B6.1 Numerical Solution of Differential Equations I

Endre Süli
Mathematical Institute, University of Oxford

November 17, 2020

Contents

Preface. The purpose of these lecture notes is to provide an introduction to computational methods for the approximate solution of ordinary differential equations (ODEs) and parabolic partial differential equations (PDEs). Only minimal prerequisites in differential and integral calculus, differential equation theory, complex analysis and linear algebra are assumed. The notes focus on the construction of numerical algorithms for ODEs and parabolic PDEs, and the mathematical analysis of their behaviour.

The notes begin with a study of well-posedness of initial-value problems for a first-order differential equations and systems of such equations. The basic ideas of discretisation and error analysis are then introduced in the case of one-step methods. This is followed by an extension of the concepts of stability and accuracy to linear multi-step methods, including a brief excursion into numerical methods for stiff systems of ODEs and symplectic methods. The final section is devoted to an overview of classical algorithms for the numerical solution of initial-boundary-value problems for the simplest parabolic equation: the linear heat equation in one space dimension.

Syllabus. Approximation of initial-value problems for ordinary differential equations: one-step methods including the explicit and implicit Euler methods, the trapezium rule method, and Runge–Kutta methods. Linear multi-step methods: consistency, zero-stability and convergence; absolute stability. Stiffness. Error control and adaptive algorithms. Symplectic methods.

Numerical solution of initial-boundary-value problems for parabolic partial differential equations: explicit and implicit methods; accuracy, stability and convergence, use of Fourier methods for analysis.

Reading List:

[1] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations.* Cambridge University Press, 2nd ed., 2009. ISBN 978-0-521-73490-5 [Chapters 1–6, 16].

[2] R. LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations.* SIAM, 2007. ISBN 978-0-898716-29-0 [Chapters 5–9].

[3] E. Süli and D.F. Mayers, *An Introduction to Numerical Analysis.* Cambridge University Press, 2006. ISBN 0-521-00794-1 [Chapter 12].

Further Reading:

[1] E. Hairer, S.P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems.* Springer-Verlag, Berlin, 1987.

[2] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations.* Wiley, New York, 1962.

[3] H.B. Keller, *Numerical Methods for Two-point Boundary Value Problems.* SIAM, Philadelphia, 1976.

[4] J.D. Lambert, *Computational Methods in Ordinary Differential Equations.* Wiley, Chichester, 1991.

[5] A.M. Stuart and A.R. Humphries, *Dynamical Systems and Numerical Analysis.* Cambridge University Press, Cambridge, 1996.

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

Note about the exercises: There will be 6 problem sheets and 6 classes, the first class being held in Week 3 of Michaelmas Term.

1 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 **Lecture 1**

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

for the real-valued function y of the real variable x, where $y' \equiv \mathrm{d}y/\mathrm{d}x$. In order to select a particular integral 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**: given two real numbers x_0 and y_0, 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**. 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 $[x_0, X_M]$ is then known at the initial time, 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 the function f, the existence and uniqueness of a solution to (1), (2) can be ensured: the result is encapsulated in the next theorem.

Theorem 1 (**Picard's Theorem**[2]) *Suppose that $f(\cdot, \cdot)$ is a continuous function of its arguments in a region U of the (x, y) plane which contains the rectangle*

$$\mathsf{R} = \{(x, y) \,:\, x_0 \le x \le X_M, \quad |y - y_0| \le Y_M\},$$

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

$$|f(x, y) - f(x, z)| \le L|y - z| \tag{3}$$

holds whenever (x, y) and (x, z) lie in the rectangle R. Finally, letting

$$M = \max\{|f(x, y)| \,:\, (x, y) \in \mathsf{R}\},$$

suppose that $M(X_M - x_0) \le Y_M$. Then, there exists a unique continuously differentiable function $x \mapsto y(x)$, defined on the closed interval $[x_0, X_M]$, which satisfies (1) and (2).

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, the interested reader is referred to any good textbook on the theory of ordinary differential equations (see, for example, P. J. Collins, *Differential and Integral Equations*, Oxford University Press, 2006). The essence of the proof is to consider the sequence of functions $\{y_n\}_{n=0}^{\infty}$, defined recursively through what is known as the *Picard Iteration*:

$$\begin{aligned} &y_0(x) \equiv y_0, \\ &y_n(x) = y_0 + \int_{x_0}^{x} f(\xi, y_{n-1}(\xi))\,\mathrm{d}\xi, \qquad n = 1, 2, \ldots, \end{aligned} \tag{4}$$

[1] Consider, for example, the initial-value problem $y' = y^{2/3}$, $y(0) = 0$; this has solutions: $y(x) \equiv 0$ and $y(x) = x^3/27$.
[2] Emile Picard (1856–1941)
[3] Rudolf Lipschitz (1832–1903)

and show, using the conditions of the theorem, that $\{y_n\}_{n=0}^{\infty}$, as a sequence of continuous functions, converges uniformly on the interval $[x_0, X_M]$ to a continuous function y defined on $[x_0, X_M]$ such that

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

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 an initial-value problem for a system of m differential equations of the form

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

where $\mathbf{y}_0 \in \mathbb{R}^m$ and $\mathbf{f} : [x_0, X_M] \times \mathbb{R}^m \to \mathbb{R}^m$. On introducing the Euclidean norm $\|\cdot\|$ on $\mathbb{R}^m$ by

$$\|v\| = \left(\sum_{i=1}^{m} |v_i|^2\right)^{1/2}, \qquad v \in \mathbb{R}^m,$$

we can state the following result.

Theorem 2 (Picard's Theorem) *Suppose that $\mathbf{f}(\cdot,\cdot)$ is a continuous function of its arguments in a region U of the $(x, \mathbf{y})$ space $\mathbf{R}^{1+m}$ which contains the parallelepiped*

$$\mathsf{R} = \{(x, \mathbf{y}) : x_0 \le x \le X_M, \quad \|\mathbf{y} - \mathbf{y}_0\| \le Y_M\},$$

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

$$\|\mathbf{f}(x, \mathbf{y}) - \mathbf{f}(x, \mathbf{z})\| \le L\|\mathbf{y} - \mathbf{z}\| \tag{6}$$

holds whenever $(x, \mathbf{y})$ and $(x, \mathbf{z})$ lie in R. Finally, letting

$$M = \max\{\|\mathbf{f}(x, \mathbf{y})\| : (x, \mathbf{y}) \in \mathsf{R}\},$$

suppose that $M(X_M - x_0) \le Y_M$. Then, there exists a unique continuously differentiable function $x \mapsto \mathbf{y}(x)$, defined on the closed interval $[x_0, X_M]$, which satisfies (5).

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

$$\left\| \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(x, \mathbf{y}) \right\| \le L \qquad \text{for all } (x, \mathbf{y}) \in \mathrm{int}(\mathsf{R}), \tag{7}$$

where $\partial \mathbf{f}/\partial \mathbf{y}$ denotes the $m \times m$ Jacobi matrix of $\mathbf{y} \in \mathbf{R}^m \mapsto \mathbf{f}(x, \mathbf{y}) \in \mathbf{R}^m$, and $\|\cdot\|$ is a matrix norm subordinate to 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 $\mathbf{f}(\mathbf{y}) = (|y_1|, \ldots, |y_m|)^{\mathrm{T}}$, with $x_0 = 0$ and $\mathbf{y}_0 = \mathbf{0}$, satisfies (6) but violates (7) because $\mathbf{y} \mapsto \mathbf{f}(\mathbf{y})$ is not differentiable at the point $\mathbf{y} = \mathbf{0}$.

