Numerical Methods in Differential Equations

based on notes written by Endre Süli (University of Oxford),
modified by Timo Sprekeler

The original notes by Endre Süli are available from
`http://people.maths.ox.ac.uk/suli/nsde.pdf` and
`http://people.maths.ox.ac.uk/suli/nspde.2021.2022.pdf`

Some good books for this course:

[1] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations.* (Cambridge University Press, second edition, 2009).

[2] B.S. Jovanović and E. Süli, *Analysis of Finite Difference Schemes for Linear Partial Differential Equations with Generalized Solutions.* (Springer, 2014).

[3] R. LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations.* (SIAM, 2007).

[4] K.W. Morton and D.F. Mayers, *Numerical Solution of Partial Differential Equations: An Introduction.* (Cambridge University Press, second edition, 2012).

last updated: April 14, 2023

Contents

Part I
Ordinary Differential Equations (ODEs)

1 Preliminaries: Picard's Theorem

Ordinary differential equations frequently occur as mathematical models in many branches of science, engineering, and economics. Unfortunately it is seldom that these equations have solutions that can be expressed in closed form, so it is common to seek approximate solutions by means of numerical methods; nowadays this can usually be achieved very inexpensively to high accuracy and with a reliable bound on the error between the analytical solution and its numerical approximation. We shall be concerned with the construction and the analysis of numerical methods for first-order differential equations of the form

$$y'(x) = f(x, y(x)) \tag{1}$$

for the real-valued function y of the variable $x \in \mathbb{R}$, where $y' := \frac{\mathrm{d}y}{\mathrm{d}x}$. In order to select a particular integral (i.e., a particular solution) from the infinite family of solution curves that constitute the general solution to (1), the differential equation will be considered in tandem with an **initial condition** (we sometimes simply write i.c.): given two real numbers $x_0, y_0 \in \mathbb{R}$, we seek a solution to (1) for $x > x_0$ such that

$$y(x_0) = y_0. \tag{2}$$

The differential equation (1) together with the initial condition (2) is called an **initial-value problem (IVP)**. The motivation for this terminology is that in applications the variable x usually plays the role of time, and the **initial value**, y_0, of the process whose evolution is modelled by the differential equation over an interval of time $x \in [x_0, X_M]$ is then known at the initial time $x = x_0$.

In general, even if $f(\cdot, \cdot)$ is a continuous function, there is no guarantee that the initial-value problem (1)–(2) possesses a unique solution.[1] Fortunately, under a further mild condition on f, the existence and uniqueness of a solution to (1)–(2) can be ensured:

Theorem 1 (**Picard's Theorem**[2]) *Suppose that $f(\cdot, \cdot)$ is a continuous function of its arguments in a region $U \subseteq \mathbb{R}^2$ which contains the rectangle*

$$\mathsf{R} := [x_0, X_M] \times [y_0 - Y_M, y_0 + Y_M],$$

where $X_M > x_0$ and $Y_M > 0$ are constants. Suppose also, that there exists a constant $L > 0$ such that

$$|f(x, z) - f(x, \tilde{z})| \leq L|z - \tilde{z}| \qquad \forall (x, z), (x, \tilde{z}) \in \mathsf{R}. \tag{3}$$

Finally, suppose that

$$M(X_M - x_0) \leq Y_M, \quad \text{where} \quad M := \max_{(x,z)\in\mathsf{R}} |f(x, z)|.$$

Then, there exists a unique continuously differentiable function $y : [x_0, X_M] \to \mathbb{R}$ satisfying (1)–(2).

Remark 1 *In the situation of Theorem 1, we have that the graph of the unique solution y lies in R, i.e., $(x, y(x)) \in \mathsf{R}$ for any $x \in [x_0, X_M]$.*

Indeed, if this were not true, then by continuity of y there exists $x_* \in (x_0, X_M)$ such that $|y(x_*) - y_0| = Y_M$ and $|y(x) - y_0| < Y_M$ for all $x \in [x_0, x_*)$. But this implies

$$|y(x_*) - y_0| \leq \int_{x_0}^{x_*} |y'(x)| \, \mathrm{d}x = \int_{x_0}^{x_*} |f(x, y(x))| \, \mathrm{d}x \leq M(x_* - x_0) < M(X_M - x_0) \leq Y_M,$$

which contradicts $|y(x_*) - y_0| = Y_M$.

[1] Example: $y'(x) = (y(x))^{\frac{2}{3}}$ for $x > 0$, and i.c. $y(0) = 0$; this has more than one solution: $y_1(x) := 0$ and $y_2(x) := \frac{1}{27}x^3$.
[2] Emile Picard (1856–1941)

The condition (3) is called a **Lipschitz condition**[3], and L is called a **Lipschitz constant** for f. We shall not dwell on the proof of Picard's Theorem; for details, see any good textbook on the theory of ODEs (see, e.g., P. J. Collins, *Differential and Integral Equations*, Oxford University Press, 2006). The essence of the proof is to consider the sequence of functions $(y_n)_{n\in\mathbb{N}_0}$, defined recursively through what is known as the **Picard Iteration**:

$$\begin{aligned} &y_0(x) \equiv y_0, \\ &y_n(x) := y_0 + \int_{x_0}^{x} f(t, y_{n-1}(t))\,\mathrm{d}t, \qquad n \in \mathbb{N} = \{1, 2, \dots\} \end{aligned} \tag{4}$$

and show, using the conditions of the theorem, that $(y_n)_{n\in\mathbb{N}_0}$, as a sequence of continuous functions, converges uniformly on the interval $[x_0, X_M]$ to a continuous function $y : [x_0, X_M] \to \mathbb{R}$ (that is, $\sup_{x\in[x_0,X_M]}|y_n(x) - y(x)| \longrightarrow 0$ as $n \to \infty$), and that y satisfies

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\,\mathrm{d}t.$$

This then implies that y is continuously differentiable on $[x_0, X_M]$ and it satisfies the differential equation (1) and the initial condition (2). The uniqueness of the solution follows from the Lipschitz condition.

Picard's Theorem has a natural extension to IVPs for systems of m differential equations of the form

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \qquad \mathbf{y}(x_0) = \mathbf{y}_0, \tag{5}$$

where $\mathbf{y}_0 \in \mathbb{R}^m$, $\mathbf{f} : [x_0, X_M] \times \mathbb{R}^m \to \mathbb{R}^m$, and we seek a solution $\mathbf{y} : [x_0, X_M] \to \mathbb{R}^m$. Introducing the Euclidean norm $\|\cdot\| : \mathbb{R}^m \to [0, \infty)$ on $\mathbb{R}^m$ by

$$\|\mathbf{u}\| := \sqrt{\sum_{i=1}^{m} |u_i|^2}, \quad \text{for } \mathbf{u} = (u_1, \dots, u_m)^{\mathrm{T}} \in \mathbb{R}^m,$$

we can state the following result.

Theorem 2 (**Picard's Theorem (version for systems)**) *Suppose that $\mathbf{f}(\cdot,\cdot)$ is a continuous function of its arguments in a region $U \subseteq \mathbb{R}^{1+m}$ which contains the set*

$$\mathsf{R} = \{(x, \mathbf{z}) \in \mathbb{R} \times \mathbb{R}^m \,:\, x \in [x_0, X_M], \quad \|\mathbf{z} - \mathbf{y}_0\| \le Y_M\},$$

where $X_M > x_0$ and $Y_M > 0$ are constants. Suppose also that there exists a constant $L > 0$ such that

$$\|\mathbf{f}(x, \mathbf{z}) - \mathbf{f}(x, \tilde{\mathbf{z}})\| \le L\|\mathbf{z} - \tilde{\mathbf{z}}\| \qquad \forall (x, \mathbf{z}), (x, \tilde{\mathbf{z}}) \in \mathsf{R}. \tag{6}$$

Finally, suppose that

$$M(X_M - x_0) \le Y_M, \quad \textit{where} \quad M := \max_{(x,\mathbf{z})\in\mathsf{R}} \|\mathbf{f}(x, \mathbf{z})\|.$$

Then, there exists a unique continuously differentiable function $\mathbf{y} : [x_0, X_M] \to \mathbb{R}^m$ which satisfies (5).

A sufficient condition for (6) is that $\mathbf{f}$ is continuous on R, differentiable at each point $(x, \mathbf{z})$ in $\mathrm{int}(\mathsf{R})$, the interior of R, and there exists an $L > 0$ such that

$$\|\partial_{\mathbf{z}}\mathbf{f}(x, \mathbf{z})\| \le L \qquad \text{for all } (x, \mathbf{z}) \in \mathrm{int}(\mathsf{R}), \tag{7}$$

where $\partial_{\mathbf{z}}\mathbf{f} := \frac{\partial \mathbf{f}}{\partial \mathbf{z}}$ denotes the $m \times m$ Jacobi matrix of the function $\mathbf{R}^m \ni \mathbf{z} \mapsto \mathbf{f}(x, \mathbf{z}) \in \mathbf{R}^m$, and $\|\cdot\| : \mathbb{R}^{m\times m} \to [0, \infty)$ is the matrix norm induced by the Euclidean vector norm on $\mathbb{R}^m$ (i.e., for

[3] Rudolf Lipschitz (1832–1903)

$A \in \mathbb{R}^{m\times m}$ have $\|A\| := \sup_{\mathbf{x}\in\mathbb{R}^m\setminus\{0\}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}$ – note the norms on the right-hand side are the Euclidean vector norm on $\mathbb{R}^m$). Indeed, when (7) holds, the Mean-Value Theorem implies that (6) is also valid. The converse of this statement is not true: the function

$$\mathbf{f} : \mathbb{R}^{1+m} \to \mathbb{R}^m, \quad \mathbf{f}(x, \mathbf{z}) := f(x, z_1, \ldots, z_m) := \begin{pmatrix} |z_1| \\ \vdots \\ |z_m| \end{pmatrix}$$

with $x_0 = 0$ and $\mathbf{y}_0 = \mathbf{0}$, satisfies (6) but violates (7) because $\mathbf{z} \mapsto \mathbf{f}(x, \mathbf{z})$ is not differentiable at $\mathbf{z} = \mathbf{0}$.

As the counter-example in the footnote 1 on page 2 indicates, the expression $|z-\tilde{z}|$ in (3) and $\|\mathbf{z}-\tilde{\mathbf{z}}\|$ in (6) cannot be replaced by expressions of the form $|z-\tilde{z}|^\alpha$ and $\|\mathbf{z}-\tilde{\mathbf{z}}\|^\alpha$, respectively, where $0 < \alpha < 1$, for otherwise the uniqueness of the solution to the corresponding initial-value problem cannot be guaranteed.

We conclude this section by introducing the notion of *stability*.

Definition 1 *We define the following notions of stability:*

*(i) A solution $\mathbf{y} = \mathbf{v}(x)$ is said to be **stable** on the interval $[x_0, X_M]$ if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $\mathbf{z} \in \mathbb{R}^m$ satisfying $\|\mathbf{v}(x_0) - \mathbf{z}\| < \delta$, a solution $\mathbf{w}$ to*

$$\mathbf{w}'(x) = \mathbf{f}(x, \mathbf{w}(x)), \qquad \mathbf{w}(x_0) = \mathbf{z} \tag{8}$$

is defined for all $x \in [x_0, X_M]$ and satisfies $\|\mathbf{v}(x) - \mathbf{w}(x)\| < \varepsilon$ for all x in $[x_0, X_M]$.

*(ii) A solution $\mathbf{y} = \mathbf{v}(x)$ which is stable on $[x_0, \infty)$ (i.e. stable on $[x_0, X_M]$ for each X_M and with δ independent of X_M) is said to be **stable in the sense of Lyapunov**.*

(iii) If in addition to (ii) there holds

$$\lim_{x\to\infty} \|\mathbf{v}(x) - \mathbf{w}(x)\| = 0,$$

*then the solution $\mathbf{y} = \mathbf{v}(x)$ is called **asymptotically stable**.*

Using this definition, we can state the following theorem.

Theorem 3 *Under the hypotheses of Picard's Theorem, the (unique) solution $\mathbf{y} = \mathbf{v}(x)$ to the initial-value problem (5) is stable on the interval $[x_0, X_M]$, (where we assume that $-\infty < x_0 < X_M < \infty$).*

Proof: For $\mathbf{z} \in \mathbb{R}^m$, let $\mathbf{w}$ be the solution to (8). First, note that integrating the DEs for $\mathbf{v}$ and $\mathbf{w}$ over the interval $[x_0, x]$ yields

$$\mathbf{v}(x) = \mathbf{v}(x_0) + \int_{x_0}^{x} \mathbf{f}(t, \mathbf{v}(t))\,\mathrm{d}t, \qquad \mathbf{w}(x) = \mathbf{z} + \int_{x_0}^{x} \mathbf{f}(t, \mathbf{w}(t))\,\mathrm{d}t \qquad \forall x \in [x_0, X_M].$$

Using triangle inequality and the fact that $\|\int_a^b \mathbf{g}(t)\,\mathrm{d}t\| \le \int_a^b \|\mathbf{g}(t)\|\,\mathrm{d}t$ for any $\mathbf{g} : [a, b] \to \mathbb{R}^m$, we have

$$\begin{aligned} \|\mathbf{v}(x) - \mathbf{w}(x)\| &\le \|\mathbf{v}(x_0) - \mathbf{z}\| + \int_{x_0}^{x} \|\mathbf{f}(t, \mathbf{v}(t)) - \mathbf{f}(t, \mathbf{w}(t))\|\,\mathrm{d}t \\ &\le \|\mathbf{v}(x_0) - \mathbf{z}\| + L\int_{x_0}^{x} \|\mathbf{v}(t) - \mathbf{w}(t)\|\,\mathrm{d}t \end{aligned} \tag{9}$$

for any $x \in [x_0, X_M]$. Setting $A(x) := \|\mathbf{v}(x) - \mathbf{w}(x)\|$ and $a := \|\mathbf{v}(x_0) - \mathbf{z}\|$, we can rewrite (9) as

$$A(x) \le a + L\int_{x_0}^{x} A(t)\,\mathrm{d}t \qquad \forall x \in [x_0, X_M]. \tag{10}$$

Multiplying (10) by e^{-Lx}, we find that

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[\mathrm{e}^{-Lx}\int_{x_0}^{x} A(t)\,\mathrm{d}t + \frac{a}{L}e^{-Lx}\right] \le 0 \qquad \forall x \in [x_0, X_M], \tag{11}$$

and hence,

$$\mathrm{e}^{-Lx}\int_{x_0}^{x} A(t)\,\mathrm{d}t + \frac{a}{L}e^{-Lx} \le \mathrm{e}^{-Lx_0}\int_{x_0}^{x_0} A(t)\,\mathrm{d}t + \frac{a}{L}e^{-Lx_0} = \frac{a}{L}\mathrm{e}^{-Lx_0} \qquad \forall x \in [x_0, X_M],$$

i.e., (multiply by Le^{Lx})

$$L\int_{x_0}^{x} A(t)\,\mathrm{d}t \le a\left(\mathrm{e}^{L(x-x_0)} - 1\right). \tag{12}$$

Substituting (12) into (10) gives

$$A(x) \le a\,\mathrm{e}^{L(x-x_0)} \qquad \forall x \in [x_0, X_M]. \tag{13}$$

The implication "(10) $\Rightarrow$ (13)" is usually referred to as **Gronwall Lemma**. Returning to our original notation, we deduce from (13) that

$$\|\mathbf{v}(x) - \mathbf{w}(x)\| \le \|\mathbf{v}(x_0) - \mathbf{z}\|e^{L(x-x_0)} \le e^{L(X_M-x_0)}\|\mathbf{v}(x_0) - \mathbf{z}\| \qquad \forall x \in [x_0, X_M]. \tag{14}$$

Thus, given $\varepsilon > 0$ as in Definition 1, we choose $\delta = \varepsilon\, e^{-L(X_M-x_0)}$ to deduce stability. $\diamond$

To conclude this section, we observe that if either $x_0 = -\infty$ or $X_M = +\infty$, the statement of Theorem 3 is *false*. For example, the trivial solution $y \equiv 0$ to the differential equation $y' = y$ is unstable on $[x_0, \infty)$ for any $x_0 > -\infty$. Let us consider the IVP

$$y'(x) = \lambda y(x), \qquad y(0) = 1, \tag{15}$$

with $\lambda \in \mathbb{R}$, which has the unique solution $y(x) = e^{\lambda x}$. Noting that for $c \in \mathbb{R}$, the IVP $w' = \lambda w$, $w(0) = c$ has the unique solution $w(x) = ce^{\lambda x}$ and hence, $|y(x) - w(x)| = |1 - c|e^{\lambda x}$. We see that y is unstable on $[0, \infty)$ when $\lambda > 0$; stable in the sense of Lyapunov when $\lambda \le 0$; and asymptotically stable for $\lambda < 0$.

Remark 2 *The stability concepts can be extended to the complex case, i.e., when f is a complex-valued function and $y_0 \in \mathbb{C}$, in which case the solution y to $y'(x) = f(x, y(x))$, $y(x_0) = y_0$ is a function from one real variable into $\mathbb{C}$ (analogously for systems). For $\lambda \in \mathbb{C}$, the solution to the IVP* (15)*, which is given by $y(x) = e^{\lambda x} = e^{(\mathrm{Re}\,\lambda)x}e^{i(\mathrm{Im}\,\lambda)x}$, is unstable for $\mathrm{Re}\,\lambda > 0$; stable in the sense of Lyapunov for $\mathrm{Re}\,\lambda \le 0$ and asymptotically stable for $\mathrm{Re}\,\lambda < 0$.*

In the next section we shall consider numerical methods for the approximate solution of the IVP (1)–(2). Since everything we shall say has a straightforward extension to the case of the system (5), for the sake of notational simplicity we shall restrict ourselves to considering a single ODE (i.e., $m = 1$). We shall suppose throughout that the function f satisfies the conditions of Picard's Theorem on the rectangle R and that the IVP has a unique solution defined on $[x_0, X_M]$, $-\infty < x_0 < X_M < \infty$. We begin by discussing one-step methods; this will be followed in subsequent sections by multi-step methods.

2 One-step methods

2.1 Euler's method and its relatives: the θ-method

The simplest example of a one-step method for the numerical solution of the IVP (1)–(2) is Euler's method.[4] Suppose that the IVP (1)–(2) is to be solved on the interval $[x_0, X_M]$. For $N \in \mathbb{N}$, we divide this interval by the $N+1$ **mesh-points**

$$x_0, \quad x_1 = x_0 + h, \quad x_2 = x_0 + 2h, \quad \cdots \quad , \quad x_N = x_0 + Nh = X_M, \qquad \left[\text{where } h := \frac{X_M - x_0}{N}\right].$$

The number $h > 0$ is called the **step size**. Now let us suppose that, for each $n \in \{0, 1, \ldots, N\}$, we seek a numerical approximation y_n to $y(x_n)$, the value of the solution at the mesh point x_n. As $y(x_0) = y_0$ is known, let us suppose that we have already calculated y_n, up to some n, $0 \leq n \leq N-1$; we define

$$y_{n+1} = y_n + hf(x_n, y_n), \qquad n \in \{0, 1, \ldots, N-1\}, \qquad y_0 = y(x_0).$$

Thus, taking in succession $n = 0, 1, \ldots, N-1$, one step at a time, the approximate values y_n at the mesh points x_n can be easily obtained. This numerical method is known as **Euler's method**.

A simple derivation of Euler's method proceeds by first integrating the differential equation (1) between two consecutive mesh points x_n and x_{n+1} to deduce that

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x, y(x))\,\mathrm{d}x, \qquad n \in \{0, 1, \ldots, N-1\}, \tag{16}$$

and then applying the numerical integration rule

$$\int_{x_n}^{x_{n+1}} g(x)\,\mathrm{d}x \approx (x_{n+1} - x_n)g(x_n) = hg(x_n),$$

called the **rectangle rule**, with $g(x) = f(x, y(x))$, to get

$$y(x_{n+1}) \approx y(x_n) + hf(x_n, y(x_n)), \qquad n \in \{0, 1, \ldots, N-1\}, \qquad y(x_0) = y_0.$$

This then motivates the definition of Euler's method.

The idea can be generalised by replacing the rectangle rule in the above derivation with a one-parameter family of integration rules of the form

$$\int_{x_n}^{x_{n+1}} g(x)\,\mathrm{d}x \approx h\left[(1-\theta)g(x_n) + \theta g(x_{n+1})\right], \tag{17}$$

with $\theta \in [0, 1]$ a parameter. By applying this in (16) with $g(x) = f(x, y(x))$ we find that

$$y(x_{n+1}) \approx y(x_n) + h\left[(1-\theta)f(x_n, y(x_n)) + \theta f(x_{n+1}, y(x_{n+1}))\right], \quad n \in \{0, 1, \ldots, N-1\}, \quad y(x_0) = y_0.$$

This motivates the following family of methods, called the **θ-method**: with y_0 supplied by (2), define

$$y_{n+1} = y_n + h\left[(1-\theta)f(x_n, y_n) + \theta f(x_{n+1}, y_{n+1})\right], \quad n \in \{0, 1, \ldots, N-1\}, \quad y_0 = y(x_0), \tag{18}$$

parametrised by $\theta \in [0, 1]$. For $\theta = 0$ we recover Euler's method. For $\theta = 1$, we obtain the method

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}), \quad n \in \{0, 1, \ldots, N-1\}, \quad y_0 = y(x_0), \tag{19}$$

[4] Leonard Euler (1707–1783)

called the **implicit Euler method** since, unlike Euler's method considered above, (19) requires the solution of an implicit equation in order to determine y_{n+1}, given y_n. In order to emphasise this difference, Euler's method is sometimes termed the **explicit Euler method**. The scheme which results for the value of $\theta = 1/2$ is also of interest: y_0 is supplied by (2) and subsequent values y_{n+1} are computed from

$$y_{n+1} = y_n + h\frac{f(x_n, y_n) + f(x_{n+1}, y_{n+1})}{2}, \qquad n \in \{0, 1, \ldots, N-1\}, \quad y_0 = y(x_0);$$

this is called the **trapezium rule method**.

Remark 3 *The trapezium rule method involves the arithmetic average of $f(x_n, y_n)$ and $f(x_{n+1}, y_{n+1})$. Another possibility would have been to evaluate f at the arithmetic averages of x_n and x_{n+1} and y_n and y_{n+1} respectively. The resulting implicit one-step method:*

$$y_{n+1} = y_n + hf\left(\frac{x_n + x_{n+1}}{2}, \frac{y_n + y_{n+1}}{2}\right), \quad n \in \{0, 1, \ldots, N-1\}, \quad y_0 = y(x_0),$$

*is called the **implicit midpoint rule**.*

The θ-method is an explicit method for $\theta = 0$ and is an implicit method for $0 < \theta \leq 1$, because in the latter case (16) requires the solution of an implicit equation for y_{n+1}. Further, for each value of the parameter $\theta \in [0, 1]$, (16) is a one-step method in the sense that to compute y_{n+1} we only use one previous value y_n. Methods which require more than one previously computed value are referred to as multi-step methods, and will be discussed later on in the notes.

In order to assess the accuracy of the θ-method for various values of the parameter θ in $[0, 1]$, we perform a numerical experiment on a simple model problem.

Example 1 *Given the initial-value problem*

$$y'(x) = x - [y(x)]^2 \quad \textit{for} \quad x \in (0, 0.4), \qquad y(0) = 0,$$

we compute an approximate solution using the θ-method, for $\theta = 0$, $\theta = 1/2$ and $\theta = 1$, using the step size $h = 0.1$. The results are shown in Table 1. In the case of the two implicit methods, corresponding to $\theta = 1/2$ and $\theta = 1$, the nonlinear equations have been solved by a fixed-point iteration.

k	x_k	y_k for $\theta = 0$	y_k for $\theta = 1/2$	y_k for $\theta = 1$
0	0	0	0	0
1	0.1	0	0.00500	0.00999
2	0.2	0.01000	0.01998	0.02990
3	0.3	0.02999	0.04486	0.05955
4	0.4	0.05990	0.07944	0.09857

Table 1: The values of the numerical solution at the mesh points

For comparison, we also compute the value of the true solution $y(x)$ at the mesh points $x_n = \frac{n}{10}$, $n \in \{0, 1, 2, 3, 4\}$. Since the solution is not available in closed form, we use a Picard iteration to calculate an accurate approximation to the true solution on $[0, 0.4]$ and call this "exact solution". Thus, we consider

$$y_0(x) \equiv 0, \qquad y_k(x) = \int_0^x \left(t - [y_{k-1}(t)]^2\right) \mathrm{d}t, \quad k \in \mathbb{N}.$$

Hence,

$$y_0(x) \equiv 0 \qquad y_1(x) = \frac{1}{2}x^2, \qquad y_2(x) = \frac{1}{2}x^2 - \frac{1}{20}x^5, \qquad y_3(x) = \frac{1}{2}x^2 - \frac{1}{20}x^5 + \frac{1}{160}x^8 - \frac{1}{4400}x^{11}.$$

By induction, one shows that

$$y(x) = \frac{1}{2}x^2 - \frac{1}{20}x^5 + \frac{1}{160}x^8 - \frac{1}{4400}x^{11} + O\left(x^{14}\right),$$

Tabulating $y_3(x)$ on the interval $[0, 0.4]$ with step size $h = 0.1$, we get the values of the "exact solution" at the mesh points shown in Table 2.

k	x_k	$y(x_k)$
0	0	0
1	0.1	0.00500
2	0.2	0.01998
3	0.3	0.04488
4	0.4	0.07949

Table 2: Values of the "exact solution" at the mesh points

The "exact solution" is in good agreement with the results obtained with $\theta = 1/2$: the error is $\leq 5 \cdot 10^{-5}$. For $\theta = 0$ and $\theta = 1$ the discrepancy between y_k and $y(x_k)$ is larger: it is $\leq 2 \cdot 10^{-2}$.

So, why is the gap between the analytical solution and its numerical approximation in this example so much larger for $\theta \neq 1/2$ than for $\theta = 1/2$? The answer is the subject of the next section.

2.2 Error analysis of the θ-method

First we have to explain what we mean by *error*. The exact solution of the IVP (1)–(2) is a function of a continuously varying argument $x \in [x_0, X_M]$, while the numerical solution y_n is only defined at the mesh points x_n, $n \in \{0, 1, \ldots, N\}$, so it is a function of a "discrete" argument. We can compare these two functions either by extending in some fashion the approximate solution from the mesh points to the whole of the interval $[x_0, X_M]$ (say by interpolating between the values y_n), or by restricting the function y to the mesh points and comparing $y(x_n)$ with y_n for $n \in \{0, 1, \ldots, N\}$. Since the first of these approaches is somewhat arbitrary because it does not correspond to any procedure performed in a practical computation, we adopt the second approach, and we define the **global error** e_n by

$$e_n := y(x_n) - y_n \quad \text{for} \quad n \in \{0, 1, \ldots, N\}.$$

We wish to investigate the decay of the global error for the θ-method with respect to the reduction of the mesh size h. We shall show in detail how this is done in the case of Euler's method ($\theta = 0$) and then quote the corresponding result in the general case ($0 \leq \theta \leq 1$).

So let us consider the explicit Euler method:

$$y_{n+1} = y_n + hf(x_n, y_n), \qquad n \in \{0, 1, \ldots, N-1\}, \qquad y_0 = y(x_0).$$

The quantity

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - f(x_n, y(x_n)), \qquad n \in \{0, 1, \ldots, N-1\}, \tag{20}$$

obtained by inserting the true solution y into the numerical method and dividing by the mesh size is referred to as the **consistency error** (or **truncation error**) of the explicit Euler method and will play a key role in the analysis. Indeed, it measures the extent to which the true solution fails to satisfy the difference equation for the explicit Euler method.

By noting that $f(x_n, y(x_n)) = y'(x_n)$ and applying Taylor's Theorem, it follows from (20) that there exists a $\xi_n \in (x_n, x_{n+1})$ such that

$$|T_n| = \frac{|y(x_{n+1}) - y(x_n) - hy'(x_n)|}{h} = \frac{\frac{1}{2}h^2|y''(\xi_n)|}{h} \leq \frac{h}{2}M_2, \quad \text{where} \quad M_2 := \max_{x\in[x_0,X_M]} |y''(x)|, \qquad (21)$$

where we have assumed that f is a sufficiently smooth function of two variables so as to ensure that $y''(x) = \frac{\mathrm{d}}{\mathrm{d}x}[f(x, y(x))]$ exists and is bounded on the interval $[x_0, X_M]$. Since from the definition of Euler's method

$$0 = \frac{y_{n+1} - y_n}{h} - f(x_n, y_n),$$

By subtracting this from (20), we deduce that

$$e_{n+1} = e_n + h[f(x_n, y(x_n)) - f(x_n, y_n)] + hT_n.$$

Thus, assuming that $|y_n - y_0| \leq Y_M$, from the Lipschitz condition (3) we get

$$|e_{n+1}| \leq |e_n| + h|f(x_n, y(x_n)) - f(x_n, y_n)| + h|T_n| \leq (1 + hL)|e_n| + h|T_n|, \qquad n \in \{0, 1, \ldots, N-1\}.$$

Now, let $T := \max_{n\in\{0,1,\ldots,N-1\}} |T_n|$; then,

$$|e_{n+1}| \leq (1 + hL)|e_n| + hT \qquad \forall n \in \{0, 1, \ldots, N-1\}.$$

This gives

$$\begin{aligned} |e_n| &\leq (1 + hL)|e_{n-1}| + hT \\ &\leq (1 + hL)\left((1 + hL)|e_{n-2}| + hT\right) + hT \\ &\vdots \\ &\leq (1 + hL)^n|e_0| + \frac{T}{L}\left[(1 + hL)^n - 1\right], \end{aligned} \qquad (22)$$

which can be made rigorous using induction. Noting that $1 + x \leq e^x \ \forall x \in \mathbb{R}$, and $nh = x_n - x_0$, we have

$$|e_n| \leq e^{nhL}|e_0| + \frac{T}{L}\left[e^{nhL} - 1\right] = e^{L(x_n-x_0)}|e_0| + \frac{T}{L}\left[\mathrm{e}^{L(x_n-x_0)} - 1\right] \qquad \forall n \in \{0, 1, \ldots, N\}.$$

This estimate, together with the bound $T \leq \frac{h}{2}M_2$, which follows from (21), yields

$$|e_n| \leq \mathrm{e}^{L(x_n-x_0)}|e_0| + h\frac{M_2}{2L}\left[\mathrm{e}^{L(x_n-x_0)} - 1\right] \qquad \forall n \in \{0, 1, \ldots, N\}. \qquad (23)$$

To conclude, we note that by an analogous argument it is possible to prove that, in the general case of the θ-method (and assuming that h is sufficiently small, i.e. that $h \in (0, h_0]$ where $\frac{1}{2} - \theta L h_0 > 0$)

$$|e_n| \quad \leq \quad \exp\left(L\frac{x_n - x_0}{1 - \theta Lh}\right)|e_0| + h\left(\left|\frac{1}{2} - \theta\right|\frac{M_2}{L} + h\frac{1 + 3\theta}{6}\frac{M_3}{L}\right)\left[\exp\left(L\frac{x_n - x_0}{1 - \theta Lh}\right) - 1\right], \qquad (24)$$

for $n \in \{0, 1, \ldots, N\}$, where $M_3 := \max_{x\in[x_0,X_M]} |y'''(x)|$. In the absence of rounding errors in the imposition of the i.c. (2) we can suppose that $e_0 = y(x_0) - y_0 = 0$. Then, we see from (24) that

$$\max_{n\in\{0,1,\ldots,N\}} |e_n| = \mathcal{O}(h^2) \quad \text{when} \quad \theta = \tfrac{1}{2}, \qquad\qquad \max_{n\in\{0,1,\ldots,N\}} |e_n| = \mathcal{O}(h) \quad \text{when} \quad \theta \in [0, 1]\backslash\{\tfrac{1}{2}\}.$$

This explains why in Tables 1 and 2 the values y_n of the numerical solution computed with the trapezium-rule method ($\theta = 1/2$) were considerably closer to the true solution $y(x_n)$ at the mesh points than those which were obtained with the explicit/implicit Euler methods ($\theta = 0/\theta = 1$).

In particular, we see from this analysis, that each time the mesh size h is halved, the consistency error and the global error are reduced by a factor of 2 when $\theta \neq 1/2$, and by a factor of 4 when $\theta = 1/2$.

While the trapezium rule method leads to more accurate approximations than the explicit Euler method, it is less convenient from the computational point of view because it requires the solution of implicit equations at each mesh point x_{n+1} to compute y_{n+1}. An attractive compromise is to use explicit Euler to compute an initial crude approximation to $y(x_{n+1})$ and then use this value within the trapezium rule to obtain a more accurate approximation for $y(x_{n+1})$: the resulting numerical method is

$$y_{n+1} = y_n + h\frac{f(x_n, y_n) + f(x_{n+1}, y_n + hf(x_n, y_n))}{2}, \quad n \in \{0, 1, \ldots, N-1\}, \quad y_0 = y(x_0),$$

and is frequently referred to as the **improved Euler method**. Clearly, it is an explicit one-step scheme, albeit of a more complicated form than the explicit Euler method. In the next section, we shall take this idea further and consider a very general class of one-step methods.

2.3 General one-step methods

Definition 2 *A* one-step method *is a function Ψ that takes the triplet $(\xi, \eta; h) \in \mathbb{R} \times \mathbb{R} \times (0, \infty)$ and a function $f(\cdot, \cdot)$, and computes an approximation $\Psi(\xi, \eta; h, f) \in \mathbb{R}$ of $y(\xi + h)$, which is the solution at $x = \xi + h$ of the IVP*

$$y'(x) = f(x, y(x)), \qquad y(\xi) = \eta. \tag{25}$$

Here, we tacitly assume that (25) has a unique solution, and therefore $y(\xi + h)$ exists. Additionally, the step size h may need to be assumed to be sufficiently small for Ψ to be well-defined.

To give two simple examples, let us consider the implicit Euler method and the explicit Euler method:

- In the case of the implicit Euler method the function Ψ is defined implicitly, by

  $$\Psi(\xi, \eta; h, f) = \eta + hf(\xi + h, \Psi(\xi, \eta; h, f)).$$

  Assuming that f satisfies a global Lipschitz condition with Lipschitz constant L (see Example 2 for the definition), one can use the Contraction Mapping Theorem to show that, given a pair $(\xi, \eta) \in \mathbb{R}^2$, and $h \in (0, 1/L)$, there exists a unique $\Psi(\xi, \eta; h, f) \in \mathbb{R}$ satisfying this implicit relationship, and therefore for such a "sufficiently small" h the function Ψ associated with the implicit Euler method is well-defined.

- In the case of the explicit Euler method the function Ψ is defined explicitly, by

  $$\Psi(\xi, \eta; h, f) = \eta + hf(\xi, \eta).$$

In the case of general explicit one-step methods, to be investigated in the next section, we have

$$\Psi(\xi, \eta; h, f) = \eta + h\Phi(\xi, \eta; h, f),$$

where $\Phi(\xi, \eta; h, f)$ can be explicitly computed (without solving implicit equations) in terms of ξ, η, h, and f. In what follows, for the sake of notational simplicity, we shall not indicate the dependence of $\Phi(\xi, \eta; h, f)$ on f, and will write $\Phi(\xi, \eta; h)$ instead. For example, in the case of the explicit Euler method $\Phi(\xi, \eta; h) = f(\xi, \eta)$, for all h.

2.4 General explicit one-step methods

A general explicit one-step method may be written in the form:

$$y_{n+1} = y_n + h\Phi(x_n, y_n; h), \qquad n \in \{0, 1, \ldots, N-1\}, \qquad y_0 = y(x_0), \tag{26}$$

where $\Phi(\cdot, \cdot; \cdot)$ is a continuous function of its variables. For example, in the case of the explicit Euler method, $\Phi(x_n, y_n; h) = f(x_n, y_n)$, while for the improved Euler method

$$\Phi(x_n, y_n; h) = \frac{f(x_n, y_n) + f(x_n + h, y_n + hf(x_n, y_n))}{2}.$$

In order to assess the accuracy of the numerical method (26), we define the **global error**, e_n, by

$$e_n := y(x_n) - y_n, \qquad n \in \{0, 1, \ldots, N\}.$$

We define the **consistency error**, T_n, of the method by

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - \Phi(x_n, y(x_n); h), \qquad n \in \{0, 1, \ldots, N-1\}. \tag{27}$$

Remark 4 *For an implicit one-step method of the form $y_{n+1} = y_n + h\Phi(x_n, y_n, y_{n+1}; h)$, the consistency error is analogously defined by $T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - \Phi(x_n, y(x_n), y(x_{n+1}); h)$ for $n \in \{0, 1, \ldots, N-1\}$.*

The next theorem provides a bound on the global error in terms of the consistency error.

Theorem 4 *Consider the general one-step method (26) where, in addition to being a continuous function of its arguments, Φ is assumed to satisfy a Lipschitz condition with respect to its second argument; namely, there exist constants $L_\Phi, h_0 > 0$ such that, for $h \in [0, h_0]$ and for the same region R as in Picard's Theorem,*

$$|\Phi(x, z; h) - \Phi(x, \tilde{z}; h)| \leq L_\Phi |z - \tilde{z}|, \qquad \forall (x, z), (x, \tilde{z}) \in \mathsf{R}. \tag{28}$$

Then, assuming that $|y_n - y_0| \leq Y_M$ for $n \in \{0, 1, \ldots, N\}$, it follows that

$$|e_n| \leq e^{L_\Phi (x_n - x_0)} |e_0| + \frac{e^{L_\Phi (x_n - x_0)} - 1}{L_\Phi} T, \qquad n \in \{0, 1, \ldots, N\}, \tag{29}$$

where $T := \max_{n \in \{0,1,\ldots,N-1\}} |T_n|$.

