\right)-\frac{1}{4}(\Delta x)^2\left(1-\left(\frac{c\Delta t}{\Delta x}\right)^2\right)(D_x^+D_x^-V^m,V^m)\\&\geq 4\sin^2\left(\frac{\pi\Delta x}{2}\right)\|V^m\|^2+\frac{1}{4}(\Delta x)^2\left(1-\left(\frac{c\,\Delta t}{\Delta x}\right)^2\right)\|D_x^-V^m]|^2\\&\geq 4\sin^2\left(\frac{\pi\Delta x}{2}\right)\|V^m\|^2+\left(1-\left(\frac{c\,\Delta t}{\Delta x}\right)^2\right)\|V^m\|^2.\end{aligned} \tag{131}$$

Using this inequality instead of (127) one can now show that $|||\cdot|||$ is a norm on $\mathcal{V}$ as long as $c\Delta t/\Delta x\leq 1$. We note however that when $c\Delta t/\Delta x=1$ the right-hand side of the inequality (131) reduces to $4\sin^2\left(\frac{\pi\Delta x}{2}\right)\|V^m\|^2$. Since $4\sin^2\left(\frac{\pi\Delta x}{2}\right)\asymp\pi^2(\Delta x)^2$ as $\Delta x\to 0$, and $\pi^2(\Delta x)^2\to 0$ as $\Delta x\to 0$, the prefactor of $\|V^m\|^2$ on the right-hand side of the inequality (131) gradually deteriorates when $c\Delta t/\Delta x=1$ and $\Delta x\to 0$. In particular when $c\Delta t/\Delta x=1$ there is a no constant $c_0>0$, independent of Δx, such that the left-hand side of the inequality (131) is bounded below by $c_0\|V^m\|^2$.

We now return to the general case when f is not identically zero, under the following assumption, which is slightly more restrictive than the CFL condition (128): **Start of optional material**

$$\exists\, c_0 \in (0,1) \quad \text{such that} \quad \frac{c\,\Delta x}{\Delta t} \le c_0, \tag{132}$$

and we focus our attention on the second term on the right-hand side of (130). For $m \in \{1,\dots,M\}$, let Z^m be the solution of the problem

$$\begin{aligned}\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right) Z_j^m &= f(x_j,t_m), \qquad j = 1,\dots,J-1.\\ Z_0^m = Z_J^m &= 0.\end{aligned} \tag{133}$$

To show the existence of a unique solution Z^m to this problem we note that (133) is in fact a system of $J-1$ linear algebraic equations for the $J-1$ unknowns $Z_1^m,\dots,Z_{J-1}^m$. Therefore (133) 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 $j \in \{1,\dots,J-1\}$ on the right-hand side of (133)) has $Z_j^m = 0$, $j = 0,\dots,J$, as its unique solution. Clearly, the homogeneous counterpart of (133) does indeed have $Z_j^m = 0$, $j = 0,\dots,J$, as a solution. The fact that this is the *unique solution* to the homogeneous counterpart of (133) follows by noting that, thanks to the inequality (127) with $V^m = Z^m$ and the assumed CFL condition (132), we have that $\|Z^m\|^2 = 0$. Therefore $Z_j^m = 0$ for all $j = 0,\dots,J$. In other words, the homogeneous counterpart of (133) has the trivial solution as its unique solution; therefore the nonhomogeneous problem (133) possesses a unique solution.

Having shown the existence of unique solution to (133), it makes sense to write

$$Z^m = \left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)^{-1} f(\cdot,t_m). \tag{134}$$

With this, we return to the second term on the right-hand side of (130) and decompose it as follows:

$$\begin{aligned}&(f(\cdot,t_m), U^{m+1} - U^{m-1}) = (f(\cdot,t_m), U^{m+1} - U^m) + (f(\cdot,t_m), U^m - U^{m-1})\\ &\quad = \left(\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right) Z^m, U^{m+1} - U^m\right) + \left(\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right) Z^m, U^m - U^{m-1}\right).\end{aligned}$$

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 $\{x_j : j = 0,\dots,J\}$ and such that $V_0 = V_J = 0$ and $W_0 = W_J = 0$, one has, by summation by parts,

$$(D_x^+ D_x^- V, W) = (V, D_x^+ D_x^- W). \tag{135}$$

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

$$\begin{aligned}&(f(\cdot,t_m), U^{m+1} - U^{m-1}) = (f(\cdot,t_m), U^{m+1} - U^m) + (f(\cdot,t_m), U^m - U^{m-1})\\ &\quad = \left(Z^m, \left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)(U^{m+1} - U^m)\right) + \left(Z^m, \left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)(U^m - U^{m-1})\right)\\ &\quad = \left(\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)(U^{m+1} - U^m), Z^m\right) + \left(\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)(U^m - U^{m-1}), Z^m\right).\end{aligned}$$

To simplify our notation, for mesh functions V, W defined on the mesh $\{x_j : j = 0,\dots,J\}$ 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 (135), $[V,W]=[W,V]$; and, by virtue of (127) and (132), 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 the stability analysis of the explicit scheme. In terms of this newly introduced notation we 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 (130) and, after dividing and multiplying by Δt, we find that

$$\mathcal{N}^2(U^m) \leq \mathcal{N}^2(U^{m-1}) + \Delta t \left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right| |[Z^m]| + \Delta t \left|\left[\frac{U^{m}-U^{m-1}}{\Delta t}\right]\right| |[Z^m]|. \tag{136}$$

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

$$\begin{aligned}\left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right|^2 + c^2 \left\|D_x^- \frac{U^{m+1}+U^m}{2}\right]\Big|^2 \leq \left|\left[\frac{U^{m}-U^{m-1}}{\Delta t}\right]\right|^2 + c^2 \left\|D_x^- \frac{U^{m}+U^{m-1}}{2}\right]\Big|^2\\ +\Delta t \left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right| |[Z^m]| + \Delta t \left|\left[\frac{U^{m}-U^{m-1}}{\Delta t}\right]\right| |[Z^m]|.\end{aligned} \tag{137}$$

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 (137):

$$\begin{aligned}\Delta t \left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right| |[Z^m]| &= \sqrt{\frac{\Delta t}{T}} \left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right| \sqrt{\Delta t\, T}\,|[Z^m]|\\ &\leq \frac{\Delta t}{T} \left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right|^2 + \frac{\Delta t\, T}{4}|[Z^m]|^2;\end{aligned}$$

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 (137) 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:

$$\begin{aligned}&\left(1-\frac{\Delta t}{T}\right)\left(\left|\left[\frac{U^{m+1}-U^m}{\Delta t}\right]\right|^2 + c^2 \left\|D_x^- \frac{U^{m+1}+U^m}{2}\right]\Big|^2\right)\\ &\quad\leq \left(1+\frac{\Delta t}{T}\right)\left(\left|\left[\frac{U^{m}-U^{m-1}}{\Delta t}\right]\right|^2 + c^2 \left\|D_x^- \frac{U^{m}+U^{m-1}}{2}\right]\Big|^2\right) + \frac{\Delta t\, T}{2}|[Z^m]|^2.\end{aligned} \tag{138}$$

By recalling the definition of $\mathcal{N}^2(U^m)$ we can rewrite (138) in the following compact form:

$$\mathcal{N}^2(U^m) \leq \frac{T+\Delta t}{T-\Delta t}\mathcal{N}^2(U^{m-1}) + \frac{T^2}{2(T-\Delta t)}\,\Delta t\,|[Z^m]|^2, \quad m = 1,\dots,M-1.$$

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

$$\mathcal{N}^2(U^m) \leq \frac{T+\Delta t}{T-\Delta t}\mathcal{N}^2(U^{m-1}) + T\,\Delta t\,|[Z^m]|^2, \quad m = 1,\dots,M-1.$$

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

$$\mathcal{N}^2(U^m) \leq \left(\frac{T+\Delta t}{T-\Delta t}\right)^m \mathcal{N}^2(U^0) + T\,\Delta t \sum_{k=1}^{m} \left(\frac{T+\Delta t}{T-\Delta t}\right)^{m-k} |[Z^k]|^2, \quad m = 1,\dots,M-1.$$

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 $m \in \{0,1,\dots,M\}$, with $\Delta t \leq \frac{T}{2}$, and one has

$$\frac{1+x}{1-x} \leq 1 + 4x \quad \forall x \in \left[0, \tfrac{1}{2}\right]$$

and

$$(1+4x)^{\frac{1}{x}} \leq \mathrm{e}^4 \quad \forall x \in \left(0, \tfrac{1}{2}\right],$$

where the second inequality follows by noting that $1 + 4x \leq \mathrm{e}^{4x}$ for all $x \geq 0$. Hence,

$$\mathcal{N}^2(U^m) \leq \mathrm{e}^4\,\mathcal{N}^2(U^0) + \mathrm{e}^4\,T \sum_{k=1}^{m} \Delta t\,|[Z^k]|^2 \quad \text{for } m = 1,\dots,M-1.$$

Finally, by recalling the definition of Z^m from (134), we deduce the following stability result for the explicit finite difference scheme under consideration.