As the counter-example in the footnote on page 1 indicates, the expression $|y - z|$ in (3) and $\|\mathbf{y} - \mathbf{z}\|$ in (6) cannot be replaced by expressions of the form $|y - z|^\alpha$ and $\|\mathbf{y} - \mathbf{z}\|^\alpha$, 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 *A solution $\mathbf{y} = \mathbf{v}(x)$ to (5) 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 $\mathbf{z}$ satisfying $\|\mathbf{v}(x_0) - \mathbf{z}\| < \delta$ the solution $\mathbf{y} = \mathbf{w}(x)$ to the differential equation $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$ satisfying the initial condition $\mathbf{w}(x_0) = \mathbf{z}$ is defined for all $x \in [x_0, X_M]$ and satisfies $\|\mathbf{v}(x) - \mathbf{w}(x)\| < \varepsilon$ for all x in $[x_0, X_M]$.*

*A solution 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**.*

Moreover, if

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

*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: Since

$$\mathbf{v}(x) = \mathbf{v}(x_0) + \int_{x_0}^{x} \mathbf{f}(\xi, \mathbf{v}(\xi))\, \mathrm{d}\xi$$

and

$$\mathbf{w}(x) = \mathbf{z} + \int_{x_0}^{x} \mathbf{f}(\xi, \mathbf{w}(\xi))\, \mathrm{d}\xi,$$

it follows that

$$\begin{aligned} \|\mathbf{v}(x) - \mathbf{w}(x)\| &\leq \|\mathbf{v}(x_0) - \mathbf{z}\| + \int_{x_0}^{x} \|\mathbf{f}(\xi, \mathbf{v}(\xi)) - \mathbf{f}(\xi, \mathbf{w}(\xi))\|\, \mathrm{d}\xi \\ &\leq \|\mathbf{v}(x_0) - \mathbf{z}\| + L \int_{x_0}^{x} \|\mathbf{v}(\xi) - \mathbf{w}(\xi)\|\, \mathrm{d}\xi. \end{aligned} \tag{8}$$

Now put $A(x) = \|\mathbf{v}(x) - \mathbf{w}(x)\|$ and $a = \|\mathbf{v}(x_0) - \mathbf{z}\|$; then, (8) can be written as

$$A(x) \leq a + L \int_{x_0}^{x} A(\xi)\, \mathrm{d}\xi, \qquad x_0 \leq x \leq X_M. \tag{9}$$

Multiplying (9) by $\exp(-Lx)$, we find that

$$\frac{\mathrm{d}}{\mathrm{d}x} \left[\mathrm{e}^{-Lx} \int_{x_0}^{x} A(\xi)\, \mathrm{d}\xi \right] \leq a \mathrm{e}^{-Lx}. \tag{10}$$

Integrating the inequality (10), we deduce that

$$\mathrm{e}^{-Lx} \int_{x_0}^{x} A(\xi)\, \mathrm{d}\xi \leq \frac{a}{L} \left(\mathrm{e}^{-Lx_0} - \mathrm{e}^{-Lx} \right),$$

that is

$$L \int_{x_0}^{x} A(\xi)\, \mathrm{d}\xi \leq a \left(\mathrm{e}^{L(x - x_0)} - 1 \right). \tag{11}$$

Now substituting (11) into (9) gives

$$A(x) \leq a \mathrm{e}^{L(x - x_0)}, \qquad x_0 \leq x \leq X_M. \tag{12}$$

The implication "(9) $\Rightarrow$ (12)" is usually referred to as the **Gronwall Lemma**. Returning to our original notation, we deduce from (12) that

$$\|\mathbf{v}(x) - \mathbf{w}(x)\| \leq \|\mathbf{v}(x_0) - \mathbf{z}\|\, \mathrm{e}^{L(x - x_0)}, \qquad x_0 \leq x \leq X_M. \tag{13}$$

Thus, given $\varepsilon > 0$ as in Definition 1, we choose $\delta = \varepsilon \exp(-L(X_M - x_0))$ 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 $x_0 > -\infty$. More generally, given the initial-value problem

$$y' = \lambda y, \qquad y(x_0) = y_0,$$

with λ a complex number, the solution $y(x) = y_0 \exp(\lambda(x - x_0))$ is unstable for $\operatorname{Re}\lambda > 0$; the solution is stable in the sense of Lyapunov for $\operatorname{Re}\lambda \leq 0$ and is asymptotically stable for $\operatorname{Re}\lambda < 0$.

In the next section we shall consider numerical methods for the approximate solution of the initial-value problem (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 ordinary differential equation corresponding to $m = 1$. We shall suppose throughout that the function f satisfies the conditions of Picard's Theorem on the rectangle R and that the initial-value problem has a unique solution defined on the interval $[x_0, X_M]$, $-\infty < x_0 < X_M < \infty$. We begin by discussing one-step methods; this will be followed in subsequent sections by the study of linear 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 initial-value problem (1), (2) is Euler's method.[4] **Lecture 2**

Euler's method. Suppose that the initial-value problem (1), (2) is to be solved on the interval $[x_0, X_M]$. We divide this interval by the **mesh-points** $x_n = x_0 + nh$, $n = 0, \ldots, N$, where $h = (X_M - x_0)/N$ and N is a positive integer. The positive real number h is called the **step size**. Now let us suppose that, for each n, we seek a numerical approximation y_n to $y(x_n)$, the value of the analytical solution at the mesh point x_n. Given that $y(x_0) = y_0$ is known, let us suppose that we have already calculated y_n, up to some n, $0 \leq n \leq N - 1$; we define

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

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

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x, y(x))\,\mathrm{d}x, \qquad n = 0, \ldots, N - 1, \tag{14}$$

and then applying the numerical integration rule

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

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

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

This then motivates the definition of Euler's method. The idea can be generalised by replacing the rectangle rule in the derivation of Euler's method with a one-parameter family of integration rules of the form

$$\int_{x_n}^{x_{n+1}} g(x)\,\mathrm{d}x \approx h\left[(1 - \theta)g(x_n) + \theta g(x_{n+1})\right], \tag{15}$$

[4] Leonard Euler (1707–1783)

with $\theta \in [0,1]$ a parameter. By applying this in (14) with $g(x) = f(x, y(x))$ we find that

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

This then motivates the introduction of the following one-parameter family of methods: with y_0 supplied by (2), define

$$y_{n+1} = y_n + h\left[(1-\theta)f(x_n, y_n) + \theta f(x_{n+1}, y_{n+1})\right], \quad n = 0, \ldots, N-1, \tag{16}$$

parametrised by $\theta \in [0,1]$; (16) is called the **θ-method**. Now, for $\theta = 0$ we recover Euler's method. For $\theta = 1$, and y_0 specified by (2), we get

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}), \quad n = 0, \ldots, N-1, \tag{17}$$

referred to as the **implicit Euler method** since, unlike Euler's method considered above, (17) requires the solution of an implicit equation in order to determine y_{n+1}, given y_n. In order to emphasize 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

$$y_{n+1} = y_n + \frac{1}{2}h\left[f(x_n, y_n) + f(x_{n+1}, y_{n+1})\right], \qquad n = 0, \ldots, N-1;$$

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

Remark 1 *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} respetively. The resulting implicit one-step method:*

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

*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 $y' = x - y^2$, $y(0) = 0$, on the interval of $x \in [0, 0.4]$, 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.*

For comparison, we also compute the value of the analytical solution $y(x)$ at the mesh points $x_n = 0.1 * n$, $n = 0, \ldots, 4$. Since the solution is not available in closed form,[5] we use a Picard iteration to

[5] Using MAPLE, we obtain the solution in terms of Bessel functions:

```
> dsolve({diff(y(x),x) + y(x)*y(x) = x, y(0)=0}, y(x));
```

$$\mathrm{y}(x) = -\frac{\sqrt{x}\left(\dfrac{\sqrt{3}\,\mathrm{BesselK}(\frac{-2}{3}, \frac{2}{3}x^{3/2})}{\pi} - \mathrm{BesselI}(\frac{-2}{3}, \frac{2}{3}x^{3/2})\right)}{\dfrac{\sqrt{3}\,\mathrm{BesselK}(\frac{1}{3}, \frac{2}{3}x^{3/2})}{\pi} + \mathrm{BesselI}(\frac{1}{3}, \frac{2}{3}x^{3/2})}$$

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

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

calculate an accurate approximation to the analytical solution on the interval $[0, 0.4]$ *and call this the "exact solution". Thus, we consider*

$$y_0(x) \equiv 0, \qquad y_k(x) = \int_0^x \left(\xi - y_{k-1}^2(\xi)\right)\, \mathrm{d}\xi, \quad k = 1, 2, \ldots.$$

Hence,

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

It is easy to prove by induction 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.*

The "exact solution" is in good agreement with the results obtained with $\theta = 1/2$*: the error is* $\leq 5{*}10^{-5}$*. For* $\theta = 0$ *and* $\theta = 1$ *the discrepancy between* y_k *and* $y(x_k)$ *is larger: it is* $\leq 3{*}10^{-2}$*. We note in conclusion that a plot of the analytical solution can be obtained, for example, by using MAPLE, by entering the following at the command line:*

```
> with(DEtools):
> DEplot(diff(y(x),x)+y(x)*y(x)=x, y(x), x=0..0.4, [[y(0)=0]],
  y=-0.1..0.1, stepsize=0.05);
```

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 to this question 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 initial-value problem (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, $n = 0, \ldots, N$, 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 $n = 0, \ldots, N$. 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 by

$$e_n = y(x_n) - y_n, \qquad n = 0, \ldots, N.$$

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 ($0 \leq \theta \leq 1$) leaving it to the reader to fill the gap.