Proof: First, note that by (26) and (27) we have that

$$y_{n+1} = y_n + h\Phi(x_n, y_n; h), \qquad y(x_{n+1}) = y(x_n) + h\Phi(x_n, y(x_n); h) + hT_n$$

for any $n \in \{0, 1, \ldots, N-1\}$. Subtracting the first equality from the second, we find that

$$e_{n+1} = e_n + h[\Phi(x_n, y(x_n); h) - \Phi(x_n, y_n; h)] + hT_n$$

for any $n \in \{0, 1, \ldots, N-1\}$. Since $(x_n, y(x_n)), (x_n, y_n) \in \mathsf{R}$, the Lipschitz condition (28) implies that

$$|e_{n+1}| \leq |e_n| + hL_\Phi |e_n| + h|T_n| \leq (1 + hL_\Phi)|e_n| + hT, \qquad n \in \{0, 1, \ldots, N-1\}.$$

Hence, analogously to (22), we find that

$$|e_n| \leq (1 + hL_\Phi)^n |e_0| + \frac{(1 + hL_\Phi)^n - 1}{L_\Phi} T, \qquad n \in \{0, 1, \ldots, N\}.$$

Noting that $1 + x \leq e^x \ \forall x \in \mathbb{R}$, and $nh = x_n - x_0$, we have

$$|e_n| \leq e^{nhL_\Phi} |e_0| + \frac{e^{nhL_\Phi} - 1}{L_\Phi} T \leq e^{L_\Phi (x_n - x_0)} |e_0| + \frac{e^{L_\Phi (x_n - x_0)} - 1}{L_\Phi} T, \qquad n \in \{0, 1, \ldots, N\},$$

which concludes the proof. $\diamond$

Let us note that the error bound (23) for Euler's explicit method is a special case of (29). We highlight the practical relevance of the error bound (29) by focusing on a particular example.

Example 2 *Consider the IVP*

$$y'(x) = \arctan(y(x)) \quad \textit{for} \quad x \in (0,1), \qquad y(0) = 1,$$

and suppose that this is solved by the explicit Euler method.

- *First, let us show that this IVP has a unique (continuously differentiable) solution* $y : [0,1] \to \mathbb{R}$. *We define* $\mathsf{R} := [x_0, X_M] \times [y_0 - Y_M, y_0 + Y_M]$ *with* $x_0 := 0$, $X_M := 1$, $y_0 := 1$, *and* $Y_M > 0$ *chosen later. The function* $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,z) := \arctan(z)$ *is continuous in* $\mathbb{R}^2$ *and satisfies*

$$|f(x,z) - f(x,\tilde{z})| \leq L|z - \tilde{z}| \qquad \forall (x,z), (x,\tilde{z}) \in [x_0, X_M] \times \mathbb{R}$$

  with $L := 1$ *(we say that* $f(\cdot,\cdot)$ *satisfies a* ***global Lipschitz condition*** *in its second argument). This follows from* $|\partial_z f(x,z)| = \frac{1}{1+z^2} \leq 1$ $\forall z \in \mathbb{R}$ *and the Mean-Value Theorem. Noting* $M := \max_{(x,z)\in\mathsf{R}} |f(x,z)| \leq \frac{\pi}{2}$, *we choose* $Y_M := \frac{\pi}{2}$ *so that* $M(X_M - x_0) \leq \frac{\pi}{2}(1-0) = Y_M$ *and deduce from Picard's theorem that there is a unique solution* $y : [0,1] \to \mathbb{R}$.

- *We apply (29) to quantify the size of the global error (note that here,* $\Phi(x,z;h) := f(x,z)$*). We take* $L_\Phi := 1$. *Let us note that as* $\Phi(\cdot,\cdot;\cdot)$ *satisfies a global Lipschitz condition in its second argument, we see from the proof of Theorem 4 that the assumption* $|y_n - y_0| \leq Y_M$ *is not needed in this case. By (29) and (21), and assuming* $e_0 = 0$, *we have that*

$$|e_n| \leq \frac{e^{x_n} - 1}{2} \left(\max_{x\in[0,1]} |y''(x)| \right) h, \qquad n \in \{0, 1, \ldots, N\}.$$

  To find a bound for $\max_{x\in[0,1]} |y''(x)|$, *we differentiate the DE to find*

$$y''(x) = \frac{\mathrm{d}}{\mathrm{d}x}(\arctan(y(x))) = \frac{y'(x)}{1 + [y(x)]^2} = \frac{\arctan(y(x))}{1 + [y(x)]^2}.$$

  We see that $|y''(x)| \leq |\arctan(x)| \leq \frac{\pi}{2}$ *for any* $x \in [0,1]$ *and hence,* $\max_{x\in[0,1]} |y''(x)| \leq \frac{\pi}{2}$. *We find*

$$|e_n| \leq \frac{\pi(e^{x_n} - 1)}{4} h \leq \frac{\pi(e-1)}{4} h, \qquad n \in \{0, 1, \ldots, N\}$$

  (note $x_n \leq X_M = 1$*). Thus, given a tolerance* $\texttt{TOL} > 0$ *specified beforehand, we can ensure that the error between the (unknown) true solution and its numerical approximation does not exceed* TOL *by choosing a step size* $h > 0$ *such that*

$$h \leq \frac{4}{\pi(e-1)} \texttt{TOL}.$$

  For such h *we shall have* $|y(x_n) - y_n| = |e_n| \leq \texttt{TOL}$ *for each* $n \in \{0, 1, \ldots, N\}$, *as required. Thus, at least in principle, we can calculate the numerical solution to arbitrarily high accuracy by choosing a sufficiently small step size. In practice, because digital computers use finite-precision arithmetic, there will always be small (but not infinitely small) pollution effects because of rounding errors; however, these can also be bounded by performing an analysis similar to the one above where* $f(x_n, y_n)$ *is replaced by its finite-precision representation.*

Returning to the general one-step method (26), we consider the choice of the function Φ. Theorem 4 suggests that if the consistency error 'approaches zero' as $h \to 0$ then the global error 'converges to zero' also (as long as $|e_0| \to 0$ when $h \to 0$). This observation motivates the following definition.

Definition 3 *The numerical method (26) is* ***consistent*** *with the ODE (1) if the consistency error defined by (27) is such that for any* $\varepsilon > 0$ *there exists* $h_\varepsilon > 0$ *for which* $|T_n| < \varepsilon$ *for all* $h \in (0, h_\varepsilon)$ *and any pair of points* $(x_n, y(x_n))$, $(x_{n+1}, y(x_{n+1}))$ *on the graph of* y.

For the general one-step method (26) we have assumed that the function $\Phi(\cdot,\cdot;\cdot)$ is continuous; also y' is a continuous function on $[x_0, X_M]$. Therefore, from (27),

$$\lim_{\substack{h\to 0,\, n\to\infty \\ x_n \to x \in [x_0, X_M]}} T_n = y'(x) - \Phi(x, y(x); 0) \qquad \forall\, x \in [x_0, X_M].$$

As $y'(x) = f(x, y(x))$, this implies that the one-step method (26) is consistent if, and only if, (we often simply write iff)

$$\Phi(x, y; 0) \equiv f(x, y), \tag{30}$$

i.e., $\Phi(x, y(x); 0) = f(x, y(x))\ \forall x \in [x_0, X_M]$. Now we are ready to state a convergence theorem for the general one-step method (26).

Theorem 5 *Suppose that the solution of the IVP (1)–(2) lies in* R *as does its approximation generated from (26) when* $h \le h_0$. *Suppose also that the function* $\Phi(\cdot,\cdot;\cdot)$ *is uniformly continuous on* $\mathsf{R} \times [0, h_0]$ *and satisfies the consistency condition (30) and the Lipschitz condition*

$$|\Phi(x, z; h) - \Phi(x, \tilde{z}; h)| \le L_\Phi |z - \tilde{z}| \qquad \forall (x, z, h), (x, \tilde{z}, h) \in \mathsf{R} \times [0, h_0]. \tag{31}$$

Then, if successive approximation sequences (y_n), *generated for* $x_n = x_0 + nh$, $n \in \{1, \ldots, N\}$, *are obtained from (26) with successively smaller values of* h, *each less than* h_0, *we have convergence of the numerical solution to the solution of the IVP in the sense that*

$$|y(x) - y_n| \longrightarrow 0 \qquad \text{as } h \to 0,\ n \to \infty,\ x_n \to x \in [x_0, X_M].$$

Proof: Suppose that $h = \frac{X_M - x_0}{N}$ where $N \in \mathbb{N}$. We shall assume that N is sufficiently large so that $h \le h_0$. Since $y(x_0) = y_0$ and therefore $e_0 = 0$, Theorem 4 implies that

$$|y(x_n) - y_n| \le \frac{\mathrm{e}^{L_\phi (X_M - x_0)} - 1}{L_\phi} \max_{m \in \{0,1,\ldots,N-1\}} |T_m|, \qquad n \in \{1, \ldots, N\}. \tag{32}$$

From the consistency condition (30) we have

$$T_n = \left[\frac{y(x_{n+1}) - y(x_n)}{h} - f(x_n, y(x_n))\right] + [\Phi(x_n, y(x_n); 0) - \Phi(x_n, y(x_n); h)].$$

According to the Mean-Value Theorem the expression in the first bracket is equal to $y'(\xi) - y'(x_n)$ for some $\xi \in [x_n, x_{n+1}]$. Since $y'(x) = f(x, y(x)) = \Phi(x, y(x); 0)$ and $\Phi(\cdot,\cdot;\cdot)$ is uniformly continuous on $\mathsf{R} \times [0, h_0]$, it follows that y' is uniformly continuous on $[x_0, X_M]$. Thus, for each $\varepsilon > 0$ there exists an $h_1(\varepsilon) > 0$ such that

$$|y'(\xi) - y'(x_n)| \le \frac{\varepsilon}{2} \qquad \text{for } h \in (0, h_1(\varepsilon)),\ n \in \{0, 1, \ldots, N-1\}.$$

Also, by the uniform continuity of Φ with respect to its third argument, there exists an $h_2(\varepsilon) > 0$ such that

$$|\Phi(x_n, y(x_n); 0) - \Phi(x_n, y(x_n); h)| \le \frac{\varepsilon}{2} \qquad \text{for } h \in (0, h_2(\varepsilon)),\ n \in \{0, 1, \ldots, N-1\}.$$

Thus, defining $h_\varepsilon := \min(h_1(\varepsilon), h_2(\varepsilon)) > 0$, we have

$$|T_n| \le \varepsilon \qquad \text{for } h \in (0, h_\varepsilon),\ n \in \{0, 1, \ldots, N-1\}.$$

Inserting this into (32) we deduce that $|y(x_n) - y_n| \to 0$ as $h \to 0$ and $n \to \infty$. Since

$$|y(x) - y_n| \le |y(x) - y(x_n)| + |y(x_n) - y_n|,$$

and the first term on the right also converges to zero as $n \to \infty$ and $x_n \to x$, by the uniform continuity of y on the interval $[x_0, X_M]$ the proof is complete. $\diamond$

We saw earlier that for Euler's method the absolute value of the consistency error T_n is bounded above by a constant multiple of the step size h, that is

$$|T_n| \leq Kh \qquad \forall\, h \in (0, h_0],$$

where K is a positive constant, independent of h. However there are other one-step methods (a class of which, called Runge–Kutta methods, will be considered below) for which we can do better. More generally, in order to quantify the asymptotic rate of decay of the consistency error as the step size h converges to zero, we introduce the following definition.

Definition 4 *The numerical method (26) is said to have* ***order of accuracy*** *p (or order of consistency p), if $p \in \mathbb{N}$ is the largest natural number such that, for any sufficiently smooth solution curve $(x, y(x))$ in R of the IVP (1)–(2) we have*

$$|T_n| = \mathcal{O}(h^p),$$

i.e., there exist constants $h_0, K > 0$ such that $|T_n| \leq Kh^p$ for all $h \in (0, h_0]$, for any pair of points $(x_n, y(x_n))$, $(x_{n+1}, y(x_{n+1}))$ on the solution curve.

Having introduced the general class of explicit one-step methods and the associated concepts of consistency and order of accuracy (or order of consistency), we now focus on a specific family: explicit Runge–Kutta methods.

2.5 Explicit Runge–Kutta methods

In the sense of Definition 4, the explicit Euler method is only first-order accurate; nevertheless, it is simple and cheap to implement because to obtain y_{n+1} from y_n we only require a single evaluation of the function f at (x_n, y_n). Runge–Kutta (RK) methods aim to achieve higher accuracy by sacrificing the efficiency of Euler's method through re-evaluating $f(\cdot, \cdot)$ at points intermediate between $(x_n, y(x_n))$ and $(x_{n+1}, y(x_{n+1}))$. The general form of the **R-stage explicit RK family** is as follows:

$$\begin{aligned} y_{n+1} &= y_n + h\Phi(x_n, y_n; h), \qquad \Phi(x, z; h) = \sum_{r=1}^{R} c_r k_r(x, z; h), \\ k_1(x, z; h) &= f(x, z), \qquad k_r(x, z; h) = f\left(x + ha_r, z + h\sum_{s=1}^{r-1} b_{rs} k_s(x, z; h)\right), \qquad r \in \{2, \ldots, R\}. \end{aligned}$$

Remark 5 *The most* ***general version of a R-stage RK method*** *is as follows:*

$$y_{n+1} = y_n + h\sum_{r=1}^{R} c_r k_r, \quad \textit{where} \quad k_r = f\left(x_n + ha_r, y_n + h\sum_{s=1}^{R} b_{rs} k_s\right) \quad \textit{for} \quad r \in \{1, \ldots, R\}.$$

If the method is not a R-stage explicit RK method, then it is called a ***R-stage implicit RK method****. The information about the coefficients of a RK method is usually displayed in the so-called* ***Butcher tableau*** $\begin{array}{c|c} a & B \\ \hline & c^T \end{array}$*, where $a = (a_1, \ldots, a_R)^{\mathrm{T}} \in \mathbb{R}^R$, $B = (b_{ij})_{1 \leq i,j \leq R} \in \mathbb{R}^{R \times R}$, $c = (c_1, \ldots, c_R)^{\mathrm{T}} \in \mathbb{R}^R$. In the case of an explicit RK method, the matrix B is strictly lower-triangular, i.e., the diagonal and superdiagonal entries of B are all equal to zero.*

For the sake of simplicity we now focus on explicit RK methods.

One-stage explicit RK methods. Suppose that $R = 1$, i.e.,

$$y_{n+1} = y_n + h\Phi(x_n, y_n; h) = y_n + c_1 h f(x_n, y_n), \quad \text{where} \quad \Phi(x, z; h) = c_1 f(x, z).$$

By the condition (30), a method from this family will be consistent iff $c_1 = 1$. The resulting one-stage explicit RK method is simply the explicit Euler method:

$$y_{n+1} = y_n + hf(x_n, y_n). \tag{33}$$

In the language of RK methods, $y_{n+1} = y_n + h\Phi(x_n, y_n; h)$ with $\Phi(x, z; h) = \sum_{r=1}^{1} c_r k_r(x, z; h)$, $c_1 = 1$ and $k_1(x, z; h) = f(x, z)$.

Remark 6 *The implicit Euler method $y_{n+1} = y_n + hf(x_{n+1}, y_{n+1})$ is an example of a one-stage implicit RK method: it can be written as $y_{n+1} = y_n + h\Phi(x_n, y_n; h)$, where $\Phi(x, z; h) = k_1(x, z; h)$ and $k_1(x, z; h) = f(x + h, z + hk_1)$ (note that, unsurprisingly, k_1 is now defined through an implicit relationship). For the sake of simplicity we shall continue to concentrate here on explicit RK methods only.*

Two-stage explicit RK methods. Next, consider the case of $R = 2$, corresponding to the following family of methods:

$$y_{n+1} = y_n + h(c_1 k_1 + c_2 k_2), \tag{34}$$

where

$$k_1 = f(x_n, y_n), \tag{35}$$
$$k_2 = f(x_n + a_2 h, y_n + b_{21} h k_1), \tag{36}$$

and where the parameters c_1, c_2, a_2 and b_{21} are to be determined.[5] Clearly (34)–(36) can be rewritten in the form (26) and therefore it is a family of one step methods. By the condition (30), a method from this family will be consistent iff $\Phi(x, y; 0) = c_1 f(x, y) + c_2 f(x + 0, y + 0) \equiv f(x, y)$, i.e., iff

$$c_1 + c_2 = 1.$$

Further conditions on the parameters are obtained by attempting to maximise the order of accuracy of the method. Let us expand the consistency error of (34)–(36) in powers of h. Let us write $f_x := \frac{\partial f}{\partial x}, f_z := \frac{\partial f}{\partial z}$ for the first-order partial derivatives of $f = f(x, z)$, and $f_{xx} := \frac{\partial^2 f}{\partial x^2}, f_{xz} := \frac{\partial^2 f}{\partial x \partial z}, f_{zz} := \frac{\partial^2 f}{\partial z^2}$ for the second-order partial derivatives of $f = f(x, z)$. We have

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - c_1 f(x_n, y(x_n)) - c_2 k_2(x_n, y(x_n); h)$$

with (using Taylor's Theorem)

$$\begin{aligned} k_2(x_n, y(x_n); h) &:= f(x_n + a_2 h, y(x_n) + b_{21} h f(x_n, y(x_n))) \\ &= \left[f + a_2 h f_x + b_{21} h f f_z + \frac{1}{2} a_2^2 h^2 f_{xx} + a_2 b_{21} h^2 f f_{xz} + \frac{1}{2} b_{21}^2 h^2 f^2 f_{zz} \right] (x_n, y(x_n)) + \mathcal{O}(h^3). \end{aligned}$$

Noting that for the first term in T_n, we have the expansion

$$\frac{y(x_{n+1}) - y(x_n)}{h} = y'(x_n) + \frac{1}{2} h y''(x_n) + \frac{1}{6} h^2 y'''(x_n) + \mathcal{O}(h^3),$$

[5] We note in passing that the explicit Euler method is a member of this family of methods, corresponding to $c_1 = 1$ and $c_2 = 0$. However we are now seeking methods that are at least second-order accurate.

and noting that $y'(x_n) = f(x_n, y(x_n)) = (c_1 + c_2)f(x_n, y(x_n))$, we deduce that

$$T_n = \frac{1}{2}hy''(x_n) + \frac{1}{6}h^2y'''(x_n) - c_2h[a_2f_x + b_{21}ff_z](x_n, y(x_n)) \\ - c_2h^2\left[\frac{1}{2}a_2^2f_{xx} + a_2b_{21}ff_{xz} + \frac{1}{2}b_{21}^2f^2f_{zz}\right](x_n, y(x_n)) + \mathcal{O}(h^3).$$

Note that for y', y'', y''' we have from the DE that

$$y'(x) = f(x, y(x)), \qquad y''(x) = f_x(x, y(x)) + y'(x)f_z(x, y(x)) = [f_x + ff_z](x, y(x)) = F_1(x, y(x)),$$

and

$$\begin{aligned} y'''(x) &= [f_{xx} + f_xf_z + ff_{xz}](x, y(x)) + y'(x)[f_{xz} + f_z^2 + ff_{zz}](x, y(x)) \\ &= [f_xf_z + ff_z^2 + f_{xx} + 2ff_{xz} + f^2f_{zz}](x, y(x)) \\ &= [f_zF_1 + F_2](x, y(x)), \end{aligned}$$

where the functions F_1, F_2 are defined as

$$F_1 := f_x + ff_z, \qquad F_2 := f_{xx} + 2ff_{xz} + f^2f_{zz}.$$

We find that

$$T_n = h\left[\frac{1}{2}F_1 - a_2c_2f_x - b_{21}c_2ff_z\right](x_n, y(x_n)) \\ + h^2\left[\frac{1}{6}f_zF_1 + \frac{1}{6}F_2 - \frac{1}{2}a_2^2c_2f_{xx} - a_2b_{21}c_2ff_{xz} - \frac{1}{2}b_{21}^2c_2f^2f_{zz}\right](x_n, y(x_n)) + \mathcal{O}(h^3).$$

It follows that $T_n = \mathcal{O}(h^2)$ for any f provided that

$$a_2c_2 = b_{21}c_2 = \frac{1}{2}, \quad c_1 + c_2 = 1,$$

or equivalently,

$$b_{21} = a_2, \quad c_2 = \frac{1}{2a_2}, \quad c_1 = 1 - \frac{1}{2a_2}.$$

This still leaves one free parameter, a_2, but no choice of the parameters will make the method generally third-order accurate.

There are two well-known examples of second-order explicit RK methods of the form (34)–(36):

a) **The modified Euler method:** In this case we take $a_2 := \frac{1}{2}$, $b_{21} := \frac{1}{2}$, $c_1 := 0$, $c_2 := 1$ to obtain

$$y_{n+1} = y_n + h\, f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}f(x_n, y_n)\right).$$

The consistency error is

$$T_n = h^2\left[\frac{1}{6}f_zF_1 + \frac{1}{24}F_2\right](x_n, y(x_n)) + \mathcal{O}(h^3).$$

b) **The improved Euler method:** In this case we take $a_2 := 1$, $b_{21} := 1$, $c_1 := \frac{1}{2}$, $c_2 := \frac{1}{2}$ to obtain

$$y_{n+1} = y_n + \frac{h}{2}\left[f(x_n, y_n) + f(x_n + h, y_n + hf(x_n, y_n))\right].$$

The consistency error is

$$T_n = h^2\left[\frac{1}{6}f_zF_1 - \frac{1}{12}F_2\right](x_n, y(x_n)) + \mathcal{O}(h^3).$$

Exercise 1 *Let $\alpha \in \mathbb{R}\backslash\{0\}$ and let $x_n = a + nh$, $n \in \{0, 1, \ldots, N\}$, be a uniform mesh on the interval $[a, b]$ of step size $h = \frac{b-a}{N}$. Consider the explicit one-step method for the numerical solution of the IVP $y'(x) = f(x, y(x))$ for $x \in (a, b)$, $y(a) = y_0$, which determines approximations y_n to the values $y(x_n)$ via*

$$y_{n+1} = y_n + h(1-\alpha)f(x_n, y_n) + h\alpha f\left(x_n + \frac{h}{2\alpha}, y_n + \frac{h}{2\alpha}f(x_n, y_n)\right), \qquad n \in \{0, 1, \ldots, N-1\}.$$

(i) Denoting the first-order partial derivatives of $f = f(x, z)$ by f_x, f_z, show that this method is consistent and that its consistency error, $T_n(h, \alpha)$, can be expressed as

$$T_n(h, \alpha) = \frac{h^2}{8\alpha}\left[\left(\frac{4}{3}\alpha - 1\right)y'''(x_n) + y''(x_n)f_z(x_n, y(x_n))\right] + \mathcal{O}(h^3).$$

(ii) This method is applied to the IVP $y'(x) = -[y(x)]^p$, $y(0) = 1$, where $p \in \mathbb{N}$. Show that 1) if $p = 1$ then there does not exist a choice $\alpha \neq 0$ for which $T_n(h, \alpha) = \mathcal{O}(h^3)$, and 2) if $p \geq 2$ then there exists $\alpha_0 \neq 0$ such that $T_n(h, \alpha_0) = \mathcal{O}(h^3)$.

SOLUTION: (i) The method is of the form $y_{n+1} = y_n + h\Phi(x_n, y_n; h)$, where the function Φ is given by

$$\Phi(x, z; h) = (1-\alpha)f(x, z) + \alpha f\left(x + \frac{h}{2\alpha}, z + \frac{h}{2\alpha}f(x, z)\right).$$

Since $\Phi(x, y; 0) \equiv f(x, y)$, the method is consistent. By definition, the consistency error is

$$T_n(h, \alpha) = \frac{y(x_{n+1}) - y(x_n)}{h} - \Phi(x_n, y(x_n); h).$$

Note that $y'(x) = f(x, y(x))$, $y''(x) = f_x(x, y(x)) + f_z(x, y(x))y'(x)$, and $y'''(x) = f_z(x, y(x))y''(x) + [f_{xx} + 2ff_{xz} + f^2 f_{zz}](x, y(x))$. Using Taylor expansion, we find

$$\begin{aligned}
&T_n(h, \alpha) \\
&= y'(x_n) + \frac{h}{2}y''(x_n) + \frac{h^2}{6}y'''(x_n) - (1-\alpha)y'(x_n) - \alpha f\left(x_n + \frac{h}{2\alpha}, y(x_n) + \frac{h}{2\alpha}y'(x_n)\right) + \mathcal{O}(h^3) \\
&= y'(x_n) + \frac{h}{2}y''(x_n) + \frac{h^2}{6}y'''(x_n) - (1-\alpha)y'(x_n) - \alpha f(x_n, y(x_n)) - \frac{h}{2}f_x(x_n, y(x_n)) - \frac{h}{2}f_z(x_n, y(x_n))y'(x_n) \\
&\quad - \frac{\alpha}{2}\left(\frac{h}{2\alpha}\right)^2 f_{xx}(x_n, y(x_n)) - \alpha\left(\frac{h}{2\alpha}\right)^2 f_{xz}(x_n, y(x_n))y'(x_n) - \frac{\alpha}{2}\left(\frac{h}{2\alpha}\right)^2 f_{zz}(x_n, y(x_n))[y'(x_n)]^2 + \mathcal{O}(h^3) \\
&= y'(x_n) - (1-\alpha)y'(x_n) - \alpha y'(x_n) + \frac{h}{2}y''(x_n) - \frac{h}{2}\left[f_x(x_n, y(x_n)) + f_z(x_n, y(x_n))y'(x_n)\right] \\
&\quad + \frac{h^2}{6}y'''(x_n) - \frac{h^2}{8\alpha}\left[f_{xx}(x_n, y(x_n)) + 2f_{xz}(x_n, y(x_n))y'(x_n) + f_{zz}(x_n, y(x_n))[y'(x_n)]^2\right] + \mathcal{O}(h^3) \\
&= \frac{h^2}{6}y'''(x_n) - \frac{h^2}{8\alpha}\left[y'''(x_n) - y''(x_n)f_z(x_n, y(x_n))\right] + \mathcal{O}(h^3) \\
&= \frac{h^2}{8\alpha}\left[\left(\frac{4}{3}\alpha - 1\right)y'''(x_n) + y''(x_n)f_z(x_n, y(x_n))\right] + \mathcal{O}(h^3).
\end{aligned}$$

(ii) IVP $y'(x) = -[y(x)]^p$, $y(0) = 1$, where $p \in \mathbb{N}$. Here, $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, z) := -z^p$, and we have $f_z(x, z) = -pz^{p-1}$. 1) If $p = 1$, then $f_z(x, z) = -1$ and we have that $y'(x) = -y(x)$, $y''(x) = -y'(x) = y(x)$, and $y'''(x) = y'(x) = -y(x)$, so that

$$T_n(h, \alpha) = \frac{h^2}{8\alpha}\left[-\left(\frac{4}{3}\alpha - 1\right)y(x_n) - y(x_n)\right] + \mathcal{O}(h^3) = -\frac{h^2}{6}y(x_n) + \mathcal{O}(h^3) = -\frac{h^2}{6}e^{-x_n} + \mathcal{O}(h^3).$$

Here, we have used that the true solution to the IVP $y'(x) = -y(x)$, $y(0) = 1$ is given by $y(x) = e^{-x}$. Since $e^{-x_n} \neq 0$, we see that there is no $\alpha \neq 0$ for which $T_n(h, \alpha) = \mathcal{O}(h^3)$.

2) Now consider $p \geq 2$. Then, $y'(x) = -[y(x)]^p$, $y''(x) = -p[y(x)]^{p-1}y'(x) = p[y(x)]^{2p-1}$, and $y'''(x) = p(2p-1)[y(x)]^{2p-2}y'(x) = -p(2p-1)[y(x)]^{3p-2}$, and therefore

$$T_n(h,\alpha) = -\frac{h^2}{8\alpha}\left[\left(\frac{4}{3}\alpha - 1\right)p(2p-1) + p^2\right][y(x_n)]^{3p-2} + \mathcal{O}(h^3).$$

Choosing α such that $(\frac{4}{3}\alpha - 1)p(2p-1) + p^2 = 0$, namely $\alpha = \frac{3p-3}{8p-4} =: \alpha_0$, gives $T_n(h,\alpha_0) = \mathcal{O}(h^3)$. $\diamond$

Three-stage explicit RK methods. Let us now suppose that $R = 3$ to illustrate the general idea. Thus, we consider the family of methods:

$$y_{n+1} = y_n + h\left[c_1k_1 + c_2k_2 + c_3k_3\right],$$

where

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + ha_2, y_n + hb_{21}k_1), \\ k_3 &= f(x_n + ha_3, y_n + hb_{31}k_1 + hb_{32}k_2). \end{aligned}$$

The method is consistent iff $\Phi(x,y;0) = c_1f(x,y) + c_2f(x+0,y+0) + c_3f(x+0,y+0+0) \equiv f(x,y)$, i.e., iff

$$c_1 + c_2 + c_3 = 1.$$

Our goal is to expand the consistency error T_n in powers of h.

Simplification: For simplicity, we assume that $f = f(x,z)$ is independent of x, i.e.,

$$f(x,z) = \tilde{f}(z)$$

for some function $\tilde{f}$ of one real variable. In this case, the ODE has the form $y'(x) = \tilde{f}(y(x))$ and is called **autonomous**. In this, case the method reads

$$y_{n+1} = y_n + h\left[c_1\tilde{k}_1 + c_2\tilde{k}_2 + c_3\tilde{k}_3\right],$$

where

$$\begin{aligned} \tilde{k}_1 &= \tilde{f}(y_n), \\ \tilde{k}_2 &= \tilde{f}(y_n + hb_{21}\tilde{k}_1), \\ \tilde{k}_3 &= \tilde{f}(y_n + hb_{31}\tilde{k}_1 + hb_{32}\tilde{k}_2). \end{aligned}$$

Now, let us expand the consistency error in powers of h. We have

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - c_1\tilde{f}(y(x_n)) - c_2\tilde{k}_2(y(x_n);h) - c_3\tilde{k}_3(y(x_n);h)$$

with (using Taylor's Theorem)

$$\begin{aligned} \tilde{k}_2(y(x_n);h) &:= \tilde{f}(y(x_n) + hb_{21}\tilde{f}(y(x_n))) \\ &= \tilde{f}(y(x_n)) + h\left[b_{21}\tilde{f}\tilde{f}'\right](y(x_n)) + h^2\left[\frac{b_{21}^2}{2}\tilde{f}^2\tilde{f}''\right](y(x_n)) + \mathcal{O}(h^3) \end{aligned}$$

and

$$\begin{aligned}
k_3(y(x_n);h) &:= \tilde{f}(y(x_n) + hb_{31}\tilde{f}(y(x_n)) + hb_{32}\tilde{k}_2(y(x_n);h)) \\
&= \tilde{f}(y(x_n)) + h\left[b_{31}\tilde{f}\tilde{f}' + b_{32}(\tilde{f} + hb_{21}\tilde{f}\tilde{f}' + \mathcal{O}(h^2))\tilde{f}'\right](y(x_n)) \\
&\qquad + h^2\left[\frac{1}{2}(b_{31}\tilde{f} + b_{32}(\tilde{f} + \mathcal{O}(h)))^2\tilde{f}''\right](y(x_n)) + \mathcal{O}(h^3) \\
&= \tilde{f}(y(x_n)) + h\left[(b_{31} + b_{32})\tilde{f}\tilde{f}'\right](y(x_n)) \\
&\qquad + h^2\left[b_{21}b_{32}\tilde{f}\tilde{f}'^2 + \frac{(b_{31} + b_{32})^2}{2}\tilde{f}^2\tilde{f}''\right](y(x_n)) + \mathcal{O}(h^3).
\end{aligned}$$

We find that

$$\begin{aligned}
\Phi(x_n, y(x_n); h) &= c_1\tilde{f}(y(x_n)) + c_2\tilde{k}_2(y(x_n);h) + c_3\tilde{k}_3(y(x_n);h) \\
&= (c_1 + c_2 + c_3)\tilde{f}(y(x_n)) + h\left[(b_{21}c_2 + (b_{31} + b_{32})c_3)\tilde{f}\tilde{f}'\right](y(x_n)) \\
&\qquad + h^2\left[b_{21}b_{32}c_3\tilde{f}\tilde{f}'^2 + \frac{b_{21}^2c_2 + (b_{31} + b_{32})^2c_3}{2}\tilde{f}^2\tilde{f}''\right](y(x_n)) + \mathcal{O}(h^3)
\end{aligned}$$

and we also have that

$$\begin{aligned}
\frac{y(x_{n+1}) - y(x_n)}{h} &= y'(x_n) + \frac{1}{2}hy''(x_n) + \frac{1}{6}h^2y'''(x_n) + \mathcal{O}(h^3) \\
&= \tilde{f}(y(x_n)) + h\left[\frac{1}{2}\tilde{f}\tilde{f}'\right](y(x_n)) + h^2\left[\frac{1}{6}\tilde{f}\tilde{f}'^2 + \frac{1}{6}\tilde{f}^2\tilde{f}''\right](y(x_n)) + \mathcal{O}(h^3).
\end{aligned}$$

Noting that $T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - \Phi(x_n, y(x_n); h)$, we conclude that we can achieve third-order accuracy, i.e., $T_n = \mathcal{O}(h^3)$, if there holds

$$\begin{aligned}
c_1 + c_2 + c_3 &= 1, \\
b_{21}c_2 + (b_{31} + b_{32})c_3 &= \frac{1}{2}, \\
b_{21}^2c_2 + (b_{31} + b_{32})^2c_3 &= \frac{1}{3}, \\
b_{21}b_{32}c_3 &= \frac{1}{6}.
\end{aligned}$$

Solving this system of four equations for the six unknowns: $c_1, c_2, c_3, b_{21}, b_{31}, b_{32}$, we obtain a two-parameter family of third-order accurate 3-stage explicit RK methods. We shall only highlight two notable examples from this family:

(i) **Heun's method** corresponds to

$$c_1 = \frac{1}{4}, \quad c_2 = 0, \quad c_3 = \frac{3}{4}, \quad b_{21} = \frac{1}{3}, \quad b_{31} = 0, \quad b_{32} = \frac{2}{3},$$

yielding

$$\begin{aligned}
y_{n+1} &= y_n + h\left(\frac{1}{4}\tilde{k}_1 + \frac{3}{4}\tilde{k}_3\right), \\
\tilde{k}_1 &= \tilde{f}(y_n), \\
\tilde{k}_2 &= \tilde{f}\left(y_n + \frac{1}{3}h\tilde{k}_1\right), \\
\tilde{k}_3 &= \tilde{f}\left(y_n + \frac{2}{3}h\tilde{k}_2\right).
\end{aligned}$$

(ii) **Standard third-order explicit RK method**. This is arrived at by selecting

$$c_1 = \frac{1}{6}, \quad c_2 = \frac{2}{3}, \quad c_3 = \frac{1}{6}, \quad b_{21} = \frac{1}{2}, \quad b_{31} = -1, \quad b_{32} = 2,$$

yielding

$$\begin{aligned} y_{n+1} &= y_n + h\left(\frac{1}{6}\tilde{k}_1 + \frac{2}{3}\tilde{k}_2 + \frac{1}{6}\tilde{k}_3\right), \\ \tilde{k}_1 &= \tilde{f}(y_n), \\ \tilde{k}_2 &= \tilde{f}\left(y_n + \frac{1}{2}h\tilde{k}_1\right), \\ \tilde{k}_3 &= \tilde{f}\left(y_n - h\tilde{k}_1 + 2h\tilde{k}_2\right). \end{aligned}$$

Four-stage explicit RK methods. For $R = 4$, an analogous argument leads to a two-parameter family of four-stage RK methods of order four. A particularly popular example from this family is:

$$y_{n+1} = y_n + \frac{1}{6}h\left(k_1 + 2k_2 + 2k_3 + k_4\right),$$

where

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_1\right), \\ k_3 &= f\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_2\right), \\ k_4 &= f(x_n + h, y_n + hk_3). \end{aligned}$$

Here k_2 and k_3 represent approximations to the derivative $y'(\cdot)$ at points on the solution curve, intermediate between $(x_n, y(x_n))$ and $(x_{n+1}, y(x_{n+1}))$, and $\Phi(x_n, y_n; h)$ is a weighted average of the k_i, $i = 1, \ldots, 4$, the weights corresponding to those of the Simpson rule method (to which the fourth-order explicit RK method reduces when $f_z \equiv 0$, i.e., when $f(x, z) = \hat{f}(x)$ for some function $\hat{f}$).

In this section, we have constructed R-stage explicit RK methods of order of accuracy $\mathcal{O}(h^R)$, $R = 1, 2, 3, 4$. It is natural to ask whether there exists an R stage method of order R for $R \geq 5$. The answer to this question is negative: in a series of papers John Butcher showed that for $R = 5, 6, 7, 8, 9$, the highest order that can be attained by an R-stage RK method is, respectively, $4, 5, 6, 6, 7$, and that for $R \geq 10$ the highest order is $\leq R - 2$.