Theorem 17 *Suppose that the CFL condition* (132) *is satisfied. Then, the explicit finite difference approximation* (122) *of the initial-boundary-value problem* (109)*, 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*

$$\mathcal{N}^2(U^m) \leq \mathrm{e}^4\,\mathcal{N}^2(U^0) + \mathrm{e}^4\,T \sum_{k=1}^{m} \Delta t \left|\left[\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)^{-1} f(\cdot, t_k)\right]\right|^2, \quad \textit{for } m = 1,\dots,M-1.$$

Consistency of the explicit scheme. We define the consistency error of the explicit scheme by

$$T_j^m := \frac{u_j^{m+1} - 2u_j^m + u_j^{m-1}}{(\Delta t)^2} - c^2 \frac{u_{j+1}^m - 2u_j^m + u_{j-1}^m}{(\Delta x)^2} - f(x_j, t_m) \quad \text{for} \quad \begin{cases} m = 1,\dots,M-1, \\ j = 1,\dots,J-1, \end{cases}$$

and

$$T_j^0 := \frac{u_j^1 - u_j^0}{\Delta t} - u_1(x_j), \quad \text{for } j = 1, \dots, J-1,$$

where $u_j^m := u(x_j, t_m)$, $j = 0, \dots, J$, $m = 0, \dots, M$. 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 $m = 1, \dots, M-1$ and $j = 1, \dots, J-1$, and

$$T_j^0 := \frac{u_j^1 - u_j^0}{\Delta t} - \frac{\partial u}{\partial t}(x_j, 0), \quad \text{for } j = 1, \dots, J-1.$$

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

$$T_j^m = \frac{1}{12}(\Delta t)^2 \frac{\partial^4 u}{\partial t^4}(x_j, \tau_m) - \frac{1}{12} c^2 (\Delta x)^2 \frac{\partial^4 u}{\partial t^4}(\xi_j, t_m), \tag{139}$$

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,

$$T_j^0 = \frac{1}{2} \Delta t \frac{\partial^2 u}{\partial t^2}(x_j, \tau_0),$$

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,

$$|T_j^m| \leq \frac{1}{12} c^2 (\Delta x)^2 M_{4x} + \frac{1}{12} (\Delta t)^2 M_{4t}, \qquad \begin{cases} j = 1, \dots, J-1, \\ m = 1, \dots, M-1, \end{cases}$$

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

$$|T_j^0| \leq \frac{1}{2} \Delta t \, M_{2t},$$

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 (123) is used instead of $(122)_3$, 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,

$$|T_j^0| \leq \frac{1}{6} (\Delta t)^2 M_{3t} + \frac{1}{24} c^2 \Delta t \, (\Delta x)^2 M_{4x}.$$

Convergence of the explicit scheme. The *global error* of the finite difference scheme (122) 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 $j=1,\dots,J-1$ and $k=1,\dots,M-1$, that

$$\mathcal{N}^2(e^m)\le \mathrm{e}^4\,\mathcal{N}^2(e^0)+\mathrm{e}^4\,T^2\max_{1\le k\le m}\left\|\left[\left(I+\frac{1}{4}c^2(\Delta t)^2D_x^+D_x^-\right)^{-1}T^k\right]\right\|^2,\quad \text{for } m=1,\dots,M-1. \tag{140}$$

It remains to bound the two terms on the right-hand side of this inequality. We note that

$$\begin{aligned}\mathcal{N}^2(e^0)&:=\left(\left(I+\frac{1}{4}c^2(\Delta t)^2D_x^+D_x^-\right)\frac{e^1-e^0}{\Delta t},\frac{e^1-e^0}{\Delta t}\right)+c^2\left\|D_x^-\frac{e^1+e^0}{2}\right]\right\|^2\\ &=\left(\left(I+\frac{1}{4}c^2(\Delta t)^2D_x^+D_x^-\right)T^0,T^0\right)+\frac{1}{4}c^2\,(\Delta t)^2\,\|D_x^-T^0]\|^2.\end{aligned}$$

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

$$\begin{aligned}&\left(\left(I+\frac{1}{4}c^2(\Delta t)^2D_x^+D_x^-\right)T^0,T^0\right)+\frac{1}{4}c^2\,(\Delta t)^2\,\|\,D_x^-T^0]\|^2\\ &\qquad=\|T^0\|^2+\frac{1}{4}c^2(\Delta t)^2(D_x^+D_x^-T^0,T^0)+\frac{1}{4}c^2\,(\Delta t)^2\|D_x^-T^0]\|^2\\ &\qquad=\|T^0\|^2-\frac{1}{4}c^2(\Delta t)^2\|D_x^-T^0]\|^2+\frac{1}{4}c^2\,(\Delta t)^2\|D_x^-T^0]\|^2\\ &\qquad=\|T^0\|^2.\end{aligned}$$

By Taylor series expansion with a remainder term we have that

$$T_j^0=\frac{u_j^1-u_j^0}{\Delta t}-\frac{\partial u}{\partial t}(x_j,0)=\frac{1}{2}\Delta t\frac{\partial^2u}{\partial t^2}(x_j,\tau_0),$$

where $\tau_0\in[0,\Delta t]$; it therefore follows that

$$|T_j^0|\le\frac{1}{2}\Delta t\,M_{2t},$$

with

$$M_{2t}:=\max_{(x,t)\in[a,b]\times[0,T]}\left|\frac{\partial^2u}{\partial t^2}(x,t)\right|.$$

Hence,

$$\|T^0\|^2=\Delta t\sum_{j=1}^{J-1}|T_j^0|^2\le\frac{1}{4}(\Delta t)^2M_{2t}^2.$$

and therefore

$$\left(\left(I+\frac{1}{4}c^2(\Delta t)^2D_x^+D_x^-\right)T^0,T^0\right)+\frac{1}{4}c^2\,(\Delta t)^2\|D_x^-T^0]\|^2=\|T_0\|^2\le\frac{1}{4}(\Delta t)^2\,M_{2t}^2=\mathcal{O}((\Delta t)^2).$$

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 (140), we proceed to bound the second term on the right-hand side of (140):

$$\mathrm{e}^4\, T^2 \max_{1\le k\le m} \left|\left[\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)^{-1} T^k\right]\right|^2.$$

Letting

$$V^k := \left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)^{-1} T^k,$$

it follows that

$$\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right) V_j^k = T_j^k \qquad \text{for } j = 1, \dots, J-1,$$

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 (127) in conjunction with the CFL condition (132), we deduce that

$$(1 - c_0^2)\|V^k\|^2 \le \left(\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)V^k, V^k\right) = |[V^k]|^2 = (T^k, V^k) \le \|T^k\|\,\|V^k\|, \tag{141}$$

and therefore

$$\|V^k\| \le (1 - c_0^2)^{-1}\|T^k\|.$$

This and the last inequality in (141) imply that

$$|[V^k]|^2 \le (1 - c_0^2)^{-1}\|T^k\|^2.$$

Hence, thanks to the definition of V^k above and (139), we have that

$$\begin{aligned}\mathrm{e}^4\, T^2 \max_{1\le k\le m} \left|\left[\left(I + \frac{1}{4}c^2(\Delta t)^2 D_x^+ D_x^-\right)^{-1} T^k\right]\right|^2 &= \mathrm{e}^4\, T^2\, |[V^k]|^2 \le \mathrm{e}^4\, T^2\, (1 - c_0^2)^{-1} \max_{1\le k\le m} \|T^k\|^2 \\ &= \mathcal{O}(((\Delta x)^2 + (\Delta t)^2)^2).\end{aligned}$$

Thus we have also bounded the second term on the right-hand side of the inequality (140); 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 $(122)_3$. If instead of $(122)_3$ one uses the more accurate initial condition (123), 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 summary then, 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 (123) of the second initial condition $(109)_3$ 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 (122). 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 $\max_{m\in\{1,\dots,M-1\}}[\mathcal{N}^2(\cdot)]^{\frac{1}{2}}$ if the second initial condition $(109)_3$ is approximated by $(122)_3$, but if one uses the more accurate approximation (123) of the second initial condition, then the explicit scheme exhibits second-order convergence with respect to both Δx and Δt in the norm $\max_{m\in\{1,\dots,M-1\}}[\mathcal{N}^2(\cdot)]^{\frac{1}{2}}$. Both of these convergence results are conditional, in the sense that they hold in the limit of Δx and Δt tending to zero provided the CFL condition $c\Delta t/\Delta x \le c_0$ holds, where $c_0 \in (0, 1)$ is a constant, independent of Δx and Δt.

6.4 Fourier analysis of the implicit and explicit finite difference schemes for the pure initial-value problem

Consider the second-order wave equation

$$u_{tt} = c^2 u_{xx}, \quad x \in \mathbb{R}, \quad t > 0,$$

subject to the initial conditions $u(x, 0) = u_0(x)$, $u_t(x, 0) = u_1(x)$ for $x \in \mathbb{R}$. We shall assume that $u_0, u_1 \in C(\mathbb{R})$ and that both u_0 and u_1 have compact support in $\mathbb{R}$ (and therefore, trivially $\|u_0\|_{\ell_2} < \infty$ and $\|u_1\|_{\ell_2} < \infty$ as well as $\|u_0\|_{L_2((-\infty,\infty))} < \infty$ and $\|u_1\|_{L_2((-\infty,\infty))} < \infty$).