So let us consider Euler's explicit method:

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

The quantity

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - f(x_n, y(x_n)), \tag{18}$$

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

By noting that $f(x_n, y(x_n)) = y'(x_n)$ and applying Taylor's Theorem, it follows from (18) that there exists a $\xi_n \in (x_n, x_{n+1})$ such that

$$T_n = \frac{1}{2} h y''(\xi_n), \tag{19}$$

where we have assumed that that f is a sufficiently smooth function of two variables so as to ensure that y'' 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 (18), 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

$$|e_{n+1}| \leq (1 + hL)|e_n| + h|T_n|, \qquad n = 0, \ldots, N-1.$$

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

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

By induction, and noting that $1 + hL \leq \mathrm{e}^{hL}$,

$$\begin{aligned} |e_n| &\leq \frac{T}{L}\left[(1 + hL)^n - 1\right] + (1 + hL)^n |e_0| \\ &\leq \frac{T}{L}\left(\mathrm{e}^{L(x_n - x_0)} - 1\right) + \mathrm{e}^{L(x_n - x_0)} |e_0|, \qquad n = 1, \ldots, N. \end{aligned}$$

This estimate, together with the bound

$$|T| \leq \frac{1}{2}hM_2, \qquad M_2 = \max_{x\in[x_0,X_M]} |y''(x)|,$$

which follows from (19), yields

$$|e_n| \leq \mathrm{e}^{L(x_n-x_0)}|e_0| + \frac{M_2 h}{2L}\left(\mathrm{e}^{L(x_n-x_0)} - 1\right), \qquad n = 0,\ldots,N. \tag{20}$$

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 $\frac{1}{2} - \theta L h_0 > 0$)

$$\begin{aligned} |e_n| \quad &\leq \quad |e_0| \exp\left(L\frac{x_n - x_0}{1-\theta Lh}\right) \\ &\quad + \frac{h}{L}\left\{\left|\frac{1}{2} - \theta\right| M_2 + \frac{1}{6}(1+3\theta)hM_3\right\}\left[\exp\left(L\frac{x_n-x_0}{1-\theta Lh}\right) - 1\right], \end{aligned} \tag{21}$$

for $n = 0,\ldots,N$, where now $M_3 = \max_{x\in[x_0,X_M]} |y'''(x)|$. In the absence of rounding errors in the imposition of the initial condition (2) we can suppose that $e_0 = y(x_0) - y_0 = 0$. Assuming that this is the case, we see from (21) that $|e_n| = \mathcal{O}(h^2)$ for $\theta = 1/2$, while for $\theta = 0$ and $\theta = 1$, and indeed for any $\theta \neq 1/2$, $|e_n| = \mathcal{O}(h)$ only. 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 analytical solution $y(x_n)$ at the mesh points than those which were obtained with the explicit and the implicit Euler methods ($\theta = 0$ and $\theta = 1$, respectively).

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 forward 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 the forward Euler method 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

$$y_{n+1} = y_n + \frac{1}{2}h\left[f(x_n, y_n) + f(x_{n+1}, y_n + hf(x_n, y_n))\right], \quad n = 0,\ldots,N-1, \quad y_0 = \text{given},$$

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

Lecture 3

Definition 2 *A* one-step method *is a function Ψ that takes the triplet $(\xi, \eta; h) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}_{>0}$ and a function f, 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 initial-value problem*

$$y'(x) = f(x, y(x)), \qquad y(\xi) = \eta. \tag{22}$$

Here, we tacitly assume that (22) 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.

For example, 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 the Lipschitz condition with Lipschitz constant L, 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 situation is simpler:

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

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

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

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 method

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

$$y_{n+1} = y_n + h\Phi(x_n, y_n; h), \quad n = 0, \ldots, N-1, \quad y_0 = y(x_0) \,[= \text{specified by (2)}], \tag{23}$$

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

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

In order to assess the accuracy of the numerical method (23), we define the **global error**, e_n, by

$$e_n = y(x_n) - y_n.$$

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

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - \Phi(x_n, y(x_n); h). \tag{24}$$

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

Theorem 4 *Consider the general one-step method (23) 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 exists a positive constant L_Φ such that, for $0 \le h \le h_0$ and for the same region R as in Picard's Theorem,*

$$|\Phi(x, y; h) - \Phi(x, z; h)| \le L_\Phi |y - z|, \qquad \text{for } (x, y),\ (x, z) \text{ in } \mathsf{R}. \tag{25}$$

Then, assuming that $|y_n - y_0| \le Y_M$, it follows that

$$|e_n| \le \mathrm{e}^{L_\Phi(x_n - x_0)}|e_0| + \left[\frac{\mathrm{e}^{L_\Phi(x_n - x_0)} - 1}{L_\Phi}\right] T, \quad n = 0, \ldots, N, \tag{26}$$

where $T = \max_{0 \le n \le N-1} |T_n|$.

PROOF: Subtracting (23) from (24) we obtain:

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

Then, since $(x_n, y(x_n))$ and (x_n, y_n) belong to R, the Lipschitz condition (25) implies that

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

That is,

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

Hence

$$\begin{aligned}
|e_1| &\leq (1 + hL_\Phi)|e_0| + hT, \\
|e_2| &\leq (1 + hL_\Phi)^2|e_0| + h[1 + (1 + hL_\Phi)]T, \\
|e_3| &\leq (1 + hL_\Phi)^3|e_0| + h[1 + (1 + hL_\Phi) + (1 + hL_\Phi)^2]T, \\
&\text{etc.} \\
|e_n| &\leq (1 + hL_\Phi)^n|e_0| + [(1 + hL_\Phi)^n - 1]T/L_\Phi.
\end{aligned}$$

Observing that $1 + hL_\Phi \leq \exp(hL_\Phi)$, we obtain (26). $\diamond$

Let us note that the error bound (20) for Euler's explicit method is a special case of (26). We highlight the practical relevance of the error bound (26) by focusing on a particular example.

Example 2 *Consider the initial-value problem $y' = \tan^{-1} y$, $y(0) = y_0$, and suppose that this is solved by the explicit Euler method. The aim of the exercise is to apply (26) to quantify the size of the associated global error; thus, we need to find L and M_2. Here $f(x, y) = \tan^{-1} y$, so by the Mean-Value Theorem*

$$|f(x, y) - f(x, z)| = \left|\frac{\partial f}{\partial y}(x, \eta)\,(y - z)\right|,$$

where η lies between y and z. In our case

$$\left|\frac{\partial f}{\partial y}\right| = |(1 + y^2)^{-1}| \leq 1,$$

and therefore $L = 1$. To find M_2 we need to obtain a bound on $|y''|$ (without actually solving the initial-value problem!). This is easily achieved by differentiating both sides of the differential equation with respect to x:

$$y'' = \frac{\mathrm{d}}{\mathrm{d}x}(\tan^{-1} y) = (1 + y^2)^{-1}\frac{\mathrm{d}y}{\mathrm{d}x} = (1 + y^2)^{-1}\tan^{-1} y.$$

Therefore $|y''(x)| \leq M_2 = \frac{1}{2}\pi$. Inserting the values of L and M_2 into (20),

$$|e_n| \leq \mathrm{e}^{x_n}|e_0| + \frac{1}{4}\pi\,(\mathrm{e}^{x_n} - 1)\,h, \quad n = 0, \ldots, N.$$

In particular if we assume that no error has been committed initially (i.e. $e_0 = 0$), we have that

$$|e_n| \leq \frac{1}{4}\pi\,(\mathrm{e}^{x_n} - 1)\,h, \quad n = 0, \ldots, N.$$

Thus, given a positive tolerance `TOL` *specified beforehand, we can ensure that the error between the (unknown) analytical solution and its numerical approximation does not exceed this tolerance by choosing a positive step size h such that*

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

For such h we shall have $|y(x_n) - y_n| = |e_n| \leq$ TOL for each $n = 0, \ldots, N$, 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 (23), 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 (23) is **consistent** with the differential equation (1) if the consistency error defined by (24) is such that for any $\varepsilon > 0$ there exists a positive $h(\varepsilon)$ for which $|T_n| < \varepsilon$ for $0 < h < h(\varepsilon)$ and any pair of points $(x_n, y(x_n))$, $(x_{n+1}, y(x_{n+1}))$ on any solution curve in* R.