2.6 Absolute stability of explicit Runge–Kutta methods

We consider the model problem

$$y'(x) = \lambda y(x), \quad y(0) = y_0, \tag{37}$$

with $\lambda \in (-\infty, 0)$ and $y_0 \neq 0$. Trivially, the true solution to this IVP

$$y(x) = y_0\, e^{\lambda x}$$

converges to 0 at an exponential rate as $x \to \infty$. The question that we wish to investigate here is under what conditions on the step size h does a RK method reproduce this behaviour. The understanding of this matter will provide useful information about the adequate selection of h in the numerical approximation of an IVP by an explicit RK method over an interval $[x_0, X_M]$ with $X_M \gg x_0$. For the sake of simplicity, we shall restrict our attention to the case of R-stage methods of order of accuracy R, with $1 \leq R \leq 4$.

Let us begin with $R = 1$. The only consistent explicit one-stage RK method is the explicit Euler method. Applying (33) to (37) yields (note that here $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, z) := \lambda z$)

$$y_{n+1} = y_n + hf(x_n, y_n) = y_n + \lambda h y_n = (1 + \bar{h})y_n, \qquad n \in \mathbb{N}_0,$$

where $\bar{h} := \lambda h$. Thus, for any $n \in \mathbb{N}_0$ we have

$$y_n = (1 + \bar{h})^n y_0.$$

Consequently, the sequence $(y_n)_{n\in\mathbb{N}_0}$ will converge to 0 iff $|1 + \bar{h}| < 1$, yielding $\bar{h} \in (-2, 0)$; for such h the explicit Euler method is said to be **absolutely stable** and the interval $(-2, 0)$ is referred to as the **interval of absolute stability** of the method.

Now consider $R = 2$ corresponding to two-stage second-order explicit RK methods:

$$y_{n+1} = y_n + h(c_1k_1 + c_2k_2),$$

where

$$k_1 = f(x_n, y_n), \qquad k_2 = f(x_n + a_2h, y_n + b_{21}hk_1)$$

with the order conditions

$$c_1 + c_2 = 1, \qquad a_2c_2 = b_{21}c_2 = \frac{1}{2}$$

for second-order accuracy. Applying this to (37) yields,

$$y_{n+1} = y_n + h(c_1\lambda y_n + c_2\lambda(y_n + b_{21}h\lambda y_n)) = \left(1 + \bar{h} + \frac{1}{2}\bar{h}^2\right) y_n, \qquad n \in \mathbb{N}_0,$$

and therefore, for any $n \in \mathbb{N}_0$ we have

$$y_n = \left(1 + \bar{h} + \frac{1}{2}\bar{h}^2\right)^n y_0.$$

Hence the method is absolutely stable iff

$$\left|1 + \bar{h} + \frac{1}{2}\bar{h}^2\right| < 1,$$

namely when $\bar{h} \in (-2, 0)$.

In the case of $R = 3$ an analogous argument shows that

$$y_{n+1} = \left(1 + \bar{h} + \frac{1}{2}\bar{h}^2 + \frac{1}{6}\bar{h}^3\right) y_n.$$

Demanding that

$$\left|1 + \bar{h} + \frac{1}{2}\bar{h}^2 + \frac{1}{6}\bar{h}^3\right| < 1$$

then yields the interval of absolute stability: $\bar{h} \in (-a, 0)$ with $a \approx 2.51$.

When $R = 4$, we have that

$$y_{n+1} = \left(1 + \bar{h} + \frac{1}{2}\bar{h}^2 + \frac{1}{6}\bar{h}^3 + \frac{1}{24}\bar{h}^4\right) y_n,$$

and the associated interval of absolute stability is $\bar{h} \in (-a, 0)$ with $a \approx 2.78$.

For $R \geq 5$ on applying the explicit RK method to the model problem (37) still results in a recursion of the form

$$y_{n+1} = A_R(\bar{h})y_n, \qquad n \in \mathbb{N}_0,$$

however, unlike the case when $R = 1, 2, 3, 4$, in addition to $\bar{h}$ the expression $A_R(\bar{h})$ also depends on the coefficients of the explicit RK method; by a convenient choice of the free parameters the associated interval of absolute stability may be maximised.

3 Linear multi-step methods

While explicit RK methods present an improvement over, e.g., the explicit Euler method in terms of accuracy, this is achieved by investing additional computational effort; in fact, RK methods require more evaluations of $f(\cdot,\cdot)$ than would seem necessary. For example, the fourth-order method involves four function evaluations per step. For comparison, by considering three consecutive points x_{n-1}, $x_n = x_{n-1}+h$, $x_{n+1} = x_{n-1}+2h$, integrating the DE between x_{n-1} and x_{n+1}, and applying Simpson's rule to approximate the resulting integral (that is, $\int_a^b g(x)\mathrm{d}x \approx \frac{b-a}{6}(g(a)+4g(\frac{a+b}{2})+g(b))$), yields

$$\begin{aligned} y(x_{n+1}) &= y(x_{n-1}) + \int_{x_{n-1}}^{x_{n+1}} f(x,y(x))\,\mathrm{d}x \\ &\approx y(x_{n-1}) + \frac{1}{3}h\left[f(x_{n-1},y(x_{n-1})) + 4f(x_n,y(x_n)) + f(x_{n+1},y(x_{n+1}))\right], \end{aligned}$$

which leads to the **Simpson rule method**

$$y_{n+1} = y_{n-1} + \frac{1}{3}h\left[f(x_{n-1},y_{n-1}) + 4f(x_n,y_n) + f(x_{n+1},y_{n+1})\right]. \tag{38}$$

In contrast with the one-step methods considered in the previous section where only a single value y_n was required to compute the next approximation y_{n+1}, here we need *two* preceding values, y_n and y_{n-1} to be able to calculate y_{n+1}, and therefore (38) is not a one-step method.

In this section we consider a class of methods of the type (38) for the numerical solution of the IVP (1)–(2), called **linear multi-step methods (LMMs)**.

Given a sequence of equally spaced mesh points (x_n) with step size h, we consider the general **linear k-step method**

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h\sum_{j=0}^{k} \beta_j f(x_{n+j},y_{n+j}), \tag{39}$$

where $\alpha_0,\alpha_1,\ldots,\alpha_k,\beta_0,\beta_1,\ldots,\beta_k \in \mathbb{R}$. In order to avoid degenerate cases, we shall assume that $\alpha_k \neq 0$ and that $\alpha_0^2+\beta_0^2 \neq 0$, i.e., α_0 and β_0 are not both equal to zero. If $\beta_k = 0$ then y_{n+k} is obtained explicitly from previous values of y_j and $f(x_j,y_j)$, and the k-step method is then said to be **explicit**. On the other hand, if $\beta_k \neq 0$ then y_{n+k} appears not only on the left-hand side but also on the right, within $f(x_{n+k},y_{n+k})$; because of this implicit dependence on y_{n+k} the method is then called **implicit**. The numerical method (39) is called *linear* because it involves only linear combinations of the $\{y_n\}$ and the $\{f(x_n,y_n)\}$; for the sake of notational simplicity, henceforth we shall write

$$f_n := f(x_n,y_n).$$

Let us give some examples of LMMs.

Example 3 *We have already seen an example of a linear 2-step method in (38). Further examples:*

a) The explicit Euler method $y_{n+1} = y_n + hf_n$ is an explicit linear one-step method. The implicit Euler method $y_{n+1} = y_n + hf_{n+1}$ is an implicit linear one-step method.

b) The trapezium rule method $y_{n+1} = y_n + \frac{h}{2}(f_{n+1}+f_n)$ is an implicit linear one-step method.

*c) The four-step **Adams**[6]**–Bashforth method***

$$y_{n+4} = y_{n+3} + \frac{h}{24}\left(55f_{n+3} - 59f_{n+2} + 37f_{n+1} - 9f_n\right)$$

[6] J. C. Adams (1819–1892)

is an explicit linear four-step method; the four-step ***Adams–Moulton method***

$$y_{n+4} = y_{n+3} + \frac{h}{720}\left(251f_{n+4} + 646f_{n+3} - 264f_{n+2} + 106f_{n+1} - 19f_n\right)$$

is an implicit linear four-step method.

The construction of general classes of LMMs, such as the (explicit) Adams–Bashforth family and the (implicit) Adams–Moulton family will be discussed in the next section.

3.1 Construction of linear multi-step methods

Let us suppose that $(u_n)_{n\in\mathbb{N}_0}$ is a sequence of real numbers. We introduce the shift operator E, the forward difference operator Δ_+ and the backward difference operator Δ_- by

$$\begin{aligned}
E &: (u_n)_{n\in\mathbb{N}_0} = (u_0, u_1, u_2, \dots) \mapsto (u_{n+1})_{n\in\mathbb{N}_0} = (u_1, u_2, \dots),\\
\Delta_+ &: (u_n)_{n\in\mathbb{N}_0} = (u_0, u_1, u_2, \dots) \mapsto (u_{n+1} - u_n)_{n\in\mathbb{N}_0} = (u_1 - u_0, u_2 - u_1, \dots),\\
\Delta_- &: (u_n)_{n\in\mathbb{N}_0} = (u_0, u_1, u_2, \dots) \mapsto (u_n - u_{n-1})_{n\in\mathbb{N}_0} = (u_0, u_1 - u_0, u_2 - u_1, \dots).
\end{aligned}$$

(Note we used the notation $u_{-1} := 0$.) Further, we define

$$E^{-1} : (u_n)_{n\in\mathbb{N}_0} = (u_0, u_1, u_2, \dots) \mapsto (u_{n-1})_{n\in\mathbb{N}_0} = (0, u_0, u_1, \dots)$$

and note that $E \circ E^{-1} = I$, where I denotes the identity map $I : (u_n)_{n\in\mathbb{N}_0} \mapsto (u_n)_{n\in\mathbb{N}_0}$. Observe that

$$\Delta_+ = E - I = E\Delta_-, \qquad \Delta_- = I - E^{-1}, \qquad E \circ (I - \Delta_-) = I.$$

Writing $u := (u_n)_{n\in\mathbb{N}_0}$, it follows that for any $k \in \mathbb{N}$ we have

$$[\Delta_+^k u]_n = [(E-I)^k u]_n = \sum_{j=0}^{k} (-1)^j \binom{k}{j} u_{n+k-j}, \quad [\Delta_-^k u]_n = [(I - E^{-1})^k u]_n = \sum_{j=0}^{k} (-1)^j \binom{k}{j} u_{n-j}.$$

(Notation: $\Delta_+^k = \Delta_+ \circ \Delta_+ \circ \cdots \circ \Delta_+$ (k times), $\Delta_+^k u$ is a sequence with entries $[\Delta_+^k u]_n$, $n \in \mathbb{N}_0$, i.e., $\Delta_+^k u = ([\Delta_+^k u]_0, [\Delta_+^k u]_1, \dots)$. Similarly for $\Delta_-^k u$.)

Now suppose that $u : \mathbb{R} \to \mathbb{R}$ is a function whose derivative exists and is integrable on $[x_0, x_n]$ for each $n \in \mathbb{N}_0$. We then define $u_n := u(x_n)$ where $x_n = x_0 + nh$ for $n \in \mathbb{N}_0$. With a slight abuse of notation, we call the resulting sequence $(u_0, u_1, u_2, \dots)$ again u as it will be clear from the context when we mean the function and when we mean the sequence.

Letting $D := \frac{\mathrm{d}}{\mathrm{d}x}$ be the differentiation-operator, by applying a Taylor series expansion we find that

$$[E^s u]_n = u(x_n + sh) = u_n + sh[Du]_n + \frac{1}{2!}(sh)^2[D^2u]_n + \cdots = \sum_{k=0}^{\infty} \frac{(sh)^k}{k!}[D^k u]_n = [\mathrm{e}^{shD}u]_n.$$

Thus, formally, $hD = \ln(E) = -\ln(I - \Delta_-)$, and therefore, again by Taylor series expansion,

$$hu'(x_n) = \left[\left(\Delta_- + \frac{1}{2}\Delta_-^2 + \frac{1}{3}\Delta_-^3 + \cdots\right)u\right]_n.$$

Now letting $u(x) = y(x)$ where y is the solution of the IVP (1)–(2) and noting $u'(x) = y'(x) = f(x, y(x))$, we find that

$$hf(x_n, y(x_n)) = \left[\left(\Delta_- + \frac{1}{2}\Delta_-^2 + \frac{1}{3}\Delta_-^3 + \cdots\right)y\right]_n.$$

By successive truncations of the infinite series on the right, we find that

$$\begin{aligned} y(x_n) - y(x_{n-1}) &\approx hf(x_n, y(x_n)), \qquad (n \geq 1) \\ \frac{3}{2}y(x_n) - 2y(x_{n-1}) + \frac{1}{2}y(x_{n-2}) &\approx hf(x_n, y(x_n)), \qquad (n \geq 2) \\ \frac{11}{6}y(x_n) - 3y(x_{n-1}) + \frac{3}{2}y(x_{n-2}) - \frac{1}{3}y(x_{n-3}) &\approx hf(x_n, y(x_n)), \qquad (n \geq 3) \end{aligned}$$

and so on. These approximate equalities give rise to a class of implicit LMMs called **backward differentiation formulae (BDF)**, the simplest of which is the implicit Euler method.

Similarly, using $E^{-1} = I - \Delta_-$ and $hD = -\ln(I - \Delta_-)$, we find

$$E^{-1}(hD) = -(I - \Delta_-)\ln(I - \Delta_-),$$

and therefore

$$hu'(x_n) = \left[\left(\Delta_- - \frac{1}{2}\Delta_-^2 - \frac{1}{6}\Delta_-^3 + \cdots\right) u\right]_{n+1}.$$

Letting, again, $u(x) = y(x)$ where y is the solution of the IVP (1)–(2) and noting $u'(x) = y'(x) = f(x, y(x))$, successive truncations of the infinite series on the right result in

$$\begin{aligned} y(x_{n+1}) - y(x_n) &\approx hf(x_n, y(x_n)), \\ \frac{1}{2}y(x_{n+1}) - \frac{1}{2}y(x_{n-1}) &\approx hf(x_n, y(x_n)), \qquad (n \geq 1) \\ \frac{1}{3}y(x_{n+1}) + \frac{1}{2}y(x_n) - y(x_{n-1}) + \frac{1}{6}y(x_{n-2}) &\approx hf(x_n, y(x_n)), \qquad (n \geq 2) \end{aligned}$$

and so on. The first of these yields the explicit Euler method, the second the so-called explicit midpoint rule, and so on.

Further methods can be created using a similar methodology. Without going into detail, one can show that

$$y(x_{n+1}) - y(x_n) \approx h\left[\left(I - \frac{1}{2}\Delta_- - \frac{1}{12}\Delta_-^2 - \frac{1}{24}\Delta_-^3 - \frac{19}{720}\Delta_-^4 - \cdots\right) y'\right]_{n+1} \tag{40}$$

and

$$y(x_{n+1}) - y(x_n) \approx h\left[\left(I + \frac{1}{2}\Delta_- + \frac{5}{12}\Delta_-^2 + \frac{3}{8}\Delta_-^3 + \frac{251}{720}\Delta_-^4 + \cdots\right) y'\right]_{n}. \tag{41}$$

Using $y'(x) = f(x, y(x))$, successive truncations of (40) yield the family of Adams–Moulton methods, while similar successive truncations of (41) gives rise to the family of Adams–Bashforth methods.

Next, we turn our attention to the analysis of LMMs and introduce the concepts of stability, consistency and convergence.

3.2 Zero-stability

As is clear from (39) we need k starting values, $y_0, \ldots, y_{k-1}$, before we can apply a linear k-step method to the IVP (1)–(2): of these, y_0 is given by the i.c. (2), but the others, $y_1, \ldots, y_{k-1}$, have to be computed by other means: say, by using a suitable RK method. The starting values will contain numerical errors and it is important to know how these will affect further approximations y_n, $n \geq k$, which are calculated by means of (39). Thus, we wish to consider the 'stability' of the numerical method with respect to 'small perturbations' in the starting conditions.

Definition 5 *A linear k-step method for the ODE $y'(x) = f(x, y(x))$ is called **zero-stable** if there exists a constant $K > 0$ such that, for any two sequences (y_n) and $(\hat{y}_n)$, which have been generated by the same formulae but with different initial data $y_0, y_1, \ldots, y_{k-1}$ and $\hat{y}_0, \hat{y}_1, \ldots, \hat{y}_{k-1}$, respectively, we have*

$$|y_n - \hat{y}_n| \leq K \max_{j \in \{0,1,\ldots,k-1\}} |y_j - \hat{y}_j| \tag{42}$$

for $n \in \{0, 1, \ldots, N\}$, and as h tends to 0.

We shall prove later (cf. the first line of the proof of Theorem 6) that whether or not a method is zero-stable can be determined by merely considering its behaviour when applied to the trivial differential equation $y'(x) = 0$, corresponding to (1) with $f \equiv 0$; it is for this reason that the kind of stability expressed in Definition 5 is called *zero* stability. While Definition 5 is expressive in the sense that it conforms with the intuitive notion of stability whereby "small perturbations at input give rise to small perturbations at output", it would be a very tedious exercise to verify the zero-stability of a LMM using Definition 5 only; thus we shall next formulate an algebraic equivalent of zero-stability, known as the root condition, which will simplify this task. Before doing so we introduce some notation.

Given the linear k-step method (39) we consider its **first characteristic polynomial**

$$\rho : \mathbb{C} \to \mathbb{C}, \qquad \rho(z) := \sum_{j=0}^{k} \alpha_j z^j,$$

and its **second characteristic polynomial**

$$\sigma : \mathbb{C} \to \mathbb{C}, \qquad \sigma(z) := \sum_{j=0}^{k} \beta_j z^j,$$

where, as before, $\alpha_k \neq 0$ and $\alpha_0^2 + \beta_0^2 \neq 0$. For $r \in (0, \infty)$ and $a \in \mathbb{C}$, we introduce the notation

$$D_r(a) := \{z \in \mathbb{C} : |z - a| < r\}, \quad \bar{D}_r(a) := \{z \in \mathbb{C} : |z - a| \leq r\}, \quad \partial D_r(a) := \{z \in \mathbb{C} : |z - a| = r\}.$$

Now we are ready to state the main result of this section.

Theorem 6 *A LMM is zero-stable for any ODE of the form (1) where f satisfies the Lipschitz condition (3), iff all zeros of its first characteristic polynomial lie inside the closed unit disc $\bar{D}_1(0)$, with any which lie on the unit circle $\partial D_1(0)$ being simple.*

The algebraic stability condition contained in this theorem, namely that the roots of the first characteristic polynomial lie in the closed unit disc and those on the unit circle are simple, is often called the **root condition.**

PROOF: (Sketch) *Necessity.* Apply the linear k-step method to the ODE $y'(x) = 0$ (i.e., $f \equiv 0$):

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_1 y_{n+1} + \alpha_0 y_n = 0. \tag{43}$$

Denoting the distinct zeros of the first characteristic polynomial ρ by $z_1, \ldots, z_S \in \mathbb{C}$, the general solution of this k-th order linear difference equation has the form

$$y_n = \sum_{s=1}^{S} p_s(n) z_s^n, \tag{44}$$

where $p_s(\cdot)$ is a polynomial of degree one less than the multiplicity of the zero. Clearly, if $|z_s| > 1$ then there are starting values for which the corresponding solutions grow like $|z_s|^n$ and if $|z_s| = 1$ and its multiplicity is $m_s > 1$ then there are solutions growing like $n^{m_s - 1}$. In either case there are solutions that grow unbounded as $n \to \infty$, i.e. as $h \to 0$ with nh fixed. Considering starting data y_0, $y_1, \ldots, y_{k-1}$ which give rise to such an unbounded solution (y_n), and starting data $\hat{y}_0 = \hat{y}_1 = \cdots = \hat{y}_{k-1} = 0$ for which the corresponding solution of (43) is $(\hat{y}_n)$ with $\hat{y}_n = 0$ for all n, we see that (42) cannot hold. To summarise, if the root condition is violated then the method is not zero-stable.

Sufficiency. The proof that the root condition is sufficient for zero-stability is long and technical, and will be omitted here. For details, see, for example, P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962. $\diamond$

Example 4 *We shall consider the methods from Example 3.*

a) Explicit Euler method: $y_{n+1} - y_n = hf_n$. *Here,* $\alpha_1 = 1$, $\alpha_0 = -1$, $\beta_1 = 0$, $\beta_0 = 1$. *Hence,* $\rho(z) = z - 1$ *which has a simple root at* $z = 1$. *Hence, the explicit Euler method is zero-stable.*

Implicit Euler method: $y_{n+1} - y_n = hf_{n+1}$. *Here,* $\alpha_1 = 1$, $\alpha_0 = -1$, $\beta_1 = 1$, $\beta_0 = 0$. *Hence,* $\rho(z) = z - 1$ *which has a simple root at* $z = 1$. *Hence, the implicit Euler method is zero-stable.*

Trapezium rule method: $y_{n+1} - y_n = h(\frac{1}{2}f_{n+1} + \frac{1}{2}f_n)$. *Here,* $\alpha_1 = 1$, $\alpha_0 = -1$, $\beta_1 = \beta_0 = \frac{1}{2}$. *Hence,* $\rho(z) = z - 1$ *which has a simple root at* $z = 1$. *Hence, the trapezium rule method is zero-stable.*

b) 4-step Adams–Bashforth method: $y_{n+4} - y_{n+3} = h\left(\frac{55}{24}f_{n+3} - \frac{59}{24}f_{n+2} + \frac{37}{24}f_{n+1} - \frac{9}{24}f_n\right)$. *Here,* $\alpha_4 = 1$, $\alpha_3 = -1$, $\alpha_2 = \alpha_1 = \alpha_0 = 0$, $\beta_4 = 0$, $\beta_3 = \frac{55}{24}$, $\beta_2 = -\frac{59}{24}$, $\beta_1 = \frac{37}{24}$, $\beta_0 = -\frac{9}{24}$. *Hence,* $\rho(z) = z^4 - z^3 = z^3(z-1)$ *which has the root* $z_1 = 0$ *with multiplicity* 3, *and the root* $z_2 = 1$ *with multiplicity* 1. *Hence, the four-step Adams–Bashforth method is zero-stable.*

4-step Adams–Moulton method: $y_{n+4} - y_{n+3} = h\left(\frac{251}{720}f_{n+4} + \frac{646}{720}f_{n+3} - \frac{264}{720}f_{n+2} + \frac{106}{720}f_{n+1} - \frac{19}{720}f_n\right)$. *Here,* $\alpha_4 = 1$, $\alpha_3 = -1$, $\alpha_2 = \alpha_1 = \alpha_0 = 0$, $\beta_4 = \frac{251}{720}$, $\beta_3 = \frac{646}{720}$, $\beta_2 = -\frac{264}{720}$, $\beta_1 = \frac{106}{720}$, $\beta_0 = -\frac{19}{720}$. *Hence,* $\rho(z) = z^4 - z^3 = z^3(z-1)$ *which has the root* $z_1 = 0$ *with multiplicity* 3, *and the root* $z_2 = 1$ *with multiplicity* 1. *Hence, the four-step Adams–Moulton method is zero-stable.*

c) Consider the three-step (sixth order accurate) LMM

$$11y_{n+3} + 27y_{n+2} - 27y_{n+1} - 11y_n = h\left(3f_{n+3} + 27f_{n+2} + 27f_{n+1} + 3f_n\right).$$

Here, $\alpha_3 = 11$, $\alpha_2 = 27$, $\alpha_1 = -27$, $\alpha_0 = -11$, $\beta_3 = \beta_2 = 27$, $\beta_1 = 27$, $\beta_0 = 3$. *Hence,* $\rho(z) = 11z^3 + 27z^2 - 27z - 11$ *has the three simple roots* $z_1 = 1$, $z_2 = -\frac{19-4\sqrt{15}}{11}$, $z_3 = -\frac{19+4\sqrt{15}}{11}$. *Since* $|z_3| = \frac{19+4\sqrt{15}}{11} > 1$, *the method is not zero-stable.*

3.3 Consistency

In this section we consider the accuracy of the linear k-step method (39). For this purpose, as in the case of one-step methods, we introduce the notion of consistency error. Thus, suppose that y is a solution of the ODE (1). Then the consistency error of (39) is defined as

$$T_n := \frac{\sum_{j=0}^{k}\left[\alpha_j y(x_{n+j}) - h\beta_j y'(x_{n+j})\right]}{h\sum_{j=0}^{k}\beta_j} = \frac{\sum_{j=0}^{k}\left[\alpha_j y(x_{n+j}) - h\beta_j y'(x_{n+j})\right]}{h\,\sigma(1)}. \tag{45}$$

Of course, the definition requires implicitly that $\sigma(1) \neq 0$ (Rk: for any convergent LMM there holds $\sigma(1) = \rho'(1) \neq 0$; see proof of Theorem 8). Again, as in the case of one-step methods, the consistency error can be thought of as the residual that is obtained by inserting the solution of the ODE into the formula (39) and scaling this residual appropriately.

Definition 6 *The numerical scheme (39) is said to be* ***consistent*** *with the ODE (1) if the consistency error defined by (45) is such that for any* $\varepsilon > 0$ *there exists an* $h_\varepsilon > 0$ *for which* $|T_n| < \varepsilon$ *for all* $h \in (0, h_\varepsilon)$ *and for any* $(k+1)$ *points* $(x_n, y(x_n)), \ldots, (x_{n+k}, y(x_{n+k}))$ *on any solution curve in* R *of the IVP (1)–(2).*

Now let us suppose that the solution to the ODE is sufficiently smooth, and let us expand $y(x_{n+j})$ and $y'(x_{n+j})$ into a Taylor series about the point x_n and substitute these expansions into the numerator in (45) to obtain

$$T_n = \frac{C_0 y(x_n) + C_1 h y'(x_n) + C_2 h^2 y''(x_n) + \cdots}{h\,\sigma(1)}, \tag{46}$$

where $\sigma(1) \neq 0$, and the constants $C_0, C_1, C_2, \cdots \in \mathbb{R}$ are given by

$$C_0 = \sum_{j=0}^{k} \alpha_j, \qquad C_q = \sum_{j=0}^{k} \frac{j^q}{q!}\alpha_j - \sum_{j=0}^{k} \frac{j^{q-1}}{(q-1)!}\beta_j \quad \text{for } q \in \mathbb{N}.$$

These constants can also be computed from the following relations:

$$\begin{aligned}
C_0 &= \rho(1), \\
C_1 &= \rho'(1) - \sigma(1), \\
2\,C_2 &= \rho'(1) - 2\sigma'(1) + \rho''(1), \\
6\,C_3 &= \rho'(1) - 3\sigma'(1) + 3\rho''(1) - 3\sigma''(1) + \rho'''(1), \\
24\,C_4 &= \rho'(1) - 4\sigma'(1) + 7\rho''(1) - 12\sigma''(1) + 6\rho'''(1) - 4\sigma'''(1) + \rho''''(1), \\
120\,C_5 &= \rho'(1) - 5\sigma'(1) + 15\rho''(1) - 35\sigma''(1) + 25\rho'''(1) - 30\sigma'''(1) + 10\rho''''(1) - 5\sigma''''(1) + \rho'''''(1), \\
&\vdots \\
q!\,C_q &= \sum_{j=1}^{q-1} \Big(S(q,j)\rho^{(j)}(1) - qS(q-1,j)\sigma^{(j)}(1) \Big) + \rho^{(q)}(1), \qquad q \in \mathbb{N}_{\geq 2}.
\end{aligned}$$

Here, $S(q,j) := \frac{1}{j!}\sum_{i=0}^{j}(-1)^i \binom{j}{i}(j-i)^q$ denote the *Stirling numbers of the second kind*. For consistency we need that $T_n \to 0$ as $h \to 0$ and this requires that $C_0 = C_1 = 0$, i.e.,

$$\rho(1) = 0 \quad \text{and} \quad \rho'(1) = \sigma(1) \neq 0.$$

Let us observe that, according to this condition, if a linear multi-step method is consistent then it has a *simple* root on the unit circle at $z = 1$; thus the root condition is not violated by this zero.

Definition 7 *The numerical method (39) is said to have* ***order of accuracy*** *p (or* ***order of consistency*** *p) if $p \in \mathbb{N}$ is the largest natural number such that for any sufficiently smooth solution curve in* R *of the IVP (1)–(2) we have*

$$|T_n| = \mathcal{O}(h^p),$$

i.e., there exist constants $h_0, K > 0$ such that $|T_n| \leq Kh^p$ for all $h \in (0, h_0)$, for any $(k+1)$ points $(x_n, y(x_n)), \ldots, (x_{n+k}, y(x_{n+k}))$ on the solution curve.

We deduce from (46) that the method is of order of accuracy p iff

$$C_0 = C_1 = \cdots = C_p = 0 \quad \text{and} \quad C_{p+1} \neq 0.$$

In this case,

$$T_n = \frac{C_{p+1}}{\sigma(1)} h^p y^{(p+1)}(x_n) + \mathcal{O}(h^{p+1});$$

the number $C_{p+1} \neq 0$ is then called the **error constant** of the method.

Exercise 2 *Construct an implicit linear two-step method of maximum order of accuracy, containing one free parameter. Determine the order of accuracy and the error constant of the method.*

SOLUTION: Taking $\alpha_0 = a$ as parameter, the method has the form

$$y_{n+2} + \alpha_1 y_{n+1} + a y_n = h(\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n),$$

with $\beta_2 \neq 0$ and $a^2 + \beta_0^2 \neq 0$. Here, $\alpha_2 = 1$, $\alpha_0 = a$. We have $\rho(z) = z^2 + \alpha_1 z + a$ and $\sigma(z) = \beta_2 z^2 + \beta_1 z + \beta_0$. Assume $\sigma(1) = \beta_0 + \beta_1 + \beta_2 \neq 0$. We see that $\rho(1) = 1 + a + \alpha_1$, $\rho'(1) = 2 + \alpha_1$, $\rho''(1) = 2$, $\sigma(1) = \beta_0 + \beta_1 + \beta_2$, $\sigma'(1) = \beta_1 + 2\beta_2$, $\sigma''(1) = 2\beta_2$, and $\rho^{(i)}(1) = \sigma^{(i)}(1) = 0$ for $i \geq 3$. We have to determine four unknowns: α_1, β_2, β_1, β_0, so we require four equations; these will be arrived at by demanding that

$$\begin{aligned} C_0 &= \rho(1) = 1 + a + \alpha_1 = 0, \\ C_1 &= \rho'(1) - \sigma(1) = 2 + \alpha_1 - \beta_0 - \beta_1 - \beta_2 = 0, \\ 2\,C_2 &= \rho'(1) - 2\sigma'(1) + \rho''(1) = 4 + \alpha_1 - 2\beta_1 - 4\beta_2 = 0, \\ 6\,C_3 &= \rho'(1) - 3\sigma'(1) + 3\rho''(1) - 3\sigma''(1) + \rho'''(1) = 8 + \alpha_1 - 3\beta_1 - 12\beta_2 = 0. \end{aligned}$$

This gives $\alpha_1 = -1 - a$, $\beta_0 = -\frac{1}{12}(1 + 5a)$, $\beta_1 = \frac{2}{3}(1 - a)$, $\beta_2 = \frac{1}{12}(5 + a)$, and the resulting method is

$$y_{n+2} - (1 + a)y_{n+1} + a y_n = \frac{h}{12}\left((5 + a)f_{n+2} + 8(1 - a)f_{n+1} - (1 + 5a)f_n\right). \tag{47}$$

Note that $\sigma(1) = \beta_0 + \beta_1 + \beta_2 = 1 - a \neq 0$ iff $a \neq 1$. We have that

$$24\,C_4 = \rho'(1) - 4\sigma'(1) + 7\rho''(1) - 12\sigma''(1) + 6\rho'''(1) - 4\sigma'''(1) + \rho''''(1) = 16 + \alpha_1 - 4\beta_1 - 32\beta_2 = -(1 + a),$$

and that

$$\begin{aligned} 120\,C_5 &= \rho'(1) - 5\sigma'(1) + 15\rho''(1) - 35\sigma''(1) + 25\rho'''(1) - 30\sigma'''(1) + 10\rho''''(1) - 5\sigma''''(1) + \rho'''''(1) \\ &= 32 + \alpha_1 - 5\beta_1 - 80\beta_2 = -\frac{1}{3}(17 + 13a). \end{aligned}$$

If $a \notin \{-1, 1\}$, then $C_4 \neq 0$, and the method (47) is third order accurate. If, on the other hand, $a = -1$, then $C_4 = 0$ and $C_5 \neq 0$ and the method (47) becomes the Simpson rule method: $y_{n+2} - y_n = \frac{h}{3}(f_{n+2} + 4f_{n+1} + f_n)$, a fourth-order accurate two-step method. The error constant is $C_4 = -\frac{1}{24}(1 + a)$ if $a \neq -1$, and $C_5 = -\frac{1}{90}$ if $a = -1$. $\diamond$

Exercise 3 *Determine all values of the parameter $b \in \mathbb{R}\backslash\{0\}$, for which the LMM*

$$y_{n+3} + (2b - 3)(y_{n+2} - y_{n+1}) - y_n = hb(f_{n+2} + f_{n+1})$$

is zero-stable. Show that there exists a value of b for which the order of accuracy is 4. Is the method convergent for this value of b? Show that if the method is zero-stable then its order of accuracy is 2.

SOLUTION: According to the root condition, this LMM is zero-stable iff all roots of its first characteristic polynomial

$$\rho(z) = z^3 + (2b - 3)(z^2 - z) - 1$$

lie in the closed unit disc $\bar{D}_1(0)$, and those on the unit circle $\partial D_1(0)$ are simple. Clearly, $z_1 = 1$ is a root of ρ and we note that

$$\rho(z) = (z - 1)\rho_1(z), \qquad \rho_1(z) := z^2 - 2(1 - b)z + 1.$$

Thus the method is zero-stable if, and only if, all roots of the polynomial ρ_1 belong to the closed unit disc, and those on the unit circle are simple and differ from $z_1 = 1$.

Suppose that the method is zero-stable. Then, it follows that $b \neq 0$ and $b \neq 2$, since these values of b correspond to double roots of ρ_1 on the unit circle, respectively, $z = 1$ and $z = -1$. Since the product of the two roots of ρ_1 is equal to 1 and neither of them is equal to ± 1, it follows that they are strictly complex; hence the discriminant of the quadratic polynomial ρ_1 is negative. Namely, $4(1 - b)^2 - 4 < 0$, i.e., $b \in (0, 2)$.

Conversely, suppose that $b \in (0, 2)$. Then the roots of ρ are

$$z_1 = 1, \qquad z_2 = (1 - b) + i\sqrt{1 - (1 - b)^2}, \qquad z_3 = (1 - b) - i\sqrt{1 - (1 - b)^2}.$$

Since $|z_2| = |z_3| = 1$, $z_2 \neq 1$, $z_3 \neq 1$, and $z_2 \neq z_3$, all roots of ρ lie on the unit circle and they are simple. Hence the method is zero-stable. To summarise, the method is zero-stable iff $b \in (0, 2)$.

Let us analyse the order of accuracy: Note that $\sigma(z) = b(z^2 + z)$. We see that $\rho(1) = 0$, $\rho'(1) = 2b$, $\rho''(1) = 4b$, $\rho'''(1) = 6$, and $\rho^{(i)}(1) = 0$ for $i \geq 4$, and we see that $\sigma(1) = 2b \neq 0$, $\sigma'(1) = 3b$, $\sigma''(1) = 2b$, and $\sigma^{(i)}(1) = 0$ for $i \geq 3$. We compute $C_0 = 0$, $C_1 = 0$, $C_2 = 0$, $C_3 = \frac{6-b}{6}$, $C_4 = \frac{36-6b}{24} = \frac{6-b}{4}$, and $C_5 = \frac{150-23b}{120}$. We find that

$$\begin{aligned} T_n &= \frac{C_0}{\sigma(1)}h^{-1}y(x_n) + \frac{C_1}{\sigma(1)}y'(x_n) + \frac{C_2}{\sigma(1)}hy''(x_n) + \frac{C_3}{\sigma(1)}h^2y^{(3)}(x_n) + \frac{C_4}{\sigma(1)}h^3y^{(4)}(x_n) + \frac{C_5}{\sigma(1)}h^4y^{(5)}(x_n) + \mathcal{O}(h^5) \\ &= \frac{6-b}{12b}h^2y^{(3)}(x_n) + \frac{6-b}{8b}h^3y^{(4)}(x_n) + \frac{150-23b}{240b}h^4y^{(5)}(x_n) + \mathcal{O}(h^5). \end{aligned}$$

If $b = 6$, then $T_n = \frac{1}{120}h^4y^{(5)}(x_n) + \mathcal{O}(h^5)$ and so the method is 4th order accurate. As $b = 6$ does not belong to the interval $(0, 2)$, we deduce that the method is *not* zero-stable for $b = 6$. Finally, since zero-stability requires $b \in (0, 2)$, in which case $\frac{6-b}{12b} \neq 0$, it follows that if the method is zero-stable then its order of accuracy is 2. $\diamond$

3.4 Convergence

The concepts of zero-stability and consistency are of great theoretical importance. However, what matters most from the practical point of view is that the numerically computed approximations y_n at the mesh-points x_n, $n \in \{0, 1, \ldots, N\}$, are close to those of the true solution $y(x_n)$ at these point, and that the **global error** $e_n = y(x_n) - y_n$ between the numerical approximation y_n and the exact solution-value $y(x_n)$ decays when the step size h is reduced. We introduce the following definition.