The implicit scheme. The implicit finite difference approximation of this initial-value problem is

$$\frac{U_j^{m+1} - 2U_j^m + U_j^{m-1}}{(\Delta t)^2} = c^2 \frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2}, \quad j \in \mathbb{Z}, \quad m = 1, 2, \dots,$$

subject to the initial conditions

$$U_j^0 := u_0(x_j), \quad \frac{U_j^1 - U_j^0}{\Delta t} := u_1(x_j), \quad j \in \mathbb{Z}.$$

Define the CFL number $\mu := c\Delta t/\Delta x$, and seek U_j^m in the form (cf. the definition of the inverse semidiscrete Fourier transform on p.56):

$$U_j^m = \frac{1}{2\pi}\int_{-\pi/\Delta x}^{\pi/\Delta x} \hat{U}^m(k)\, \mathrm{e}^{\imath k x_j}\, \mathrm{d}k, \quad j \in \mathbb{Z}, \quad m \ge 0.$$

After inserting this into the finite difference scheme and noting that the semidiscrete Fourier transform and its inverse are one-to-one mappings, we deduce that, for all $k \in [-\pi/\Delta x, \pi/\Delta x]$,

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

Therefore (by Euler's formula and simple trigonometry) we have that, for all $k \in [-\pi/\Delta x, \pi/\Delta x]$,

$$\left(1 + 4\mu^2 \sin^2\frac{k\Delta x}{2}\right)\hat{U}^{m+1}(k) - 2\hat{U}^m(k) + \hat{U}^{m-1}(k) = 0, \quad m \ge 1.$$

For each fixed $k \in [-\pi/\Delta x, \pi/\Delta x]$, this is a second-order difference equation of the form

$$\alpha z_{m+1} + \beta z_m + \gamma z_{m-1} = 0, \quad m \ge 1, \tag{142}$$

where $\alpha = 1 + 4\mu^2 \sin^2 \frac{k\Delta x}{2}$, $\beta = -2$, $\gamma = 1$, and $z_m = \hat{U}^m(k)$. We seek a (nontrivial) solution to (142) in the form $z_m = \lambda^m$, with $\lambda \neq 0$ to be found. By plugging $z_m = \lambda^m$ into (142) we obtain:

$$\lambda^{m-1}(\alpha\lambda^2 + \beta\lambda + \gamma) = 0, \quad m \geq 1.$$

As $\lambda \neq 0$, it follows that $\alpha\lambda^2 + \beta\lambda + \gamma = 0$. If $\lambda_{1,2}$ are the roots of this quadratic equation and $\lambda_1 \neq \lambda_2$, then the general solution of (142) is of the form

$$z_m = A\lambda_1^m + B\lambda_2^m, \quad m \geq 0,$$

where A and B are independent of m and are to be found from z_0 and z_1 (by setting $m = 0$ and $m = 1$ and solving a system of linear algebraic equations for A and B). If on the other hand $\lambda_1 = \lambda_2 = \lambda$ (repeated root), then the general solution of (142) is of the form

$$z_m = A\lambda^m + Bm\lambda^m,$$

with A and B independent of m, to be determined from z_0 and z_1 (again, by setting $m = 0$ and $m = 1$ and solving a system of linear algebraic equation for A and B).

In our case, because the roots of the quadratic depend on the wave number $k \in [-\pi/\Delta x, \pi/\Delta x]$ we shall write $\lambda_{1,2}(k)$ instead of $\lambda_{1,2}$ to emphasize this fact. The roots are then as follows:

$$\lambda_{1,2}(k) = \frac{1 \pm \imath\mu \left|\sin \frac{k\Delta x}{2}\right|}{1 + 4\mu^2 \sin^2 \frac{k\Delta x}{2}} \quad \text{for all } k \in [-\pi/\Delta x, \pi/\Delta x].$$

Obviously $|\lambda_{1,2}(k)| \leq 1$; also, $\lambda_1(k) = \lambda_2(k) = 1$ if, and only if, $k = 0$.

We note in passing that the definition of *practical stability* of finite difference approximations of the pure initial-value problem for the second-order wave equation is precisely that $|\lambda_{1,2}(k)| \leq 1$ for all $k \in [-\pi/\Delta x, \pi/\Delta x]$. Thus we have shown that the implicit scheme under consideration is unconditionally practically stable.

As things stand, it is unclear however what, in anything, the requirement that $|\lambda_{1,2}(k)| \leq 1$ for all $k \in [-\pi/\Delta x, \pi/\Delta x]$ has to do with "stability" of the finite difference scheme. The aim of the discussion that will now follow is therefore to explain the kind of bound on the ℓ_2 norm $\|U^m\|_{\ell_2}$ of the sequence of numerical approximations U^m, $m = 2, 3, \ldots$, where $t_m = m\Delta t$, generated by the finite difference scheme, in terms of the ℓ_2 norms of the initial data u_0 and u_1, that practical stability thus defined then implies. Incidentally, by applying a similar technique to the initial-value problem under consideration we shall derive an analogous bound on of the $L_2((-\infty, \infty))$ norm $\|u(\cdot, t)\|_{L_2((-\infty,\infty))}$ of the exact solution u. This analogy of the bound on $\|U^m\|_{\ell_2}$ with the bound on $\|u(\cdot, t)\|_{L_2((-\infty,\infty))}$ then serves as a justification (as was the case for the initial-value problem for the heat equation) for the use of the terminology "practical stability".

We start with the derivation of the bound on $\|U^m\|_{\ell_2}$. It follows from the discussion above concerning the form of the general solution to a second-order difference equation that

$$\hat{U}^m(k) = A(\lambda_1(k))^m + B(\lambda_2(k))^m \quad \text{for } k \neq 0 \text{ and all } m \geq 0,$$

with A and B to be determined. On the other hand, when $k = 0$ (in which case $\lambda_1(0) = \lambda_2(0) = \lambda = 1$), we have that

$$\hat{U}^m(0) = A\,1^m + Bm\,1^m,$$

with A and B (possibly different from the A and B above) to be determined.

Next, we determine A and B by using the prescribed initial conditions for the finite difference scheme, first in the case of $k \neq 0$ and then in the case of $k = 0$.

(1) First consider the case when $k \neq 0$. Then, with $m = 0$ and $m = 1$, respectively, we have that (for the sake of simplicity of the notation we shall now write λ_1, λ_2 instead of $\lambda_1(k), \lambda_2(k)$, respectively):

$$\hat{U}^0(k) = A + B, \quad \hat{U}^1(k) = \lambda_1 A + \lambda_2 B.$$

We solve this linear system for A and B and obtain

$$A = \frac{\lambda_2 \hat{U}^0(k) - \hat{U}^1(k)}{\lambda_2 - \lambda_1} \quad \text{and} \quad B = \frac{\hat{U}^1(k) - \lambda_1 \hat{U}^0(k)}{\lambda_2 - \lambda_1}.$$

Recalling that $U_j^1 = U_j^0 + \Delta t\, u_1(x_j)$ it follows that $\hat{U}^1(k) = \hat{U}^0(k) + \Delta t\, \widehat{u_1}(k)$. Thus,

$$A = \frac{\lambda_2 - 1}{\lambda_2 - \lambda_1}\hat{U}^0(k) - \frac{\Delta t\, \widehat{u_1}(k)}{\lambda_2 - \lambda_1} \quad \text{and} \quad B = \frac{1 - \lambda_1}{\lambda_2 - \lambda_1}\hat{U}^0(k) + \frac{\Delta t\, \widehat{u_1}(k)}{\lambda_2 - \lambda_1}.$$

Hence, for $m \geq 2$, we have that

$$\hat{U}^m(k) = \frac{\lambda_2 - 1}{\lambda_2 - \lambda_1}\hat{U}^0(k)\lambda_1^m + \frac{1 - \lambda_1}{\lambda_2 - \lambda_1}\hat{U}^0(k)\lambda_2^m + \frac{\Delta t\, \widehat{u_1}(k)}{\lambda_2 - \lambda_1}(\lambda_2^m - \lambda_1^m). \tag{143}$$

Note that, for all $m \geq 4$ and all $k \in [-\pi/\Delta x, \pi/\Delta x] \setminus \{0\}$,

$$\begin{aligned}
\left|\frac{\lambda_2 - 1}{\lambda_2 - \lambda_1}\right| |\lambda_1|^m = \left|\frac{1 - \lambda_1}{\lambda_2 - \lambda_1}\right| |\lambda_2|^m &= \frac{1}{2}\left(1 + 16\mu^2 \sin^2 \frac{k\Delta x}{2}\right)^{\frac{1}{2}} \left[\frac{1 + \mu^2 \sin^2 \frac{k\Delta x}{2}}{(1 + 4\mu^2 \sin^2 \frac{k\Delta x}{2})^2}\right]^{\frac{m}{2}} \\
&\leq \frac{1}{2}\left(1 + 16\mu^2 \sin^2 \frac{k\Delta x}{2}\right)^{\frac{1}{2}} \left[\frac{1}{1 + 4\mu^2 \sin^2 \frac{k\Delta x}{2}}\right]^{\frac{m}{2}} \\
&= \frac{1}{2}\left[\frac{1 + 16\mu^2 \sin^2 \frac{k\Delta x}{2}}{(1 + 4\mu^2 \sin^2 \frac{k\Delta x}{2})^m}\right]^{\frac{1}{2}} \leq \frac{1}{2}.
\end{aligned}$$