For the general one-step method (23) 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 (24),

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

As $y'(x) = f(x, y(x))$, this implies that the one-step method (23) is consistent if, and only if,

$$\Phi(x, y; 0) \equiv f(x, y). \tag{27}$$

Now we are ready to state a convergence theorem for the general one-step method (23).

Theorem 5 *Suppose that the solution of the initial-value problem (1), (2) lies in* R *as does its approximation generated from (23) when $h \leq h_0$. Suppose also that the function $\Phi(\cdot, \cdot; \cdot)$ is uniformly continuous on $\mathsf{R} \times [0, h_0]$ and satisfies the consistency condition (27) and the Lipschitz condition*

$$|\Phi(x, y; h) - \Phi(x, z; h)| \leq L_\Phi |y - z| \qquad \text{on } \mathsf{R} \times [0, h_0]. \tag{28}$$

Then, if successive approximation sequences (y_n), generated for $x_n = x_0 + nh$, $n = 1, 2, \ldots, N$, are obtained from (23) with successively smaller values of h, each less than h_0, we have convergence of the numerical solution to the solution of the initial-value problem in the sense that

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

PROOF: Suppose that $h = (X_M - x_0)/N$ where N is a positive integer. We shall assume that N is sufficiently large so that $h \leq h_0$. Since $y(x_0) = y_0$ and therefore $e_0 = 0$, Theorem 4 implies that

$$|y(x_n) - y_n| \leq \left[\frac{\mathrm{e}^{L_\phi (X_M - x_0)} - 1}{L_\phi}\right] \max_{0 \leq m \leq n-1} |T_m|, \qquad n = 1, \ldots, N. \tag{29}$$

From the consistency condition (27) 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 $y'(\xi) - y'(x_n)$, where $\xi \in [x_n, x_{n+1}]$. Since $y'(\cdot) = f(\cdot, y(\cdot)) = \Phi(\cdot, y(\cdot); 0)$ 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 $h_1(\varepsilon)$ such that

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

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

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

Thus, defining $h(\varepsilon) = \min(h_1(\varepsilon), h_2(\varepsilon))$, we have

$$|T_n| \leq \varepsilon \qquad \text{for } h < h(\varepsilon),\ n = 0, 1, \ldots, N-1.$$

Inserting this into (29) we deduce that $|y(x_n) - y_n| \to 0$ as $h \to 0$ and $n \to \infty$. Since

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

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

$$|T_n| \leq Kh \qquad \text{for } 0 < h \leq h_0,$$

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 (23) is said to have* ***order of accuracy*** *p (or order of consistency p), if p is the largest positive integer such that, for any sufficiently smooth solution curve $(x, y(x))$ in R of the initial-value problem (1), (2), there exist constants K and h_0 such that*

$$|T_n| \leq Kh^p \qquad \textit{for } 0 < h \leq h_0$$

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 Euler's 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 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 Runge–Kutta family** is as follows:

Lecture 4

$$\begin{aligned}
y_{n+1} &= y_n + h\Phi(x_n, y_n; h), \\
\Phi(x, y; h) &= \sum_{r=1}^{R} c_r k_r, \\
k_1 &= f(x, y), \\
k_r &= f\left(x + ha_r, y + h\sum_{s=1}^{r-1} b_{rs} k_s\right), \quad r = 2, \ldots, R, \\
a_r &= \sum_{s=1}^{r-1} b_{rs}, \quad r = 2, \ldots, R.
\end{aligned} \tag{30}$$

$a = Be$	B
	c^{T}

where $e = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^{R-1}$.

Figure 1: Butcher tableau of a Runge–Kutta method: $a \in \mathbb{R}^{R-1}$, $B \in \mathbb{R}^{(R-1)\times(R-1)}$, $c \in \mathbb{R}^R$. In the case of an explicit Runge–Kutta method $B \in \mathbb{R}^{(R-1)\times(R-1)}$ is a strictly lower-triangular matrix, i.e. the diagonal and superdiagonal entries of B are all equal to zero. For the sake of simplicity we focus on explicit Runge–Kutta methods only.

In compressed form, the information about the coefficients of a Runge–Kutta method is usually displayed in the so-called Butcher tableau shown in Fig. 1.

One-stage explicit Runge–Kutta methods. Suppose that $R = 1$. Then, the resulting one-stage explicit Runge–Kutta method is simply Euler's explicit method:

$$y_{n+1} = y_n + hf(x_n, y_n). \tag{31}$$

Thus, in the language of Runge–Kutta methods, $y_{n+1} = y_n + h\Phi(x_n, y_n; h)$ with $\Phi(x, y; h) = \sum_{r=1}^{1} c_r k_r$, $c_1 = 1$ and $k_1 = f(x, y)$.

Remark 2 *The implicit Euler method $y_{n+1} = y_n + hf(x_{n+1}, y_{n+1})$ is an example of a one-stage implicit Runge-Kutta method: it can be written as $y_{n+1} = y_n + h\Phi(x_n, y_n; h)$, where $\Phi(x, y; h) = k_1$ and $k_1 = f(x + h, y + hk_1)$ (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 Runge–Kutta methods only.*

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

$$y_{n+1} = y_n + h(c_1 k_1 + c_2 k_2), \tag{32}$$

where

$$k_1 = f(x_n, y_n), \tag{33}$$
$$k_2 = f(x_n + a_2 h, y_n + b_{21} h k_1), \tag{34}$$

and where the parameters c_1, c_2, a_2 and b_{21} are to be determined.[6] Clearly (32)–(34) can be rewritten in the form (23) and therefore it is a family of one step methods. By the condition (27), a method from this family will be consistent if, and only if,

$$c_1 + c_2 = 1.$$

Further conditions on the parameters are obtained by attempting to maximise the order of accuracy of the method. Indeed, expanding the consistency error of (32)–(34) in powers of h, after some algebra we obtain

$$\begin{aligned} T_n &= \frac{1}{2}hy''(x_n) + \frac{1}{6}h^2 y'''(x_n) \\ &\quad -c_2 h[a_2 f_x + b_{21} f_y f] - c_2 h^2 \left[\frac{1}{2}a_2^2 f_{xx} + a_2 b_{21} f_{xy} f + \frac{1}{2}b_{21}^2 f_{yy} f^2\right] + \mathcal{O}(h^3). \end{aligned}$$

[6] We note in passing that Euler's explicit 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.

Here we have used the abbreviations $f = f(x_n, y(x_n))$, $f_x = \frac{\partial f}{\partial x}(x_n, y(x_n))$, etc. On noting that $y'' = f_x + f_y f$, it follows that $T_n = \mathcal{O}(h^2)$ for any f provided that

$$a_2 c_2 = b_{21} c_2 = \frac{1}{2},$$

which implies that if $b_{21} = a_2$, $c_2 = 1/(2a_2)$ and $c_1 = 1 - 1/(2a_2)$ then the method is second-order accurate; while this still leaves one free parameter, a_2, it is easy to see that no choice of the parameters will make the method generally third-order accurate. There are two well-known examples of second-order explicit Runge–Kutta methods of the form (32), (34):

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

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

b) **The improved Euler method:** This is arrived at by choosing $a_2 = 1$ which gives

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

For these two methods it is easily verified by Taylor series expansion that the consistency error is of the form, respectively,

$$\begin{aligned} T_n &= \frac{1}{6}h^2\left[f_y F_1 + \frac{1}{4}F_2\right] + \mathcal{O}(h^3), \\ T_n &= \frac{1}{6}h^2\left[f_y F_1 - \frac{1}{2}F_2\right] + \mathcal{O}(h^3), \end{aligned}$$

where

$$F_1 = f_x + f f_y \quad \text{and} \quad F_2 = f_{xx} + 2f f_{xy} + f^2 f_{yy}.$$

The family (32)–(34) is referred to as the class of explicit two-stage explicit Runge–Kutta methods.