Definition 8 *The LMM (39) is said to be* ***convergent*** *if, for all IVPs (1)–(2) subject to the hypotheses of Theorem 1, we have that*

$$\lim_{\substack{h\to 0 \\ nh=x-x_0}} y_n = y(x) \tag{48}$$

holds for all $x \in [x_0, X_M]$ and for all solutions $\{y_n\}_{n=0}^N$ of the difference equation (39) with ***consistent starting conditions****, i.e. with starting conditions $y_s = \eta_s(h)$, $s \in \{0, 1, \ldots, k-1\}$, for which $\lim_{h\to 0} \eta_s(h) = y_0$, $s \in \{0, 1, \ldots, k-1\}$.*

We emphasise here that Definition 8 requires that (48) holds *not only* for those sequences $\{y_n\}_{n=0}^N$ which have been generated from (39) using *exact* starting values $y_s = y(x_s)$, $s = 0, 1, \ldots, k-1$, but also for all sequences $\{y_n\}_{n=0}^N$ whose starting values $\eta_s(h)$ tend to the correct value, y_0, as $h \to 0$. This assumption is made as in practice, exact starting values are usually not available and have to be computed numerically.

In the remainder of this section we shall investigate the interplay between zero-stability, consistency and convergence; the section culminates in Dahlquist's Equivalence Theorem which, under some technical assumptions, states that for a consistent LMM zero-stability is necessary and sufficient for convergence.

3.4.1 Necessary conditions for convergence

In this section we show that both zero-stability and consistency are necessary for convergence.

Theorem 7 *A necessary condition for the convergence of the LMM (39) is that it is zero-stable.*

Proof: Let us suppose that the LMM (39) is convergent; we need to show that it is then zero-stable. We consider the IVP $y'(x) = 0$, $y(0) = 0$, on the interval $[0, X_M]$, $X_M > 0$, whose solution is $y \equiv 0$. Applying (39) to this problem yields the difference equation

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = 0. \tag{49}$$

Since the method is assumed to be convergent, for any $x \in [0, X_M]$, we have that

$$\lim_{\substack{h\to 0 \\ nh=x}} y_n = 0, \tag{50}$$

for all solutions of (49) satisfying $y_s = \eta_s(h)$, $s \in \{0, 1, \ldots, k-1\}$, where

$$\lim_{h\to 0} \eta_s(h) = 0, \qquad s \in \{0, 1, \ldots, k-1\}. \tag{51}$$

Let $z = r\mathrm{e}^{i\phi}$ with $r \geq 0$, $\phi \in [0, 2\pi)$ be a root of the first characteristic polynomial ρ. Then, the numbers

$$y_n = hr^n \cos(n\phi)$$

define a solution to (49) satisfying (51). Indeed, using that $\sum_{j=0}^k \alpha_j r^j \cos(j\phi) = \mathrm{Re}[\rho(re^{i\phi})] = 0$ and $\sum_{j=0}^k \alpha_j r^j \sin(j\phi) = \mathrm{Im}[\rho(re^{i\phi})] = 0$, we find

$$\sum_{j=0}^k \alpha_j y_{n+j} = \sum_{j=0}^k \alpha_j h r^{n+j} \cos((n+j)\phi) = hr^n \left[\cos(n\phi) \sum_{j=0}^k \alpha_j r^j \cos(j\phi) - \sin(n\phi) \sum_{j=0}^k \alpha_j r^j \sin(j\phi)\right] = 0.$$

We observe that if $\phi \notin \{0, \pi\}$, then

$$\frac{y_n^2 - y_{n+1}y_{n-1}}{\sin^2(\phi)} = h^2 r^{2n} \frac{\cos^2(n\phi) - \cos((n+1)\phi)\cos((n-1)\phi)}{\sin^2(\phi)} = h^2 r^{2n}.$$

Since the left-hand side of this identity converges to 0 as $h \to 0$, $n \to \infty$, $nh = x$, we must have

$$\lim_{n\to\infty} \left(\frac{x}{n}\right)^2 r^{2n} = 0$$

for any $x \in [0, X_M]$. This implies that $r \in [0, 1]$ (recall $r \geq 0$). In other words, we have proved that any root of the first characteristic polynomial of (39) lies in the closed unit disc $\bar{D}_1(0)$.

Next we prove that any root of the first characteristic polynomial of (39) that lies on the unit circle $\partial D_1(0)$ must be *simple*. Assume, instead, that $z = r\mathrm{e}^{i\phi}$, is a *multiple* root of $\rho(z)$, with $|z| = r = 1$ and $\phi \in [0, 2\pi)$. We shall prove below that this contradicts our assumption that the method (49) is convergent. Similarly to before, we see that

$$y_n = \sqrt{h}\, n\, r^n \cos(n\phi) = \sqrt{h}\, n \cos(n\phi) \tag{52}$$

defines a solution to (49), where we have used $r = 1$ in the second equality. This satisfies (51) as for any $s \in \{0, 1, \ldots, k-1\}$ we have

$$|\eta_s(h)| = |y_s| \leq \sqrt{h}\, s \leq \sqrt{h}(k-1) \to 0 \quad \text{as} \quad h \to 0.$$

If $\phi \in \{0, \pi\}$, it follows from (52) with $nh = x$ that

$$|y_n| = \sqrt{x}\sqrt{n}, \tag{53}$$

and we deduce from (53) that $\lim_{n\to\infty, nh=x} |y_n| = \infty$ whenever $x \neq 0$, which contradicts (50). If, on the other hand, $\phi \notin \{0, \pi\}$, then

$$\frac{z_n^2 - z_{n+1}z_{n-1}}{\sin^2(\phi)} = 1, \tag{54}$$

where $z_n = \frac{1}{n\sqrt{h}} y_n = \frac{\sqrt{h}}{x} y_n$. Since, by (50), z_n converges to 0 as $h \to 0$, $n \to \infty$, $nh = x$, it follows that the left-hand side of (54) converges to 0 as $h \to 0$, $n \to \infty$, $nh = x$, which is a contradiction (as the right-hand side converges to 1).

To summarise, we have proved that all roots of the first characteristic polynomial ρ of the LMM (39) lie in the closed unit disc $\bar{D}_1(0)$, and those which belong to the unit circle $\partial D_1(0)$ are simple. In view of Theorem 6, the LMM is zero-stable. $\diamond$

Theorem 8 *A necessary condition for the convergence of the LMM (39) is that it is consistent.*

PROOF: Let us suppose that the LMM (39) is convergent; we need to show that it is then consistent.

Let us first show that $C_0 = 0$. We consider the IVP $y'(x) = 0$, $y(0) = 1$, on the interval $[0, X_M]$, $X_M > 0$, whose solution is $y \equiv 1$. Applying (39) to this problem yields the difference equation

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = 0. \tag{55}$$

We supply "exact" starting values for the numerical method; namely, $y_s = 1$, $s \in \{0, 1, \ldots, k-1\}$. Given that by hypothesis the method is convergent, we deduce that

$$\lim_{\substack{h \to 0 \\ nh = x}} y_n = 1. \tag{56}$$

Since in the present case y_n is independent of the choice of h, (56) is equivalent to saying that

$$\lim_{n \to \infty} y_n = 1. \tag{57}$$

Passing to the limit $n \to \infty$ in (55), we deduce that

$$C_0 = \rho(1) = \sum_{j=0}^{k} \alpha_j = 0. \tag{58}$$

In order to show that $C_1 = 0$, we now consider the IVP $y'(x) = 1$, $y(0) = 0$, on the interval $[0, X_M]$, $X_M > 0$, whose solution is $y(x) = x$. The difference equation (39) now becomes

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j, \tag{59}$$

where $X_M - x_0 = X_M - 0 = Nh$ and $n \in \{0, 1, \ldots, N-k\}$. For a convergent method every solution of (59) satisfying

$$\lim_{h \to 0} \eta_s(h) = 0, \qquad s \in \{0, 1, \ldots k-1\}, \tag{60}$$

where $y_s = \eta_s(h)$, $s \in \{0, 1, \ldots k-1\}$, must also satisfy

$$\lim_{\substack{h \to 0 \\ nh = x}} y_n = x. \tag{61}$$

Since according to Theorem 7 zero-stability is necessary for convergence, we may take it for granted that the first characteristic polynomial ρ of the method does not have a multiple root on the unit circle $\partial D_1(0)$; therefore

$$\rho'(1) = \sum_{j=1}^{k} j\alpha_j \neq 0.$$

Let the sequence $\{y_n\}_{n=0}^{N}$ be defined by $y_n = Knh$, where

$$K = \frac{\sigma(1)}{\rho'(1)} = \frac{\sum_{j=0}^{k} \beta_j}{\sum_{j=1}^{k} j\alpha_j}; \tag{62}$$

this sequence clearly satisfies (60) and is a solution of (59) as

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = hK \sum_{j=0}^{k} \alpha_j (n+j) = KnhC_0 + Kh\rho'(1) = h\sigma(1) = h \sum_{j=0}^{k} \beta_j.$$

Furthermore, (61) implies that

$$x = y(x) = \lim_{\substack{h \to 0 \\ nh = x}} y_n = \lim_{\substack{h \to 0 \\ nh = x}} Knh = Kx$$

for any $x \in [0, X_M]$, and therefore $K = 1$. Hence, from (62), we obtain that

$$C_1 = \rho'(1) - \sigma(1) = 0;$$

equivalently, $\rho'(1) = \sigma(1)$. $\diamond$

3.4.2 Sufficient conditions for convergence

Theorem 9 (**Dahlquist**) *For a LMM that is consistent with the ODE (1) where f is assumed to satisfy a Lipschitz condition, and starting with consistent initial data, zero-stability is necessary and sufficient for convergence. Moreover if the solution y has continuous derivatives of order $(p+1)$ and consistency error $\mathcal{O}(h^p)$, then the global error $e_n = y(x_n) - y_n$ is also $\mathcal{O}(h^p)$, i.e. the method is p-th order convergent.*

According to Dahlquist's theorem, if a LMM is not zero-stable its global error cannot be made arbitrarily small by taking the mesh size h sufficiently small for any sufficiently accurate initial data. In fact, if the root condition is violated then there exists a solution to the LMM which will grow by an arbitrarily large factor in a fixed interval of x, however accurate the starting conditions are. This highlights the importance of the concept of zero-stability and indicates its relevance in practical computations.

3.5 Maximum order of accuracy of a zero-stable linear multi-step method

Let us suppose that we have already chosen the coefficients α_j, $j \in \{0, 1, \ldots, k\}$, of the k-step method (39). The question we shall be concerned with in this section is how to choose the coefficients β_j, $j \in \{0, 1, \ldots, k\}$, so that the order of accuracy of the resulting method (39) is as high as possible.

In view of Theorem 9 we shall only be interested in consistent methods, so it is natural to assume that the first and second characteristic polynomials ρ and σ associated with (39) satisfy $\rho(1) = 0$, $\rho'(1) - \sigma(1) = 0$, with $\sigma(1) \neq 0$.

By inspection, the linear k-step method (39) has $2k + 2$ coefficients: α_j, β_j, $j \in \{0, 1, \ldots, k\}$, of which α_k is taken to be 1 by normalisation. This leaves us with $2k + 1$ free parameters if the method is implicit and $2k$ free parameters if the method is explicit (because in the latter case β_k is fixed to have value 0). According to (46), if the method is required to have order p, the $p + 1$ linear relationships $C_0 = 0, C_1 = 0, \ldots, C_p = 0$ involving α_j, β_j, $j \in \{0, 1, \ldots, k\}$, must be satisfied. Thus, in the case of the implicit method, we can impose $p + 1 = 2k + 1$ linear constraints $C_0 = 0$, $C_1 = 0$, $\ldots$, $C_{2k} = 0$ to determine the unknown constants, yielding a method of order $p = 2k$. Similarly, in the case of an explicit method, the highest order we can expect is $p = 2k - 1$. Unfortunately, there is no guarantee that such methods will be zero-stable. Indeed, in a paper published in 1956 Dahlquist proved that there is *no* consistent, zero-stable k-step method whose order exceeds $k + 2$. Therefore the maximum orders $2k$ and $2k - 1$ cannot be attained without violating the condition of zero-stability when $k \geq 3$. We formalise these facts in the next theorem.

Theorem 10 *There is no zero-stable linear k-step method whose order exceeds $k + 1$ if k is odd or $k + 2$ if k is even.*

Definition 9 *A zero-stable linear k-step method of order $k+2$ is said to be an* ***optimal method****.*

Let us note that it can be shown that all roots of the first characteristic polynomial ρ associated with an optimal LMM have modulus 1.

Example 5 *As $k+2 = 2k$ iff $k = 2$ and the Simpson rule method is the only zero-stable linear 2-step method of maximum order (see Exercise 2), we deduce that the Simpson rule method is the only zero-stable LMM which is both of maximum order ($2k = 4$) and optimal ($k+2 = 4$).*

Optimal methods have certain disadvantages in terms of their stability properties; we shall return to this question later on in the notes. Linear k-step methods for which the first characteristic polynomial has the form $\rho(z) = z^k - z^{k-1}$ are called **Adams methods**. Explicit Adams methods are referred to as **Adams–Bashforth methods**, while implicit Adams methods are termed **Adams–Moulton methods**. Linear k-step methods for which $\rho(z) = z^k - z^{k-2}$ are called **Nyström methods** if explicit and **Milne–Simpson methods** if implicit. All these methods are zero-stable.

3.6 Absolute stability of linear multi-step methods

Up to now we have been concerned with the stability and accuracy properties of LMMs in the asymptotic limit of $h \to 0$, $n \to \infty$, nh fixed. However, it is of practical significance to investigate the performance of methods in the case of $h > 0$ fixed and $n \to \infty$. Specifically, we would like to ensure that when applied to an IVP whose solution decays to zero as $x \to \infty$, the LMM exhibits a similar behaviour, for $h > 0$ fixed and $x_n = x_0 + nh \to \infty$. The canonical model problem with exponentially decaying solution is

$$y'(x) = \lambda y(x), \quad x > 0, \qquad y(0) = y_0, \tag{63}$$

where $\lambda < 0$ and $y_0 \neq 0$. Indeed, the true solution is $y(x) = y_0\, e^{\lambda x}$ and therefore, $\lim_{x\to\infty} y(x) = 0$. Let us recall from Section 1 that the solution is asymptotically stable.

Now consider the linear k-step method (39) and apply it to the model problem (63). Noting that $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, z) = \lambda z$, this yields the linear difference equation

$$0 = \sum_{j=0}^{k} \left(\alpha_j y_{n+j} - h\beta_j f(x_{n+j}, y_{n+j})\right) = \sum_{j=0}^{k} \left(\alpha_j - h\lambda\beta_j\right) y_{n+j}.$$

Since the general solution y_n to this homogeneous difference equation can be expressed as a linear combination of powers of roots of the associated characteristic polynomial

$$\pi(z; \bar{h}) = \rho(z) - \bar{h}\,\sigma(z), \qquad z \in \mathbb{C}, \qquad (\bar{h} := \lambda h) \tag{64}$$

it follows that y_n will converge to zero for $h > 0$ fixed and $n \to \infty$ iff all roots of $\pi(z; \bar{h})$ have modulus less than 1. The kth degree polynomial $\pi(z; \bar{h})$ defined by (64) is called the **stability polynomial** of the linear k-step method with first and second characteristic polynomials ρ and σ, respectively.

Definition 10 *The LMM (39) is called* ***absolutely stable*** *for a given $\bar{h}$ iff for that $\bar{h}$ all the roots $r_s = r_s(\bar{h})$ of the stability polynomial $z \mapsto \pi(z; \bar{h})$ defined by (64) satisfy $|r_s| < 1$, $s \in \{1, \ldots, k\}$. Otherwise, the method is called* ***absolutely unstable****. An interval $(\alpha, \beta) \subset \mathbb{R}$ is called the* ***interval of absolute stability*** *if it is the largest open interval with the property that the method is absolutely stable for all $\bar{h} \in (\alpha, \beta)$. If the method is absolutely unstable for all $\bar{h}$, it is said to have* ***no interval of absolute stability****.*

Since for $\lambda > 0$ the solution of (63) exhibits exponential growth, it is reasonable to expect that a consistent and zero-stable (and, therefore, convergent) LMM will have a similar behaviour for $h > 0$ sufficiently small, and will be therefore absolutely unstable for small $\bar{h} = \lambda h > 0$. According to the next theorem, this is indeed the case.

Theorem 11 *Every consistent and zero-stable LMM is absolutely unstable for small positive* $\bar{h}$.

PROOF: As the method is consistent, there exists $p \in \mathbb{N}$ such that $C_0 = C_1 = \cdots = C_p = 0$ and $C_{p+1} \neq 0$. From the problem sheets, we know that then

$$\pi(\mathrm{e}^{\bar{h}}; \bar{h}) = \sum_{q=0}^{\infty} \bar{h}^q C_q = \sum_{q=p+1}^{\infty} C_q \bar{h}^q = \mathcal{O}(\bar{h}^{p+1}). \tag{65}$$

On the other hand, noting that the polynomial $\pi(z; \bar{h})$ can be written in the factorised form

$$\pi(z; \bar{h}) = (\alpha_k - \bar{h}\beta_k) \prod_{s=1}^{k} (z - r_s)$$

where $r_s = r_s(\bar{h})$, $s \in \{1, \ldots, k\}$, signify the roots of $z \mapsto \pi(z; \bar{h})$, we deduce that

$$\pi(\mathrm{e}^{\bar{h}}; \bar{h}) = (\alpha_k - \bar{h}\beta_k) \prod_{s=1}^{k} (e^{\bar{h}} - r_s(\bar{h})). \tag{66}$$

As $\bar{h} \to 0$, $\alpha_k - \bar{h}\beta_k \to \alpha_k \neq 0$ and $r_s(\bar{h}) \to \zeta_s$, $s \in \{1, \ldots, k\}$, where ζ_s, $s \in \{1, \ldots, k\}$, are the roots of ρ. Since, by assumption, the method is consistent, 1 is a root of ρ; furthermore, by zero-stability 1 is a simple root of ρ. Let us suppose, for the sake of definiteness that it is ζ_1 that is equal to 1. Then, $\zeta_s \neq 1$ for $s \neq 1$ and therefore

$$\lim_{\bar{h} \to 0} (\mathrm{e}^{\bar{h}} - r_s(\bar{h})) = 1 - \zeta_s \neq 0, \qquad s \neq 1.$$

We deduce from (66) that the only factor of $\pi(\mathrm{e}^{\bar{h}}; \bar{h})$ that converges to 0 as $\bar{h} \to 0$ is $\mathrm{e}^{\bar{h}} - r_1(\bar{h})$ (the other factors converge to nonzero constants). Now, by (65), $\pi(e^{\bar{h}}; \bar{h}) = \mathcal{O}(\bar{h}^{p+1})$, so it follows that

$$e^{\bar{h}} - r_1(\bar{h}) = \mathcal{O}(\bar{h}^{p+1}).$$

Thus we have shown that $r_1(\bar{h}) = e^{\bar{h}} + \mathcal{O}(\bar{h}^{p+1})$. This implies that $r_1(\bar{h}) > 1 + \frac{1}{2}\bar{h}$ for small positive $\bar{h}$. That completes the proof. $\diamond$

According to the definition adopted in the previous section, an optimal k-step method is a zero-stable linear k-step method of order $k + 2$. Recall that all roots of the first characteristic polynomial of an optimal k-step method lie on the unit circle. It can be shown that an optimal LMM has no interval of absolute stability.

3.6.1 General methods for locating the interval of absolute stability

In this section we shall describe two methods for identifying the endpoints of the interval of absolute stability. The first of these is based on the Schur criterion, the second on the Routh–Hurwitz criterion.

The Schur criterion

Consider the polynomial

$$\phi : \mathbb{C} \to \mathbb{C}, \quad \phi(z) = c_k z^k + c_{k-1} z^{k-1} + \cdots + c_1 z + c_0,$$

with $c_0, c_1, \ldots, c_k \in \mathbb{C}$ and $c_k \neq 0$, $c_0 \neq 0$. The polynomial ϕ is said to be a **Schur polynomial** if each of its roots r_s, satisfies $|r_s| < 1$, i.e., $r_s \in D_1(0)$ for all $s \in \{1, \ldots, k\}$.

Let us consider the polynomial

$$\hat{\phi} : \mathbb{C} \to \mathbb{C}, \quad \hat{\phi}(z) = \bar{c}_0 z^k + \bar{c}_1 z^{k-1} + \cdots + \bar{c}_{k-1} z + \bar{c}_k,$$

where $\bar{c}_j$ denotes the complex conjugate of c_j for $j \in \{1, \ldots, k\}$. Further, let us define

$$\phi_1 : \mathbb{C} \to \mathbb{C}, \quad \phi_1(z) = \frac{\hat{\phi}(0)\phi(z) - \phi(0)\hat{\phi}(z)}{z}.$$

The following key result is stated without proof.

Theorem 12 (**Schur's Criterion**) *The polynomial ϕ is a Schur polynomial iff $|\hat{\phi}(0)| > |\phi(0)|$ and ϕ_1 is a Schur polynomial.*

Exercise 4 *Use Schur's criterion to determine the interval of absolute stability of the LMM*

$$y_{n+2} - y_n = \frac{h}{2}\left(f_{n+1} + 3f_n\right).$$

SOLUTION: The first and second characteristic polynomials of the method are

$$\rho(z) = z^2 - 1, \qquad \sigma(z) = \frac{1}{2}(z+3).$$

Therefore the stability polynomial is

$$\pi(z; \bar{h}) = \rho(z) - \bar{h}\sigma(z) = z^2 - \frac{1}{2}\bar{h}z - \left(1 + \frac{3}{2}\bar{h}\right).$$

Suppose $1 + \frac{3}{2}\bar{h} \neq 0$, i.e., $\bar{h} \neq -\frac{2}{3}$ so that we can apply the Schur criterion. We have

$$\hat{\pi}(z; \bar{h}) = -\left(1 + \frac{3}{2}\bar{h}\right) z^2 - \frac{1}{2}\bar{h}z + 1.$$

Clearly, $|\hat{\pi}(0; \bar{h})| = 1 > |1 + \frac{3}{2}\bar{h}| = |\pi(0; \bar{h})|$ iff $\bar{h} \in (-\frac{4}{3}, 0)$. For such $\bar{h}$, the polynomial

$$\pi_1(z; \bar{h}) = \frac{\hat{\pi}(0; \bar{h})\pi(z; \bar{h}) - \pi(0; \bar{h})\hat{\pi}(z; \bar{h})}{z} = -\frac{1}{2}\bar{h}\left(2 + \frac{3}{2}\bar{h}\right)(3z + 1)$$

has the unique root $r_1 = -\frac{1}{3}$ and is, therefore, a Schur polynomial. We deduce from Schur's criterion that $z \mapsto \pi(z; \bar{h})$, $\bar{h} \neq -\frac{2}{3}$, is a Schur polynomial iff $\bar{h} \in (-\frac{4}{3}, 0)$. Finally, for $\bar{h} = -\frac{2}{3}$, we see that $\pi(z; -\frac{2}{3}) = z(z + \frac{1}{3})$ is Schur polynomial. We conclude that the interval of absolute stability of the method is $(-\frac{4}{3}, 0)$. $\diamond$

The Routh–Hurwitz criterion

Consider the mapping

$$m : D_1(0) \to \mathbb{C}^-, \quad m(z) := \frac{z-1}{z+1},$$

where $\mathbb{C}^- := \{z \in \mathbb{C} : \mathrm{Re}(z) < 0\}$. Note that m is a bijection and its inverse is given by

$$m^{-1} : \mathbb{C}^- \to D_1(0), \quad m^{-1}(z) := \frac{1+z}{1-z}.$$

Consider

$$\pi\left(\frac{1+z}{1-z}; \bar{h}\right) = \rho\left(\frac{1+z}{1-z}\right) - \bar{h}\sigma\left(\frac{1+z}{1-z}\right).$$

By multiplying this with $(1-z)^k$, we obtain a polynomial of the form

$$(1-z)^k \left[\pi\left(\frac{1+z}{1-z}; \bar{h}\right)\right] = a_0 z^k + a_1 z^{k-1} + \cdots + a_k. \tag{67}$$

The roots of the stability polynomial $z \mapsto \pi(z; \bar{h})$ lie inside $D_1(0)$ iff the roots of the polynomial (67) lie in $\mathbb{C}^-$ and $a_0 \neq 0$.

Theorem 13 (**Routh–Hurwitz Criterion**) *The roots of a polynomial $P : \mathbb{C} \to \mathbb{C}$, $P(z) := a_0 z^k + a_1 z^{k-1} + \cdots + a_k$ with $a_0, \ldots, a_k \in \mathbb{R}$ and $a_0 > 0$ lie in $\mathbb{C}^-$ iff all leading principal minors of the matrix*

$$H := \begin{bmatrix} a_1 & a_3 & a_5 & \cdots & a_{2k-1} \\ a_0 & a_2 & a_4 & \cdots & a_{2k-2} \\ 0 & a_1 & a_3 & \cdots & a_{2k-3} \\ 0 & a_0 & a_2 & \cdots & a_{2k-4} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & a_k \end{bmatrix} \in \mathbb{R}^{k \times k}$$

are positive, where we set $a_j := 0$ if $j > k$. In particular, for $k \in \{2, 3, 4\}$, the necessary and sufficient conditions for ensuring that all roots of P lie in $\mathbb{C}^-$ are as follows:

a) for $k = 2$: $a_1 > 0$, $a_2 > 0$;

b) for $k = 3$: $a_1 > 0$, $a_2 > 0$, $a_3 > 0$, $a_1 a_2 - a_3 a_0 > 0$;

c) for $k = 4$: $a_1 > 0$, $a_2 > 0$, $a_3 > 0$, $a_4 > 0$, $a_1 a_2 a_3 - a_0 a_3^2 - a_4 a_1^2 > 0$.

Exercise 5 *Use the Routh–Hurwitz criterion to determine the interval of absolute stability of the LMM from the previous exercise.*

SOLUTION: We have the stability polynomial

$$\pi(z; \bar{h}) = z^2 - \frac{1}{2}\bar{h}z - \left(1 + \frac{3}{2}\bar{h}\right).$$

We compute

$$P(z) := (1-z)^2 \left[\pi\left(\frac{1+z}{1-z}; \bar{h}\right)\right] = -\bar{h}z^2 + (4 + 3\bar{h})z - 2\bar{h} =: a_0 z^2 + a_1 z + a_2.$$

The roots of the stability polynomial $z \mapsto \pi(z; \bar{h})$ lie inside $D_1(0)$ iff the roots of P lie in $\mathbb{C}^-$ and $a_0 \neq 0$. So, we are unstable for $\bar{h} = 0$. For $\bar{h} \neq 0$, we use the Routh–Hurwitz criterion:

Case $\bar{h} < 0$: Then, applying Theorem 13 a) to P, we find that all roots of P lie in $\mathbb{C}^-$ iff $4 + 3\bar{h} > 0$ and $-2\bar{h} > 0$, i.e., iff $\bar{h} \in (-\frac{4}{3}, 0)$.
Case $\bar{h} > 0$: Then, applying Theorem 13 a) to $-P(z) = \bar{h}z^2 - (4 + 3\bar{h})z + 2\bar{h}$, we find that all roots of P lie in $\mathbb{C}^-$ iff all roots of $-P$ lie in $\mathbb{C}^-$ iff $-(4 + 3\bar{h}) > 0$ and $2\bar{h} > 0$, which is impossible.
Altogether, the interval of absolute stability is $(-\frac{4}{3}, 0)$. $\diamond$

We conclude this section by listing the intervals of absolute stability $(\alpha, 0)$ of k-step Adams–Bashforth and Adams–Moulton methods, for $k = 1, 2, 3, 4$. We shall also supply the orders p^* and p and error constants C_{p^*+1} and C_{p+1}, respectively, of these methods. The verification is left as exercise.

k-step Adams–Bashforth (explicit) methods:

(1) $k = 1$, $p^* = 1$, $C_{p^*+1} = \frac{1}{2}$, $\alpha = -2$,

$$y_{n+1} - y_n = h f_n;$$

(2) $k = 2$, $p^* = 2$, $C_{p^*+1} = \frac{5}{12}$, $\alpha = -1$,

$$y_{n+2} - y_{n+1} = \frac{h}{2}(3f_{n+1} - f_n);$$

(3) $k = 3$, $p^* = 3$, $C_{p^*+1} = \frac{3}{8}$, $\alpha = -\frac{6}{11}$,

$$y_{n+3} - y_{n+2} = \frac{h}{12}(23f_{n+2} - 16f_{n+1} + 5f_n);$$

(4) $k = 4$, $p^* = 4$, $C_{p^*+1} = \frac{251}{720}$, $\alpha = -\frac{3}{10}$,

$$y_{n+4} - y_{n+3} = \frac{h}{24}(55f_{n+3} - 59f_{n+2} + 37f_{n+1} - 9f_n).$$

k-step Adams–Moulton (implicit) methods:

(1) $k = 1$, $p = 2$, $C_{p+1} = -\frac{1}{12}$, $\alpha = -\infty$,

$$y_{n+1} - y_n = \frac{h}{2}(f_{n+1} + f_n);$$

(2) $k = 2$, $p = 3$, $C_{p+1} = -\frac{1}{24}$, $\alpha = -6$,

$$y_{n+2} - y_{n+1} = \frac{h}{12}(5f_{n+2} + 8f_{n+1} - f_n);$$

(3) $k = 3$, $p = 4$, $C_{p+1} = -\frac{19}{720}$, $\alpha = -3$,

$$y_{n+3} - y_{n+2} = \frac{h}{24}(9f_{n+3} + 19f_{n+2} - 5f_{n+1} + f_n);$$

(4) $k = 4$, $p = 5$, $C_{p+1} = -\frac{27}{1440}$, $\alpha = -\frac{90}{49}$,

$$y_{n+4} - y_{n+3} = \frac{h}{720}(251f_{n+4} + 646f_{n+3} - 264f_{n+2} + 106f_{n+1} - 19f_n).$$

We notice that the k-step Adams–Moulton (implicit) method has a larger interval of absolute stability and smaller error constant than the k-step Adams–Bashforth (explicit) method.

4 Stiff problems

Let us consider an IVP for a *system* of m ODEs of the form

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \qquad \mathbf{y}(x_0) = \mathbf{y}_0, \tag{68}$$

where $\mathbf{y}(x) = (\mathbf{y}_1(x), \ldots, \mathbf{y}_m(x))^{\mathrm{T}}$. A linear k-step method for the numerical solution of (68) has the form

$$\sum_{j=0}^{k} \alpha_j \mathbf{y}_{n+j} = h \sum_{j=0}^{k} \beta_j \mathbf{f}_{n+j}, \tag{69}$$

where $\mathbf{f}_{n+j} := \mathbf{f}(x_{n+j}, \mathbf{y}_{n+j})$. Let us suppose, for simplicity, that $\mathbf{f}(x, \mathbf{y}) = A\mathbf{y} + \mathbf{b}$ where $A \in \mathbb{C}^{m\times m}$ and $\mathbf{b} \in \mathbb{C}^m$; then (69) becomes

$$\sum_{j=0}^{k} (\alpha_j I_m - h\beta_j A)\mathbf{y}_{n+j} = h\sigma(1)\mathbf{b}, \tag{70}$$

where $\sigma(1) = \sum_{j=0}^{k} \beta_j \neq 0$ and I_m is the $m \times m$ identity matrix. Let us suppose that the eigenvalues $\lambda_1, \ldots, \lambda_m \in \mathbb{C}$ of A are distinct. Then, there exists an invertible matrix $H \in \mathbb{C}^{m\times m}$ such that

$$H^{-1}AH = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{pmatrix} =: \operatorname{diag}(\lambda_1, \ldots, \lambda_m) =: \Lambda. \tag{71}$$

Let us define $\mathbf{z}_{n+j} := H^{-1}\mathbf{y}_{n+j}$ for $j \in \{0, 1, \ldots, k\}$, and $\mathbf{c} := h\sigma(1)H^{-1}\mathbf{b}$. Then, (70) can be rewritten as

$$\sum_{j=0}^{k} (\alpha_j I_m - h\beta_j \Lambda)\mathbf{z}_{n+j} = \sum_{j=0}^{k} (\alpha_j H^{-1} - h\beta_j \Lambda H^{-1})\mathbf{y}_{n+j} = H^{-1} \sum_{j=0}^{k} (\alpha_j I_m - h\beta_j A)\mathbf{y}_{n+j} = \mathbf{c}, \tag{72}$$

or, in component-wise form,

$$\sum_{j=0}^{k} (\alpha_j - h\beta_j \lambda_i) z_{n+j,i} = c_i,$$

where $z_{n+j,i}$ and c_i, $i \in \{1, \ldots, m\}$, are the components of $\mathbf{z}_{n+j}$ and $\mathbf{c}$ respectively. Each of these m equations is completely decoupled from the other $m - 1$ equations. Thus we are now in the framework of Section 3 where we considered LMMs for a single ODE. However, there is a new feature here: as the numbers λ_i, $i \in \{1, \ldots, m\}$, are eigenvalues of the matrix A, they need not be real. As a consequence the parameter $\bar{h} = h\lambda$, where λ is any of the m eigenvalues, can be complex. This leads to the following modification of our earlier definition of absolute stability (cf. Section 2.6 and Definition 10).

Definition 11 *A linear k-step method is said to be **absolutely stable** in an open set $\mathcal{R}_A \subseteq \mathbb{C}$ if, for all $\bar{h} \in \mathcal{R}_A$, all roots r_s, $s \in \{1, \ldots, k\}$, of the stability polynomial $z \mapsto \pi(z; \bar{h})$ associated with the method, and defined by (64), satisfy $|r_s| < 1$. The largest such set $\mathcal{R}_A$ is called the **region of absolute stability** of the method.*

Clearly, the interval of absolute stability of a LMM is a subset of its region of absolute stability.

Exercise 6 *a) Find the region of absolute stability of the explicit Euler method.*

b) For $\lambda \in \mathbb{C}^-$ with $|\lambda| \gg 1$, consider the second-order differential equation

$$y''(x) + (1-\lambda)y'(x) - \lambda y(x) = 0, \qquad y(0) = 1, \quad y'(0) = -\lambda - 2.$$

1) Setting $\mathbf{y}(x) := (y(x), y'(x))^{\mathrm{T}}$, rewrite this problem as a first-order system

$$\mathbf{y}'(x) = A\mathbf{y}(x), \qquad \mathbf{y}(0) = \mathbf{y}_0$$

for some $A \in \mathbb{C}^{2\times 2}$ and $\mathbf{y}_0 \in \mathbb{C}^2$.

2) Solve this problem and show that $\mathbf{y}(x) \to (0,0)^{\mathrm{T}}$ as $x \to \infty$. Find the solution y to the original problem.

3) Now, the explicit Euler method is applied to this first-order system, i.e., $\mathbf{y}_{n+1} = \mathbf{y}_n + hA\mathbf{y}_n$ for $n \in \mathbb{N}_0$. What choice of the step size $h \in (0,1)$ will guarantee absolute stability in the sense that $\mathbf{y}_n \to 0$ as $n \to \infty$?

SOLUTION: a) For the explicit Euler method we have $\rho(z) = z - 1$ and $\sigma(z) = 1$, so that

$$\pi(z; \bar{h}) = \rho(z) - \bar{h}\sigma(z) = z - (1 + \bar{h}).$$

This has the root $r := 1 + \bar{h}$. Hence the region of absolute stability is

$$\mathcal{R}_A = \{\bar{h} \in \mathbb{C} : \, |1 + \bar{h}| < 1\} = D_1(-1),$$

which is an the open disc with radius 1 centred at -1.

b) 1) Writing $\mathbf{y} = (y, y')^{\mathrm{T}}$, the IVP for the given second-order differential equation can be recast as

$$\mathbf{y}'(x) = A\mathbf{y}(x), \qquad \mathbf{y}(0) = \mathbf{y}_0,$$

where

$$A = \begin{pmatrix} 0 & 1 \\ \lambda & \lambda - 1 \end{pmatrix} \quad \text{and} \quad \mathbf{y}_0 = \begin{pmatrix} 1 \\ -\lambda - 2 \end{pmatrix}.$$

2) The eigenvalues of A are $\lambda_1 := -1$, $\lambda_2 := \lambda$, and the vectors $\mathbf{v}_1 := (1, -1)^{\mathrm{T}}$, $\mathbf{v}_2 := (1, \lambda)^{\mathrm{T}}$ are corresponding eigenvectors. The general solution to the system is

$$\mathbf{y}(x) = c_1 e^{\lambda_1 x}\mathbf{v}_1 + c_2 e^{\lambda_2 x}\mathbf{v}_2 = \begin{pmatrix} c_1 e^{-x} + c_2 e^{\lambda x} \\ -c_1 e^{-x} + \lambda c_2 e^{\lambda x} \end{pmatrix}.$$

From the initial condition, we see that $\mathbf{y}(0) = (c_1 + c_2, -c_1 + \lambda c_2)^{\mathrm{T}} = (1, -\lambda - 2)^{\mathrm{T}}$, which yields $c_1 = 2$, $c_2 = -1$. Hence, the solution is given by

$$\mathbf{y}(x) = \begin{pmatrix} 2e^{-x} - e^{\lambda x} \\ -2e^{-x} - \lambda e^{\lambda x} \end{pmatrix}$$

and we see from the assumptions on λ that $\mathbf{y}(x) \to (0,0)^{\mathrm{T}}$ as $x \to \infty$. Note that we have found the solution (and its derivative) to the original problem:

$$y(x) = 2e^{-x} - e^{\lambda x}, \qquad y'(x) = -2e^{-x} - \lambda e^{\lambda x}.$$

3) Explicit Euler for this system has the form

$$\mathbf{y}_{n+1} = (I_2 + hA)\mathbf{y}_n, \qquad n \in \mathbb{N}_0$$

with $\mathbf{y}_0 = (1, -\lambda - 2)^{\mathrm{T}}$. Consider the matrix $M := I_2 + hA$ whose eigenvalues are $m_1 := 1 - h$, $m_2 := 1 + \lambda h$, and corresponding eigenvectors are $\mathbf{v}_1 := (1, -1)^{\mathrm{T}}$, $\mathbf{v}_2 := (1, \lambda)^{\mathrm{T}}$. We find that M is diagonalisable:

$$M = SDS^{-1}, \qquad S := \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{1+\lambda^2}} \\ -\frac{1}{\sqrt{2}} & \frac{\lambda}{\sqrt{1+\lambda^2}} \end{pmatrix}, \qquad D := \begin{pmatrix} 1-h & 0 \\ 0 & 1+\lambda h \end{pmatrix}.$$

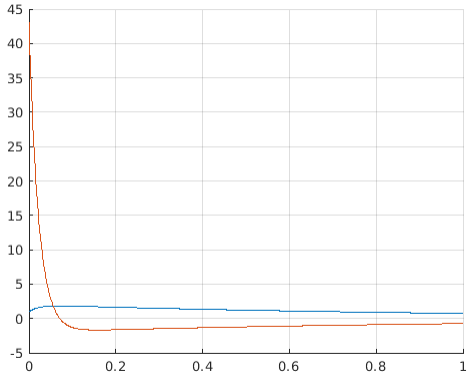

Figure 1: The functions $y(x) = 2e^{-x} - e^{\lambda x}$ (blue) and $y'(x) = -2e^{-x} - \lambda e^{\lambda x}$ (red) for $\lambda = -45$.