Concerning the excluded values $m = 2, 3$, by plotting the function

$$x \in \mathbb{R} \mapsto \frac{1}{2}(1 + 16x^2)^{\frac{1}{2}} \left[\frac{1 + x^2}{(1 + 4x^2)^2}\right]^{\frac{m}{2}}$$

we see that this is, for both $m = 2$ and $m = 3$, again, bounded by $1/2$. Thus, for all $m \geq 2$,

$$\left|\frac{\lambda_2 - 1}{\lambda_2 - \lambda_1}\right| |\lambda_1|^m = \left|\frac{1 - \lambda_1}{\lambda_2 - \lambda_1}\right| |\lambda_2|^m \leq \frac{1}{2}.$$

Hence, and because for $|\lambda_{1,2}| \leq 1$ we have that

$$\left|\frac{\lambda_2^m - \lambda_1^m}{\lambda_2 - \lambda_1}\right| = |\lambda_2^{m-1} + \lambda_2^{m-2}\lambda_1 + \cdots + \lambda_2\lambda_1^{m-2} + \lambda_1^{m-1}| \leq m,$$

it follows that

$$|\hat{U}^m(k)| \leq \frac{1}{2}\,|\hat{U}^0(k)| + \frac{1}{2}\,|\hat{U}^0(k)| + m\Delta t\, |\widehat{u_1}(k)|.$$

Since $\hat{U}^0(k) = \widehat{u_0}(k)$, this then implies that

$$|\hat{U}^m(k)| \leq |\widehat{u_0}(k)| + m\Delta t\, |\widehat{u_1}(k)| \quad \text{for all } k \in [-\pi/\Delta x, \pi/\Delta x] \setminus \{0\} \text{ and all } m \geq 2.$$

(2) Now consider the case when $k = 0$ (in which case $\lambda_1 = \lambda_2 = \lambda = 1$). Then,

$$\hat{U}^0(0) = A + B \cdot 0, \quad \hat{U}^1(0) = A + B.$$

Thus, $A = \hat{U}^0(k)$ and $B = \hat{U}^1(0) - \hat{U}^0(0) = \Delta t\, \hat{U}^0(0)$. Consequently,

$$\hat{U}^m(0) = A\lambda^m + Bm\,\lambda^m = \hat{U}^0(0) + (\hat{U}^1(0) - \hat{U}^0(0))m = \widehat{u_0}(0) + m\Delta t\, \widehat{u_1}(0).$$

Hence,

$$|\hat{U}^m(0)| \le |\widehat{u_0}(0)| + m\Delta t\, |\widehat{u_1}(0)| \quad \text{for } k = 0 \text{ and all } m \ge 2.$$

Combining the bounds on $|\hat{U}^m(k)|$ for the cases $k \neq 0$ and $k = 0$ thus obtained we therefore have that

$$|\hat{U}^m(k)| \le |\widehat{u_0}(k)| + m\Delta t\, |\widehat{u_1}(k)| \quad \text{for all } k \in [-\pi/\Delta x, \pi/\Delta x] \text{ and all } m \ge 2.$$

By the triangle inequality we then deduce that

$$\|\hat{U}^m\|_{L_2((-\pi/\Delta x,\pi/\Delta x))} \le \|\widehat{u_0}\|_{L_2((-\pi/\Delta x,\pi/\Delta x))} + t_m\, \|\widehat{u_1}\|_{L_2((-\pi/\Delta x,\pi/\Delta x))}$$

for all $m \ge 2$, where $t_m := m\Delta t$. Multiplying this inequality by $1/\sqrt{2\pi}$ and using the (discrete) Parseval identity (cf. Lemma 12 on p.56) it follows that

$$\boxed{\|U^m\|_{\ell_2} \le \|u_0\|_{\ell_2} + t_m \|u_1\|_{\ell_2} \quad \text{for all } m \ge 2.} \tag{144}$$

Remark: We note in passing that if the precise forms of $\lambda_{1,2}(k)$ are not taken into account, and we only use that $|\lambda_{1,2}(k)| \le 1$ for all $k \in [-\pi/\Delta x, \pi/\Delta x]$ and $\lambda_1(0) = \lambda_2(0) = 1$, then we obtain a cruder stability inequality because, instead of being bounded by 1/2, we can then only deduce that

$$\left|\frac{\lambda_2 - 1}{\lambda_2 - \lambda_1}\right| |\lambda_1|^m = \left|\frac{1 - \lambda_1}{\lambda_2 - \lambda_1}\right| |\lambda_2|^m \le \frac{1}{2}\left(1 + 16\mu^2 \sin^2 \frac{k\Delta x}{2}\right)^{\frac{1}{2}} \le \frac{1}{2}(1 + 16\mu^2)^{\frac{1}{2}}.$$

Thus, instead of (144) we then end up with the bound

$$\boxed{\|U^m\|_{\ell_2} \le (1 + 16\mu^2)^{\frac{1}{2}}\, \|u_0\|_{\ell_2} + t_m \|u_1\|_{\ell_2} \quad \text{for all } m \ge 2.} \tag{145}$$

For comparison, we Fourier transform the wave equation with respect to x (we shall abuse the notation used above and will write $\hat{\cdot}$ in this calculation to denote the Fourier transform rather than the semidiscrete Fourier transform denoted by $\hat{\cdot}$ above) and solve the resulting ordinary differential equation

$$\hat{u}_{tt}(\xi, t) + c^2\xi^2\hat{u}(\xi, t) = 0 \quad \text{with the initial conditions} \quad \hat{u}(\xi, 0) = \hat{u}_0(\xi) \quad \text{and} \quad \hat{u}_t(\xi, 0) = \hat{u}_1(\xi),$$

with $\xi \in \mathbb{R}$ treated as a parameter, to find that

$$\hat{u}(\xi, t) = \cos(c\xi t)\, \hat{u}_0(\xi) + \frac{\sin(c\xi t)}{c\xi t}\, t\hat{u}_1(\xi).$$

Hence, and because $|\sin(s)/s| \le 1$ for all $s \in \mathbb{R} \setminus \{0\}$ and $\lim_{s\to 0} \sin(s)/s = 1$, it follows that

$$|\hat{u}(\xi, t)| \le |\hat{u}_0(\xi)| + t|\hat{u}_1(\xi)|.$$

This then implies by the triangle inequality that

$$\|\hat{u}(\cdot, t)\|_{L_2((-\infty,\infty))} \le \|\hat{u}_0(\cdot)\|_{L_2((-\infty,\infty))} + t\|\hat{u}_1(\cdot)\|_{L_2((-\infty,\infty))}.$$

Multiplying by $1/\sqrt{2\pi}$ and using Parseval's identity for the Fourier transform on $\mathbb{R}$ it follows that

$$\|u(\cdot, t)\|_{L_2((-\infty,\infty))} \le \|u_0(\cdot)\|_{L_2((-\infty,\infty))} + t\|u_1(\cdot)\|_{L_2((-\infty,\infty))}, \quad t > 0. \tag{146}$$

It is instructive to compare this with the inequality (144) obtained for the numerical method. Clearly, the form of the stability inequality (144) is the same as the stability inequality (146) for the exact solution of the initial-value problem. This then justifies the use of the terminology "practical stability".

The explicit scheme. Next we shall perform a similar analysis for the explicit scheme:

$$\frac{U_j^{m+1} - 2U_j^m + U_j^{m-1}}{(\Delta t)^2} = c^2 \frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2}, \quad j \in \mathbb{Z}, \quad m = 1, 2, \ldots,$$

subject to the initial conditions

$$U_j^0 := u_0(x_j), \quad \frac{U_j^1 - U_j^0}{\Delta t} := u_1(x_j), \quad j \in \mathbb{Z}.$$

We shall suppose in what follows that $|\mu| < 1$. The corresponding expression in Fourier space is therefore

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

Equivalently,

$$\hat{U}^{m+1}(k) - 2\left(1 - 2\mu^2 \sin^2 \frac{k\Delta x}{2}\right)\hat{U}^m(k) + \hat{U}^{m-1}(k) = 0, \quad m \geq 1.$$

The corresponding quadratic characteristic equation is

$$\lambda^2 - 2\left(1 - 2\mu^2 \sin^2 \frac{k\Delta x}{2}\right)\lambda + 1 = 0.$$

When $k = 0$ this has the repeated root $\lambda_1(0) = \lambda_2(0) = \lambda = 1$. If on the other hand $k \neq 0$, then

$$\begin{aligned}\lambda_{1,2}(k) &= \frac{2\left(1 - 2\mu^2 \sin^2 \frac{k\Delta x}{2}\right) \pm \sqrt{4\left(1 - 2\mu^2 \sin^2 \frac{k\Delta x}{2}\right)^2 - 4}}{2} \\ &= \left(1 - 2\mu^2 \sin^2 \frac{k\Delta x}{2}\right) \pm \imath\sqrt{1 - \left(1 - 2\mu^2 \sin^2 \frac{k\Delta x}{2}\right)^2}\end{aligned}$$

Clearly, $|\lambda_1(k)| = |\lambda_2(k)| = 1$ for all $k \in [-\pi/\Delta x, \pi/\Delta x]$.