Exercise 1 *Let α be a nonzero real number and let $x_n = a + nh$, $n = 0, \ldots, N$, be a uniform mesh on the interval $[a, b]$ of step size $h = (b-a)/N$. Consider the explicit one-step method for the numerical solution of the initial-value problem $y' = f(x, y)$, $y(a) = y_0$, which determines approximations y_n to the values $y(x_n)$ from the recurrence relation*

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

Show that this method is consistent and that its consistency error, $T_n(h, \alpha)$, can be expressed as

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

This numerical method is applied to the initial-value problem $y' = -y^p$, $y(0) = 1$, where p is a positive integer. Show that if $p = 1$ then $T_n(h, \alpha) = \mathcal{O}(h^2)$ for every nonzero real number α. Show also that if $p \geq 2$ then there exists a nonzero real number α_0 such that $T_n(h, \alpha_0) = \mathcal{O}(h^3)$.

SOLUTION: Let us define

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

Then the numerical method can be rewritten as

$$y_{n+1} = y_n + h\Phi(x_n, y_n; h).$$

Since

$$\Phi(x, y; 0) = f(x, y),$$

the method is consistent. By definition, the consistency error is

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

We shall perform a Taylor expansion of $T_n(h, \alpha)$ to show that it can be expressed in the desired form. Indeed,

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

as required.

Now let us apply the method to $y' = -y^p$, with $p \geq 1$. If $p = 1$, then $y''' = -y'' = y' = -y$, so that

$$T_n(h, \alpha) = -\frac{h^2}{6}y(x_n) + \mathcal{O}(h^3).$$

As $y(x_n) = \mathrm{e}^{-x_n} \neq 0$, it follows that

$$T_n(h, \alpha) = \mathcal{O}(h^2)$$

for all (nonzero) α.

Finally, suppose that $p \geq 2$. Then,

$$y'' = -py^{p-1}y' = py^{2p-1}$$

and

$$y''' = p(2p-1)y^{2p-2}y' = -p(2p-1)y^{3p-2},$$

and therefore

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

Choosing α such that

$$\left(\frac{4}{3}\alpha - 1\right)p(2p-1) + p^2 = 0,$$

namely

$$\alpha = \alpha_0 = \frac{3p-3}{8p-4},$$

gives

$$T_n(h, \alpha_0) = \mathcal{O}(h^3).$$

We note in passing that for $p > 1$ the exact solution of the initial-value problem $y' = -y^p$, $y(0) = 1$, is $y(x) = [(p-1)x + 1]^{1/(1-p)}$. $\diamond$

Three-stage explicit Runge–Kutta 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

$$\begin{aligned} k_1 &= f(x, y), \\ k_2 &= f(x + ha_2, y + hb_{21}k_1), \\ k_3 &= f(x + ha_3, y + hb_{31}k_1 + hb_{32}k_2), \\ a_2 &= b_{21}, \quad a_3 = b_{31} + b_{32}. \end{aligned}$$

Writing $b_{21} = a_2$ and $b_{31} = a_3 - b_{32}$ in the definitions of k_2 and k_3 respectively and expanding k_2 and k_3 into Taylor series about the point (x, y) yields:

$$\begin{aligned} k_2 &= f + ha_2(f_x + k_1f_y) + \frac{1}{2}h^2a_2^2(f_{xx} + 2k_1f_{xy} + k_1^2f_{yy}) + \mathcal{O}(h^3) \\ &= f + ha_2(f_x + ff_y) + \frac{1}{2}h^2a_2^2(f_{xx} + 2ff_{xy} + f^2f_{yy}) + \mathcal{O}(h^3) \\ &= f + ha_2F_1 + \frac{1}{2}h^2a_2^2F_2 + \mathcal{O}(h^3), \end{aligned}$$

where

$$F_1 = f_x + ff_y \quad \text{and} \quad F_2 = f_{xx} + 2ff_{xy} + f^2f_{yy},$$

and

$$\begin{aligned} k_3 &= f + h\left\{a_3f_x + \left[(a_3 - b_{32})k_1 + b_{32}k_2\right]f_y\right\} \\ &\quad + \frac{1}{2}h^2\left\{a_3^2f_{xx} + 2a_3\left[(a_3 - b_{32})k_1 + b_{32}k_2\right]f_{xy}\right. \\ &\quad \left. + \left[(a_3 - b_{32})k_1 + b_{32}k_2\right]^2 f_{yy}\right\} + \mathcal{O}(h^3) \\ &= f + ha_3F_1 + h^2\left(a_2b_{32}F_1f_y + \frac{1}{2}a_3^2F_2\right) + \mathcal{O}(h^3). \end{aligned}$$

Substituting these expressions for k_2 and k_3 into (30) with $R = 3$ we find that

$$\begin{aligned} \Phi(x, y, h) &= (c_1 + c_2 + c_3)f + h(c_2a_2 + c_3a_3)F_1 \\ &\quad + \frac{1}{2}h^2\left[2c_3a_2b_{32}F_1f_y + \left(c_2a_2^2 + c_3a_3^2\right)F_2\right] + \mathcal{O}(h^3). \end{aligned} \tag{35}$$

We match this with the Taylor series expansion:

$$\begin{aligned} \frac{y(x+h) - y(x)}{h} &= y'(x) + \frac{1}{2}hy''(x) + \frac{1}{6}h^2y'''(x) + \mathcal{O}(h^3) \\ &= f + \frac{1}{2}hF_1 + \frac{1}{6}h^2\left(F_1f_y + F_2\right) + \mathcal{O}(h^3). \end{aligned}$$

This yields:

$$\begin{aligned} c_1 + c_2 + c_3 &= 1, \\ c_2 a_2 + c_3 a_3 &= \frac{1}{2}, \\ c_2 a_2^2 + c_3 a_3^2 &= \frac{1}{3}, \\ c_3 a_2 b_{32} &= \frac{1}{6}. \end{aligned}$$

Solving this system of four equations for the six unknowns: $c_1, c_2, c_3, a_2, a_3, b_{32}$, we obtain a two-parameter family of 3-stage explicit Runge–Kutta methods. We shall only highlight two notable examples from this family:

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

$$c_1 = \frac{1}{4}, \quad c_2 = 0, \quad c_3 = \frac{3}{4}, \quad a_2 = \frac{1}{3}, \quad a_3 = \frac{2}{3}, \quad b_{32} = \frac{2}{3},$$

yielding

$$\begin{aligned} y_{n+1} &= y_n + \frac{1}{4} h \left(k_1 + 3k_3 \right), \\ k_1 &= f(x_n, y_n), \\ k_2 &= f\left(x_n + \frac{1}{3} h, y_n + \frac{1}{3} h k_1 \right), \\ k_3 &= f\left(x_n + \frac{2}{3} h, y_n + \frac{2}{3} h k_2 \right). \end{aligned}$$

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

$$c_1 = \frac{1}{6}, \quad c_2 = \frac{2}{3}, \quad c_3 = \frac{1}{6}, \quad a_2 = \frac{1}{2}, \quad a_3 = 1, \quad b_{32} = 2,$$

yielding

$$\begin{aligned} y_{n+1} &= y_n + \frac{1}{6} h \left(k_1 + 4k_2 + k_3 \right), \\ k_1 &= f(x_n, y_n), \\ k_2 &= f\left(x_n + \frac{1}{2} h, y_n + \frac{1}{2} h k_1 \right), \\ k_3 &= f\left(x_n + h, y_n - h k_1 + 2h k_2 \right). \end{aligned}$$

Four-stage explicit Runge–Kutta methods. For $R = 4$, an analogous argument leads to a two-parameter family of four-stage Runge–Kutta 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

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

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 Runge–Kutta method reduces when $\frac{\partial f}{\partial y} \equiv 0$).