Hence, $S^{-1}\mathbf{y}_{n+1} = S^{-1}M\mathbf{y}_n = DS^{-1}\mathbf{y}_n$ for $n \in \mathbb{N}_0$, and therefore $S^{-1}\mathbf{y}_n = D^n S^{-1}\mathbf{y}_0$. Now, $\lim_{n\to\infty} \mathbf{y}_n = \mathbf{0}$ iff $\lim_{n\to\infty} S^{-1}\mathbf{y}_n = \mathbf{0}$, iff $(D^n)_{n\in\mathbb{N}}$ converges to the zero matrix, iff $|1-h| < 1$ and $|1+h\lambda| < 1$. Since we only consider $h \in (0,1)$, the first of these two requirements is satisfied and we deduce that the requirement for absolute stability is that $|1+\lambda h| < 1$, i.e., $\lambda h \in D_1(-1)$.

The graphs of the functions y and y' are depicted in Figure 1 for $\lambda = -45$. Note that, if $x \in [0,\infty)$ is thought of as time, y' exhibits a fast transition near $x = 0$ while y is slowly varying for $x > 0$ and y' is slowly varying for $x > \frac{1}{45}$. Despite the fact that over the interval $(\frac{1}{45}, \infty)$ both y and y' are 'slowly varying', we are forced to use a small step size of $h < \frac{2}{45}$ to ensure absolute stability. $\diamond$

In the example the y' component of the solution $\mathbf{y} = (y, y')$ exhibited two vastly different time scales; in addition, the fast transition (which occurs between $x = 0$ and $x \approx \frac{1}{-\lambda}$ for $\lambda \in \mathbb{R}_{<0}$) has negligible effect on the solution so its accurate resolution does not appear to be important for obtaining an overall accurate solution. Nevertheless, in order to ensure the stability of the numerical method under consideration, the mesh size h is forced to be exceedingly small, $h < -2\frac{\mathrm{Re}(\lambda)}{|\lambda|^2}$, smaller than an accurate approximation of the solution for $x \gg 1/|\lambda|$ would necessitate. Systems of ODEs which exhibit this behaviour are generally referred to as **stiff systems**. We refrain from formulating a rigorous definition of stiffness. Indeed, stiffness of an ODE is a concept that lacks a rigorous definition.[7] A historic and pragmatic 'definition' by Curtis and Hirschfelder[8] (adapted to our setting) reads: stiff equations are equations where the implicit Euler method works significantly better than the explicit Euler method. The idea behind this definition is that for a 'stiff system' stability of the explicit Euler method requires the choice of a very small step size, much smaller than the one required by accuracy.

4.1 Stability of numerical methods for stiff systems

In order to motivate the various definitions of stability which occur in this section, we begin with a simple example. Consider the implicit Euler method for the IVP

$$y'(x) = \lambda y(x), \qquad y(0) = y_0,$$

where $\lambda \in \mathbb{C}$. The stability polynomial of the method is $\pi(z;\bar{h}) = \rho(z) - \bar{h}\sigma(z)$ where $\bar{h} = \lambda h$, $\rho(z) = z-1$ and $\sigma(z) = z$. Since the only root of the stability polynomial is $z = \frac{1}{1-\bar{h}}$, we deduce that the method has the region of absolute stability

$$\mathcal{R}_A = \{\bar{h} \in \mathbb{C} : |1-\bar{h}| > 1\} = \mathbb{C}\backslash\bar{D}_1(1).$$

In particular $\mathcal{R}_A$ includes $\mathbb{C}^-$. The next definition is due to Dahlquist (1963).

[7] See G. Söderlind, L. Jay, and M. Calvo, *Stiffness 1952–2012: Sixty years in search of a definition.* BIT Numerical Mathematics, June 2015 55(2), 531–558.

[8] *Integration of stiff equations.* Proceedings of the National Academy of Sciences, March 1, 1952 38 (3) 235–243.

Definition 12 *A LMM is called A-stable if its region of absolute stability $\mathcal{R}_A$ is such that $\mathbb{C}^- \subseteq \mathcal{R}_A$.*

For example, according to the discussion preceding Definition 12, the implicit Euler method is A-stable. As the next theorem by Dahlquist (1963) shows, Definition 12 is unfortunately far too restrictive.

Theorem 14

(i) No explicit LMM is A-stable.

(ii) The order of accuracy of an A-stable implicit LMM cannot exceed 2.

(iii) The second-order accurate A-stable LMM with smallest error constant is the trapezium rule method.

This motivates us to consider the following, less restrictive notion of stability, due to Widlund (1967).

Definition 13 *A LMM is called $A(\alpha)$-**stable**, $\alpha \in (0, \frac{\pi}{2})$, if its region of absolute stability $\mathcal{R}_A$ is such that $W_\alpha \subseteq \mathcal{R}_A$, where W_α denotes the infinite open wedge*

$$W_\alpha = \{\bar{h} \in \mathbb{C} : \arg(\bar{h}) \in (\pi - \alpha, \pi + \alpha)\}.$$

*A LMM is called $A(0)$-**stable** if it is $A(\alpha)$-stable for some $\alpha \in (0, \frac{\pi}{2})$. A LMM is called A_0-stable if $\mathcal{R}_A$ includes the negative real axis in the complex plane.*

Let us note in connection with this definition that if $\lambda \in \mathbb{C}^-$ for a given λ then $\bar{h} = \lambda h$ either lies inside the wedge W_α or outside W_α for *all* positive h. Consequently, if all eigenvalues λ of the matrix A (cf. the sentence starting two lines above equation (71)) happen to lie in some wedge W_α then an $A(\alpha)$-stable method can be used for the numerical solution of the IVP (68) without any restrictions on the step size h. In particular, if all eigenvalues of A are real and negative, then an $A(0)$-stable method can be used. The next theorem can be regarded the analogue of Theorem 14 for the case of $A(\alpha)$ and $A(0)$-stability.

Theorem 15

(i) No explicit LMM is $A(0)$-stable.

(ii) The only $A(0)$-stable linear k-step method whose order exceeds k is the trapezium rule method.

(iii) For each $\alpha \in [0, \frac{\pi}{2})$ there exist $A(\alpha)$-stable linear k-step methods of order p for which $k = p = 3$ and $k = p = 4$.

A different way of loosening the concept of A-stability was proposed by Gear (1969). The motivation behind it is the fact that for a typical stiff problem the eigenvalues of the matrix A which produce the fast transients all lie to the left of a line $\{\bar{h} \in \mathbb{C} : \mathrm{Re}(\bar{h}) = -a\}$, $a > 0$, in the complex plane, while those that are responsible for the slow transients are clustered around zero.

Definition 14 *A LMM is said to be* **stiffly stable** *if there exist $a, c > 0$ such that its region of absolute stability $\mathcal{R}_A$ is such that $\mathcal{R}_A \supseteq \mathcal{R}_1 \cup \mathcal{R}_2$ where*

$$\mathcal{R}_1 = \{\bar{h} \in \mathbb{C} : \mathrm{Re}(\bar{h}) \in (-\infty, -a)\},$$
$$\mathcal{R}_2 = \{\bar{h} \in \mathbb{C} : \mathrm{Re}(\bar{h}) \in [-a, 0),\ \ \mathrm{Im}(\bar{h}) \in [-c, c]\}.$$

It is clear that stiff stability implies $A(\alpha)$-stability with $\alpha = \arctan(\frac{c}{a})$. More generally, we have the following chain of implications:

$$A\text{-stability} \Rightarrow \text{stiff-stability} \Rightarrow A(\alpha)\text{-stability} \Rightarrow A(0)\text{-stability} \Rightarrow A_0\text{-stability}.$$

In the next section we shall consider LMMs which are particularly well suited for the numerical solution of stiff systems of ODEs.

4.2 Backward differentiation methods for stiff systems

Consider a LMM with stability polynomial $\pi(z;\bar{h}) = \rho(z) - \bar{h}\sigma(z)$. If the method is $A(\alpha)$-stable or stiffly stable, the roots $r(\bar{h})$ of $\pi(\cdot;\bar{h})$ lie in $D_1(0)$ when $\bar{h}$ is real and $\bar{h} \to -\infty$. Hence,

$$0 = \lim_{\bar{h}\to-\infty} \frac{\rho(r(\bar{h}))}{\bar{h}} = \lim_{\bar{h}\to-\infty} \sigma(r(\bar{h})) = \sigma\left(\lim_{\bar{h}\to-\infty} r(\bar{h})\right);$$

in other words, the roots of $\pi(\cdot;\bar{h})$ approach those of σ. Thus it is natural to choose σ in such a way that its roots lie within the unit disk. Indeed, a particularly simple choice would be to take $\sigma(z) = \beta_k z^k$; the resulting class of **backward differentiation formulae (BDF)** (see Section 3.1 for construction) has the general form:

$$\sum_{j=0}^{k} \alpha_j \mathbf{y}_{n+j} = h\beta_k \mathbf{f}_{n+k},$$

where the coefficients α_j and β_k are given in Table 3 which also displays the value of a in the definition of stiff-stability, the angle α from the definition of $A(\alpha)$-stability, the order p of the method and the corresponding error constant C_{p+1} for $k \in \{1,\dots,6\}$. Backward differentiation methods with $k \geq 7$ of the kind considered here are *not* zero-stable and are therefore irrelevant from the practical point of view.

k	α_6	α_5	α_4	α_3	α_2	α_1	α_0	β_k	p	C_{p+1}	a_{min}	α_{max}
1						1	-1	1	1	$-\frac{1}{2}$	0	90^o
2					1	$-\frac{4}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	2	$-\frac{2}{9}$	0	90^o
3				1	$-\frac{18}{11}$	$\frac{9}{11}$	$-\frac{2}{11}$	$\frac{6}{11}$	3	$-\frac{3}{22}$	0.1	88^o
4			1	$-\frac{48}{25}$	$\frac{36}{25}$	$-\frac{16}{25}$	$\frac{3}{25}$	$\frac{12}{25}$	4	$-\frac{12}{125}$	0.7	73^o
5		1	$-\frac{300}{137}$	$\frac{300}{137}$	$-\frac{200}{137}$	$\frac{75}{137}$	$-\frac{12}{137}$	$\frac{60}{137}$	5	$-\frac{10}{137}$	2.4	52^o
6	1	$-\frac{360}{147}$	$\frac{450}{147}$	$-\frac{400}{147}$	$\frac{225}{147}$	$-\frac{72}{147}$	$\frac{10}{147}$	$\frac{60}{147}$	6	$-\frac{20}{343}$	6.1	19^o

Table 3: Coefficients, order, error constant and stability parameters for backward differentiation methods

4.3 Adaptivity for stiff problems

Ideally, we would like to compute an approximate solution of the following IVP for a system of first-order ODEs:

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \qquad \mathbf{y}(x_0) = \mathbf{y}_0, \tag{73}$$

for all $x \in [x_0, X_M]$, and make sure that this approximation is accurate up to a certain (absolute/relative) precision. In addition, we would like to achieve such a precision in the fastest/cheapest way possible. How should this be done? We shall describe two attempts, the first attempt being conceptually simpler, while the second attempt being the one preferred in practice for reasons which we shall explain.

4.3.1 First attempt

A simple strategy could be to:

1. choose a one-step method of order p;
2. choose $N \in \mathbb{N}$ and compute the approximate solution $\{\mathbf{y}_n\}_{n=0}^{N}$ using the step size $h = \frac{X_M - x_0}{N}$;
3. choose a large natural number $\tilde{N} \in \mathbb{N}$ with $\tilde{N} > N$ and compute the approximate solution $\{\tilde{\mathbf{y}}_n\}_{n=0}^{\tilde{N}}$ using the step size $\tilde{h} = \frac{X_M - x_0}{\tilde{N}}$.

This way, we obtain two approximations $\mathbf{y}_N$ and $\tilde{\mathbf{y}}_{\tilde{N}}$ of $\mathbf{y}(X_M)$, which we may use to estimate the error. To be more precise, we may use the (computable) difference $\|\tilde{\mathbf{y}}_{\tilde{N}} - \mathbf{y}_N\|$ to estimate the (noncomputable) error $\|\mathbf{y}(X_M) - \mathbf{y}_N\|$. If $\|\tilde{\mathbf{y}}_{\tilde{N}} - \mathbf{y}_N\|$ is smaller than a target absolute tolerance `TOL`, then we finish the computation. Otherwise, we

1. increase N so that $N > \tilde{N}$;
2. compute the approximate solution $\{\mathbf{y}_n\}_{n=0}^{N}$ using $h = \frac{X_M - x_0}{N}$;
3. check whether $\|\tilde{\mathbf{y}}_{\tilde{N}} - \mathbf{y}_N\| < \texttt{TOL}$.

If $\|\tilde{\mathbf{y}}_{\tilde{N}} - \mathbf{y}_N\| < \texttt{TOL}$, then we finish the computation. Otherwise, we select a new $\tilde{N}$ such that $\tilde{N} > N$, and compute $\{\tilde{\mathbf{y}}_n\}_{n=0}^{\tilde{N}}$ using the step size $\tilde{h} = \frac{X_M - x_0}{\tilde{N}}$. This procedure is repeated until convergence (alternating N and $\tilde{N}$). The following argument suggests that the (computable) difference $\|\tilde{\mathbf{y}}_{\tilde{N}} - \mathbf{y}_N\|$ can be used to estimate the error $\|\mathbf{y}(X_M) - \mathbf{y}_N\|$.

The idea to use $\|\tilde{\mathbf{y}}_{\tilde{N}} - \mathbf{y}_N\|$ to estimate $\|\mathbf{y}(X_M) - \mathbf{y}_N\|$ is based on the following calculations. Let us assume that $\tilde{N} > N$, and define $\alpha := \frac{\tilde{h}}{h} = \frac{N}{\tilde{N}} < 1$. For h sufficiently small, we have

$$\|\tilde{\mathbf{y}}_{\tilde{N}} - \mathbf{y}_N\| \leq \|\tilde{\mathbf{y}}_{\tilde{N}} - \mathbf{y}(X_M)\| + \|\mathbf{y}(X_M) - \mathbf{y}_N\| \leq C(\tilde{h}^p + h^p) = (1 + \alpha^p)Ch^p$$

for some constant $C > 0$, and thus,

$$\|\mathbf{y}(X_M) - \mathbf{y}_N\| \leq \|\mathbf{y}(X_M) - \tilde{\mathbf{y}}_{\tilde{N}}\| + \|\tilde{\mathbf{y}}_{\tilde{N}} - \mathbf{y}_N\| \leq C\tilde{h}^p + (1 + \alpha^p)Ch^p = \alpha^p\big(Ch^p\big) + (1 + \alpha^p)\big(Ch^p\big).$$

For $\alpha < 1$, $\alpha^p \ll 1 + \alpha^p$ (in relative terms). Therefore, the term $\|\mathbf{y}(X_M) - \tilde{\mathbf{y}}_{\tilde{N}}\|$ has a minor contribution, and $\|\tilde{\mathbf{y}}_{\tilde{N}} - \mathbf{y}_N\|$ may be used to estimate $\|\mathbf{y}(X_M) - \mathbf{y}_N\|$.

This first adaptive strategy could deliver an accurate solution, but it is likely to be computationally inefficient, because whenever the target tolerance is not met we need to compute another solution from scratch on a finer computational mesh over the entire interval $[x_0, X_M]$ (i.e. a global mesh-refinement needs to be performed – a new numerical approximation has to be computed on a globally refined mesh).

4.3.2 Second attempt

To improve efficiency, we can try to control the consistency error for each mesh point x_n. Indeed, Theorem 4 states that the global error is bounded by the maximum of the consistency error up to a constant factor (however, note the exponential term in the constant factor!). Therefore, the hope is that we may compute a sufficiently accurate solution by choosing a suitable h or, better still, by adapting the step size locally, that is, by selecting a suitable h_n for every x_n to control the local size of the consistency error.

To estimate the consistency error at $x = x_n$, in addition to the one step method

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\boldsymbol{\Phi}(x_n, \mathbf{y}_n; h) =: \boldsymbol{\Psi}(x_n, \mathbf{y}_n; h), \qquad n \in \{0, 1, \ldots, N-1\}$$

of order p being used, we consider an additional one-step method

$$\tilde{\mathbf{y}}_{n+1} = \tilde{\mathbf{y}}_n + h\tilde{\mathbf{\Phi}}(x_n, \tilde{\mathbf{y}}_n; h) =: \tilde{\mathbf{\Psi}}(x_n, \tilde{\mathbf{y}}_n; h), \qquad n \in \{0, 1, \ldots, N-1\}$$

of order $\tilde{p}$, with $\tilde{p} > p$, and we compute

$$\mathrm{ERR}(x_n; h) := \|\tilde{\mathbf{\Psi}}(x_n, \mathbf{y}_n; h) - \mathbf{\Psi}(x_n, \mathbf{y}_n; h)\|. \tag{74}$$

The idea behind using (74) to estimate the consistency error T_n is that, if the error has been controlled from x_0 up until x_n, for some $n \geq 1$, then the difference between $\mathbf{y}(x_n)$ and $\mathbf{y}_n$ is "negligible", and therefore $\mathbf{y}_n$ can be assumed to be equal to $\tilde{\mathbf{y}}_n$ (both being "equal" to $\mathbf{y}(x_n)$). Hence,

$$\begin{aligned} hT_n &= \mathbf{y}(x_{n+1}) - \mathbf{\Psi}(x_n, \mathbf{y}(x_n); h) \\ &= \mathbf{y}(x_{n+1}) - \tilde{\mathbf{\Psi}}(x_n, \mathbf{y}(x_n); h) + \tilde{\mathbf{\Psi}}(x_n, \mathbf{y}(x_n); h) - \mathbf{\Psi}(x_n, \mathbf{y}(x_n); h) \\ &\approx \mathbf{y}(x_{n+1}) - \tilde{\mathbf{\Psi}}(x_n, \mathbf{y}(x_n); h) + \tilde{\mathbf{\Psi}}(x_n, \mathbf{y}_n; h) - \mathbf{\Psi}(x_n, \mathbf{y}_n; h) \\ &\approx Ch^{\tilde{p}+1} + \tilde{\mathbf{\Psi}}(x_n, \mathbf{y}_n; h) - \mathbf{\Psi}(x_n, \mathbf{y}_n; h). \end{aligned} \tag{75}$$

Since the left-hand side of (75) is of the order $\mathcal{O}(h \cdot h^p) = \mathcal{O}(h^{p+1})$ and $\tilde{p} > p$, it follows that the term $\approx Ch^{\tilde{p}+1}$ on the right-hand side is "negligible" compared to the "leading term" $\tilde{\mathbf{\Psi}}(x_n, \mathbf{y}_n; h) - \mathbf{\Psi}(x_n, \mathbf{y}_n; h)$. Hence, $hT_n \approx \tilde{\mathbf{\Psi}}(x_n, \mathbf{y}_n; h) - \mathbf{\Psi}(x_n, \mathbf{y}_n; h)$.

Summing up, the locally adaptive strategy proceeds as follows: at every step x_n

1. select an initial local step size h_n;
2. compute $\mathrm{ERR}(x_n; h_n)$;
3. if $\mathrm{ERR}(x_n; h_n) < \texttt{TOL}$, set $\mathbf{y}_{n+1} = \mathbf{\Psi}(x_n, \mathbf{y}_n; h_n)$; otherwise, choose a smaller h_n and go to step 2.

To make this algorithm more efficient, it is common to increase the step h_n every time this step has been accepted, that is, to select βh_n for a suitable $\beta > 1$.

Remark 7 *Let* $\texttt{TOL} > 0$ *be a target absolute error tolerance and let* $\mathrm{ERR}(x_n; h_n) < \texttt{TOL}$. *Then, the "optimal"* β *is*

$$\beta = \beta_n = \left(\frac{\texttt{TOL}}{\mathrm{ERR}(x_n; h_n)}\right)^{\frac{1}{p+1}}. \tag{76}$$

Indeed, if $\mathrm{ERR}(x_n; h_n) < \texttt{TOL}$, *we could have chosen a larger* h_n *and still satisfied the tolerance criterion. Let* β_n *be such that* $\mathrm{ERR}(x_n, \beta_n h_n) = \texttt{TOL}$, *so that* $\beta_n h_n$ *is the ideal step size. Then, we deduce* (76), *because*

$$\mathrm{ERR}(x_n; \beta_n h_n) \approx C(\beta_n h_n)^{p+1} = \beta_n^{p+1} C h_n^{p+1} \approx \beta_n^{p+1} \mathrm{ERR}(x_n; h_n).$$

To further improve the efficiency of this adaptive algorithm, it is convenient to use embedded RK methods, which limit the number of function evaluations.

Definition 15 *Two RK methods are* embedded *if they use the same stages. The Butcher tableau of two embedded RK methods can be written as*

$$\begin{array}{c|c} \mathbf{a} & \mathbf{B} \\ \hline & \mathbf{c}_2^{\mathrm{T}} \\ & \mathbf{c}_1^{\mathrm{T}} \end{array}, \quad \textit{where} \quad \begin{array}{c|c} \mathbf{a} & \mathbf{B} \\ \hline & \mathbf{c}_2^{\mathrm{T}} \end{array} \quad \textit{and} \quad \begin{array}{c|c} \mathbf{a} & \mathbf{B} \\ \hline & \mathbf{c}_1^{\mathrm{T}} \end{array}$$

are the Butcher tableaus of the two RK methods, respectively.

Example 6 *The Heun–Euler method has the Butcher tableau*

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \\ & 1 & 0 \end{array}, \quad \text{where} \quad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array} \quad \text{and} \quad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1 & 0 \end{array}$$

are the Butcher tableaus of Heun's method $y_{n+1} = y_n + \frac{h}{2}(f(x_n, y_n) + f(x_n + h, y_n + hf(x_n, y_n)))$ *and the explicit Euler method* $y_{n+1} = y_n + hf(x_n, y_n)$*, respectively.*

Example 7 MATLAB *integrators for ODEs (such as the functions* `ode45`*,* `ode23`*, etc.) are based on embedded RK methods.*[9]

[9]See L. F. Shampine and M. W. Reichelt, *The MATLAB ODE suite* (1997).

Part II
Partial Differential Equations (PDEs)

Partial differential equations (PDEs) 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.

This second part of the course is devoted to the construction and the mathematical analysis of the conceptually simplest class of numerical techniques, finite difference (FD) methods, for the approximate solution of elliptic and parabolic PDEs, 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.

5 Preliminaries: Function spaces

The accuracy of a numerical method for the approximate solution of PDEs depends on its capability to represent the important qualitative features of the true 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 this second part of the course, focusing on spaces of continuous functions, spaces of integrable functions, and Sobolev spaces.

5.1 Spaces of continuous functions

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.

An n-tuple $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}_0^n$ is called a **multi-index**. The number $|\alpha| := \alpha_1 + \cdots + \alpha_n \in \mathbb{N}_0$ is called the length of the multi-index $\alpha = (\alpha_1, \dots, \alpha_n)$. We denote $(0, \dots, 0) \in \mathbb{N}_0^n$ by $\mathbf{0}$; clearly $|\mathbf{0}| = 0$. We define

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

Example. Let $u : \mathbb{R}^3 \to \mathbb{R}$, $u(x) := u(x_1, x_2, x_3) := x_1^3 x_2^3 x_3^3$. Then, we have the following:

- For $\alpha := (1, 2, 3)$, we have $D^\alpha u(x) = \partial^6_{x_1 x_2 x_2 x_3 x_3 x_3} u(x) = 108 x_1^2 x_2$.
- For $\alpha := (0, 1, 0)$, we have $D^\alpha u(x) = \partial_{x_2} u(x) = 3 x_1^3 x_2^2 x_3^3$.
- For $\alpha := (2, 0, 0)$, we have $D^\alpha u(x) = \partial^2_{x_1 x_1} u(x) = 6 x_1 x_2^3 x_3^3$.
- We have $\sum_{\substack{\alpha \in \mathbb{N}_0^3, \\ |\alpha| = 3}} D^\alpha u = \partial^3_{x_1 x_1 x_1} u + \partial^3_{x_1 x_1 x_2} u + \partial^3_{x_1 x_1 x_3} u + \partial^3_{x_1 x_2 x_2} u + \partial^3_{x_1 x_3 x_3} u + \partial^3_{x_2 x_2 x_2} u + \partial^3_{x_1 x_2 x_3} u + \partial^3_{x_2 x_2 x_3} u + \partial^3_{x_2 x_3 x_3} u + \partial^3_{x_3 x_3 x_3} u$.

◇

Let Ω be an open set in $\mathbb{R}^n$, and let $k \in \mathbb{N}_0$. We denote by $C^k(\Omega)$ the set of all continuous real-valued functions $u : \Omega \to \mathbb{R}$ such that $D^\alpha u$ is continuous on Ω for all $\alpha = (\alpha_1, \dots, \alpha_n)$ with $|\alpha| \le k$.

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

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

where $x := (x_1, \ldots, x_n)$. 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 $k = 1$,

$$\|u\|_{C^1(\overline{\Omega})} = \sum_{|\alpha| \leq 1} \sup_{x \in \Omega} |D^\alpha u(x)| = \sup_{x \in \Omega} |u(x)| + \sum_{j=1}^{n} \sup_{x \in \Omega} \left|\partial_{x_j} u(x)\right|.$$

We write $C^\infty(\Omega) := \bigcap_{k=0}^{\infty} C^k(\Omega)$ and $C^\infty(\overline{\Omega}) := \bigcap_{k=0}^{\infty} C^k(\overline{\Omega})$.

Example. Let $n = 1$, and consider the open interval $\Omega := (0, 1) \subset \mathbb{R}$. The function $u : \Omega \to \mathbb{R}$, $u(x) := \frac{1}{x}$ belongs to $C^k(\Omega)$ for all $k \geq 0$. 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 $k \geq 0$. $\diamond$

The **support** of a function $u \in C(\Omega)$, denoted $\mathrm{supp}(u)$, is defined as the closure in Ω of the set $\{x \in \Omega : u(x) \neq 0\}$, i.e.,

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

In other words, $\mathrm{supp}(u)$ is the smallest closed subset of Ω such that $u = 0$ in $\Omega \backslash \mathrm{supp}(u)$.

Example. Let $w : \mathbb{R}^n \to \mathbb{R}$ be the function given by

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

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

We denote by $C_c^k(\Omega)$ the set of all $u \in C^k(\Omega)$ such that $\mathrm{supp}(u) \subset \Omega$ and $\mathrm{supp}(u)$ is bounded (or equivalently, the set of all functions $u \in C^k(\Omega)$ such that $\mathrm{supp}(u) \subset \Omega$ and $\mathrm{supp}(u)$ is compact). Let

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

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

5.2 Spaces of integrable functions

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

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

Here, $x := (x_1, \ldots, x_n)$ and $\mathrm{d}x := \mathrm{d}x_1 \cdots \mathrm{d}x_n$. Functions which are equal almost everywhere on Ω (i.e., equal, except on a set of measure zero) are identified with each other. $L^p(\Omega)$ is equipped with the norm

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

A particularly important case is $p = 2$; then,

$$\|u\|_{L^2(\Omega)} = \sqrt{\int_\Omega |u(x)|^2 \, \mathrm{d}x}.$$

The space $L^2(\Omega)$ can be equipped with an inner product

$$(u, v)_{L^2(\Omega)} := \int_\Omega u(x)v(x) \, \mathrm{d}x.$$

Clearly, $\|u\|_{L^2(\Omega)} = \sqrt{(u, u)_{L^2(\Omega)}}$.

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 (Cauchy–Schwarz inequality) *Let $u, v \in L^2(\Omega)$. Then,*

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

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

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

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., $\left|2(u, v)_{L^2(\Omega)}\right|^2 - 4\|u\|^2_{L^2(\Omega)}\|v\|^2_{L^2(\Omega)} \le 0$, and hence the desired inequality holds. □

Corollary 1 (Triangle inequality) *Let $u, v \in L^2(\Omega)$. 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

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

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 $(\cdot, \cdot)_{L^2(\Omega)}$ (and the associated norm $\|u\|_{L^2(\Omega)} = \sqrt{(u, u)_{L^2(\Omega)}}$) 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 $\|u\|_X := \sqrt{(u, u)_X}$) is called a Hilbert space if, whenever $(u_m)_{m\in\mathbb{N}}$ 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 $\lim_{m\to\infty} \|u_m - u\|_X = 0$ (i.e., (u_m) converges to u in the norm of X).

5.3 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 $u \in C^k(\Omega)$; then we have the following integration-by-parts formula:

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

We note here that all integrals on $\partial\Omega$ that arise in the course of partial integration, based on the divergence theorem,[10] have vanished because $v \in C_c^\infty(\Omega)$. 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

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

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 $\Omega := \mathbb{R}$, and suppose that we wish to determine the first weak derivative of the function $u : \Omega \to \mathbb{R}$, $u(x) := (1 - |x|)_+$. Here, for a number $y \in \mathbb{R}$, we write $y_+ := \max\{y, 0\}$ to denote the nonnegative part of y. 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 $v \in C_c^\infty(\Omega)$, we have that

$$\begin{aligned}
\int_{-\infty}^{\infty} u(x)\, v'(x)\, \mathrm{d}x &= \int_{-\infty}^{\infty} (1 - |x|)_+\, v'(x)\, \mathrm{d}x \\
&= \int_{-1}^{0} (1 + x)\, v'(x)\, \mathrm{d}x + \int_0^1 (1 - x)\, v'(x)\, \mathrm{d}x \\
&= -\int_{-1}^{0} v(x)\, \mathrm{d}x + [(1 + x)v(x)]_{x=-1}^{x=0} + \int_0^1 v(x)\, \mathrm{d}x + [(1 - x)v(x)]_{x=0}^{x=1} \\
&= -\left(\int_{-1}^{0} v(x)\, \mathrm{d}x - \int_0^1 v(x)\, \mathrm{d}x \right) = -\int_{-\infty}^{\infty} w(x)\, v(x)\, \mathrm{d}x,
\end{aligned}$$

where

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

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

Now we are ready to give a precise definition of a Sobolev space. Let $k \in \mathbb{N}_0$. We define (with D^α denoting a weak derivative of order $|\alpha|$)

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

[10] Observe that

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

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 $\mathrm{div}(0, \ldots, 0, uv, 0, \ldots, 0) = \partial_{x_i}(uv)$ and $(0, \ldots, 0, uv, 0, \ldots, 0) \cdot \nu = uv\, \nu_i$.

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

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

and the inner product

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

With this inner product, $H^k(\Omega)$ is a Hilbert space. Letting

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

we can write

$$\|u\|_{H^k(\Omega)} = \sqrt{\sum_{j=0}^{k} |u|^2_{H^j(\Omega)}}.$$

The map $u \mapsto |u|_{H^k(\Omega)}$ 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 $u = 0$).

We will frequently use the spaces $H^1(\Omega)$ and $H^2(\Omega)$:

$$H^1(\Omega) := \left\{ u \in L^2(\Omega) : \partial_{x_j} u \in L^2(\Omega) \ \ \forall j \in \{1,\dots,n\} \right\},$$

$$\|u\|_{H^1(\Omega)} := \sqrt{\|u\|^2_{L^2(\Omega)} + \sum_{j=1}^{n} \|\partial_{x_j} u\|^2_{L^2(\Omega)}}, \qquad |u|_{H^1(\Omega)} := \sqrt{\sum_{j=1}^{n} \|\partial_{x_j} u\|^2_{L^2(\Omega)}}$$

Similarly,

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

$$\|u\|_{H^2(\Omega)} := \sqrt{\|u\|^2_{L^2(\Omega)} + \sum_{j=1}^{n} \|\partial_{x_j} u\|^2_{L^2(\Omega)} + \sum_{i,j=1}^{n} \|\partial^2_{x_i x_j} u\|^2_{L^2(\Omega)}}, \qquad |u|_{H^2(\Omega)} := \sqrt{\sum_{i,j=1}^{n} \|\partial^2_{x_i x_j} u\|^2_{L^2(\Omega)}}.$$

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 $\partial\Omega$, (strictly speaking, trace of u being zero on) the boundary of the set Ω. We shall use this space when considering a PDE 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 $\partial\Omega$). Then, there exists a constant $c_\star > 0$, depending only on Ω, such that for any $u \in H^1_0(\Omega)$ there holds*

$$\|u\|^2_{L^2(\Omega)} \le c_\star \sum_{i=1}^{n} \|\partial_{x_i} u\|^2_{L^2(\Omega)}. \tag{77}$$

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. For $u \in H_0^1(\Omega)$, we have

$$u(x_1,x_2) = u(a,x_2) + \int_a^{x_1} \partial_{x_1} u(\xi,x_2)\,\mathrm{d}\xi = \int_a^{x_1} \partial_{x_1} u(\xi,x_2)\,\mathrm{d}\xi.$$

Thus, by the Cauchy–Schwarz inequality,

$$\begin{aligned}
\int_\Omega |u(x_1,x_2)|^2 \,\mathrm{d}x_1\,\mathrm{d}x_2 &= \int_a^b \int_c^d \left| \int_a^{x_1} \partial_{x_1} u(\xi,x_2)\,\mathrm{d}\xi \right|^2 \mathrm{d}x_2\,\mathrm{d}x_1 \\
&\le \int_a^b \int_c^d (x_1-a) \left(\int_a^{x_1} |\partial_{x_1} u(\xi,x_2)|^2 \,\mathrm{d}\xi \right) \mathrm{d}x_2\,\mathrm{d}x_1 \\
&\le \left(\int_a^b (x_1-a)\,\mathrm{d}x_1 \right) \left(\int_c^d \int_a^b |\partial_{x_1} u(\xi,x_2)|^2 \,\mathrm{d}\xi\,\mathrm{d}x_2 \right) \\
&= \frac{1}{2}(b-a)^2 \int_\Omega |\partial_{x_1} u(x_1,x_2)|^2 \,\mathrm{d}x_1\,\mathrm{d}x_2.
\end{aligned}$$

Analogously, one shows that

$$\int_\Omega |u(x_1,x_2)|^2 \,\mathrm{d}x_1\,\mathrm{d}x_2 \le \frac{1}{2}(d-c)^2 \int_\Omega |\partial_{x_2} u(x_1,x_2)|^2 \,\mathrm{d}x_1\,\mathrm{d}x_2.$$

By combining the two inequalities, we obtain

$$\int_\Omega |u(x_1,x_2)|^2 \,\mathrm{d}x_1\,\mathrm{d}x_2 \le c_\star \int_\Omega \left(|\partial_{x_1} u(x_1,x_2)|^2 + |\partial_{x_2} u(x_1,x_2)|^2 \right) \mathrm{d}x_1\,\mathrm{d}x_2$$

with $c_\star > 0$ given by $\frac{1}{c_\star} = \frac{2}{(b-a)^2} + \frac{2}{(d-c)^2}$. □

6 Introduction to the theory of finite difference (FD) schemes

6.1 Elliptic boundary-value problems

We will start by focusing on **boundary-value problems (BVPs)** for elliptic PDEs. Elliptic equations are typified by the **Laplace equation**[11]

$$\Delta u = 0,$$

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 PDE

$$-\sum_{i,j=1}^n \partial_{x_j}(a_{ij}\partial_{x_i} u) + \sum_{i=1}^n b_i \partial_{x_i} u + cu = f \quad \text{in } \Omega, \tag{78}$$

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

$$a_{ij} \in C^1(\overline{\Omega}), \quad b_i \in C(\overline{\Omega}), \quad c, f \in C(\overline{\Omega})$$

[11] Recall that in n space dimensions the Laplace operator Δ is defined by $\Delta u := \mathrm{div}(\nabla u) = \partial^2_{x_1 x_1} u + \cdots + \partial^2_{x_n x_n} u$.

for $i, j \in \{1, \ldots, n\}$, and additionally

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \tilde{c}|\xi|^2 \quad \forall x \in \overline{\Omega},\ \xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n \tag{79}$$

for some constant $\tilde{c} > 0$ independent of x and ξ. The condition (79) is usually referred to as **uniform ellipticity** and (78) is called an **elliptic equation**. In the case of Poisson's equation, for example, $a_{ij} \equiv \delta_{ij}$ for $i, j \in \{1, \ldots, n\}$ (and also $b_i \equiv 0$ for $i \in \{1, \ldots, n\}$ and $c \equiv 0$), and the ellipticity condition is therefore trivially satisfied, with $\tilde{c} = 1$.