(1) First consider the case when $k \neq 0$. Then the roots are distinct and therefore, as in the case of the implicit scheme studied above,

$$\hat{U}^m(k) = \frac{\lambda_2 - 1}{\lambda_2 - \lambda_1}\hat{U}^0(k)\lambda_1^m + \frac{1 - \lambda_1}{\lambda_2 - \lambda_1}\hat{U}^0(k)\lambda_2^m + \frac{\Delta t\,\widehat{u_1}(k)}{\lambda_2 - \lambda_1}(\lambda_2^m - \lambda_1^m). \tag{147}$$

Let as write, for the sake of brevity, $S := \mu^2 \sin^2 \frac{k\Delta x}{2}$. Hence,

$$\frac{\lambda_2 - 1}{\lambda_2 - \lambda_1} = \frac{-2S - \imath\sqrt{1 - (1 - 2S)^2}}{-2\imath\sqrt{1 - (1 - 2S)^2}} = \frac{-S - \imath\sqrt{S(1 - S)}}{-2\imath\sqrt{S(1 - S)}}.$$

Because $S \leq \mu^2 < 1$ for all $k \in [-\pi/\Delta x, \pi/\Delta x]$, this then implies that

$$\left|\frac{\lambda_2 - 1}{\lambda_2 - \lambda_1}\right|^2 = \frac{S^2 + S(1 - S)}{4S(1 - S)} = \frac{1}{4(1 - S)} \leq \frac{1}{4(1 - \mu^2)}.$$

Therefore,

$$\left|\frac{\lambda_2 - 1}{\lambda_2 - \lambda_1}\right| \leq \frac{1}{2\sqrt{1 - \mu^2}}.$$

Analogously,

$$\left|\frac{1-\lambda_1}{\lambda_2-\lambda_1}\right| \le \frac{1}{2\sqrt{1-\mu^2}}.$$

It then follows from (147) that for all $k \in [-\pi/\Delta x, \pi/\Delta x] \setminus \{0\}$ we have that

$$|\hat{U}^m(k)| \le \frac{1}{\sqrt{1-\mu^2}}\,|\widehat{u_0}(k)| + m\Delta t\,|\widehat{u_1}(k)| \quad \text{for all } m \ge 2. \tag{148}$$

(2) When $k = 0$, we have a repeated root $\lambda_1(0) = \lambda_2(0) = \lambda = 1$ and therefore, in the same way as in the case of the implicit scheme considered above,

$$|\hat{U}^m(0)| \le |\widehat{u_0}(0)| + m\Delta t\,|\widehat{u_1}(0)| \quad \text{for all } m \ge 2.$$

Therefore, by combining this with (148), analogously as in the case of the implicit scheme, we arrive at the stability inequality

$$\boxed{\|U^m\|_{\ell_2} \le \frac{1}{\sqrt{1-\mu^2}}\|u_0\|_{\ell_2} + t_m\|u_1\|_{\ell_2} \quad \text{for all } m \ge 2,}$$

but, in contrast with the implicit scheme, now only under the assumed CFL condition $|\mu| < 1$, where $\mu := c\Delta t/\Delta x$.

End of optional material

6.5 First-order hyperbolic equations: initial-boundary-value problem and energy estimate

Let Ω be a bounded open set in $\mathbb{R}^n$, $n \ge 1$, with boundary $\Gamma = \partial\Omega$, and let $T > 0$. In $Q = \Omega \times (0, T]$, we consider the initial boundary-value problem **Lecture 15**

$$\frac{\partial u}{\partial t} + \sum_{i=1}^{n} b_i(x)\,\frac{\partial u}{\partial x_i} + c(x,t)u = f(x,t), \qquad x \in \Omega, \;\; t \in (0,T], \tag{149}$$

$$u(x,t) = 0, \qquad x \in \Gamma_-, \;\; t \in [0,T], \tag{150}$$

$$u(x,0) = u_0(x) \qquad x \in \bar{\Omega}, \tag{151}$$

where

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

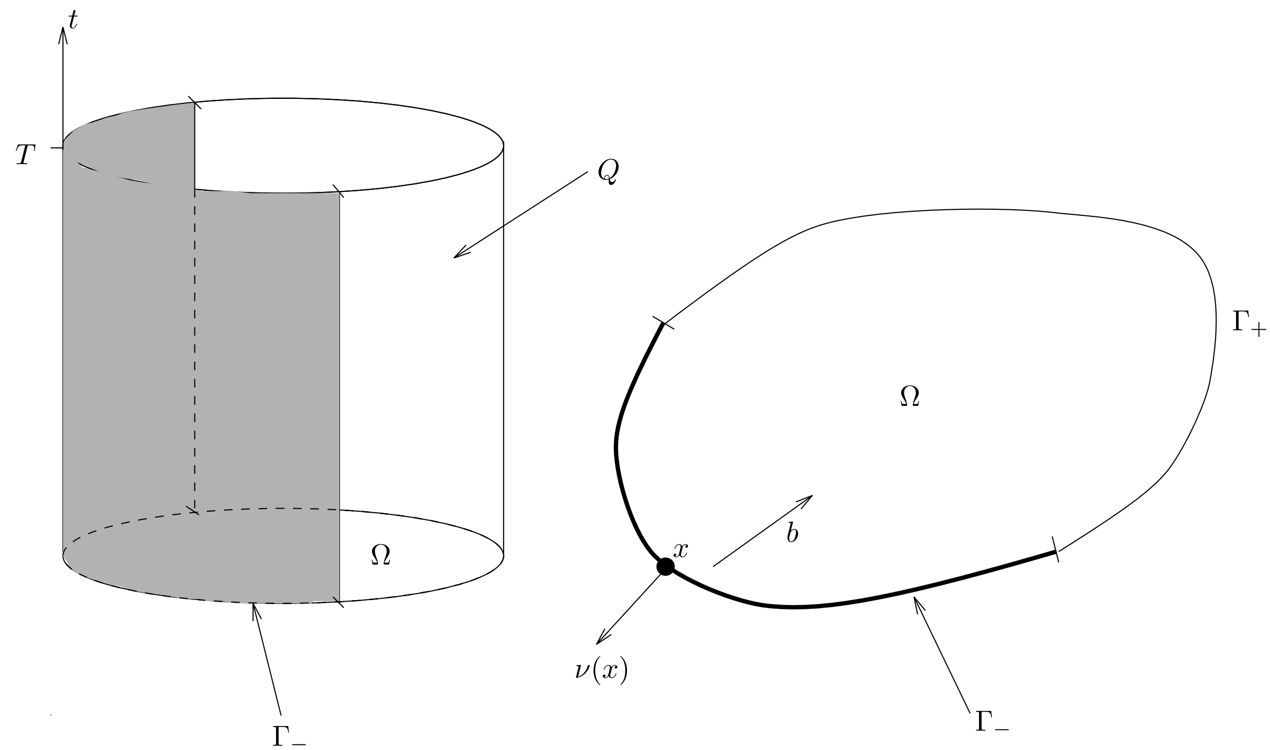
$b = (b_1, \dots, b_n)$ and $\nu(x)$ denotes the unit outward normal to Γ at $x \in \Gamma$. Γ_- will be called the *inflow boundary*. Its complement, $\Gamma_+ = \Gamma\backslash\Gamma_-$, will be referred to as the *outflow boundary*. It is important to note that unlike elliptic equations where a boundary condition is prescribed on the whole of $\partial\Omega$, and parabolic equations and second-order hyperbolic equations, such as the wave equation considered in the previous section, where a boundary condition is specified on the whole of $\Gamma \times [0,T] = \partial\Omega \times [0,T]$, in the initial boundary-value problem for the first-order hyperbolic equation stated above, a boundary condition is only imposed on part of the boundary, namely on $\Gamma_- \times [0,T]$; — else, the problem may have no solution, or if a solution exists continuous dependence of the solution on the data may fail to hold.

We shall assume that

$$b_i \in C^1(\overline{\Omega}), \;\; i = 1, \dots, n, \tag{152}$$

$$c \in C(\overline{Q}), \;\; f \in L_2(Q), \tag{153}$$

$$u_0 \in L_2(\Omega). \tag{154}$$



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, $f \in C^1(\overline{Q})$, $u_0 \in C^1(\overline{\Omega})$) 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 (152), a unique solution still exists, but the proof of this result is beyond the scope of these notes. We shall therefore assume henceforth that the initial-boundary-value problem (149)–(151) has a unique ('sufficiently smooth') solution, and consider the behaviour of the solution as it evolves as a function of time, t, from the given initial datum u_0.