In this section, we have constructed R-stage explicit Runge–Kutta 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 Runge–Kutta 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

It is instructive to consider the model problem **Lecture 5**

$$y' = \lambda y, \quad y(0) = y_0 \ (\neq 0), \tag{36}$$

with λ real and *negative*. Trivially, the analytical solution to this initial value problem, $y(x) = y_0 \exp(\lambda x)$, 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 Runge–Kutta 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 initial-value problem by an explicit Runge–Kutta 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 explicit one-stage Runge–Kutta method is Euler's explicit method. Applying (31) to (36) yields:

$$y_{n+1} = (1 + \bar{h}) y_n, \quad n \geq 0,$$

where $\bar{h} := \lambda h$. Thus,

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

Consequently, the sequence $\{y_n\}_{n=0}^{\infty}$ will converge to 0 if, and only if, $|1 + \bar{h}| < 1$, 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 Runge–Kutta methods:

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

where

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

with

$$c_1 + c_2 = 1, \qquad a_2 c_2 = b_{21} c_2 = \frac{1}{2}.$$

Applying this to (36) yields,

$$y_{n+1} = \left(1 + \bar{h} + \frac{1}{2}\bar{h}^2\right) y_n, \qquad n \geq 0,$$

and therefore

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

Hence the method is absolutely stable if, and only if,

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

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: $\bar{h} \in (-2.51, 0)$.

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 $\bar{h} \in (-2.78, 0)$.

For $R \geq 5$ on applying the explicit Runge–Kutta method to the model problem (36) still results in a recursion of the form

$$y_{n+1} = A_R(\bar{h}) y_n, \qquad n \geq 0,$$

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 Runge–Kutta method; by a convenient choice of the free parameters the associated interval of absolute stability may be maximised. For further results in this direction, the reader is referred to the book of J.D. Lambert.

3 Linear multi-step methods

While explicit Runge–Kutta methods present an improvement over Euler's method in terms of accuracy, this is achieved by investing additional computational effort; in fact, Runge–Kutta 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 differential equation between x_{n-1} and x_{n+1}, and applying Simpson's rule to approximate the resulting integral 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 method

$$y_{n+1} = y_{n-1} + \frac{1}{3}h\left[f(x_{n-1}, y_{n-1}) + 4f(x_n, y_n) + f(x_{n+1}, y_{n+1})\right]. \tag{37}$$

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 (37) is not a one-step method.

In this section we consider a class of methods of the type (37) for the numerical solution of the initial-value problem (1), (2), called **linear multi-step methods**.

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

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(x_{n+j}, y_{n+j}), \tag{38}$$

where the coefficients $\alpha_0, \ldots, \alpha_k$ and $\beta_0, \ldots, \beta_k$ are real constants. In order to avoid degenerate cases, we shall assume that $\alpha_k \neq 0$ and that 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 (38) 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 instead of $f(x_n, y_n)$.

Example 3 *We have already seen an example of a linear 2-step method in (37); here we present further examples of linear multi-step methods.*

a) Euler's method is a trivial case: it is an explicit linear one-step method. The ***implicit Euler method***

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1})$$

is an implicit linear one-step method.

b) The ***trapezium method****, given by*

$$y_{n+1} = y_n + \frac{1}{2}h[f_{n+1} + f_n]$$

is also an implicit linear one-step method.

c) The four-step ***Adams[7]–Bashforth method***

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

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

$$y_{n+4} = y_{n+3} + \frac{1}{24}h[9f_{n+4} + 19f_{n+3} - 5f_{n+2} - 9f_{n+1}]$$

is an implicit linear four-step method.

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

3.1 Construction of linear multi-step methods

Start of optional material

Let us suppose that u_n, $n = 0, 1, \ldots$, is a sequence of real numbers. We introduce the shift operator E, the forward difference operator Δ_+ and the backward difference operator Δ_- by

$$E \ : \ u_n \mapsto u_{n+1}, \qquad \Delta_+ \ : \ u_n \mapsto (u_{n+1} - u_n), \qquad \Delta_- \ : \ u_n \mapsto (u_n - u_{n-1}).$$

Further, we note that E^{-1} exists and is given by $E^{-1} \ : \ u_{n+1} \mapsto u_n$. Since

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

where I signifies the identity operator, it follows that, for any positive integer k,

$$\Delta_+^k u_n = (E - I)^k u_n = \sum_{j=0}^{k} (-1)^j \begin{pmatrix} k \\ j \end{pmatrix} u_{n+k-j}$$

[7] J. C. Adams (1819–1892)

and

$$\Delta_-^k u_n = (I - E^{-1})^k u_n = \sum_{j=0}^{k} (-1)^j \left(\begin{array}{c} k \\ j \end{array} \right) u_{n-j}.$$

Now suppose that u is a real-valued function defined on $\mathbb{R}$ whose derivative exists and is integrable on $[x_0, x_n]$ for each $n \geq 0$, and let u_n denote $u(x_n)$ where $x_n = x_0 + nh$, $n = 0, 1, \ldots$, are equally spaced points on the real line. Letting D denote $\mathrm{d}/\mathrm{d}x$, by applying a Taylor series expansion we find that

$$\begin{aligned} (E^s u)_n = u(x_n + sh) &= u_n + sh(Du)_n + \frac{1}{2!}(sh)^2 (D^2 u)_n + \cdots \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} ((shD)^k u)_n = (\mathrm{e}^{shD} u)_n, \end{aligned}$$

and hence

$$E^s = \mathrm{e}^{shD}.$$

Thus, formally,

$$hD = \ln E = -\ln(I - \Delta_-),$$

and therefore, again by Taylor series expansion,

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

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

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

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

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

and so on. These approximate equalities give rise to a class of implicit linear multi-step methods called **backward differentiation formulae**, the simplest of which is Euler's implicit method.

Similarly,

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

and therefore

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

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

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

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

Next we derive additional identities which will allow us to construct further classes of linear multi-step methods. Let us define

$$D^{-1}u(x_n) = u(x_0) + \int_{x_0}^{x_n} u(\xi)\,\mathrm{d}\xi,$$

and observe that

$$(E-I)D^{-1}u(x_n) = \int_{x_n}^{x_{n+1}} u(\xi)\,\mathrm{d}\xi.$$

Now,

$$\begin{aligned}(E-I)D^{-1} &= \Delta_+ D^{-1} = E\Delta_- D^{-1} = hE\Delta_-(hD)^{-1} \\ &= -hE\Delta_-\left[\ln(I-\Delta_-)\right]^{-1}. \end{aligned} \tag{39}$$

Furthermore,

$$\begin{aligned}(E-I)D^{-1} &= E\Delta_- D^{-1} = \Delta_- E D^{-1} = \Delta_-(DE^{-1})^{-1} = h\Delta_-(hDE^{-1})^{-1} \\ &= -h\Delta_-\left[(I-\Delta_-)\ln(I-\Delta_-)\right]^{-1}. \end{aligned} \tag{40}$$

Letting, again, $u(x) = y(x)$ where y is the solution of the initial-value problem (1), (2), noting that $u'(x) = y'(x) = f(x, y(x))$ and using (39) and (40) we deduce that

$$\begin{aligned} y(x_{n+1}) - y(x_n) &= \int_{x_n}^{x_{n+1}} y'(\xi)\,\mathrm{d}\xi = (E-I)D^{-1}y'(x_n) = (E-I)D^{-1}f(x_n, y(x_n)) \\ &= \begin{cases} -hE\Delta_-\left[\ln(I-\Delta_-)\right]^{-1} f(x_n, y(x_n)), \\ -h\Delta_-\left[(I-\Delta_-)\ln(I-\Delta_-)\right]^{-1} f(x_n, y(x_n)). \end{cases} \end{aligned} \tag{41}$$

By expanding $\ln(I-\Delta_-)$ into a Taylor series on the right-hand side of (41) we find that

$$y(x_{n+1}) - y(x_n) \approx h\left[I - \frac{1}{2}\Delta_- - \frac{1}{12}\Delta_-^2 - \frac{1}{24}\Delta_-^3 - \frac{19}{720}\Delta_-^4 - \cdots\right] f(x_n, y(x_n)) \tag{42}$$

and

$$y(x_{n+1}) - y(x_n) \approx h\left[I + \frac{1}{2}\Delta_- + \frac{5}{12}\Delta_-^2 + \frac{3}{8}\Delta_-^3 + \frac{251}{720}\Delta_-^4 + \cdots\right] f(x_n, y(x_n)). \tag{43}$$

Successive truncations of (42) yield the family of Adams–Moulton methods, while similar successive truncations of (43) gives rise to the family of Adams–Bashforth methods. Next, we turn our attention to the analysis of linear multi-step methods and introduce the concepts of stability, consistency and convergence.