We note that equation (78) can equivalently be written as

$$-\mathrm{div}(A\nabla u) + \mathbf{b} \cdot \nabla u + cu = f \quad \text{in } \Omega,$$

where $A(x) = (a_{ij}(x))_{1 \leq i,j \leq n}$ and $\mathbf{b}(x) = (b_1(x), \ldots, b_n(x))^{\mathrm{T}}$ (note we use the notation $v \cdot w := v^{\mathrm{T}} w$ for the Euclidean inner product of two vectors $v, w \in \mathbb{R}^n$). Recall that the divergence of a vector field $\mathbf{p}(x) = (p_1(x), \ldots, p_n(x))^{\mathrm{T}}$ is defined as $\mathrm{div}(\mathbf{p}) := \sum_{i=1}^{n} \partial_{x_i} p_i$. Then, the uniform ellipticity condition (79) reads

$$(A(x)\xi) \cdot \xi \geq \tilde{c}|\xi|^2 \quad \forall x \in \overline{\Omega},\ \xi \in \mathbb{R}^n.$$

The equation (78) is supplemented with one of the following **boundary conditions (b.c.)**:

- $u = g$ on $\partial\Omega$ (**Dirichlet b.c.**); (if $g \equiv 0$, this b.c. is called **homogeneous Dirichlet b.c.**)
- $\partial_\nu u = g$ 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 $\partial_\nu u := \nabla u \cdot \nu$ (**Neumann b.c.**);
- $\partial_\nu u + \sigma u = g$ on $\partial\Omega$, where $\sigma(x) \geq 0$ on $\partial\Omega$ (**Robin b.c.**).

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 imposed on $\partial\Omega_1$ and a Neumann boundary condition on $\partial\Omega_2$). The study of such mixed BVPs is beyond the scope of this course.

6.1.1 Two solution concepts: classical and weak solutions

We begin by considering the **homogeneous Dirichlet BVP**

$$-\mathrm{div}(A\nabla u) + \mathbf{b} \cdot \nabla u + cu = f \qquad \text{in } \Omega, \tag{80}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{81}$$

where $A = A(x) = (a_{ij}(x))_{1 \leq i,j \leq n}$, $\mathbf{b} = \mathbf{b}(x) = (b_1(x), \ldots, b_n(x))^{\mathrm{T}}$, $c = c(x)$ and $f = f(x)$ are as in (79).

A function $u \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfying (80)–(81) pointwise is called a **classical solution** of this problem. The theory of PDEs tells us that (80)–(81) has a unique classical solution, provided that a_{ij}, b_i, c, f and $\partial\Omega$ are sufficiently smooth. However, in many applications one has to consider BVPs where these smoothness requirements are violated, and for such problems the classical theory of PDEs is inappropriate. Take, e.g., Poisson's equation on the cube $\Omega = (-1, 1)^n$ in $\mathbb{R}^n$, subject to a homogeneous Dirichlet b.c.:

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega, \qquad \text{where} \quad f(x) := \mathrm{sgn}(\tfrac{1}{2} - |x|), \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{$*$}$$

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 $\mathrm{sgn}(\frac{1}{2} - |x|)$ is not a continuous function on Ω.

In order to overcome the limitations of the classical theory of PDEs and to be able to deal with PDEs 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 (80)–(81). Then, for any $v \in C_c^1(\Omega)$,

$$-\int_\Omega \operatorname{div}(A\nabla u)\, v \,\mathrm{d}x + \int_\Omega \mathbf{b}\cdot\nabla u\, v \,\mathrm{d}x + \int_\Omega cu\, v \,\mathrm{d}x = \int_\Omega f\, v \,\mathrm{d}x.$$

Integration by parts (divergence theorem) in the first integral and noting that $v = 0$ on $\partial\Omega$, we obtain

$$\int_\Omega (A\nabla u)\cdot\nabla v \,\mathrm{d}x + \int_\Omega \mathbf{b}\cdot\nabla u\, v \,\mathrm{d}x + \int_\Omega cu\, v \,\mathrm{d}x = \int_\Omega f\, v \,\mathrm{d}x \qquad \forall\, v \in C_c^1(\Omega). \tag{82}$$

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 $\partial_{x_i} u \in L^2(\Omega)$ for $i \in \{1,\dots,n\}$. Thus, remembering that u has to satisfy a homogeneous Dirichlet b.c. on $\partial\Omega$, it is natural to seek u in the space $H_0^1(\Omega)$ instead, where, as in Section 5.3,

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

Therefore, we consider the following problem: find u in $H_0^1(\Omega)$ such that (82) holds.

We note that $C_c^1(\Omega) \subset H_0^1(\Omega)$, and observe that when $u \in H_0^1(\Omega)$ and $v \in H_0^1(\Omega)$, (instead of $v \in C_c^1(\Omega)$), the expressions on the left-hand side and right-hand side of (82) are both still meaningful. This motivates the following definition.

Definition 16 *Let $a_{ij} \in C(\overline{\Omega})$ for $i,j \in \{1,\dots,n\}$, $b_i \in C(\overline{\Omega})$ for $i \in \{1,\dots,n\}$, $c \in C(\overline{\Omega})$, and let $f \in L^2(\Omega)$. Let $A : \overline{\Omega} \to \mathbb{R}^{n\times n}$, $A(x) = (a_{ij}(x))_{1\le i,j\le n}$, and $\mathbf{b} : \overline{\Omega} \to \mathbb{R}^n$, $\mathbf{b}(x) = (b_1(x),\dots,b_n(x))^{\mathrm{T}}$. A function $u \in H_0^1(\Omega)$ satisfying*

$$\int_\Omega (A\nabla u)\cdot\nabla v \,\mathrm{d}x + \int_\Omega \mathbf{b}\cdot\nabla u\, v \,\mathrm{d}x + \int_\Omega cu\, v \,\mathrm{d}x = \int_\Omega f\, v \,\mathrm{d}x \qquad \forall\, v \in H_0^1(\Omega) \tag{83}$$

is called a **weak solution** *of (80)–(81). (All partial derivatives should be understood as weak derivatives.)*

6.1.2 Existence and uniqueness of weak solutions

One can show the following existence and uniqueness result for weak solutions:

Theorem 16 (Existence and uniqueness of weak solutions) *Suppose that $a_{ij} \in C(\overline{\Omega})$ for $i,j \in \{1,\dots,n\}$, $b_i \in C^1(\overline{\Omega})$ for $i \in \{1,\dots,n\}$, $c \in C(\overline{\Omega})$, $f \in L^2(\Omega)$. Let $A : \overline{\Omega} \to \mathbb{R}^{n\times n}$, $A(x) = (a_{ij}(x))_{1\le i,j\le n}$, and $\mathbf{b} : \overline{\Omega} \to \mathbb{R}^n$, $\mathbf{b}(x) = (b_1(x),\dots,b_n(x))^{\mathrm{T}}$. Assume that (79) holds, and assume that $c - \frac{1}{2}\operatorname{div}(\mathbf{b}) \ge 0$ in $\overline{\Omega}$. Then, the BVP (80)–(81) possesses a unique weak solution $u \in H_0^1(\Omega)$. In addition, there exists a constant $c_0 > 0$ such that*

$$\|u\|_{H^1(\Omega)} \le \frac{1}{c_0}\|f\|_{L^2(\Omega)}. \tag{84}$$

Now we return to our earlier example $(*)$, which has been shown to have no classical solution. However, by applying Theorem 80 with $a_{ij} \equiv 1$ for $i = j$, $a_{ij} \equiv 0$ for $i \ne j$, $1 \le i,j \le n$ (i.e., $A \equiv I_n$), $b_i \equiv 0$ for $i = 1,\dots,n$ (i.e., $\mathbf{b} \equiv 0$), $c \equiv 0$, $f(x) = \operatorname{sgn}(\frac{1}{2} - |x|)$, and $\Omega = (-1,1)^n$, we see that (79) holds with $\tilde{c} = 1$ and the inequality $c - \frac{1}{2}\operatorname{div}(\mathbf{b}) \ge 0$ is trivially satisfied. Thus $(*)$ has a unique weak solution $u \in H_0^1(\Omega)$. The key tool in proving the existence and uniqueness of a weak solution is the Lax–Milgram theorem:

Theorem 17 (Lax–Milgram) *Let V be a real Hilbert space with norm $\|\cdot\|_V$. Let $a : V \times V \to \mathbb{R}$ and $l : V \to \mathbb{R}$ be maps with the following properties:*

- *l is linear and a is bilinear, i.e., $v \mapsto a(v,w)$ is linear for any fixed w, and $w \mapsto a(v,w)$ is linear for any fixed v,*
- *$\exists\, c_0 > 0$ such that $a(v,v) \geq c_0 \|v\|_V^2 \ \forall v \in V$ (coercivity of a),*
- *$\exists\, c_1 \geq 0$ such that $|a(v,w)| \leq c_1 \|v\|_V \|w\|_V \ \forall v, w \in V$ (boundedness of a),*
- *$\exists\, c_2 \geq 0$ such that $|l(v)| \leq c_2 \|v\|_V \ \forall v \in V$ (boundedness of l).*

Then, there exists a unique $u \in V$ such that $a(u,v) = l(v) \ \forall v \in V$.

Example 8 *Let $\Omega := (0,1)$ and let $f \in L^2(\Omega)$. Let $p : \overline{\Omega} \to \mathbb{R}$, $p(x) := 2e^x$. Consider the problem*

$$-(pu')' = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega.$$

We demonstrate how the Lax–Milgram theorem can be used to show that this problem has a unique weak solution $u \in H_0^1(\Omega)$, i.e., that there exists a unique $u \in H_0^1(\Omega)$ such that

$$\int_0^1 p\, u'\, v' \,\mathrm{d}x = \int_0^1 f\, v \,\mathrm{d}x \qquad \forall v \in H_0^1(\Omega).$$

Step 1: Define a Hilbert space V with norm $\|\cdot\|_V$, a bilinear map $a : V \times V \to \mathbb{R}$, and a linear map $l : V \to \mathbb{R}$ such that $u \in V$ is a weak solution iff $a(u,v) = l(v) \ \forall v \in V$.

We consider the Hilbert space $V := H_0^1(\Omega)$ with norm $\|\cdot\|_V := \|\cdot\|_{H^1(\Omega)}$. We define $a : V \times V \to \mathbb{R}$ and $l : V \to \mathbb{R}$ by

$$a(v,w) := \int_0^1 p\, v'\, w' \,\mathrm{d}x, \qquad l(v) := \int_0^1 f\, v \,\mathrm{d}x$$

for $v, w \in V$. Note that a is bilinear, l is linear, and $u \in V$ is a weak solution iff $a(u,v) = l(v) \ \forall v \in V$.
Step 2: Show coercivity of a, i.e., that $\exists c_0 > 0$ such that $a(v,v) \geq c_0 \|v\|_V^2 \ \forall v \in V$.

We set $c_0 := \frac{2}{1+c_\star} > 0$, where $c_\star > 0$ is the constant from the Poincaré–Friedrichs inequality (Lemma 2). Using that $p(x) = 2e^x \geq 2$ for all $x \in [0,1]$, we have

$$a(v,v) = \int_0^1 p\, |v'|^2 \,\mathrm{d}x \geq 2 \int_0^1 |v'|^2 \,\mathrm{d}x \geq \frac{2}{1+c_\star} \left(\int_0^1 |v'|^2 \,\mathrm{d}x + \int_0^1 |v|^2 \,\mathrm{d}x \right) = \frac{2}{1+c_\star} \|v\|_{H^1(\Omega)}^2 = c_0 \|v\|_V^2$$

for any $v \in V$, where we have used the Poincaré–Friedrichs inequality $\int_0^1 |v|^2 \,\mathrm{d}x \leq c_\star \int_0^1 |v'|^2 \,\mathrm{d}x \ \forall v \in V$.
Step 3: We show boundedness of a, i.e., that $\exists c_1 \geq 0$ such that $|a(v,w)| \leq c_1 \|v\|_V \|w\|_V \ \forall v, w \in V$.

We set $c_1 := 2e > 0$. Using that $|p(x)| = 2e^x \leq 2e$ for all $x \in [0,1]$, and using the Cauchy–Schwarz inequality, we have that

$$\begin{aligned} |a(v,w)| \leq \int_0^1 |p|\, |v'|\, |w'| \,\mathrm{d}x &\leq 2e \int_0^1 |v'|\, |w'| \,\mathrm{d}x = 2e (|v'|, |w'|)_{L^2(\Omega)} \\ &\leq 2e \|v'\|_{L^2(\Omega)} \|w'\|_{L^2(\Omega)} \leq 2e \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)} = c_1 \|v\|_V \|w\|_V \end{aligned}$$

for any $v, w \in V$, where we have used in the final inequality that there holds $\|v'\|_{L^2(\Omega)} \leq \|v\|_{H^1(\Omega)}$ for any $v \in H^1(\Omega)$ (and hence, in particular, for any $v \in V$ as $V \subset H^1(\Omega)$).
Step 4: We show boundedness of l, i.e., that $\exists c_2 \geq 0$ such that $|l(v)| \leq c_2 \|v\|_V \ \forall v \in V$.

We set $c_2 := \|f\|_{L^2(\Omega)} \geq 0$. Using the Cauchy–Schwarz inequality, we have for any $v \in V$ that

$$|l(v)| = \left|(f,v)_{L^2(\Omega)}\right| \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} \|v\|_{H^1(\Omega)} = c_2 \|v\|_V,$$

where we have used in the final inequality that there holds $\|v\|_{L^2(\Omega)} \leq \|v\|_{H^1(\Omega)}$ for any $v \in H^1(\Omega)$ (and hence, in particular, for any $v \in V$ as $V \subset H^1(\Omega)$).

Altogether, by the Lax–Milgram theorem there exists a unique $u \in V$ such that $a(u,v) = l(v)$ for all $v \in V$, i.e., there exists a unique weak solution $u \in V$ to the given problem. In addition, we find that $c_0\|u\|^2_{H^1(\Omega)} \le a(u,u) = l(u) \le c_2\|u\|_{H^1(\Omega)} = \|f\|_{L^2(\Omega)}\|u\|_{H^1(\Omega)}$, i.e.,

$$\|u\|_{H^1(\Omega)} \le \frac{1}{c_0}\|f\|_{L^2(\Omega)} = \frac{1+c_\star}{2}\|f\|_{L^2(\Omega)}.$$

Remark. Theorem 16 implies that the weak formulation of the elliptic BVP (80)–(81) is **well-posed in the sense of Hadamard**; that is, for each $f \in L^2(\Omega)$ there exists a unique (weak) solution $u \in H^1_0(\Omega)$, and "small" changes in f give rise to "small" changes in the corresponding solution u. The latter property follows by noting that if u_1 and u_2 are weak solutions in $H^1_0(\Omega)$ of (80)–(81) corresponding to right-hand sides f_1 and f_2 in $L^2(\Omega)$, respectively, then $u_1 - u_2$ is the weak solution in $H^1_0(\Omega)$ of (80)–(81) corresponding to the right-hand side $f_1 - f_2 \in L^2(\Omega)$. Thus, by virtue of (84),

$$\|u_1 - u_2\|_{H^1(\Omega)} \le \frac{1}{c_0}\|f_1 - f_2\|_{L^2(\Omega)}, \tag{85}$$

and the required continuous dependence of the solution of the BVP on the right-hand side follows. $\diamond$

6.1.3 Maximum principle

The maximum principle is a key property of elliptic equations. Under suitable sign-conditions imposed on the source term (i.e., the right-hand side f) 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.

We consider the BVP

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega,$$

where $\Omega \subset \mathbb{R}^n$ is a bounded open set, $f \in C(\Omega)$ and $g \in C(\partial\Omega)$. Our goal is to show that if $f(x) \le 0$ for all $x \in \Omega$, and if $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is a classical solution to the above BVP, then the maximum value of u over $\overline{\Omega}$ is attained on the boundary $\partial\Omega$ of the domain, i.e.,

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

This is known as the **maximum principle**.

Proof for the case $f < 0$

Suppose 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 BVP, i.e., $-\Delta u(x) = f(x)$ for all $x \in \Omega$ and $u(x) = g(x)$ for all $x \in \partial\Omega$. We prove that the maximum value of u is then attained on $\partial\Omega$. Suppose otherwise, that u attains its maximum value at some interior point $x_0 \in \Omega$. Then,

$$\partial_{x_i} u(x_0) = 0, \quad \partial^2_{x_i x_i} u(x_0) \le 0 \qquad \forall i \in \{1,\dots,n\}.$$

Hence, $-\Delta u(x_0) = -\sum_{i=1}^n \partial^2_{x_i x_i} u(x_0) \ge 0$, which contradicts the assumption that $f(x) < 0$ for all $x \in \Omega$. The maximum value of u must therefore be attained on $\partial\Omega$, i.e., $\max_{x\in\overline{\Omega}} u(x) = \max_{x\in\partial\Omega} u(x)$.

Proof for the case $f \le 0$

Let us now prove the maximum principle 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

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

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,

$$\max_{x\in\partial\Omega} u(x) = \max_{x\in\partial\Omega}\left[v(x) - \frac{\varepsilon}{2n}|x|^2\right] \geq \max_{x\in\partial\Omega} v(x) - \max_{x\in\partial\Omega}\left[\frac{\varepsilon}{2n}|x|^2\right] = \max_{x\in\overline{\Omega}} v(x) - \frac{\varepsilon}{2n}\max_{x\in\partial\Omega}|x|^2.$$

As $v(x) = u(x) + \frac{\varepsilon}{2n}|x|^2 \geq u(x)$ for all $x \in \overline{\Omega}$, we find $\max_{x\in\partial\Omega} u(x) \geq \max_{x\in\overline{\Omega}} u(x) - \frac{\varepsilon}{2n}\max_{x\in\partial\Omega}|x|^2$ for all $\varepsilon > 0$. Since the expression on the left-hand side of this inequality is independent of ε, as is the first term on the right-hand side, by passing to the limit $\varepsilon \searrow 0$, we deduce that $\max_{x\in\partial\Omega} u(x) \geq \max_{x\in\overline{\Omega}} u(x)$. As $\partial\Omega \subset \overline{\Omega}$, trivially $\max_{x\in\overline{\Omega}} u(x) \geq \max_{x\in\partial\Omega} u(x)$. Therefore, we have $\max_{x\in\overline{\Omega}} u(x) = \max_{x\in\partial\Omega} u(x)$.

Remark 8 (Minimum principle when $f \geq 0$) *Analogously, if $-\Delta u = f$ in Ω, $u = g$ on $\partial\Omega$, and $f \geq 0$ in Ω, then $-u$ is the solution of the PDE $-\Delta(-u) = -f \leq 0$. Therefore $-u$ attains its maximum value on the boundary $\partial\Omega$ of the domain Ω. Equivalently, u attains its minimum value on $\partial\Omega$, i.e.,*

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

*This is known as the **minimum principle**.*

6.2 Methodology of FD schemes

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

$$\begin{aligned} \mathcal{L}u &= f \qquad \text{in } \Omega, \\ \mathcal{B}u &= g \qquad \text{on } \Gamma := \partial\Omega, \end{aligned} \tag{86}$$

where $\mathcal{L} : u \mapsto \mathcal{L}u$ is a linear partial differential operator, and $\mathcal{B} : u \mapsto \mathcal{B}u$ is a linear operator which specifies the b.c.. For example,

$$\mathcal{L}u := -\mathrm{div}(A\nabla u) + \mathbf{b}\cdot\nabla u + cu,$$

and $\mathcal{B}u := u$ (Dirichlet b.c.), or $\mathcal{B}u := \partial_\nu u$ (Neumann b.c.).

In general, it is impossible to determine the solution of the BVP (86) in closed form. Thus, the goal of this chapter is to describe a simple and general numerical technique for the approximate solution of (86), called the **finite difference (FD) method**. The construction of a FD scheme consists of two basic steps: first, the computational domain is approximated by a finite set of points, called the FD mesh, and second, the derivatives appearing in the PDE (and, possibly also in the b.c.) are approximated by divided differences (difference quotients) on the FD 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 x_i): the smaller $\max_{1\leq i\leq n} h_i$ 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 b.c. in a similar fashion. This yields the FD scheme

$$\begin{aligned} \mathcal{L}_h U(x) &= f_h(x), \qquad x \in \Omega_h, \\ \mathcal{B}_h U(x) &= g_h(x), \qquad x \in \Gamma_h, \end{aligned} \tag{87}$$

where f_h and g_h are suitable approximations of f and g, respectively. Now (87) is a system of linear algebraic equations involving the values of U at the mesh-points, and can be solved by, e.g., Gaussian elimination, 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 FD schemes:

- the first, and more fundamental, is the problem of approximation, that is, whether (87) approximates the BVP (86) 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.
- the second problem concerns the effective solution of the discrete problem (87) using techniques from numerical linear algebra.

In this course, our focus is on the first of these two problems. (See MA4230 for an introduction to numerical linear algebra).

6.3 FD approximation of a two-point boundary-value problem

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

$$\begin{aligned} -u'' + cu &= f \quad \text{in } (0,1), \\ u(0) = 0, \quad u(1) &= 0, \end{aligned} \tag{88}$$

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

6.3.1 Construction of a FD scheme

The first step in the construction of a FD scheme for this BVP is to define the mesh. Let $N \in \mathbb{N}$ with $N \geq 2$, and let $h := \frac{1}{N}$ be the mesh-size; the mesh-points are $x_i := ih$ for $i \in \{0, 1, \ldots, N\}$. Formally, $\Omega_h := \{x_1, \ldots, x_{N-1}\}$ 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,

$$\begin{aligned} u(x_{i+1}) = u(x_i + h) &= u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u'''(x_i) + \mathcal{O}(h^4), \\ u(x_{i-1}) = u(x_i - h) &= u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u'''(x_i) + \mathcal{O}(h^4), \end{aligned}$$

so that

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

and

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

The operators D_x^+ and D_x^- are called the **forward/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

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

Thus, we replace the second derivative u'' in the DE at a mesh point x_i by the second divided difference $D_x^+ D_x^- u(x_i)$; hence,

$$\begin{aligned} -D_x^+ D_x^- u(x_i) + c(x_i)u(x_i) &\approx f(x_i), \quad i \in \{1, \ldots, N-1\}, \\ u(x_0) = 0, \quad u(x_N) &= 0. \end{aligned} \tag{89}$$

Now (89) 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:

$$\begin{gathered} -D_x^+ D_x^- U_i + c(x_i)U_i = -\frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} + c(x_i)U_i = f(x_i), \quad i \in \{1, \ldots, N-1\}, \\ U_0 = 0, \;\; U_N = 0. \end{gathered} \tag{90}$$

This is, in fact, a system of $N-1$ linear algebraic equations for the $N-1$ unknowns $U_1, \ldots, U_{N-1}$. 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 $A \in \mathbb{R}^{(N-1)\times(N-1)}$ is the symmetric tridiagonal matrix displayed above, and $U := (U_1, \ldots, U_{N-1})^{\mathrm{T}} \in \mathbb{R}^{N-1}$ and $F := (f(x_1), \ldots, f(x_{N-1}))^{\mathrm{T}} \in \mathbb{R}^{N-1}$.

6.3.2 Existence and uniqueness of solutions to the FD scheme

We begin the analysis of the FD scheme (90) by showing that it has a unique solution. It suffices to show that the matrix A is invertible.

Remark 9 *If $c(x) > 0$ $\forall x \in \overline{\Omega}$, then A is strictly diagonally dominant, i.e.,*

$$|a_{ii}| > \sum_{j \in \{1,\ldots,N-1\} \setminus \{i\}} |a_{ij}| \qquad \forall i \in \{1, \ldots, N-1\},$$

and hence, A is invertible and the FD scheme (90) possesses a unique solution U.

In the more general case $c \geq 0$ in $\overline{\Omega}$, we will prove invertibility of A by developing a technique which we shall, in subsequent sections, extend to the FD approximation of PDEs. The purpose of this section is to introduce the key ideas through the FD approximation (89) of the simple two-point BVP (88).

For this purpose, we introduce, for two functions V and W defined at the interior mesh-points $x_1, \ldots, x_{N-1}$, 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 $(v, w)_{L^2((0,1))} := \int_0^1 v(x)w(x)\,\mathrm{d}x$.

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 BVP (88) satisfies the homogeneous b.c. $u(0) = u(1) = 0$ at the end-points of the interval $[0, 1]$:

$$\int_0^1 (-u''(x) + c(x)u(x))\, u(x)\,\mathrm{d}x = \int_0^1 |u'(x)|^2\,\mathrm{d}x + \int_0^1 c(x)|u(x)|^2\,\mathrm{d}x \geq \int_0^1 |u'(x)|^2\,\mathrm{d}x, \tag{91}$$

thanks to the assumption that $c(x) \geq 0$ for all $x \in [0, 1]$. Thus, if e.g. f is identically zero on $[0, 1]$, then so is $-u'' + cu$, and thanks to the inequality (91) 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 zero. In other words, the only solution to the homogeneous BVP (i.e., the BVP with $f \equiv 0$) is the function $u \equiv 0$. For the FD approximation of the BVP, if we could show by an analogous argument that the homogeneous system of linear algebraic equations corresponding to $f(x_i) = 0$, $i \in \{1, \ldots, N-1\}$, has the trivial solution $U_i = 0$, $i \in \{0, \ldots, N\}$, 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 $(-u'', u)_{L^2((0,1))} = (u', u')_{L^2((0,1))} = \|u'\|^2_{L^2((0,1))}$ satisfied by the function u, obeying the homogeneous b.c. $u(0) = u(1) = 0$, used in (91) above.

Lemma 3 *Suppose that V is a function defined at the mesh-points x_i, $i \in \{0, 1, \ldots, N\}$, and $V_0 = V_N = 0$; then,*

$$(-D_x^+ D_x^- V, V)_h = \sum_{i=1}^{N} h \left| D_x^- V_i \right|^2 . \tag{92}$$

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

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

where in the third equality, we shifted the index in the first summation and used $V_0 = V_N = 0$. □

Now, let V be as in the above lemma and note that as $c(x) \geq 0$ for all $x \in [0,1]$, we have that

$$(AV, V)_h = (-D_x^+ D_x^- V + cV, V)_h = (-D_x^+ D_x^- V, V)_h + (cV, V)_h \geq \sum_{i=1}^{N} h \left| D_x^- V_i \right|^2 . \tag{93}$$

Thus, if $AV = 0$ for some $V = (V_1, \ldots, V_{N-1})^{\mathrm{T}} \in \mathbb{R}^{N-1}$, then $D_x^- V_i = 0$ for all $i \in \{1, \ldots, N\}$; because $V_0 = V_N = 0$, this implies that $V_i = 0$ for all $i \in \{0, 1, \ldots, N\}$. Hence $AV = 0$ iff $V = 0$. It follows that A is invertible, and thereby (90) has a unique solution, $U = A^{-1} F$. We record this result in the next theorem.

Theorem 18 *Suppose that $c, f \in C([0,1])$, and $c(x) \geq 0$ for all $x \in [0,1]$; then, the FD scheme (90) possesses a unique solution U.*

We note in passing that, by Theorem 16, the BVP (88) has a unique (weak) solution under the hypotheses on c and f asserted in Theorem 18.

Remark 10 *We used the symbol A to denote the matrix of the system of linear equations that arises from the FD approximation as well as the FD 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 FD mesh $\overline{\Omega}_h$ with the understanding that $U_0 = U_N = 0$. For notational simplicity we shall continue to use these conventions throughout: i.e., we shall use the same notation for matrices and FD operators, and we shall use the same notation for vectors and mesh functions defined over FD meshes. It will be clear from the context which of the two interpretations of the same symbol is intended.*

6.3.3 Stability, consistency, and convergence

Next, we investigate the approximation properties of the FD scheme (90). A key ingredient in our analysis is the fact that the scheme (90) is stable (or discretely well-posed) in the sense that "small" perturbations in the data result in "small" perturbations in the FD solution. Effectively, we prove a discrete version of (84). Let us define the **discrete L^2-norm** $\|\cdot\|_h$ and the **discrete Sobolev norm** $\|\cdot\|_{1,h}$ by

$$\|U\|_h := \sqrt{(U,U)_h} = \sqrt{\sum_{i=1}^{N-1} h|U_i|^2}, \qquad \|U\|_{1,h} := \sqrt{\|U\|_h^2 + \|D_x^- U]|_h^2},$$

where $\|V]|_h := \sqrt{\sum_{i=1}^{N} h\,|V_i|^2}$ is the norm induced by the inner product $(V,W]_h := \sum_{i=1}^{N} hV_iW_i$.

Using this notation, the inequality (93) can be rewritten as

$$(AV,V)_h \geq \|D_x^- V]|_h^2. \tag{94}$$

In fact, by employing a discrete version of the Poincaré–Friedrichs inequality (77), stated in Lemma 4 below, we shall be able to prove that $(AV,V)_h \geq c_0\|V\|_{1,h}^2$, where $c_0 > 0$ is a constant independent of h.

Lemma 4 (Discrete Poincaré–Friedrichs inequality) *Let V be a function defined on the FD mesh $\{x_i := ih : i \in \{0,\dots,N\}\}$, where $h := \frac{1}{N}$ and $N \in \mathbb{N}_{\geq 2}$, and such that $V_0 = V_N = 0$; then, there exists a constant $c_\star > 0$, independent of V and h, such that, for all such V,*

$$\|V\|_h^2 \leq c_\star \|D_x^- V]|_h^2. \tag{95}$$

Proof. Using the definition of $D_x^- V_i$ and the Cauchy–Schwarz inequality, we have

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

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

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

We find the claimed inequality (95) holds with $c_\star = \frac{1}{2}$. □

Using the inequality (95) to bound the right-hand side of the inequality (94) from below we obtain

$$(AV,V)_h \geq \frac{1}{c_\star}\|V\|_h^2. \tag{96}$$

Adding the inequality (94) to the inequality (96) we arrive at the inequality

$$(AV,V)_h \geq \frac{1}{1+c_\star}\left(\|V\|_h^2 + \|D_x^- V]|_h^2\right) = c_0\|V\|_{1,h}^2, \tag{97}$$

where $c_0 := \frac{1}{1+c_\star}$. Now the stability of the FD scheme (90) easily follows.

Theorem 19 *The scheme (90) is stable in the sense that*

$$\|U\|_{1,h} \leq \frac{1}{c_0}\|f\|_h. \tag{98}$$

PROOF. From the inequality (97) and the definition (90) of the FD scheme we have that

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

and hence the inequality (98). □

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

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

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

$$Ae_i = Au(x_i) - AU_i = Au(x_i) - f(x_i) = -D_x^+D_x^-u(x_i) + c(x_i)u(x_i) - f(x_i) = u''(x_i) - D_x^+D_x^-u(x_i)$$

for $i \in \{1,\ldots,N-1\}$. Thus,

$$Ae_i = \varphi_i, \quad i \in \{1,\ldots,N-1\}, \qquad\qquad e_0 = e_N = 0 \tag{99}$$

where $\varphi_i := Au(x_i) - f(x_i) = u''(x_i) - D_x^+D_x^-u(x_i)$ is the **consistency error** (or **truncation error**). By applying the inequality (98) to the FD scheme (99), we obtain

$$\|u-U\|_{1,h} = \|e\|_{1,h} \leq \frac{1}{c_0}\|\varphi\|_h. \tag{100}$$

It remains to estimate $\|\varphi\|_h$. We showed in Section 6.3.1 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 constant $C > 0$, independent of h, such that $|\varphi_i| \leq Ch^2$ for $h > 0$ sufficiently small. Consequently,

$$\|\varphi\|_h = \sqrt{\sum_{i=1}^{N-1} h\,|\varphi_i|^2} \leq Ch^2. \tag{101}$$

By combining the inequalities (100) and (101) it follows that

$$\|u-U\|_{1,h} \leq \frac{C}{c_0}h^2. \tag{102}$$

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

$$\varphi_i = u''(x_i) - D_x^+D_x^-u(x_i) = u''(x_i) - \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} = -\frac{h^2}{12}u^{(4)}(\xi_i)$$

for some $\xi_i \in (x_{i-1}, x_{i+1})$. Thus,

$$|\varphi_i| \leq \frac{h^2}{12}\max_{x\in[0,1]}\left|u^{(4)}(x)\right| = \frac{h^2}{12}\|u^{(4)}\|_{C([0,1])},$$

and hence, we can take $C = \frac{1}{12}\|u^{(4)}\|_{C([0,1])}$ in inequality (101). Recalling that $c_0 = \frac{1}{1+c_\star}$ and $c_\star = \frac{1}{2}$, we deduce that $c_0 = \frac{2}{3}$. Substituting the values of the constants C and c_0 into (102) it follows that

$$\|u-U\|_{1,h} \leq \frac{h^2}{8}\|u^{(4)}\|_{C([0,1])}.$$

Thus we have proved the following result.

Theorem 20 *Let $f, c \in C([0,1])$ with $c(x) \geq 0$ for all $x \in [0,1]$, and suppose that the corresponding (weak) solution of the BVP (88) belongs to $C^4([0,1])$; then*

$$\|u-U\|_{1,h} \leq \frac{h^2}{8}\|u^{(4)}\|_{C([0,1])}. \tag{103}$$

6.4 Key steps of a general error analysis for FD approximations of elliptic PDEs

The analysis of the simple FD scheme (90) contains the key steps of a general error analysis for FD approximations of (elliptic) PDEs:

(1) The first step is to prove the stability of the scheme in an appropriate mesh-dependent norm (see inequality (98), for example). A typical stability result for the general FD scheme (87) is

$$|||U|||_{\Omega_h} \leq C_1(\|f_h\|_{\Omega_h} + \|g_h\|_{\Gamma_h}), \tag{104}$$

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 $C_1 > 0$ is a 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 FD scheme (88) $\varphi_{\Gamma_h} = 0$, 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 (86), we say that the scheme (87) is **consistent.** If $p \in \mathbb{N}$ is the largest natural number such that

$$\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h} = \mathcal{O}(h^p)$$

as $h \to 0$, for all sufficiently smooth u, the scheme is said to have **order of accuracy** (or **order of consistency**) p.

The FD scheme (87) is said to provide a **convergent** approximation to the solution u of the BVP (86) in the norm $|||\cdot|||_{\Omega_h}$, if

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

If $q \in \mathbb{N}$ is the largest natural number such that

$$|||u - U|||_{\Omega_h} = \mathcal{O}(h^q)$$

as $h \to 0$, then the scheme is said to have **order of convergence** q. From these definitions we deduce the following fundamental theorem.

Theorem 21 *Suppose that the FD scheme (87), involving linear FD operators $\mathcal{L}_h$ and $\mathcal{B}_h$, is stable (i.e., the inequality (104) holds for all f_h and g_h) and that the scheme is a consistent approximation of the BVP (86); then the FD scheme (87) is a convergent approximation of the BVP (86), and the order of convergence q is not smaller than the order of accuracy p.*

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

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

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 consistency of order p of the FD scheme. □

Thus, *stability* and *consistency* imply *convergence*. This abstract result is at the heart of the convergence analysis of FD approximations of PDEs.

7 FD approximation of elliptic problems

In Section 6 we presented a detailed error analysis for a FD approximation of a two-point BVP. Here we shall carry out a similar analysis for the model problem

$$\begin{aligned} -\Delta u + cu &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{105}$$

where $\Omega := (0,1)^2$, c is a continuous function on $\overline{\Omega}$, i.e., $c \in C(\overline{\Omega})$, and $c(x,y) \geq 0 \ \forall (x,y) \in \overline{\Omega}$. For the function f, we assume that f is a continuous on $\overline{\Omega}$, i.e., $f \in C(\overline{\Omega})$.

We assume that the unique weak solution u to this BVP belongs to $C^4(\overline{\Omega})$ (in particular, u is a classical solution).