We make the additional hypothesis:

$$c(x,t) - \frac{1}{2}\sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i}(x) \geq 0, \quad x \in \bar{\Omega}, \ \ t \in [0,T]. \tag{155}$$

By taking the inner product in $L_2(\Omega)$ of the equation (149) with $u(\cdot,t)$, performing partial integration and noting the boundary condition (150), we obtain:

$$\left(\frac{\partial u}{\partial t}(\cdot,t), u(\cdot,t)\right) + \left(c(\cdot,t) - \frac{1}{2}\sum_{i=1}^{n}\frac{\partial b_i}{\partial x_i}(\cdot), u^2(\cdot,t)\right) + \frac{1}{2}\int_{\Gamma_+}\left[\sum_{i=1}^{n} b_i(x)\nu_i(x)\right] u^2(x,t)\,\mathrm{d}s(x) = (f(\cdot,t), u(\cdot,t)), \tag{156}$$

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 (155) 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 (156) that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2 + \frac{1}{2}\int_{\Gamma_+}\left[\sum_{i=1}^n b_i(x)\nu_i(x)\right] u^2(x,t)\,\mathrm{d}s(x) \le (f(\cdot,t), u(\cdot,t)).$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned}(f(\cdot,t), u(\cdot,t)) &\le \|f(\cdot,t)\|\,\|u(\cdot,t)\| \\ &\le \frac{1}{2}\|f(\cdot,t)\|^2 + \frac{1}{2}\|u(\cdot,t)\|^2,\end{aligned}$$

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 inequality can be rewritten as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{-t}\|u(\cdot,t)\|^2\right) + \mathrm{e}^{-t}\int_{\Gamma_+}\left[\sum_{i=1}^n b_i(x)\nu_i(x)\right] u^2(x,t)\,\mathrm{d}s \le \mathrm{e}^{-t}\|f(\cdot,t)\|^2, \quad t\in[0,T].$$

By integrating this inequality with respect to t and noting the initial condition (151), we have that

$$\begin{aligned}\mathrm{e}^{-t}\|u(\cdot,t)\|^2 &+ \int_0^t \mathrm{e}^{-\tau}\int_{\Gamma_+}\left[\sum_{i=1}^n b_i(x)\nu_i(x)\right] u^2(x,\tau)\,\mathrm{d}s(x)\,\mathrm{d}\tau \\ &\le \|u_0\|^2 + \int_0^t \mathrm{e}^{-\tau}\|f(\cdot,\tau)\|^2\,\mathrm{d}\tau, \quad t\in[0,T].\end{aligned}$$

It therefore follows that

$$\begin{aligned}\|u(\cdot,t)\|^2 &+ \int_0^t \mathrm{e}^{t-\tau}\int_{\Gamma_+}\left[\sum_{i=1}^n b_i(x)\nu_i(x)\right] u^2(x,\tau)\,\mathrm{d}s(x)\,\mathrm{d}\tau \\ &\le \mathrm{e}^{t}\|u_0\|^2 + \int_0^t \mathrm{e}^{t-\tau}\|f(\cdot,\tau)\|^2\,\mathrm{d}\tau, \quad t\in[0,T]. \qquad (157)\end{aligned}$$

This, so called, energy inequality expresses the continuous dependence of the solution to (149)–(151) 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 (149)–(151), then $u := u_1 - u_2$ also solves (149)–(151), with $f \equiv 0$ and $u_0 \equiv 0$. Thus, by (157), $\|u(\cdot,t)\| = 0$, $t \in [0,T]$ and therefore $u \equiv 0$, i.e., $u_1 \equiv u_2$. The inequality (157) also reveals the importance of imposing a boundary condition on $\Gamma_- \times [0,T]$ only. On $\Gamma_+ \times [0,T]$, where $\Gamma_+ := \Gamma \setminus \Gamma_-$, the outflow part of $\Gamma = \partial\Omega$, the solution is 'controlled' by the data: the initial datum u_0, the source term f and the boundary condition on $\Gamma_- \times [0,T]$ (the latter does not appear explicitly in (157) because we assumed a zero boundary datum on $\Gamma_- \times [0,T]$). Note that the integrand in the second term on the left-hand side of (157) is nonnegative thanks to the definition of Γ_+.

Let us consider a particularly important case when

$$c \equiv 0, \quad f \equiv 0, \quad \text{and} \;\; \operatorname{div} b = \sum_{i=1}^{n} \frac{\partial b_i}{\partial x_i} \equiv 0,$$

where $b(x) = (b_1(x), \dots, b_n(x))$. Then, thanks to the identity (156), we have that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|^2 + \frac{1}{2}\int_{\Gamma_+} [b(x)\cdot\nu(x)]\, u^2(x,t)\,\mathrm{d}s(x) = 0,$$

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 can be viewed as an identity expressing 'conservation of energy' for the initial-boundary-value problem (149)–(151).

6.6 Explicit finite difference approximation

In this section we focus on a special case of the problem stated in the previous section, and describe a simple explicit finite difference scheme for the numerical solution of the constant-coefficient hyperbolic equation in one space dimension:

$$\frac{\partial u}{\partial t} + b\frac{\partial u}{\partial x} = f(x,t), \qquad x \in (0,1), \;\; t \in (0,T], \tag{158}$$

subject to the boundary and initial conditions

$$u(x,t) = 0, \qquad x \in \Gamma_-, \;\; t \in [0,T], \tag{159}$$

$$u(x,0) = u_0(x), \qquad x \in [0,1]. \tag{160}$$

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

$$u(0,t) = 0, \qquad t \in [0,T]. \tag{161}$$

To construct a finite difference approximation of (158)–(161) 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

$$x_j := j\Delta x, \;\; j = 0,\dots,J, \quad t_m := m\,\Delta t, \;\; m = 0,\dots,M.$$

At the mesh-point (x_j, t_m), (158) is approximated by the explicit finite difference scheme

$$\frac{U_j^{m+1} - U_j^m}{\Delta t} + b\,D_x^- U_j^m = f(x_j, t_m), \qquad j = 1,\dots,J, \quad m = 0,\dots,M-1, \tag{162}$$

subject to the boundary and initial condition, respectively:

$$U_0^m := 0, \qquad m = 0,\dots,M, \tag{163}$$

$$U_j^0 := u_0(x_j), \qquad j = 0,\dots,J. \tag{164}$$

Equivalently, this can be written as follows:

$$U_j^{m+1} = (1-\mu)U_j^m + \mu U_{j-1}^m + \Delta t\, f(x_j, t_m), \qquad \begin{cases} j = 1, \dots, J, \\ m = 0, \dots, M-1, \end{cases}$$

in conjunction with

$$\begin{aligned} U_0^m &:= 0, & m &= 0, \dots, M, \\ U_j^0 &:= u_0(x_j), & j &= 0, \dots, J, \end{aligned}$$

where

$$\mu := \frac{b\Delta t}{\Delta x};$$

μ is called the CFL (or Courant–Friedrichs–Lewy) number. The explicit finite difference scheme (162) is frequently called the *first-order upwind scheme*.

We shall explore the stability of this scheme in the discrete maximum norm. Suppose that $0 \le \mu \le 1$; then

$$\begin{aligned} \left|U_j^{m+1}\right| &\le (1-\mu)\left|U_j^m\right| + \mu\left|U_{j-1}^m\right| + \Delta t\left|f(x_j, t_m)\right| \\ &\le (1-\mu)\max_{0\le j\le J}\left|U_j^m\right| + \mu \max_{1\le j\le J+1}\left|U_{j-1}^m\right| + \Delta t \max_{0\le j\le J}\left|f(x_j, t_m)\right| \\ &= \max_{0\le j\le J}\left|U_j^m\right| + \Delta t \max_{0\le j\le J}\left|f(x_j, t_m)\right|. \end{aligned}$$

Thus we have that

$$\max_{0\le j\le J}\left|U_j^{m+1}\right| \le \max_{0\le j\le J}\left|U_j^m\right| + \Delta t \max_{0\le j\le J}\left|f(x_j, t_m)\right|.$$

Let us define the mesh-dependent norm

$$\|U\|_\infty := \max_{0\le j\le J}|U_j|;$$

then

$$\|U^{m+1}\|_\infty \le \|U^m\|_\infty + \Delta t\|f(\cdot, t_m)\|_\infty, \quad m = 0, \dots, M-1.$$

Summing through m, we get

$$\max_{1\le k\le M}\|U^k\|_\infty \le \|U^0\|_\infty + \sum_{m=0}^{M-1}\Delta t\|f(\cdot, t_m)\|_\infty, \tag{165}$$

which expresses the stability of the finite difference scheme (162)–(164) under the condition

$$0 \le \mu = \frac{b\Delta t}{\Delta x} \le 1. \tag{166}$$

Thus we have proved that the finite difference scheme (162)–(164) is conditionally stable, the condition being that the CFL number, μ, is in the interval $[0,1]$.