End of optional material

3.2 Zero-stability

Lecture 6

As is clear from (38) we need k starting values, $y_0, \ldots, y_{k-1}$, before we can apply a linear k-step method to the initial-value problem (1), (2): of these, y_0 is given by the initial condition (2), but the others, $y_1, \ldots, y_{k-1}$, have to be computed by other means: say, by using a suitable Runge–Kutta method. At any rate, 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 (38). 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 ordinary differential equation $y' = f(x, y)$ is said to be* ***zero-stable*** *if there exists a constant K 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*

$$|y_n - \hat{y}_n| \leq K \max\{|y_0 - \hat{y}_0|, |y_1 - \hat{y}_1|, \ldots, |y_{k-1} - \hat{y}_{k-1}|\} \tag{44}$$

for $x_n \leq X_M$*, 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 $y' = 0$, corresponding to (1) with $f(x, y) \equiv 0$; 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 linear multi-step method 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 (38) we consider its **first** and **second characteristic polynomial**, respectively

$$\begin{aligned} \rho(z) &= \sum_{j=0}^{k} \alpha_j z^j, \\ \sigma(z) &= \sum_{j=0}^{k} \beta_j z^j, \end{aligned}$$

where, as before, we assume that

$$\alpha_k \neq 0, \quad \alpha_0^2 + \beta_0^2 \neq 0.$$

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

Theorem 6 *A linear multi-step method is zero-stable for any ordinary differential equation of the form (1) where* f *satisfies the Lipschitz condition (3), if, and only if, its first characteristic polynomial has zeros inside the closed unit disc, with any which lie on the unit circle 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: *Necessity.* Consider the linear k-step method, applied to $y' = 0$:

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_1 y_{n+1} + \alpha_0 y_n = 0. \tag{45}$$

The general solution of this kth order linear difference equation has the form

$$y_n = \sum_s p_s(n) z_s^n, \tag{46}$$

where z_s is a zero of the first characteristic polynomial $\rho(z)$ and the polynomial $p_s(\cdot)$ has 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 (45) is $(\hat{y}_n)$ with $\hat{y}_n = 0$ for all n, we see that (44) 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) The explicit and implicit Euler methods have first characteristic polynomial $\rho(z) = z-1$ with simple root $z = 1$, so both methods are zero-stable. The same is true of the trapezium method.

b) The Adams–Bashforth and Adams–Moulton methods considered in Example 3 have the same first characteristic polynomial, $\rho(z) = z^3(z-1)$, and therefore both methods are zero-stable.

c) The three-step (sixth order accurate) linear multi-step method

$$11y_{n+3} + 27y_{n+2} - 27y_{n+1} - 11y_n = 3h[f_{n+3} + 9f_{n+2} + 9f_{n+1} + f_n]$$

is not *zero-stable. Indeed, the associated first characteristic polynomial $\rho(z) = 11z^3+27z^2-27z-11$ has roots at $z_1 = 1$, $z_2 \approx -0.3189$, $z_3 \approx -3.1356$, so $|z_3| > 1$.*

3.3 Consistency

In this section we consider the accuracy of the linear k-step method (38). For this purpose, as in the case of one-step methods, we introduce the notion of consistency error. Thus, suppose that $y(x)$ is a solution of the ordinary differential equation (1). Then the consistency error of (38) is defined as follows:

$$T_n = \frac{\sum_{j=0}^{k} [\alpha_j y(x_{n+j}) - h\beta_j y'(x_{n+j})]}{h\sum_{j=0}^{k} \beta_j}. \tag{47}$$

Of course, the definition requires implicitly that $\sigma(1) = \sum_{j=0}^{k} \beta_j \neq 0$. 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 differential equation into the formula (38) and scaling this residual appropriately (in this case dividing through by $h\sum_{j=0}^{k} \beta_j$) so that T_n resembles $y' - f(x, y(x))$.

Definition 6 *The numerical scheme (38) is said to be* ***consistent*** *with the differential equation (1) if the consistency error defined by (47) is such that for any $\varepsilon > 0$ there exists an $h(\varepsilon)$ for which*

$$|T_n| < \varepsilon \quad \text{for } 0 < h < h(\varepsilon),$$

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 initial-value problem (1), (2).*

Now let us suppose that the solution to the differential equation 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 (47) to obtain

$$T_n = \frac{1}{h\sigma(1)}[C_0 y(x_n) + C_1 h y'(x_n) + C_2 h^2 y''(x_n) + \cdots], \tag{48}$$

where $\sigma(1) \neq 0$,

$$\begin{aligned}
C_0 &= \sum_{j=0}^{k} \alpha_j, \\
C_1 &= \sum_{j=1}^{k} j\alpha_j - \sum_{j=0}^{k} \beta_j, \\
C_2 &= \sum_{j=1}^{k} \frac{j^2}{2!}\alpha_j - \sum_{j=1}^{k} j\beta_j, \\
&\text{etc.} \\
C_q &= \sum_{j=1}^{k} \frac{j^q}{q!}\alpha_j - \sum_{j=1}^{k} \frac{j^{q-1}}{(q-1)!}\beta_j.
\end{aligned}$$

For consistency we need that $T_n \to 0$ as $h \to 0$ and this requires that $C_0 = 0$ and $C_1 = 0$; as $C_0 = \rho(1)$ and $C_1 = \rho'(1) - \sigma(1)$, in terms of the characteristic polynomials ρ and σ this consistency requirement can be restated in compact form as

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

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 (38) is said to have* ***order of accuracy*** *p (or* ***order of consistency*** *p) if p is the largest positive integer such that, for any sufficiently smooth solution curve in* R *of the initial-value problem (1), (2), there exist constants K and h_0 such that*

$$|T_n| \leq Kh^p \quad \textit{for } 0 < h \leq h_0$$

for any $(k+1)$ points $(x_n, y(x_n)), \ldots, (x_{n+k}, y(x_{n+k}))$ on the solution curve.

Thus we deduce from (48) that the method is of order of accuracy (or order of consistency) p if, and only if,

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

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 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 $\alpha_2 = 1$, $\alpha_0 = a$, $\beta_2 \neq 0$. 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 the constants

$$\begin{aligned}
C_0 &= \alpha_0 + \alpha_1 + \alpha_2, \\
C_1 &= \alpha_1 + 2 - (\beta_0 + \beta_1 + \beta_2), \\
C_q &= \frac{1}{q!}(\alpha_1 + 2^q \alpha_2) - \frac{1}{(q-1)!}(\beta_1 + 2^{q-1}\beta_2), \qquad q = 2, 3,
\end{aligned}$$

appearing in (48) are all equal to zero, because we wish to maximise the order of accuracy of the method. Thus,

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

Hence,

$$\begin{aligned}
&\alpha_1 = -1 - a, \\
&\beta_0 = -\frac{1}{12}(1 + 5a), \quad \beta_1 = \frac{2}{3}(1 - a), \quad \beta_2 = \frac{1}{12}(5 + a),
\end{aligned}$$

and the resulting method is

$$y_{n+2} - (1+a)y_{n+1} + ay_n = \frac{1}{12}h\left[(5+a)f_{n+2} + 8(1-a)f_{n+1} - (1+5a)f_n\right]. \tag{49}$$

Further,

$$\begin{aligned} C_4 &= \frac{1}{4!}(\alpha_1 + 16) - \frac{1}{3!}(\beta_1 + 8\beta_2) = -\frac{1}{4!}(1+a), \\ C_5 &= \frac{1}{5!}(\alpha_1 + 32) - \frac{1}{4!}(\beta_1 + 16\beta_2) = -\frac{1}{3\cdot 5!}(17+13a). \end{aligned}$$

If $a \neq -1$ then $C_4 \neq 0$, and the method (49) is third order accurate. If, on the other hand, $a = -1$, then $C_4 = 0$ and $C_5 \neq 0$ and the method (49) becomes the Simpson rule method: a fourth-order accurate two-step method. The error constant is:

$$C_4 = -\frac{1}{4!}(1+a), \qquad a \neq -1, \tag{50}$$

$$C_5 = -\frac{4}{3\cdot 5!}, \qquad a = -1. \tag{51}$$

$\diamond$

Exercise 3 *Determine all values of the real parameter b, $b \neq 0$, for which the linear multi-step method*

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

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

SOLUTION: According to the root condition, this linear multi-step method is zero-stable if, and only if, all roots of its first characteristic polynomial

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

belong to the closed unit disc, and those on the unit circle are simple.