Remark 11 *In this course, we do not consider the more general case where f is merely in $L^2(\Omega)$, and where a classical solution does not exist; see the original notes by Endre Süli for this. This gives rise to technical difficulties: in particular, we cannot use a Taylor series expansion to estimate the size of the consistency error.*

The first step in the construction of the FD approximation of (105) is to define the mesh. Let $N \in \mathbb{N}_{\geq 2}$ and let $h := \frac{1}{N}$; the mesh-points are (x_i, y_j), $i,j \in \{0,1,\ldots,N\}$, where $x_i := ih$, $y_j := jh$. These mesh-points form the mesh

$$\overline{\Omega}_h := \{(x_i, y_j) \,:\, i,j \in \{0,1,\ldots,N\}\} \subset \overline{\Omega}.$$

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

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

and the set of boundary mesh-points $\Gamma_h := \overline{\Omega}_h \setminus \Omega_h$. Analogously to (90), the FD scheme is

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) + c(x_i,y_j)U_{i,j} &= f(x_i,y_j) \qquad \text{for } (x_i,y_j) \in \Omega_h, \\ U &= 0 \qquad \text{on } \Gamma_h. \end{aligned} \tag{106}$$

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

$$-\left[\frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} + \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{h^2}\right] + c(x_i,y_j)U_{i,j} = f(x_i,y_j), \quad i,j \in \{1,\ldots,N-1\}, \tag{107}$$

$$U_{i,j} = 0 \quad \text{if } i \in \{0,N\} \text{ or if } j \in \{0,N\}. \tag{108}$$

For each i and j, $i,j \in \{1,\ldots,N-1\}$, the FD equation (107) 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 (107)–(108) as a system of linear algebraic equations

$$AU = F, \tag{109}$$

where now

$$\begin{aligned} U &= (U_{11},\ldots,U_{1,N-1},U_{21},\ldots,U_{2,N-1},\ldots\ldots,U_{N-1,1},\ldots,U_{N-1,N-1})^{\mathrm{T}}, \\ F &= (F_{11},\ldots,F_{1,N-1},F_{21},\ldots,F_{2,N-1},\ldots\ldots,F_{N-1,1},\ldots,F_{N-1,N-1})^{\mathrm{T}}, \end{aligned}$$

and $A \in \mathbb{R}^{(N-1)^2 \times (N-1)^2}$ is a 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 FD stencil shown in Fig. 2, while the sparsity structure of A is depicted in Fig. 3.

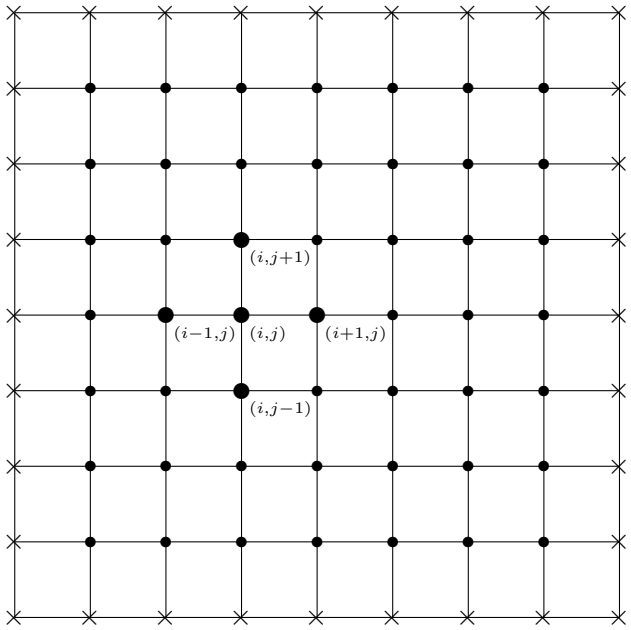

Figure 2: The mesh $\Omega_h(\cdot)$, the boundary mesh $\Gamma_h(\times)$, and a typical five-point difference stencil.

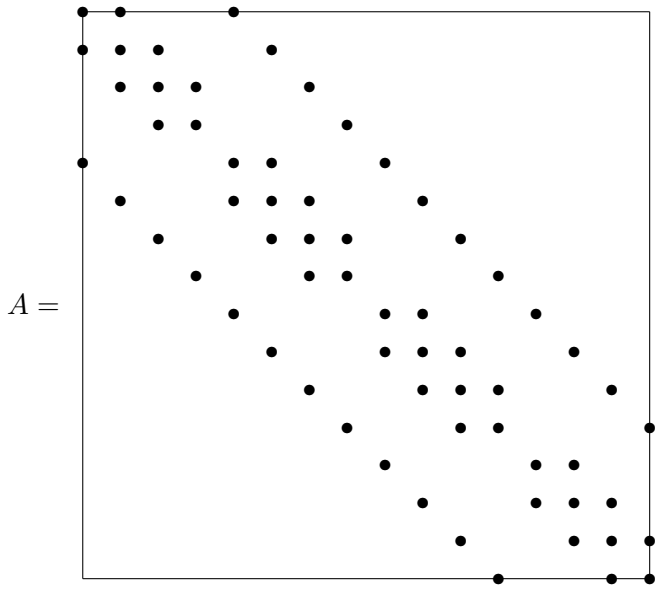

Figure 3: The sparsity structure of the matrix $A \in \mathbb{R}^{(N-1)^2 \times (N-1)^2}$ (illustration for $N = 5$).

7.1 Existence and uniqueness, stability, consistency, and convergence

Next we show that (106) has a unique solution. We proceed analogously as in Section 6. 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 $(v,w)_{L^2(\Omega)} := \int_\Omega v(x,y)\, w(x,y)\,\mathrm{d}x\,\mathrm{d}y$. 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,*

$$(-D_x^+ D_x^- V, V)_h + (-D_y^+ D_y^- V, V)_h = \sum_{i=1}^{N}\sum_{j=1}^{N-1} h^2 |D_x^- V_{i,j}|^2 + \sum_{i=1}^{N-1}\sum_{j=1}^{N} h^2 |D_y^- V_{i,j}|^2. \tag{110}$$

Proof. The identity (110) is a direct consequence of (92) and the analogous identity for $-D_y^+ D_y^-$. □

Returning to the analysis of the FD scheme (106), we shall now proceed in much the same way as in the univariate case considered in the previous section. We note that, since $c \geq 0$ on $\overline{\Omega}$, by (110) we have

$$\begin{aligned}(AV,V)_h &= (-D_x^+ D_x^- V - D_y^+ D_y^- V + cV, V)_h \\ &= (-D_x^+ D_x^- V, V)_h + (-D_y^+ D_y^- V, V)_h + (cV,V)_h \\ &\geq \sum_{i=1}^{N}\sum_{j=1}^{N-1} h^2 |D_x^- V_{i,j}|^2 + \sum_{i=1}^{N-1}\sum_{j=1}^{N} h^2 |D_y^- V_{i,j}|^2,\end{aligned} \tag{111}$$

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 6, that A is an invertible matrix. Indeed, if $AV = 0$, then (111) yields

$$\begin{aligned} D_x^- V_{i,j} &= \frac{V_{i,j} - V_{i-1,j}}{h} = 0, \quad i \in \{1,\ldots,N\}, j \in \{1,\ldots,N-1\}; \\ D_y^- V_{i,j} &= \frac{V_{i,j} - V_{i,j-1}}{h} = 0, \quad i \in \{1,\ldots,N-1\}, j \in \{1,\ldots,N\}. \end{aligned}$$

Since $V = 0$ on Γ_h, these imply that $V = 0$. Thus $AV = 0$ iff $V = 0$. Hence A is invertible, and $U = A^{-1}F$ is the unique solution of (106). Thus the solution of the FD scheme (106) may be found by solving the system of linear algebraic equations (109).

In order to prove the stability of the FD scheme (106), we introduce (similarly as in the univariate case) the mesh–dependent norms

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

where

$$\|D_x^- U]|_x := \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N-1} h^2 |D_x^- U_{i,j}|^2}, \qquad \|D_y^- U]|_y := \sqrt{\sum_{i=1}^{N-1}\sum_{j=1}^{N} h^2 |D_y^- U_{i,j}|^2}.$$

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

$$\|u\|_{H^1(\Omega)} := \sqrt{\|u\|_{L^2(\Omega)}^2 + \|\partial_x u\|_{L^2(\Omega)}^2 + \|\partial_y u\|_{L^2(\Omega)}^2}.$$

With this new notation, the inequality (111) can be rewritten in the following compact form:

$$(AV,V)_h \geq \|D_x^- V]|_x^2 + \|D_y^- V]|_y^2. \tag{112}$$

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

$$\|V\|_h^2 \leq c_* \left(\|D_x^- V]|_x^2 + \|D_y^- V]|_y^2 \right). \tag{113}$$

Proof. The inequality (113) is a straightforward consequence of its univariate counterpart (95). It follows from (95) that, for each fixed $j \in \{1, \ldots, N-1\}$,

$$\sum_{i=1}^{N-1} h|V_{i,j}|^2 \leq \frac{1}{2} \sum_{i=1}^{N} h|D_x^- V_{i,j}|^2. \tag{114}$$

Analogously, for each fixed $i \in \{1, \ldots, N-1\}$,

$$\sum_{j=1}^{N-1} h|V_{i,j}|^2 \leq \frac{1}{2} \sum_{j=1}^{N} h|D_y^- V_{i,j}|^2. \tag{115}$$

We first multiply (114) by h and sum through $j \in \{1, \ldots, N-1\}$, then multiply (115) by h and sum through $i \in \{1, \ldots, N-1\}$, and then add these two inequalities to obtain $2\|V\|_h^2 \leq \frac{1}{2}(\|D_x^- V]|_x^2 + \|D_y^- V]|_y^2)$. Hence we arrive at (113) with $c_* = \frac{1}{4}$. □

Now the inequalities (112) and (113) imply that $(AV, V)_h \geq \frac{1}{c_*}\|V\|_h^2$. Finally, combining this inequality with (112) and recalling the definition of the norm $\|\cdot\|_{1,h}$, we obtain

$$(AV, V)_h \geq c_0 \|V\|_{1,h}^2, \quad \text{where} \quad c_0 := \frac{1}{1 + c_*}. \tag{116}$$

Using the inequality (116) we can now prove the stability of the FD scheme (106).

Theorem 22 *The FD scheme (106) is stable in the sense that*

$$\|U\|_{1,h} \leq \frac{1}{c_0} \|f\|_h. \tag{117}$$

Proof. The proof of this inequality is identical to that of the stability inequality (98) in the univariate case. From (116) and (106) we have that

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

and hence we arrive at the desired inequality (117). □

Having established stability of the FD scheme (106), we turn to the question of its accuracy. We define the **global error** e by

$$e_{i,j} := u(x_i, y_j) - U_{i,j}, \quad i, j \in \{0, 1, \ldots, N\}.$$

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 for $i, j \in \{1, \ldots, N-1\}$ that

$$\begin{aligned} Ae_{i,j} = Au(x_i, y_j) - f(x_i, y_j) &= \Delta u(x_i, y_j) - (D_x^+ D_x^- u(x_i, y_j) + D_y^+ D_y^- u(x_i, y_j)) \\ &= \left[\partial_{xx}^2 u(x_i, y_j) - D_x^+ D_x^- u(x_i, y_j)\right] + \left[\partial_{yy}^2 u(x_i, y_j) - D_y^+ D_y^- u(x_i, y_j)\right] \\ &= -\frac{h^2}{12} \partial_{xxxx}^4 u(\xi_i, y_j) - \frac{h^2}{12} \partial_{yyyy}^4 u(x_i, \eta_j) \end{aligned}$$

for some $\xi_i \in (x_{i-1}, x_{i+1})$, $\eta_j \in (y_{j-1}, y_{j+1})$. We define the **consistency error** (or **truncation error**) of the FD scheme (106) by

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

Then, by the calculations above,

$$\varphi_{i,j} = -\frac{h^2}{12}\left(\partial^4_{xxxx} u(\xi_i, y_j) + \partial^4_{yyyy} u(x_i, \eta_j)\right)$$

for $i, j \in \{1, \ldots, N-1\}$, and

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

Thanks to the stability result (117), we therefore have that

$$\|u - U\|_{1,h} = \|e\|_{1,h} \leq \frac{1}{c_0}\|\varphi\|_h. \tag{118}$$

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 (118). Indeed, by noting that

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

we deduce that the consistency error φ satisfies

$$\|\varphi\|_h \leq \frac{h^2}{12}\left(\left\|\partial^4_{xxxx} u\right\|_{C(\overline{\Omega})} + \left\|\partial^4_{yyyy} u\right\|_{C(\overline{\Omega})}\right). \tag{119}$$

Finally, (118) and (119) yield the following result.

Theorem 23 *Let $f, c \in C(\overline{\Omega})$ with $c(x, y) \geq 0$ for all $(x, y) \in \overline{\Omega}$, and suppose that the corresponding weak solution u of the BVP (105) belongs to $C^4(\overline{\Omega})$; then,*

$$\|u - U\|_{1,h} \leq \frac{5h^2}{48}\left(\left\|\partial^4_{xxxx} u\right\|_{C(\overline{\Omega})} + \left\|\partial^4_{yyyy} u\right\|_{C(\overline{\Omega})}\right). \tag{120}$$

Proof. Recall that $c_0 = \frac{1}{1+c_*}$ and $c_* = \frac{1}{4}$, so that $\frac{1}{c_0} = \frac{5}{4}$, and combine (118) and (119). □

According to this result, the five-point difference scheme (106) for the BVP (105) is second-order convergent, provided that u is sufficiently smooth. As in the univariate case, we have deduced second-order convergence of the FD scheme from its stability and its second-order consistency, under the assumption that the true solution u satisfies $u \in C^4(\overline{\Omega})$.

7.2 Nonaxiparallel domains and nonuniform meshes

We have carried out an error analysis of FD schemes for the PDE $-\Delta u + cu = f$ on a square domain Ω. The error analysis of FD schemes for more general elliptic equations would proceed along similar lines. Consider, e.g.,

$$-\left[\partial_x(a_1 \partial_x u) + \partial_y(a_2 \partial_y u)\right] + b_1 \partial_x u + b_2 \partial_y u + cu = f$$

on the unit square $\Omega := (0,1)^2$ in $\mathbb{R}^2$. We approximate this PDE by

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

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

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

Similarly, let $k_{j+1} := y_{j+1} - y_j$, $k_j := y_j - y_{j-1}$, and let $\hbar_j := \frac{1}{2}(k_{j+1} + k_j)$. Let

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

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 $D_x^+D_x^- + D_y^+D_y^-$.

Consider, e.g., the Dirichlet problem

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

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

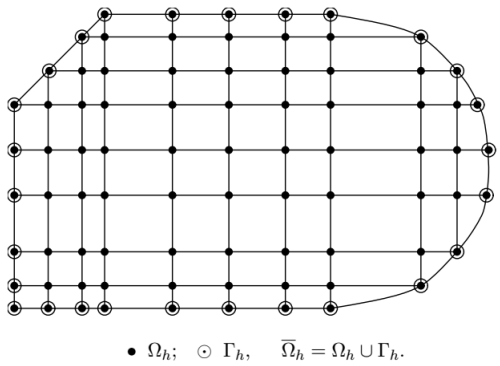

Figure 4: Nonuniform mesh $\overline{\Omega}_h$.

The FD approximation of this BVP is

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

or equivalently,

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

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

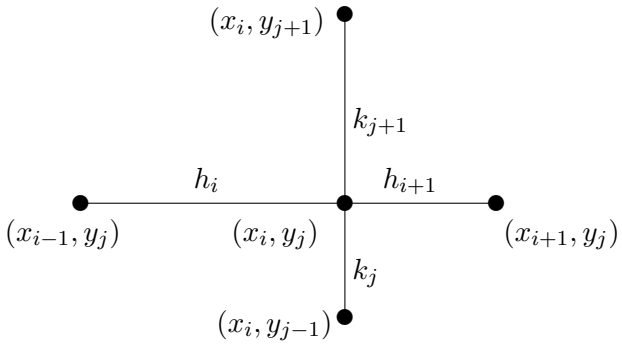

Figure 5: Five-point stencil on a nonuniform mesh.

7.3 The discrete maximum principle

Our objective is to construct a FD approximation of the elliptic BVP $-\Delta u = f$ in Ω, $u = g$ on $\partial\Omega$, and show that a discrete counterpart of the maximum principle satisfied by the function u holds for its FD approximation U. For simplicity, we 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

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

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 7.2. The FD 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

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) &= f(x_i, y_j) \qquad \text{in } \Omega_h, \\ U_{i,j} &= g(x_i, y_j) \qquad \text{on } \Gamma_h. \end{aligned} \tag{121}$$

Equivalently,

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

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

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

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

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

Note, however, that the expressions on the two sides of this 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 $f(x_i, y_j) \le 0$ for all $(x_i, y_j) \in \Omega_h$. We define the auxiliary mesh function V by

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

Hence,

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

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

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

As $V_{i,j} \ge U_{i,j}$ for $(x_i, y_j) \in \overline{\Omega}_h$, it follows $\max_{(x_i,y_j)\in\Gamma_h} U_{i,j} \ge \max_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j} - \frac{\varepsilon}{4}\max_{(x_i,y_j)\in\Gamma_h}(x_i^2+y_j^2)$ for all $\varepsilon > 0$. By passing to the limit $\varepsilon \searrow 0$ it follows that $\max_{(x_i,y_j)\in\Gamma_h} U_{i,j} \ge \max_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j}$. As $\Gamma_h \subset \overline{\Omega}_h$, trivially $\max_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j} \ge \max_{(x_i,y_j)\in\Gamma_h} U_{i,j}$, and therefore we deduce from these two inequalities that if $f(x_i, y_j) \le 0$ 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 $f(x_i, y_j) \ge 0$ for all $(x_i, y_j) \in \Omega_h$, then the **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 FD scheme (121) with respect to perturbations in the boundary data.

7.4 Stability in the discrete maximum norm

Consider the FD scheme (121) on the nonuniform mesh formulated in Section 7.2. Our first result asserts the existence of a solution to (121) as well as its uniqueness.

Lemma 7 *The FD scheme* (121) *has a unique solution.*

Proof. We note that (121) is 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 FD scheme (121) is therefore equivalent to showing that the matrix of the linear system

is invertible. The matrix of the linear system associated with (121) is invertible iff 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 FD scheme (121) with $f(x_i, y_j) = 0$ for all $(x_i, y_j) \in \Omega_h$ and $g(x_i, y_j) = 0$ for all $(x_i, y_j) \in \Gamma_h$ has the trivial solution as its only solution, i.e., that $U_{i,j} = 0$ for all $(x_i, y_j) \in \overline{\Omega}_h$. Let us therefore consider

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) &= 0 \qquad && \text{in } \Omega_h, \\ U_{i,j} &= 0 && \text{on } \Gamma_h. \end{aligned} \tag{122}$$

The existence of a solution to (122) 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 (122), we have $0 = \max_{(x_i,y_j)\in\overline{\Omega}_h} U_{i,j}$, 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 $U = 0$. This then implies the existence of a unique solution to (121). $\square$

We are now ready to embark on the analysis of the stability of the scheme (121) with respect to perturbations in the boundary data. Consider the mesh functions $U^{(1)}$ and $U^{(2)}$, which satisfy, respectively:

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j}^{(1)} + D_y^+ D_y^- U_{i,j}^{(1)}) &= f(x_i, y_j) \qquad && \text{in } \Omega_h, \\ U_{i,j}^{(1)} &= g^{(1)}(x_i, y_j) && \text{on } \Gamma_h \end{aligned} \tag{123}$$

and

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j}^{(2)} + D_y^+ D_y^- U_{i,j}^{(2)}) &= f(x_i, y_j) \qquad && \text{in } \Omega_h, \\ U_{i,j}^{(2)} &= g^{(2)}(x_i, y_j) && \text{on } \Gamma_h \end{aligned} \tag{124}$$

for given boundary data $g^{(1)}$ and $g^{(2)}$. Let $U := U^{(1)} - U^{(2)}$ and $g := g^{(1)} - g^{(2)}$. Then, by subtracting (124) from (123) we find that U solves

$$\begin{aligned} -(D_x^+ D_x^- U_{i,j} + D_y^+ D_y^- U_{i,j}) &= 0 \qquad && \text{in } \Omega_h, \\ U_{i,j} &= g(x_i, y_j) && \text{on } \Gamma_h. \end{aligned} \tag{125}$$

By the discrete maximum principle we have from (125) that

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

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

$$U_{i,j} \le \max_{(x_i,y_j)\in\Gamma_h} |g(x_i, y_j)|. \tag{126}$$

It follows from (125) that $-U$ solves

$$\begin{aligned} -(D_x^+ D_x^- (-U)_{i,j} + D_y^+ D_y^- (-U)_{i,j}) &= 0 \qquad && \text{in } \Omega_h, \\ (-U)_{i,j} &= -g(x_i, y_j) && \text{on } \Gamma_h, \end{aligned} \tag{127}$$

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

$$-U_{i,j} = (-U)_{i,j} \le \max_{(x_i,y_j)\in\Gamma_h} |-g(x_i, y_j)| = \max_{(x_i,y_j)\in\Gamma_h} |g(x_i, y_j)| \tag{128}$$

for all $(x_i, y_j) \in \overline{\Omega}_h$. By combining (126) and (128) we have the inequality $|U_{i,j}| \le \max_{(x_i,y_j)\in\Gamma_h} |g(x_i, y_j)|$ for all $(x_i, y_j) \in \overline{\Omega}_h$, and hence,

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

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

$$\max_{(x_i,y_j)\in\overline{\Omega}_h} |U_{i,j}^{(1)} - U_{i,j}^{(2)}| \leq \max_{(x_i,y_j)\in\Gamma_h} |g^{(1)}(x_i,y_j) - g^{(2)}(x_i,y_j)|. \tag{129}$$

The inequality (129) expresses continuous dependence of the solution U to the FD 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).

8 FD approximation of parabolic problems

This section is concerned with the construction and mathematical analysis of FD methods for the numerical solution of parabolic equations.

8.1 The heat equation

As a simple yet representative model problem we shall focus on the **heat equation (or diffusion equation)** in one space dimension: Seek a function $u = u(x,t)$ satisfying

$$\partial_t u = \partial_{xx}^2 u, \tag{130}$$

which we shall consider for $x \in \mathbb{R}$ and $t \in (0,\infty)$, subject to the initial condition (i.c.)

$$u(x,0) = u_0(x) \quad \text{for } \ x \in \mathbb{R}, \tag{131}$$

where $u_0 : \mathbb{R} \to \mathbb{R}$ is a given function, called an **initial datum**.

The solution of this IVP can be expressed explicitly in terms of the initial datum u_0. As the expression for the solution of the IVP provides helpful insight into the behaviour of solutions of parabolic PDEs, 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 IVP (130)–(131).

We recall that the Fourier transform of a function $v : \mathbb{R} \to \mathbb{C}$ is defined by

$$[\mathcal{F}v](\xi) := \hat{v}(\xi) := \int_{-\infty}^{\infty} v(x)\, e^{-ix\xi}\, \mathrm{d}x, \qquad \xi \in \mathbb{R}.$$

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 PDE (130) we obtain

$$\int_{-\infty}^{\infty} \partial_t u(x,t)\, e^{-ix\xi}\, \mathrm{d}x = \int_{-\infty}^{\infty} \partial_{xx}^2 u(x,t)\, e^{-ix\xi}\, \mathrm{d}x.$$

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

$$\partial_t \hat{u}(\xi,t) = (i\xi)^2 \hat{u}(\xi,t) = -\xi^2 \hat{u}(\xi,t),$$

where $\hat{u}(\xi,t) := \int_{-\infty}^{\infty} u(x,t)\, e^{-ix\xi}\, \mathrm{d}x$ is the Fourier transform of u with respect to the x-variable. Then, we see that $\hat{u}(\xi,t) = e^{-t\xi^2}\hat{u}(\xi,0) = e^{-t\xi^2}\hat{u}_0(\xi)$, and therefore $u(x,t) = \mathcal{F}^{-1}\left(\xi \mapsto e^{-t\xi^2}\hat{u}_0(\xi)\right)$, where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform defined by

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

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

$$u(x,t) = \int_{-\infty}^{\infty} w(x-y,t) u_0(y)\, \mathrm{d}y, \quad \text{where} \quad w(x,t) := \frac{1}{\sqrt{4\pi t}}\, e^{-\frac{x^2}{4t}}.$$

The function w is called the **heat kernel**. So, finally,

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-\frac{(x-y)^2}{4t}} u_0(y)\, \mathrm{d}y. \tag{132}$$

This formula gives an explicit expression for the solution of the heat equation (130) in terms of the initial datum u_0. Because $w(x,t) > 0$ for all $(x,t) \in \mathbb{R} \times (0,\infty)$, and $\int_{-\infty}^{\infty} w(y,t)\,\mathrm{d}y = 1$ for all $t \in (0,\infty)$, we deduce from (132) that if u_0 is a bounded continuous function, then

$$\sup_{x\in\mathbb{R}}|u(x,t)| \le \sup_{x\in\mathbb{R}}|u_0(x)| \qquad \forall t \in (0,\infty). \tag{133}$$

In other words, the 'largest' and 'smallest' values of $x \mapsto u(x,t)$ at $t > 0$ cannot exceed those of u_0. A similar bound on the 'magnitude' of the solution at future times in terms of the 'magnitude' of the initial datum can be obtained in the L^2-norm. We will show 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 IVP for the heat equation.

We now extend our definition of the space $L^2(\Omega)$ to complex-valued functions. In particular, we let $L^2(\mathbb{R})$ be the set of all (measurable) functions $u : \mathbb{R} \to \mathbb{C}$ for which $\int_{-\infty}^{\infty}|u(x)|^2\mathrm{d}x < \infty$, and we define

$$\|u\|_{L^2(\mathbb{R})} := \sqrt{\int_{-\infty}^{\infty}|u(x)|^2\mathrm{d}x} \quad \text{for } u \in L^2(\mathbb{R}).$$

Functions which are equal almost everywhere are identified with each other.

Lemma 8 (Parseval's identity) *Let $u \in L^2(\mathbb{R})$. Then, $\hat{u} \in L^2(\mathbb{R})$ and there holds*

$$\|u\|_{L^2(\mathbb{R})} = \frac{1}{\sqrt{2\pi}}\,\|\hat{u}\|_{L^2(\mathbb{R})}.$$

Proof. We begin by observing that

$$\begin{aligned}\int_{-\infty}^{\infty}\hat{u}(\xi)\,v(\xi)\,\mathrm{d}\xi &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}u(x)\,e^{-ix\xi}\,\mathrm{d}x\right)v(\xi)\,\mathrm{d}\xi \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}v(\xi)\,e^{-ix\xi}\,\mathrm{d}\xi\right)u(x)\,\mathrm{d}x = \int_{-\infty}^{\infty}u(x)\,\hat{v}(x)\,\mathrm{d}x,\end{aligned} \tag{134}$$

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

$$v(t) := \overline{\hat{u}(t)} = 2\pi[\mathcal{F}^{-1}\overline{u}](t), \qquad t \in \mathbb{R}.$$

Then, the left-hand side in (134) becomes $\int_{-\infty}^{\infty}\hat{u}(\xi)\,v(\xi)\,\mathrm{d}\xi = \int_{-\infty}^{\infty}|\hat{u}(\xi)|^2\mathrm{d}\xi = \|\hat{u}\|^2_{L^2(\mathbb{R})}$ and the right-hand side in (134) becomes $\int_{-\infty}^{\infty}u(x)\,\hat{v}(x)\,\mathrm{d}x = 2\pi\int_{-\infty}^{\infty}|u(x)|^2\,\mathrm{d}x = 2\pi\|u\|^2_{L^2(\mathbb{R})}$, giving the desired result. Here, we have used that

$$\hat{v}(x) = \int_{-\infty}^{\infty}v(t)\,e^{-ixt}\,\mathrm{d}t = \int_{-\infty}^{\infty}\overline{\hat{u}(t)}\,e^{-ixt}\,\mathrm{d}t = 2\pi\left(\overline{\frac{1}{2\pi}\int_{-\infty}^{\infty}\hat{u}(t)\,e^{ixt}\,\mathrm{d}t}\right) = 2\pi\overline{[\mathcal{F}^{-1}\hat{u}](x)} = 2\pi\overline{u(x)}$$

for any $x \in \mathbb{R}$. □

Returning to the equation (130), we thus have by Parseval's identity that

$$\|u(\cdot,t)\|_{L^2(\mathbb{R})} = \frac{1}{\sqrt{2\pi}}\,\|\hat{u}(\cdot,t)\|_{L^2(\mathbb{R})} \qquad \forall t \in (0,\infty),$$

and therefore

$$\|u(\cdot,t)\|_{L^2(\mathbb{R})} = \frac{1}{\sqrt{2\pi}}\,\|\xi \mapsto e^{-t\xi^2}\hat{u}_0(\xi)\|_{L^2(\mathbb{R})} \le \frac{1}{\sqrt{2\pi}}\,\|\hat{u}_0\|_{L^2(\mathbb{R})} = \|u_0\|_{L^2(\mathbb{R})} \qquad \forall t \in (0,\infty).$$

Thus we have shown that

$$\|u(\cdot,t)\|_{L^2(\mathbb{R})} \leq \|u_0\|_{L^2(\mathbb{R})} \qquad \forall t \in (0,\infty). \tag{135}$$

This is a useful result as it can be used to deduce stability of the solution of the equation (130) with respect to perturbations of the initial datum in a sense which we shall now explain. Suppose that $u_0, \tilde{u}_0 \in L^2(\mathbb{R})$ and denote by u and $\tilde{u}$ the solutions to (130) resulting from the initial datum 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 (135), we have that

$$\|u(\cdot,t) - \tilde{u}(\cdot,t)\|_{L^2(\mathbb{R})} \leq \|u_0 - \tilde{u}_0\|_{L^2(\mathbb{R})} \qquad \forall t \in (0,\infty). \tag{136}$$

This inequality implies continuous dependence of the solution on the initial datum: small perturbations in u_0 in the $L^2(\mathbb{R})$-norm will result in small perturbations in the associated solution $u(\cdot,t)$ in the $L^2(\mathbb{R})$-norm for all $t \in (0,\infty)$. The inequality (135) is therefore a relevant property, which we shall try to mimic with our numerical approximations of the equation (130).

8.2 FD approximation of the heat equation

We take our computational domain to be

$$\{(x,t) \,|\, x \in \mathbb{R}, t \in [0,T]\},$$

where $T > 0$ is a given final time. We then consider a FD mesh with spacing $\Delta x > 0$ in the x-direction and spacing $\Delta t := \frac{T}{M}$ in the t-direction, with $M \in \mathbb{N}$, and we approximate the partial derivatives appearing in the PDE using divided differences as follows. Let $x_j := j\Delta x$ and $t_m := m\Delta t$, and note that

$$\partial_t u(x_j,t_m) \approx \frac{u(x_j,t_{m+1}) - u(x_j,t_m)}{\Delta t}, \qquad \partial_{xx}^2 u(x_j,t_m) \approx \frac{u(x_{j+1},t_m) - 2u(x_j,t_m) + u(x_{j-1},t_m)}{(\Delta x)^2}.$$

This then motivates us to approximate the heat equation (130) at the point (x_j,t_m) by the following numerical method, called the **explicit Euler scheme**:

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

$$U_j^0 := u_0(x_j), \quad j \in \mathbb{Z}.$$

Equivalently, we can write this as

$$U_j^{m+1} = U_j^m + \mu(U_{j+1}^m - 2U_j^m + U_{j-1}^m), \quad j \in \mathbb{Z}, \quad m \in \{0,1,\ldots,M-1\},$$

$$U_j^0 := u_0(x_j), \quad j \in \mathbb{Z},$$

where $\mu = \frac{\Delta t}{(\Delta x)^2}$. Thus, U_j^{m+1} can be explicitly calculated, for all $j \in \mathbb{Z}$, from the values U_{j+1}^m, U_j^m, 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**:

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

$$U_j^0 := u_0(x_j), \qquad j \in \mathbb{Z}.$$

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-method is defined as follows:

$$\frac{U_j^{m+1}-U_j^m}{\Delta t} = (1-\theta)\,\frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2} + \theta\,\frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2}, \quad j \in \mathbb{Z},\ m \in \{0,1,\ldots,M-1\},$$
$$U_j^0 := u_0(x_j), \qquad j \in \mathbb{Z},$$

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 only the spatial partial derivative in the heat equation, resulting in the following IVP for a system of ODEs:

$$\frac{\mathrm{d}U_j(t)}{\mathrm{d}t} = \frac{U_{j+1}(t) - 2U_j(t) + U_{j-1}(t)}{(\Delta x)^2}, \qquad j \in \mathbb{Z},$$
$$U_j(0) := u_0(x_j), \qquad j \in \mathbb{Z}.$$

This is called a **spatially semi-discrete approximation**, because no discretization with respect to the temporal variable t has taken place. Because an IVP for the heat equation is considered for $x \in \mathbb{R}$, the spatially semidiscrete approximation consists of an infinite system of ODEs. Had the range of x been limited to a bounded interval (a,b) of the real line instead, and had, in conjunction with the i.c., boundary conditions been supplied at $x = a$ and $x = b$, spatial semi-discretization of such an initial-boundary-value problem (IBVP) for the heat equation would have resulted in a system consisting of a finite number of ODEs, coupled to algebraic equations 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, e.g., 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**.

8.2.1 Accuracy of the θ-scheme

Our aim in this section is to assess the accuracy of the θ-scheme for the IVP for the heat equation. The consistency error of the θ-scheme is defined by

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

for $j \in \mathbb{Z}$ and $m \in \{0,1,\ldots,M-1\}$, 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 choose the point $(x_j, t_{m+\frac{1}{2}})$, where $t_{m+\frac{1}{2}} := t_m + \frac{\Delta t}{2}$, we have that

$$u_j^{m+1} = u\left(x_j, t_{m+\frac{1}{2}} + \frac{\Delta t}{2}\right) = \left[u + \frac{\Delta t}{2}u_t + \frac{(\Delta t)^2}{8}u_{tt} + \frac{(\Delta t)^3}{48}u_{ttt} + \frac{(\Delta t)^4}{384}u_{tttt}\right](x_j, t_{m+\frac{1}{2}}) + \mathcal{O}((\Delta t)^5),$$
$$u_j^{m} = u\left(x_j, t_{m+\frac{1}{2}} - \frac{\Delta t}{2}\right) = \left[u - \frac{\Delta t}{2}u_t + \frac{(\Delta t)^2}{8}u_{tt} - \frac{(\Delta t)^3}{48}u_{ttt} + \frac{(\Delta t)^4}{384}u_{tttt}\right](x_j, t_{m+\frac{1}{2}}) + \mathcal{O}((\Delta t)^5).$$

Therefore, we have that

$$\frac{u_j^{m+1}-u_j^m}{\Delta t}=\left[u_t+\frac{(\Delta t)^2}{24}u_{ttt}\right](x_j,t_{m+\frac{1}{2}})+\mathcal{O}((\Delta t)^4).$$

Next, we use Taylor expansion to find

$$\begin{aligned}
\frac{u_{j+1}^m-2u_j^m+u_{j-1}^m}{(\Delta x)^2}&=\frac{u(x_{j+1},t_{m+\frac{1}{2}}-\frac{\Delta t}{2})-2u(x_j,t_{m+\frac{1}{2}}-\frac{\Delta t}{2})+u(x_{j-1},t_{m+\frac{1}{2}}-\frac{\Delta t}{2})}{(\Delta x)^2}\\
&=\frac{u(x_j+\Delta x,t_{m+\frac{1}{2}})-2u(x_j,t_{m+\frac{1}{2}})+u(x_j-\Delta x,t_{m+\frac{1}{2}})}{(\Delta x)^2}\\
&\quad-\frac{\Delta t}{2}\frac{u_t(x_j+\Delta x,t_{m+\frac{1}{2}})-2u_t(x_j,t_{m+\frac{1}{2}})+u_t(x_j-\Delta x,t_{m+\frac{1}{2}})}{(\Delta x)^2}\\
&\quad+\frac{(\Delta t)^2}{8}\frac{u_{tt}(x_j+\Delta x,t_{m+\frac{1}{2}})-2u_{tt}(x_j,t_{m+\frac{1}{2}})+u_{tt}(x_j-\Delta x,t_{m+\frac{1}{2}})}{(\Delta x)^2}\\
&\quad+\cdots\\
&=\left[u_{xx}+\frac{(\Delta x)^2}{12}u_{xxxx}+\frac{(\Delta x)^4}{360}u_{xxxxxx}+\mathcal{O}((\Delta x)^6)\right](x_j,t_{m+\frac{1}{2}})\\
&\quad-\frac{\Delta t}{2}\left[u_{xxt}+\frac{(\Delta x)^2}{12}u_{xxxxt}+\frac{(\Delta x)^4}{360}u_{xxxxxxt}+\mathcal{O}((\Delta x)^6)\right](x_j,t_{m+\frac{1}{2}})\\
&\quad+\frac{(\Delta t)^2}{8}\left[u_{xxtt}+\frac{(\Delta x)^2}{12}u_{xxxxtt}+\frac{(\Delta x)^4}{360}u_{xxxxxxtt}+\mathcal{O}((\Delta x)^6)\right](x_j,t_{m+\frac{1}{2}})+\mathcal{O}((\Delta t)^3).
\end{aligned}$$

Similarly, we find

$$\begin{aligned}
\frac{u_{j+1}^{m+1}-2u_j^{m+1}+u_{j-1}^{m+1}}{(\Delta x)^2}&=\left[u_{xx}+\frac{(\Delta x)^2}{12}u_{xxxx}+\frac{(\Delta x)^4}{360}u_{xxxxxx}+\mathcal{O}((\Delta x)^6)\right](x_j,t_{m+\frac{1}{2}})\\
&\quad+\frac{\Delta t}{2}\left[u_{xxt}+\frac{(\Delta x)^2}{12}u_{xxxxt}+\frac{(\Delta x)^4}{360}u_{xxxxxxt}+\mathcal{O}((\Delta x)^6)\right](x_j,t_{m+\frac{1}{2}})\\
&\quad+\frac{(\Delta t)^2}{8}\left[u_{xxtt}+\frac{(\Delta x)^2}{12}u_{xxxxtt}+\frac{(\Delta x)^4}{360}u_{xxxxxxtt}+\mathcal{O}((\Delta x)^6)\right](x_j,t_{m+\frac{1}{2}})+\mathcal{O}((\Delta t)^3)
\end{aligned}$$

and hence,

$$\begin{aligned}
&(1-\theta)\frac{u_{j+1}^m-2u_j^m+u_{j-1}^m}{(\Delta x)^2}+\theta\frac{u_{j+1}^{m+1}-2u_j^{m+1}+u_{j-1}^{m+1}}{(\Delta x)^2}\\
&\quad=\left[u_{xx}+\frac{(\Delta x)^2}{12}u_{xxxx}+\frac{(\Delta x)^4}{360}u_{xxxxxx}+\mathcal{O}((\Delta x)^6)\right](x_j,t_{m+\frac{1}{2}})\\
&\qquad+(2\theta-1)\frac{\Delta t}{2}\left[u_{xxt}+\frac{(\Delta x)^2}{12}u_{xxxxt}+\frac{(\Delta x)^4}{360}u_{xxxxxxt}+\mathcal{O}((\Delta x)^6)\right](x_j,t_{m+\frac{1}{2}})\\
&\qquad+\frac{(\Delta t)^2}{8}\left[u_{xxtt}+\frac{(\Delta x)^2}{12}u_{xxxxtt}+\frac{(\Delta x)^4}{360}u_{xxxxxxtt}+\mathcal{O}((\Delta x)^6)\right](x_j,t_{m+\frac{1}{2}})+\mathcal{O}((\Delta t)^3).
\end{aligned}$$

Altogether, we have that

$$\begin{aligned}
T_j^m&=\left[u_t-u_{xx}-\frac{(\Delta x)^2}{12}u_{xxxx}-\frac{(\Delta x)^4}{360}u_{xxxxxx}+\mathcal{O}((\Delta x)^6)\right](x_j,t_{m+\frac{1}{2}})\\
&\quad+(2\theta-1)\frac{\Delta t}{2}\left[-u_{xxt}-\frac{(\Delta x)^2}{12}u_{xxxxt}-\frac{(\Delta x)^4}{360}u_{xxxxxxt}+\mathcal{O}((\Delta x)^6)\right](x_j,t_{m+\frac{1}{2}})\\
&\quad+\frac{(\Delta t)^2}{8}\left[\frac{1}{3}u_{ttt}-u_{xxtt}-\frac{(\Delta x)^2}{12}u_{xxxxtt}-\frac{(\Delta x)^4}{360}u_{xxxxxxtt}+\mathcal{O}((\Delta x)^6)\right](x_j,t_{m+\frac{1}{2}})+\mathcal{O}((\Delta t)^3).
\end{aligned}$$

Finally, using that $u_{xx} = u_t$ as u is a solution to the heat equation, we arrive at the following final expansion for the consistency error:

$$\begin{aligned} T_j^m = &\left[-\frac{(\Delta x)^2}{12}u_{tt} - \frac{(\Delta x)^4}{360}u_{ttt} + \mathcal{O}((\Delta x)^6)\right](x_j, t_{m+\frac{1}{2}}) \\ &+ (2\theta - 1)\frac{\Delta t}{2}\left[-u_{tt} - \frac{(\Delta x)^2}{12}u_{ttt} - \frac{(\Delta x)^4}{360}u_{tttt} + \mathcal{O}((\Delta x)^6)\right](x_j, t_{m+\frac{1}{2}}) \\ &+ \frac{(\Delta t)^2}{8}\left[-\frac{2}{3}u_{ttt} - \frac{(\Delta x)^2}{12}u_{tttt} - \frac{(\Delta x)^4}{360}u_{ttttt} + \mathcal{O}((\Delta x)^6)\right](x_j, t_{m+\frac{1}{2}}) + \mathcal{O}((\Delta t)^3). \end{aligned}$$

Hence, we see that

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

Thus, in particular, the explicit and implicit Euler schemes have consistency error $T_j^m = \mathcal{O}((\Delta x)^2 + \Delta t)$ while the Crank–Nicolson scheme has consistency error $T_j^m = \mathcal{O}((\Delta x)^2 + (\Delta t)^2)$.