It is possible to show that the scheme (162)–(164) 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

$$\begin{aligned}(U^m, D_x^- U^m] &= \sum_{j=1}^{J} \Delta x\, U_j^m \frac{U_j^m - U_{j-1}^m}{\Delta x} \\ &= \frac{1}{2}\sum_{j=1}^{J}\{(U_j^m)^2 - (U_{j-1}^m)^2\} + \frac{\Delta x}{2}\sum_{j=1}^{J} \Delta x \left(\frac{U_j^m - U_{j-1}^m}{\Delta x}\right)^2 \qquad (167) \\ &= \frac{1}{2}(U_J^m)^2 + \frac{\Delta x}{2}\|D_x^- U^m]|^2.\end{aligned}$$

In addition, since

$$U_j^m = \frac{U_j^{m+1} + U_j^m}{2} - \frac{U_j^{m+1} - U_j^m}{2}, \quad m = 0, \ldots, M-1,$$

we have that

$$\left(\frac{U^{m+1} - U^m}{\Delta t}, U^m\right] = \frac{1}{2\Delta t}\left(\|U^{m+1}]|^2 - \|U^m]|^2\right) - \frac{\Delta t}{2}\left\|\frac{U^{m+1} - U^m}{\Delta t}\right]\Big|^2, \qquad m = 0, \ldots, M-1. \quad (168)$$

Thus, by taking the $(\cdot, \cdot]$-inner product of (162) with U^m and using (167) and (168), we find that

$$\begin{aligned}&\|U^{m+1}]|^2 + \Delta t\, b (U_J^m)^2 + b\,\Delta x\,\Delta t\,\|D_x^- U^m]|^2 - \|U^m]|^2 \\ &\quad - (\Delta t)^2\left\|\frac{U^{m+1} - U^m}{\Delta t}\right]\Big|^2 = 2\Delta t\,(f^m, U^m], \qquad m = 0, \ldots, M-1. \qquad (169)\end{aligned}$$

First suppose that $f \equiv 0$; then,

$$\frac{U^{m+1} - U^m}{\Delta t} = -b\, D_x^- U^m,$$

and by substituting this into the last term on the left-hand side of the equality (169) we have 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, we have that

$$\|U^k]|^2 + \sum_{m=0}^{k-1} \Delta t\, b\, |U_J^m|^2 + b\,\Delta x\,(1 - \mu)\sum_{m=0}^{k-1} \Delta t\,\|D_x^- U^m]|^2 = \|U^0]|^2, \quad k = 1, \ldots, M, \qquad (170)$$

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 (157), and expresses 'conservation of energy' in the discrete sense. More generally, for $0 \le \mu \le 1$, (170) 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

$$\left\| \frac{U^{m+1} - U^m}{\Delta t} \right]\Big|^2 = \|f^m - bD_x^- U^m]|^2 \le \{\|f^m]| + b\|D_x^- U^m]|\}^2$$
$$\le \left(1 + \frac{1}{\epsilon'}\right) \|f^m]|^2 + (1+\epsilon')b^2 \|D_x^- U^m]|^2, \quad \epsilon' > 0,$$

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 the equality (169) that

$$\|U^{m+1}]|^2 + \Delta t\, b\, |U_n^m|^2 + b\, \Delta x\, \Delta t \left[1 - (1+\epsilon')\frac{b\Delta t}{\Delta x}\right] \|D_x^- U^m]|^2$$
$$\le \Delta t \left[\left(1 + \frac{1}{\epsilon'}\right)\Delta t + 1\right] \|f^m]|^2 + (1 + \Delta t)\|U^m]|^2.$$

Letting $\epsilon = 1 - 1/(1+\epsilon') \in (0,1)$ and assuming that

$$0 \le \mu = \frac{b\,\Delta t}{\Delta x} \le 1 - \epsilon,$$

we have, for $m = 0, \ldots, M-1$, that

$$\|U^{m+1}]|^2 + \Delta t\, b\, |U_J^m|^2 \le \|U^m]|^2 + \Delta t \left(1 + \frac{\Delta t}{\epsilon}\right) \|f^m]|^2 + \Delta t \|U^m]|^2.$$

Upon summation of this inequality over $m = 0, \ldots, k-1$, we deduce that

$$\|U^k]|^2 + \left(\sum_{m=0}^{k-1} \Delta t\, b\, |U_J^m|^2\right) \le \|U^0]|^2 + \left(1 + \frac{\Delta t}{\epsilon}\right) \sum_{m=0}^{k-1} \Delta t\, \|f^m]|^2 + \sum_{m=0}^{k-1} \Delta t\, \|U^m]|^2 \qquad (171)$$

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 nonnegative real numbers such that the sequence (c_k) is nondecreasing 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 the inequality (171) with

$$
\begin{aligned}
a_k &:= \|U^k]|^2, \quad k \geq 0,\\
b_k &:= \sum_{m=0}^{k-1} \Delta t\, b\, |U_J^m|^2, \quad k \geq 1; \;\; b_0 = 0,\\
c_k &:= \|U^0]|^2 + \left(1 + \frac{\Delta t}{\epsilon}\right) \sum_{m=0}^{k-1} \Delta t\, \|f^m]|^2, \quad k \geq 1; \;\; c_0 = \|U^0]|^2,\\
d_k &:= \Delta t, \quad k = 1, 2, \ldots, M,
\end{aligned}
$$

we obtain,

$$
\|U^k]|^2 + \sum_{m=0}^{k-1} \Delta t\, b\, |U_J^m|^2 \leq \mathrm{e}^{t_k} \left(\|U^0]|^2 + \left(1 + \frac{\Delta t}{\epsilon}\right) \sum_{m=0}^{k-1} \Delta t \|f^m]|^2 \right), \quad k = 1, \ldots, M,
$$

where $t_k := k\Delta t$. Hence we deduce stability of the scheme, in the sense that

$$
\max_{1\leq k\leq M} \left(\|U^k]|^2 + \sum_{m=0}^{k-1} \Delta t\, b\, |U_J^m|^2 \right) \leq \mathrm{e}^{T} \left(\|U^0]|^2 + \left(1 + \frac{\Delta t}{\epsilon}\right) \sum_{m=0}^{M-1} \Delta t \|f^m]|^2 \right). \tag{172}
$$

An error bound for the difference scheme (162)–(164) is easily derived from its stability. For the sake of simplicity we shall focus on the error analysis of the scheme in the $\|\cdot\|_\infty$ norm, which we shall deduce from the stability of the scheme in the $\|\cdot\|_\infty$ norm for $\mu \in [0, 1]$.

We define the global error, e_j^m, and the consistency error, T_j^m, of the scheme, respectively, by

$$
\begin{aligned}
e_j^m &:= u(x_j, t_m) - U_j^m,\\
T_j^m &:= \frac{u(x_j, t_{m+1}) - u(x_j, t_m)}{\Delta t} + bD_x^- u(x_j, t_m) - f(x_j, t_m).
\end{aligned}
$$

It is easily seen that

$$
\begin{aligned}
\frac{e_j^{m+1} - e_j^m}{\Delta t} + bD_x^- e_j^m = T_j^m, &\qquad j = 1, \ldots, J, \quad m = 0, \ldots, M-1,\\
e_0^m = 0, &\qquad m = 0, \ldots, M,\\
e_j^0 = 0, &\qquad j = 0, \ldots, J.
\end{aligned}
$$

By virtue of the stability inequality established in the first part of this section we have that, for $\mu \in [0, 1]$,

$$
\max_{1\leq m\leq M} \|e^m\|_\infty \leq \sum_{k=0}^{M-1} \Delta t \|T^m\|_\infty. \tag{173}
$$

By Taylor series expansion of T_j^m about the point (x_j, t_m) it follows that

$$
T_j^m = \frac{1}{2}\Delta t \frac{\partial^2 u}{\partial t^2}(x_j, \tau^m) + \frac{1}{2} b\, \Delta x\, \frac{\partial^2 u}{\partial x^2}(\xi_j, t_m), \quad \tau^m \in (t_m, t_{m+1}), \quad \xi_j \in (x_{j-1}, x_j),
$$

and therefore also

$$
\left|T_j^m\right| \leq \frac{1}{2}(\Delta t M_{2t} + b\, \Delta x\, M_{2x}),
$$

where

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

By defining $\mathcal{M} := \max(M_{2t}, M_{2x})$, we have that

$$\left|T_j^m\right| \leq \frac{1}{2}\mathcal{M}(\Delta t + b\,\Delta x) \quad (= \mathcal{O}(\Delta x + \Delta t)). \tag{174}$$

Thus, by (173), we arrive at the error bound

$$\max_{1\leq m\leq M} \|u^m - U^m\|_\infty \leq \frac{1}{2}T\mathcal{M}(\Delta t + b\,\Delta x),$$

where $u^m := u(\cdot, t_m)$ and $u_j^m := u(x_j, t_m)$. Therefore the scheme (162)–(164) is first-order convergent with respect to both Δx and Δt.

Analogously, using the stability result (171) in the discrete L_2-norm $\|\cdot]|$, (174) implies that

$$\max_{1\leq m\leq M} \|u^m - U^m]| \leq c_\epsilon^\star \cdot (\Delta t + b\,\Delta x),$$

where $c_\epsilon^\star = \frac{1}{2}\mathrm{e}^{T/2}(1 + T/\epsilon)^{1/2}T^{1/2}\mathcal{M}$.

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 finite difference schemes on nonuniform meshes. We shall however continue to operate in the univariate setting and discuss, instead, a different extension of the problem considered here: a scalar *nonlinear* first-order hyperbolic partial differential equation in one space dimension.

6.7 Finite difference approximation of scalar nonlinear hyperbolic conservation laws

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 the nonlinear partial differential equation **Lecture 16**

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}f(u) = 0 \quad \text{for } (x,t) \in \mathbb{R} \times (0,\infty), \tag{175}$$

where $u = u(x,t)$, 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 that $f(0) = f'(0) = 0$, and $f''(s) \geq 0$ for all $s \in \mathbb{R}$. Under these hypotheses f' is a nondecreasing function, whereby $f'(s) \geq 0$ for all $s \geq 0$. We shall assume further that $|f'(s)| \leq f'(|s|)$ 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 $u \in C^1(\mathbb{R} \times [0,T])$ to the initial-value problem exists, then thanks to the chain rule the equation (175) can be rewritten as

$$\frac{\partial u}{\partial t} + f'(u)\frac{\partial u}{\partial x} = 0 \quad \text{for } (x,t) \in \mathbb{R} \times (0,T]. \tag{176}$$

Motivated by the construction of the first-order upwind scheme in the previous section, we decompose $f'(u)$ into its nonnegative and nonpositive parts, as follows:

$$f'(u) = [f'(u)]_+ + [f'(u)]_-,$$

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 (176) as follows:

$$\frac{\partial u}{\partial t} + [f'(u)]_+ \frac{\partial u}{\partial x} + [f'(u)]_- \frac{\partial u}{\partial x} = 0 \quad \text{for } (x,t) \in \mathbb{R} \times (0,T]. \tag{177}$$

We approximate (177) by the following finite difference scheme

$$\begin{aligned} \frac{U_j^{m+1} - U_j^m}{\Delta t} + [f'(U_j^m)]_+ D_x^- U_j^m + [f'(U_j^m)]_- D_x^+ U_j^m = 0, \qquad j \in \mathbb{Z}, \quad m = 0, \dots, M-1, \\ U_j^0 := u_0(x_j), \quad j \in \mathbb{Z}, \end{aligned} \tag{178}$$

where $\Delta t = T/M$, $M \geq 1$, and $\mathbb{Z}$ is the set of all integers.

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 (165) (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 (176) under consideration), in the sense that

$$\max_{1 \leq k \leq M} \|U^k\|_\infty \leq \|U^0\|_\infty, \tag{179}$$

where now $\|V\|_\infty := \max_{j \in \mathbb{Z}} |V_j|$.

To this end, we rewrite $(178)_1$ as follows:

$$\begin{aligned} U_j^{m+1} &= U_j^m - \frac{[f'(U_j^m)]_+ \Delta t}{\Delta x}(U_j^m - U_{j-1}^m) - \frac{[f'(U_j^m)]_- \Delta t}{\Delta x}(U_{j+1}^m - U_j^m) \\ &= \left(1 - \frac{\Delta t}{\Delta x}\left([f'(U_j^m)]_+ - [f'(U_j^m)]_-\right)\right) U_j^m + \frac{[f'(U_j^m)]_+ \Delta t}{\Delta x} U_{j-1}^m + \frac{-[f'(U_j^m)]_- \Delta t}{\Delta x} U_{j+1}^m \\ &= \left(1 - \frac{|f'(U_j^m)|\, \Delta t}{\Delta x}\right) U_j^m + \frac{[f'(U_j^m)]_+ \Delta t}{\Delta x} U_{j-1}^m + \frac{-[f'(U_j^m)]_- \Delta t}{\Delta x} U_{j+1}^m \end{aligned} \tag{180}$$

for all $j \in \mathbb{Z}$ and all $m = 0, \dots, M-1$. Suppose that the following CFL condition holds:

$$\frac{f'(\|U^0\|_\infty)\, \Delta t}{\Delta x} \leq 1. \tag{181}$$

Suppose further, as an inductive hypothesis, that, for some $m \geq 0$,

$$\frac{f'(\|U^k\|_\infty)\, \Delta t}{\Delta x} \leq 1 \qquad \text{for all } k = 0, \dots, m. \tag{182}$$

Thanks to (181) this inductive hypothesis is satisfied for $m = 0$. Suppose, for the inductive step, that (182) has already been shown to hold for some $m \geq 0$. Because of the assumptions imposed on the function f, we have that $|f'(U_j^m)| \leq f'(|U_j^m|) \leq f'(\|U^m\|_\infty)$ for all $j \in \mathbb{Z}$. It then follows from (182) with $k = m$ that

$$\frac{|f'(U_j^m)|\, \Delta t}{\Delta x} \leq 1 \qquad \text{for all } j \in \mathbb{Z},$$

and then (180) implies that

$$
\begin{aligned}
|U_j^{m+1}| &\le \left(1 - \frac{|f'(U_j^m)|\,\Delta t}{\Delta x}\right)|U_j^m| + \frac{[f'(U_j^m)]_+\,\Delta t}{\Delta x}|U_{j-1}^m| + \frac{-[f'(U_j^m)]_-\Delta t}{\Delta x}|U_{j+1}^m| \\
&\le \left(1 - \frac{|f'(U_j^m)|\,\Delta t}{\Delta x}\right)\|U^m\|_\infty + \frac{[f'(U_j^m)]_+\,\Delta t}{\Delta x}\|U^m\|_\infty + \frac{-[f'(U_j^m)]_-\,\Delta t}{\Delta x}\|U^m\|_\infty \\
&= \left(1 - \frac{|f'(U_j^m)|\,\Delta t}{\Delta x}\right)\|U^m\|_\infty + \frac{|f'(U_j^m)|\,\Delta t}{\Delta x}\|U^m\|_\infty = \|U^m\|_\infty
\end{aligned}
$$

for all $j \in \mathbb{Z}$. Therefore,

$$
\|U^{m+1}\|_\infty \le \|U^m\|_\infty. \tag{183}
$$

To complete the inductive step it remains to show that (182) holds with m replaced by $m+1$. By (183) and the fact that f' is nondecreasing imply that

$$
\frac{f'(\|U^{m+1}\|_\infty)\,\Delta t}{\Delta x} \le \frac{f'(\|U^m\|_\infty)\,\Delta t}{\Delta x} \le 1. \tag{184}
$$

The inequality (184) shows that (182) holds with m replaced by $m+1$, which then completes the inductive step. Thus we have shown that, under the CFL condition (181),

$$
\|U^{m+1}\|_\infty \le \|U^m\|_\infty \le \cdots \le \|U^0\|_\infty \qquad \text{for all } m = 0, 1, \ldots, M-1, \tag{185}
$$

which completes the proof of the assertion that the sequence $\{U_j^m\}_{j\in\mathbb{Z},\, 0\le m\le M}$ of finite difference approximations generated by the scheme is bounded; in particular (179) has been shown to hold.

Assuming that u has continuous and bounded second partial derivatives with respect to x and t defined on $\mathbb{R} \times [0, T]$, it can be shown that

$$
\max_{1\le m\le M} \|u^m - U^m\|_\infty = \mathcal{O}(\Delta x + \Delta t),
$$

as in the case of the linear first-order hyperbolic equation considered in the previous section, but we shall not include the proof of this result here. One of the main difficulties in proving such an error bound is that now, unlike the linear first-order hyperbolic equation where a bound such as (185) would, thanks to the linearity of the finite difference scheme, automatically imply the stability of the scheme, in the case of the nonlinear partial differential equation considered here this is not the case: if $\{U_j^m\}$ and $\{V_j^m\}$ are two sequences of numerical solutions generated by the scheme from initial data $\{U_j^0\}$ and $\{V_j^0\}$ the inequality (185) does not automatically imply that

$$
\|U^{m+1} - V^{m+1}\|_\infty \le \|U^m - V^m\|_\infty \le \cdots \le \|U^0 - V^0\|_\infty \qquad \text{for all } m = 0, 1, \ldots, M-1,
$$

which then complicates the convergence analysis of the finite difference scheme. A further technical complication is that, given a smooth initial function u_0 it need not be true that the solution u remains a smooth functions of x an t over the whole of $\mathbb{R} \times [0, T]$; there may be a time $t_* \in (0, T)$ at which the function $x \in \mathbb{R} \mapsto u(x, t_*)$ becomes discontinuous. Then the partial differential equation (175) no longer makes sense in the form in which it is stated, and a suitable weak formulation of the problem needs to be considered instead. The mathematical analysis of numerical approximations of weak solutions to nonlinear hyperbolic conservation laws, such as (175), is beyond the scope of these lecture notes; for further details in this direction we refer the reader to the book by R. LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations,* SIAM, 2007.