8.3 Practical stability of FD schemes

In order to be able to replicate the stability property (135) at the discrete level, we require an appropriate notion of stability. We shall say that a FD scheme for the heat equation is **(practically) stable in the ℓ^2 norm**, if

$$\|U^m\|_{\ell^2} \leq \|U^0\|_{\ell^2} \qquad \forall m \in \{1, \ldots, M\},$$

where

$$\|U^m\|_{\ell^2} := \sqrt{\Delta x \sum_{j=-\infty}^{\infty} |U_j^m|^2}.$$

We shall use the semidiscrete Fourier transform, defined below, to explore the stability of the FD 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 17 *The semidiscrete Fourier transform of a function U defined on the infinite mesh with mesh-points $x_j = j\Delta x$, $j \in \mathbb{Z}$, is defined by*

$$\hat{U}(k) := \Delta x \sum_{j=-\infty}^{\infty} U_j\, e^{-ikx_j}, \qquad k \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right].$$

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

Definition 18 *Let $\hat{U}$ be defined on the interval $[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}]$. The inverse semidiscrete Fourier transform of $\hat{U}$ is defined by*

$$U_j := \frac{1}{2\pi}\int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} \hat{U}(k)\, e^{ikx_j}\,\mathrm{d}k, \qquad j \in \mathbb{Z},$$

where $x_j = j\Delta x$ for $j \in \mathbb{Z}$.

We then have the following discrete Parseval's identity. The proof is very similar to the proof of Lemma 8 and left as an exercise.

Lemma 9 (Discrete Parseval's identity) *Let*

$$\|U\|_{\ell^2} := \sqrt{\Delta x \sum_{j=-\infty}^{\infty} |U_j|^2} \quad \textit{and} \quad \|\hat{U}\|_{L^2((-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}))} := \sqrt{\int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} |\hat{U}(k)|^2 \,\mathrm{d}k}.$$

If $\|U\|_{\ell^2}$ *is finite, then so is* $\|\hat{U}\|_{L^2((-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}))}$, *and*

$$\frac{1}{\sqrt{2\pi}}\|\hat{U}\|_{L^2((-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}))} = \|U\|_{\ell^2}.$$

With all technical prerequisites in place, we are now ready to discuss the stability of the various FD 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 $\mu := \frac{\Delta t}{(\Delta x)^2} \leq 1$), while the implicit Euler scheme will be shown to be unconditionally practically stable.

8.3.1 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 (130). By inserting

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

into the explicit Euler scheme we deduce that

$$\begin{aligned}\frac{1}{2\pi}\int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} e^{ikj\Delta x}\,\frac{\hat{U}^{m+1}(k)-\hat{U}^m(k)}{\Delta t}\,\mathrm{d}k &= \frac{1}{2\pi}\int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} \frac{e^{ik(j+1)\Delta x}-2e^{ikj\Delta x}+e^{ik(j-1)\Delta x}}{(\Delta x)^2}\,\hat{U}^m(k)\,\mathrm{d}k \\ &= \frac{1}{2\pi}\int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} e^{ikj\Delta x}\,\frac{e^{ik\Delta x}-2+e^{-ik\Delta x}}{(\Delta x)^2}\,\hat{U}^m(k)\,\mathrm{d}k.\end{aligned}$$

By comparing the left-hand side with the right-hand side we deduce that the two integrands are identically equal,[12] and therefore

$$\hat{U}^{m+1}(k) = \hat{U}^m(k) + \mu(e^{ik\Delta x}-2+e^{-ik\Delta x})\hat{U}^m(k), \quad \text{where} \quad \mu := \frac{\Delta t}{(\Delta x)^2}$$

for all **wave numbers** $k \in \left[-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}\right]$. The number μ is called the **CFL number** (after Richard Courant, Kurt Friedrichs, and Hans Lewy, who first performed an analysis of this kind).[13] We thus deduce that

$$\hat{U}^{m+1}(k) = \lambda(k)\hat{U}^m(k), \quad \text{where} \quad \lambda(k) := 1+\mu(e^{ik\Delta x}-2+e^{-ik\Delta x}).$$

We call $\lambda(k)$ the **amplification factor**. By the discrete Parseval identity (Lemma 9) we have that

$$\begin{aligned}\|U^{m+1}\|_{\ell^2} &= \frac{1}{\sqrt{2\pi}}\|\hat{U}^{m+1}\|_{L^2((-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}))} = \frac{1}{\sqrt{2\pi}}\,\|\lambda\hat{U}^m\|_{L^2((-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}))} \\ &\leq \frac{1}{\sqrt{2\pi}}\max_{k\in\left[-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}\right]}|\lambda(k)|\;\|\hat{U}^m\|_{L^2((-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}))} = \max_{k\in\left[-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}\right]}|\lambda(k)|\;\|U^m\|_{\ell^2}.\end{aligned}$$

[12] This is a consequence of the fact that the semidiscrete Fourier transform and its inverse are injective mappings.

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

In order to mimic the bound (135) we would like to ensure that

$$\|U^{m+1}\|_{\ell^2} \leq \|U^m\|_{\ell^2} \qquad \forall m \in \{0, 1, \ldots, M-1\}.$$

Thus we demand that $\max_{k \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right]} |\lambda(k)| \leq 1$, i.e.,

$$\max_{k \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right]} \left|1 + \mu(e^{ik\Delta x} - 2 + e^{-ik\Delta x})\right| \leq 1.$$

Using Euler's formula $e^{i\varphi} = \cos(\varphi) + i\sin(\varphi)$ and the trigonometric identity $1 - \cos(\varphi) = 2\sin^2(\frac{\varphi}{2})$ we can restate this as follows:

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

This holds iff $\mu \leq \frac{1}{2}$. Thus we have shown the following result.

Theorem 24 (Practical stability of explicit Euler) *Suppose that U_j^m is the solution of the explicit Euler scheme*

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

$$U_j^0 := u_0(x_j), \qquad j \in \mathbb{Z},$$

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

$$\|U^m\|_{\ell^2} \leq \|U^0\|_{\ell^2} \qquad \forall m \in \{1, \ldots, M\}. \tag{137}$$

In other words the explicit Euler scheme is **conditionally practically stable**, the condition for stability being that $\mu = \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2}$. One can also show that if $\mu > \frac{1}{2}$, then (137) will fail. In other words, once Δx has been chosen, one must choose Δt so that $\frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2}$ in order to ensure that (137) holds.

8.3.2 Stability analysis of the implicit Euler scheme

We shall now perform a similar analysis for the implicit Euler scheme for the heat equation (130), which is defined as follows:

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

$$U_j^0 := u_0(x_j), \qquad j \in \mathbb{Z}.$$

Equivalently,

$$U_j^{m+1} - \mu(U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}) = U_j^m, \quad j \in \mathbb{Z}, \quad m \in \{0, 1, \ldots, M-1\},$$

$$U_j^0 := u_0(x_j), \qquad j \in \mathbb{Z},$$

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, $\max_{k \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right]} |\lambda(k)| \leq 1$ for all values of $\mu = \frac{\Delta t}{(\Delta x)^2}$. Thus we have the following result.

Theorem 25 (Practical stability of implicit Euler) *Suppose that U_j^m is the solution of the implicit Euler scheme*

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

$$U_j^0:=u_0(x_j),\qquad j\in\mathbb{Z}.$$

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

$$\|U^m\|_{\ell^2}\le\|U^0\|_{\ell^2}\qquad\forall m\in\{1,\dots,M\}. \tag{138}$$

In other words, the implicit Euler scheme is **unconditionally practically stable**, meaning that (138) holds without any restrictions on Δx and Δt.

8.4 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*.

Definition 19 *We say that a FD scheme for the heat equation on the time interval $[0,T]$ is **von Neumann stable** in the ℓ^2 norm, if there exists a constant $C=C(T)>0$ such that*

$$\|U^m\|_{\ell^2}\le C\|U^0\|_{\ell^2}\qquad\forall m\in\left\{1,\dots,M=\frac{T}{\Delta t}\right\},$$

where

$$\|U^m\|_{\ell^2}:=\sqrt{\Delta x\sum_{j=-\infty}^{\infty}|U_j^m|^2}.$$

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 $m=1,2,\dots$, von Neumann stability only makes sense on finite time intervals $[0,T]$ and for the limited range of $0\le m\le\frac{T}{\Delta t}$.

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

Lemma 10 *Suppose that the semidiscrete Fourier transform of the solution $\{U_j^m\}_{j=-\infty}^{\infty}$, $m\in\{0,1,\dots,\frac{T}{\Delta t}\}$, of a FD scheme for the heat equation satisfies*

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

and there exists a constant $C_0\ge0$ such that

$$|\lambda(k)|\le1+C_0\Delta t\qquad\forall\,k\in\left[-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}\right].$$

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

$$\begin{aligned}\|U^{m+1}\|_{\ell^2}&=\frac{1}{\sqrt{2\pi}}\|\hat{U}^{m+1}\|_{L^2((-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}))}=\frac{1}{\sqrt{2\pi}}\,\|\lambda\hat{U}^m\|_{L^2((-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}))}\\&\le\frac{1}{\sqrt{2\pi}}\max_{k\in[-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}]}|\lambda(k)|\;\|\hat{U}^m\|_{L^2((-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}))}=\max_{k\in[-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}]}|\lambda(k)|\;\|U^m\|_{\ell^2}.\end{aligned}$$

Hence, $\|U^{m+1}\|_{\ell^2} \leq (1 + C_0\Delta t)\|U^m\|_{\ell^2}$ for all $m \in \{0, 1, \ldots, M-1\}$. Therefore,

$$\|U^m\|_{\ell^2} \leq (1 + C_0\Delta t)^m \|U^0\|_{\ell^2} \qquad \forall m \in \{1, \ldots, M\}.$$

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

$$\|U^m\|_{\ell^2} \leq e^{C_0 T} \|U^0\|_{\ell^2} \qquad \forall m \in \{1, \ldots, M\},$$

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

8.5 Initial-boundary-value problems for parabolic problems

When a parabolic PDE is considered on a bounded spatial domain, one needs to impose boundary conditions (b.c.) at the boundary of the domain. Here we shall concentrate on the simplest case, when a Dirichlet b.c. 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 (IBVP) for the heat equation:

$$\partial_t u = \partial_{xx}^2 u, \qquad x \in (a, b), \quad t \in (0, T],$$

subject to the initial condition (i.c.)

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

and the following Dirichlet b.c. at $x = a$ and $x = b$:

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

We assume that the b.c. is compatible with the i.c. in the sense that $A(0) = u_0(a)$ and $B(0) = u_0(b)$.

Remark 12 *We note in passing that the Neumann IBVP for the heat equation is*

$$\partial_t u = \partial_{xx}^2 u, \qquad x \in (a, b), \quad t \in (0, T],$$

subject to the i.c.

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

and the Neumann b.c.

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

An example of a mixed Dirichlet–Neumann IBVP for the heat equation is

$$\partial_t u = \partial_{xx}^2 u, \qquad x \in (a, b), \quad t \in (0, T],$$

subject to the i.c.

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

and the mixed Dirichlet–Neumann b.c.

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

8.5.1 θ-scheme for the Dirichlet IBVP

Our aim in this section is to construct a numerical approximation of the Dirichlet IBVP based on the θ-scheme. Let $\Delta x := \frac{b-a}{J}$ and $\Delta t := \frac{T}{M}$, and define

$$x_j := a + j\Delta x, \quad j \in \{0, 1, \ldots, J\}, \qquad t_m := m\Delta t, \quad m \in \{0, 1, \ldots, M\}.$$

We approximate the Dirichlet IBVP 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 $j \in \{1, \ldots, J-1\}$, $m \in \{0, 1, \ldots, M-1\}$,

$$U_j^0 := u_0(x_j), \qquad j \in \{0, 1, \ldots, J\},$$

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

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

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

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

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

The matrix form of this system of linear equations is therefore the following. We consider the symmetric tridiagonal matrix

$$\mathcal{A} := \begin{bmatrix} -2 & 1 & & & 0 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 0 & & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{(J-1)\times(J-1)}.$$

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

$$(\mathcal{I} - \theta\mu\mathcal{A})\mathbf{U}^{m+1} = (\mathcal{I} + (1-\theta)\mu\mathcal{A})\mathbf{U}^m + \theta\mu\mathbf{F}^{m+1} + (1-\theta)\mu\mathbf{F}^m, \qquad m \in \{0, 1, \ldots, M-1\},$$

where $\mathbf{U}^{m+1} := (U_1^{m+1}, \ldots, U_{J-1}^{m+1})^{\mathrm{T}} \in \mathbb{R}^{J-1}$, $\mathbf{U}^m := (U_1^m, \ldots, U_{J-1}^m)^{\mathrm{T}} \in \mathbb{R}^{J-1}$ and

$$\mathbf{F}^{m+1} := (A(t_{m+1}), 0, \ldots, 0, B(t_{m+1}))^{\mathrm{T}} \in \mathbb{R}^{J-1}, \qquad \mathbf{F}^m := (A(t_m), 0, \ldots, 0, B(t_m))^{\mathrm{T}} \in \mathbb{R}^{J-1}.$$

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

8.5.2 The discrete maximum principle

We now try to prove a bound for the θ-scheme in the discrete maximum norm, analogous to (133) satisfied by the solution of the heat equation. Recall that the CFL number is defined by $\mu := \frac{\Delta t}{(\Delta x)^2}$.

Theorem 26 (Discrete maximum principle for the θ-scheme) *The θ-scheme for the Dirichlet IBVP for the heat equation, with $\theta \in [0,1]$ and $(1-\theta)\mu \leq \frac{1}{2}$, yields a sequence of numerical approximations $\{U_j^m\}_{0\leq j\leq J; 0\leq m\leq M}$ satisfying*

$$\min\{U^0_{\min}, U_0^{\min}, U_J^{\min}\} \leq U_j^m \leq \max\{U^0_{\max}, U_0^{\max}, U_J^{\max}\} \qquad \forall j \in \{0,1,\ldots,J\},\, m \in \{0,1,\ldots,M\}$$

where $U^0_{\min} := \min\{U_0^0, U_1^0, \ldots, U_J^0\}$, $U_0^{\min} := \min\{U_0^0, U_0^1, \ldots, U_0^M\}$, $U_J^{\min} := \min\{U_J^0, U_J^1, \ldots, U_J^M\}$, and $U^0_{\max} := \max\{U_0^0, U_1^0, \ldots, U_J^0\}$, $U_0^{\max} := \max\{U_0^0, U_0^1, \ldots, U_0^M\}$, $U_J^{\max} := \max\{U_J^0, U_J^1, \ldots, U_J^M\}$.

Proof: We rewrite the θ-scheme as

$$(1+2\theta\mu)\, U_j^{m+1} = \theta\mu\left(U_{j+1}^{m+1} + U_{j-1}^{m+1}\right) + (1-\theta)\mu\left(U_{j+1}^m + U_{j-1}^m\right) + (1-2(1-\theta)\mu)\, U_j^m, \tag{139}$$

and note that, by hypothesis, $\theta\mu \geq 0$, $(1-\theta)\mu \geq 0$, and $1-2(1-\theta)\mu \geq 0$. Suppose that U attains its maximum value at an interior mesh-point (x_{j_0}, t_{m_0+1}) for some $j_0 \in \{1,\ldots,J-1\}$, $m_0 \in \{0,\ldots,M-1\}$. We define $U^\star := \max\{U_{j_0+1}^{m_0+1}, U_{j_0-1}^{m_0+1}, U_{j_0+1}^{m_0}, U_{j_0-1}^{m_0}, U_{j_0}^{m_0}\}$. Then,

$$(1+2\theta\mu)\, U_{j_0}^{m_0+1} \leq 2\theta\mu U^\star + 2(1-\theta)\mu U^\star + (1-2(1-\theta)\mu)\, U^\star = (1+2\theta\mu)\, U^\star, \tag{140}$$

and therefore, $U_{j_0}^{m_0+1} \leq U^\star$. We also have $U^\star \leq U_{j_0}^{m_0+1}$ as $U_{j_0}^{m_0+1}$ is assumed to be the overall maximum value. Hence, $U_{j_0}^{m_0+1} = U^\star$. Thus the maximum value is also attained at the points neighbouring (x_{j_0}, t_{m_0+1}) present in the scheme.[14]

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. □

We have just proved that when $\mu(1-\theta) \leq \frac{1}{2}$ the θ-scheme satisfies the discrete maximum principle. Clearly, this condition is more demanding than the ℓ^2-stability condition which requires $\mu(1-2\theta) \leq \frac{1}{2}$ when $\theta \in [0, \frac{1}{2})$ (see problem sheets). For example, the Crank–Nicolson scheme ($\theta = \frac{1}{2}$) is unconditionally stable in the ℓ^2 norm, yet it only satisfies the discrete maximum principle when $\mu = \frac{\Delta t}{(\Delta x)^2} \leq 1$. 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 $\mu(1-\theta) \leq \frac{1}{2}$. 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 $\mu \leq \frac{1}{2}$), for $\theta \in (0, \frac{1}{2}]$ the inequality $\mu(1-\theta) \leq \frac{1}{2}$ is more restrictive than $\mu(1-2\theta) \leq \frac{1}{2}$.

[14] To see that the maximum value $U_{j_0}^{m_0+1} = U^*$ is attained at *each* of the points neighbouring (x_{j_0}, t_{m_0+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 (139); (b) if $\theta = 1$ then U_{j+1}^m and U_{j-1}^m are absent from the right-hand side of (139); (c) if $2(1-\theta)\mu = 1$, then U_j^m is absent from the right-hand side of (139), 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 (139). 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 $U_{j_0+1}^{m_0+1}$, $U_{j_0-1}^{m_0+1}$, $U_{j_0+1}^{m_0}$, $U_{j_0-1}^{m_0}$, $U_{j_0}^{m_0}$ were strictly smaller than $U_{j_0}^{m_0+1} = U^*$, then, by returning to the transition from (139) to (140), we would deduce (140) from (139), but now with the $\leq$ symbol in (140) replaced by $<$, which would then imply that $U_{j_0}^{m_0+1} < U^*$. This would, however, contradict the equality $U_{j_0}^{m_0+1} = U^*$ which we have already proved.

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

We close our discussion of FD schemes for the heat equation (130) in one space-dimension with the convergence analysis of the θ-scheme for the Dirichlet IBVP. We begin by rewriting the scheme as follows:

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

for $j \in \{1,\ldots,J-1\}$ and $m \in \{0,1,\ldots,M-1\}$. The scheme is considered subject to the i.c.

$$U_j^0 := u_0(x_j), \qquad j \in \{0,1,\ldots,J\},$$

and the b.c.

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

The **consistency error** for the θ-scheme is given 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

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

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

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

Note that $e_0^{m+1} = e_J^{m+1} = e_j^0 = 0$ for all $j \in \{0,1,\ldots,J\}$ and $m \in \{0,1,\ldots,M-1\}$, and we have that

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

We define $E^m := \max\{|e_0^m|,|e_1^m|,\ldots,|e_J^m|\}$ and $T^m := \max\{|T_0^m|,|T_1^m|,\ldots,|T_J^m|\}$. As, by hypothesis, $\theta\mu \geq 0$, $(1-\theta)\mu \geq 0$, and $1-2(1-\theta)\mu \geq 0$, we find that

$$\begin{aligned}(1+2\theta\mu)E^{m+1} &\leq 2\theta\mu E^{m+1} + 2(1-\theta)\mu E^m + (1-2(1-\theta)\mu)\,E^m + (\Delta t)T^m\\ &= 2\theta\mu E^{m+1} + E^m + (\Delta t)T^m\end{aligned}$$

for any $m \in \{0,1,\ldots,M-1\}$. Hence, $E^{m+1} \leq E^m + (\Delta t)T^m$ for any $m \in \{0,1,\ldots,M-1\}$. As $E^0 = 0$, we find that

$$\begin{aligned}E^m &\leq E^{m-1} + (\Delta t)T^{m-1}\\ &\leq E^{m-2} + (\Delta t)T^{m-2} + (\Delta t)T^{m-1}\\ &\;\;\vdots\\ &\leq (\Delta t)\left(T^0 + T^1 + \cdots + T^{m-1}\right) \leq m(\Delta t)\max_{i\in\{0,\ldots,m-1\}} T^i \leq T\max_{i\in\{0,\ldots,m-1\}} T^i\end{aligned}$$

for any $m \in \{1,\ldots,M\}$. This implies that

$$\max_{m\in\{0,1,\ldots,M\}}\;\max_{j\in\{0,1,\ldots,J\}} |e_j^m| \leq T\max_{i\in\{0,1,\ldots,M-1\}} T^i.$$

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

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

It therefore follows that

$$\max_{m\in\{0,1,\ldots,M\}} \max_{j\in\{0,1,\ldots,J\}} |e_j^m| = \begin{cases} \mathcal{O}\left((\Delta x)^2 + (\Delta t)^2\right) & \text{if } \theta = 1/2, \\ \mathcal{O}\left((\Delta x)^2 + \Delta t\right) & \text{if } \theta \neq 1/2. \end{cases}$$

The results developed in this section can be extended to multidimensional axiparallel domains, such as rectangular or L-shaped domains in two space-dimensions whose edges are parallel with the coordinate 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, FD 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 to e.g., R. LeVeque, Finite Difference Methods for Ordinary and Partial Differential Equations. SIAM, 2007; or to K.W. Morton and D.F. Mayers, Numerical Solution of Partial Differential Equations: An Introduction, 2nd Edition, CUP, 2005.

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

8.6 FD approximation of parabolic equations in two space-dimensions

On an open rectangle $\Omega := (a, b) \times (c, d)$ in $\mathbb{R}^2$, we consider the heat equation

$$\partial_t u = \partial_{xx}^2 u + \partial_{yy}^2 u, \qquad (x, y) \in \Omega, \;\; t \in (0, T],$$

subject to the i.c.

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

and the Dirichlet b.c.

$$u(x, y, t) = B(x, y, t), \qquad (x, y) \in \partial\Omega, \;\; t \in (0, T].$$

We begin by considering the explicit Euler FD approximation of this problem.

8.6.1 The explicit and implicit Euler schemes

Let

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

Let $\Delta x := \frac{b-a}{J_x}$, $\Delta y := \frac{d-c}{J_y}$, $\Delta t := \frac{T}{M}$, and define

$$\begin{aligned} x_i &:= a + i\Delta x, & i &\in \{0, 1, \ldots, J_x\}, \\ y_j &:= c + j\Delta y, & j &\in \{0, 1, \ldots, J_y\}, \\ t_m &:= m\Delta t, & m &\in \{0, 1, \ldots, M\}. \end{aligned}$$

The explicit Euler FD approximation of the heat equation on the space-time domain $\overline{\Omega} \times [0, T]$ is

$$\frac{U_{i,j}^{m+1} - U_{i,j}^m}{\Delta t} = \frac{\delta_x^2 U_{i,j}^m}{(\Delta x)^2} + \frac{\delta_y^2 U_{i,j}^m}{(\Delta y)^2}, \tag{141}$$

for $i \in \{1, \ldots, J_x - 1\}$, $j \in \{1, \ldots, J_y - 1\}$, $m \in \{0, 1, \ldots, M-1\}$, subject to the i.c.

$$U_{i,j}^0 := u_0(x_i, y_j), \quad i \in \{0, 1, \ldots, J_x\}, \;\; j \in \{0, 1, \ldots, J_y\},$$

and the b.c.

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

The implicit Euler FD approximation of the heat equation on the space-time domain $\overline{\Omega} \times [0, T]$ is defined analogously, with (141) replaced by

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

8.6.2 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 $\Delta x := \frac{b-a}{J_x}$, $\Delta y := \frac{d-c}{J_y}$, and $\Delta t := \frac{T}{M}$. For $\theta \in [0, 1]$, consider the FD 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 $i \in \{1, \ldots, J_x - 1\}$, $j \in \{1, \ldots, J_y - 1\}$, $m \in \{0, 1, \ldots, M-1\}$, subject to the i.c.

$$U_{i,j}^0 := u_0(x_i, y_j), \quad i \in \{0, 1, \ldots, J_x\}, \;\; j \in \{0, 1, \ldots, J_y\},$$

and the b.c.

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

Practical stability

The practical stability of the θ-scheme (in the absence of a b.c. now, i.e., for the pure IVP rather than the IBVP) in the ℓ^2 norm is easily assessed by inserting

$$U_{i,j}^m = \frac{1}{(2\pi)^2} \int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} \int_{-\frac{\pi}{\Delta y}}^{\frac{\pi}{\Delta y}} e^{\imath(k_x i \Delta x + k_y j \Delta y)} \hat{U}^m(k_x, k_y) \, \mathrm{d}k_y \, \mathrm{d}k_x.$$

(Here, $\imath$ denotes the complex number, and i the index from $U_{i,j}^m$). We deduce that

$$\begin{aligned}\frac{\hat{U}^{m+1}(k_x, k_y) - \hat{U}^m(k_x, k_y)}{\Delta t} = (1-\theta)&\left(\frac{-4\sin^2\left(\frac{k_x \Delta x}{2}\right)}{(\Delta x)^2} + \frac{-4\sin^2\left(\frac{k_y \Delta y}{2}\right)}{(\Delta y)^2}\right)\hat{U}^m(k_x, k_y) \\ &+ \theta\left(\frac{-4\sin^2\left(\frac{k_x \Delta x}{2}\right)}{(\Delta x)^2} + \frac{-4\sin^2\left(\frac{k_y \Delta y}{2}\right)}{(\Delta y)^2}\right)\hat{U}^{m+1}(k_x, k_y)\end{aligned}$$

for all $(k_x, k_y) \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right] \times \left[-\frac{\pi}{\Delta y}, \frac{\pi}{\Delta y}\right]$. Writing $\mu_x := \frac{\Delta t}{(\Delta x)^2}$ and $\mu_y := \frac{\Delta t}{(\Delta y)^2}$, we find that

$$\hat{U}^{m+1}(k_x, k_y) = \lambda(k_x, k_y)\hat{U}^m(k_x, k_y) \qquad \forall (k_x, k_y) \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right] \times \left[-\frac{\pi}{\Delta y}, \frac{\pi}{\Delta y}\right],$$

where the amplification factor is given by

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

for $(k_x,k_y)\in\left[-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}\right]\times\left[-\frac{\pi}{\Delta y},\frac{\pi}{\Delta y}\right]$. For practical stability in the ℓ^2 norm, we require that

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

Note that $\lambda(k_x,k_y)\le 1$ without any restriction on μ_x,μ_y. Hence, the scheme is practically stable iff $\lambda(k_x,k_y)\ge -1\ \forall (k_x,k_y)\in\left[-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}\right]\times\left[-\frac{\pi}{\Delta y},\frac{\pi}{\Delta y}\right]$, which holds iff

$$(1-2\theta)\left[\mu_x \sin^2\left(\frac{k_x\Delta x}{2}\right)+\mu_y \sin^2\left(\frac{k_y\Delta y}{2}\right)\right]\le\frac{1}{2} \qquad \forall (k_x,k_y)\in\left[-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x}\right]\times\left[-\frac{\pi}{\Delta y},\frac{\pi}{\Delta y}\right],$$

i.e., iff

$$(1-2\theta)(\mu_x+\mu_y)\le\frac{1}{2}.$$

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:

$$\mu_x+\mu_y=\Delta t\left(\frac{1}{(\Delta x)^2}+\frac{1}{(\Delta y)^2}\right)\le\frac{1}{2}.$$

Discrete maximum principle

Under a suitable condition the θ-scheme for the IBVP also satisfies a discrete maximum principle. To see this, we rewrite the θ-scheme as

$$\begin{aligned}(1+2\theta(\mu_x+\mu_y))U^{m+1}_{i,j} &= (1-2(1-\theta)(\mu_x+\mu_y))U^m_{i,j}\\ &\quad+(1-\theta)\mu_x(U^m_{i+1,j}+U^m_{i-1,j})+(1-\theta)\mu_y(U^m_{i,j+1}+U^m_{i,j-1})\\ &\quad+\theta\mu_x(U^{m+1}_{i+1,j}+U^{m+1}_{i-1,j})+\theta\mu_y(U^{m+1}_{i,j+1}+U^{m+1}_{i,j-1}),\end{aligned}$$

for $i\in\{1,\dots,J_x-1\}$, $j\in\{1,\dots,J_y-1\}$, $m\in\{0,1,\dots,M-1\}$, subject to the i.c.

$$U^0_{i,j}:=u_0(x_i,y_j),\quad i\in\{0,1,\dots,J_x\},\ \ j\in\{0,1,\dots,J_y\},$$

and the b.c.

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

Theorem 27 *Suppose that*

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

Then, the θ-scheme satisfies the following discrete maximum principle:

$$\min\{U^0_{\min},U^{\min}_{\partial}\}\le U^m_{i,j}\le\max\{U^0_{\max},U^{\max}_{\partial}\}$$

for all $i\in\{0,1,\dots,J_x\}$, $j\in\{0,1,\dots,J_y\}$, $m\in\{0,1,\dots,M\}$, where

$$U^0_{\min}:=\min\{U^0_{i,j}\,|\,i\in\{0,1,\dots,J_x\},j\in\{0,1,\dots,J_y\}\},\quad U^{\min}_{\partial}:=\min\{U^m_{i,j}\,|\,(x_i,y_j)\in\partial\Omega,m\in\{0,1,\dots,M\}\},$$
$$U^0_{\max}:=\max\{U^0_{i,j}\,|\,i\in\{0,1,\dots,J_x\},j\in\{0,1,\dots,J_y\}\},\quad U^{\max}_{\partial}:=\max\{U^m_{i,j}\,|\,(x_i,y_j)\in\partial\Omega,m\in\{0,1,\dots,M\}\}.$$

PROOF: The proof proceeds by a straightforward modification of the proof of the discrete maximum principle for the θ-scheme in one space-dimension. $\square$

In summary, then, for

$$(1-\theta)(\mu_x+\mu_y) \leq \frac{1}{2}$$

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

$$(1-2\theta)(\mu_x+\mu_y) \leq \frac{1}{2}.$$

For example, the Crank–Nicolson scheme ($\theta = \frac{1}{2}$) is unconditionally practically stable in the ℓ^2 norm, but for the discrete maximum principle to hold we had to assume that

$$\mu_x+\mu_y = \frac{\Delta t}{(\Delta x)^2} + \frac{\Delta t}{(\Delta y)^2} \leq 1.$$

Error analysis

We close our discussion by returning to the θ-scheme for the IBVP, 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 $i \in \{1,\ldots,J_x-1\}$, $j \in \{1,\ldots,J_y-1\}$, $m \in \{0,1,\ldots,M-1\}$, subject to the i.c.

$$U_{i,j}^0 := u_0(x_i,y_j), \quad i \in \{0,1,\ldots,J_x\}, \;\; j \in \{0,1,\ldots,J_y\},$$

and the b.c.

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

Suppose further that

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

The consistency error of the θ-scheme is defined as

$$T_{i,j}^m := \frac{u_{i,j}^{m+1}-u_{i,j}^m}{\Delta t} - (1-\theta)\left(\frac{\delta_x^2 u_{i,j}^m}{(\Delta x)^2} + \frac{\delta_y^2 u_{i,j}^m}{(\Delta y)^2}\right) - \theta\left(\frac{\delta_x^2 u_{i,j}^{m+1}}{(\Delta x)^2} + \frac{\delta_y^2 u_{i,j}^{m+1}}{(\Delta y)^2}\right),$$

where we write $u_{i,j}^m := u(x_i,y_j,t_m)$. By performing Taylor series expansions, one can deduce that

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

It follows from the definition of the consistency error $T_{i,j}^m$ for the θ-scheme that

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

for $i \in \{1, \dots, J_x - 1\}$, $j \in \{1, \dots, J_y - 1\}$, $m \in \{0, 1, \dots, M-1\}$. We define the **global error** as

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

Then, $e_{i,j}^0 = 0$ for any $i \in \{0, 1, \dots, J_x\}$, $j \in \{0, 1, \dots, J_y\}$, and also $e_{i,j}^m = 0$ for any $(x_i, y_j) \in \partial\Omega$, $m \in \{1, \dots, M\}$. Further,

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

Let us define $E^m := \max_{i,j} |e_{i,j}^m|$ and $T^m := \max_{i,j} |T_{i,j}^m|$. As by hypothesis $1 - 2(1-\theta)(\mu_x + \mu_y) \geq 0$, we have that

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

Hence,

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

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

$$\begin{aligned}E^m &\leq E^{m-1} + (\Delta t)T^{m-1} \\ &\leq E^{m-2} + (\Delta t)T^{m-2} + (\Delta t)T^{m-1} \\ &\vdots \\ &\leq (\Delta t)\left(T^0 + T^1 + \cdots + T^{m-1}\right) \leq m(\Delta t) \max_{l \in \{0,\dots,m-1\}} T^l \leq T \max_{l \in \{0,\dots,m-1\}} T^l\end{aligned}$$

for any $m \in \{1, \dots, M\}$. This implies that

$$\max_{m \in \{0,1,\dots,M\}} \max_{i \in \{0,1,\dots,J_x\}, j \in \{0,1,\dots,J_y\}} |e_{i,j}^m| \leq T \max_{l \in \{0,1,\dots,M-1\}} T^l.$$

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

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

We conclude that, under the assumption $(1-\theta)(\mu_x + \mu_y) \leq \frac{1}{2}$, there holds

$$\max_{m \in \{0,1,\dots,M\}} \max_{i \in \{0,1,\dots,J_x\}, j \in \{0,1,\dots,J_y\}} |e_{i,j}^m| = \begin{cases} \mathcal{O}\left((\Delta x)^2 + (\Delta y)^2 + (\Delta t)^2\right) & \text{if } \theta = 1/2, \\ \mathcal{O}\left((\Delta x)^2 + (\Delta y)^2 + \Delta t\right) & \text{if } \theta \neq 1/2. \end{cases}$$