Clearly, $\rho(1) = 0$; upon dividing $\rho(z)$ by $z-1$ we see that $\rho(z)$ can be written in the following factorised form:

$$\rho(z) = (z-1)\left(z^2 - 2(1-b)z + 1\right) \equiv (z-1)\rho_1(z).$$

Thus the method is zero-stable if, and only if, all roots of the polynomial $\rho_1(z)$ belong to the closed unit disc, and those on the unit circle are simple and differ from 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 $\rho_1(z)$ on the unit circle, respectively, $z = 1$ and $z = -1$. Since the product of the two roots of $\rho_1(z)$ 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 $\rho_1(z)$ is negative. Namely,

$$4(1-b)^2 - 4 < 0.$$

In other words, $b \in (0,2)$.

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

$$z_1 = 1, \qquad z_{2/3} = 1 - b + \imath\sqrt{1-(b-1)^2}.$$

Since $|z_{2/3}| = 1$, $z_{2/3} \neq 1$ and $z_2 \neq z_3$, all roots of $\rho(z)$ lie on the unit circle and they are simple. Hence the method is zero-stable.

To summarise, the method is zero-stable if, and only if, $b \in (0,2)$.

In order to analyse the order of accuracy of the method we note that upon Taylor series expansion its consistency error can be written in the form

$$T_n = \frac{1}{2b}\left[\left(1 - \frac{b}{6}\right)h^2 y'''(x_n) + \frac{1}{4}(6-b)h^3 y^{IV}(x_n) + \frac{1}{120}(150 - 23b)h^4 y^V(x_n) + \mathcal{O}(h^5)\right].$$

If $b = 6$, then $T_n = \mathcal{O}(h^4)$ 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 $b = 6$.

Since zero-stability requires $b \in (0,2)$, in which case $1 - \frac{b}{6} \neq 0$, it follows that if the method is zero-stable then $T_n = \mathcal{O}(h^2)$. $\diamond$

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, $n = 0, \ldots, N$, are close to those of the analytical 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. In order to formalise the desired behaviour, we introduce the following definition. **Lecture 7**

Definition 8 *The linear multistep method (38) is said to be* ***convergent*** *if, for all initial-value problems (1), (2) subject to the hypotheses of Theorem 1, we have that*

$$\lim_{\substack{h\to 0 \\ nh=x-x_0}} y_n = y(x) \tag{52}$$

holds for all $x \in [x_0, X_M]$ and for all solutions $\{y_n\}_{n=0}^N$ of the difference equation (38) with ***consistent starting conditions****, i.e. with starting conditions $y_s = \eta_s(h)$, $s = 0, 1, \ldots, k-1$, for which $\lim_{h\to 0} \eta_s(h) = y_0$, $s = 0, 1, \ldots, k-1$.*

We emphasize here that Definition 8 requires that (52) holds *not only* for those sequences $\{y_n\}_{n=0}^N$ which have been generated from (38) 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 the $h \to 0$. This assumption is made because 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 linear multi-step method 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 linear multi-step method (38) is that it be zero-stable.*

Proof: Let us suppose that the linear multi-step method (38) is convergent; we wish to show that it is then zero-stable.

We consider the initial-value problem $y' = 0$, $y(0) = 0$, on the interval $[0, X_M]$, $X_M > 0$, whose solution is, trivially, $y(x) \equiv 0$. Applying (38) to this problem yields the difference equation

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n = 0. \tag{53}$$

Since the method is assumed to be convergent, for any $x > 0$, we have that

$$\lim_{\substack{h\to 0 \\ nh=x}} y_n = 0, \tag{54}$$

for all solutions of (53) satisfying $y_s = \eta_s(h)$, $s = 0, \ldots, k-1$, where

$$\lim_{h\to 0} \eta_s(h) = 0, \qquad s = 0, 1, \ldots, k-1. \tag{55}$$

Let $z = re^{i\phi}$, be a root of the first characteristic polynomial $\rho(z)$; $r \geq 0$, $0 \leq \phi < 2\pi$. It is an easy matter to verify then that the numbers

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

define a solution to (53) satisfying (55). If $\phi \neq 0$ and $\phi \neq \pi$, then

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

Since the left-hand side of this identity converges to 0 as $h \to 0$, $n \to \infty$, $nh = x$, the same must be true of the right-hand side; therefore,

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

This implies that $r \leq 1$. In other words, we have proved that any root of the first characteristic polynomial of (38) lies in the closed unit disc.

Next we prove that any root of the first characteristic polynomial of (38) that lies on the unit circle must be *simple*. Assume, instead, that $z = re^{i\phi}$, is a *multiple* root of $\rho(z)$, with $|z| = 1$ (and therefore $r = 1$) and $0 \leq \phi < 2\pi$. We shall prove below that this contradicts our assumption that the method (53) is convergent. It is easy to check that the numbers

$$y_n = h^{1/2} n r^n \cos(n\phi) \tag{56}$$

define a solution to (53), which satisfies (55) for

$$|\eta_s(h)| = |y_s| \leq h^{1/2} s \leq h^{1/2}(k-1), \qquad s = 0, \ldots k-1.$$

If $\phi = 0$ or $\phi = \pi$, it follows from (56) with $h = x/n$ that

$$|y_n| = x^{1/2} n^{1/2} r^n. \tag{57}$$

Since, by assumption, $|z| = 1$ (and therefore $r = 1$), we deduce from (57) that $\lim_{n\to\infty} |y_n| = \infty$, which contradicts (54).

If, on the other hand, $\phi \neq 0$ and $\phi \neq \pi$, then

$$\frac{z_n^2 - z_{n+1}z_{n-1}}{\sin^2 \phi} = r^{2n}, \tag{58}$$

where $z_n = n^{-1}h^{-1/2}y_n = h^{1/2}x^{-1}y_n$. Since, by (54), $\lim_{n\to\infty} z_n = 0$, it follows that the left-hand side of (58) converges to 0 as $n \to \infty$. But then the same must be true of the right-hand side of (58); however, the right-hand side of (58) cannot converge to 0 as $n \to \infty$, since $|r| = 1$ (and hence $r = 1$). Thus, again, we have reached a contradiction.

To summarise, we have proved that all roots of the first characteristic polynomial $\rho(z)$ of the linear multi-step method (38) lie in the unit disc $|z| \leq 1$, and those which belong to the unit circle $|z| = 1$ are simple. By virtue of Theorem 6 the linear multi-step method is zero-stable. $\diamond$

Theorem 8 *A necessary condition for the convergence of the linear multi-step method (38) is that it be consistent.*

PROOF: Let us suppose that the linear multi-step method (38) is convergent; we wish to show that it is then consistent.

Let us first show that $C_0 = 0$. We consider the initial-value problem $y' = 0$, $y(0) = 1$, on the interval $[0, X_M]$, $X_M > 0$, whose solution is, trivially, $y(x) \equiv 1$. Applying (38) to this problem yields the difference equation

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n = 0. \tag{59}$$

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

$$\lim_{\substack{h\to 0 \\ nh=x}} y_n = 1. \tag{60}$$

Since in the present case y_n is independent of the choice of h, (60) is equivalent to saying that

$$\lim_{n\to\infty} y_n = 1. \tag{61}$$

Passing to the limit $n \to \infty$ in (59), we deduce that

$$\alpha_k + \alpha_{k-1} + \cdots + \alpha_0 = 0. \tag{62}$$

Recalling the definition of C_0, (62) is equivalent to $C_0 = 0$ (i.e. $\rho(1) = 0$).

In order to show that $C_1 = 0$, we now consider the initial-value problem $y' = 1$, $y(0) = 0$, on the interval $[0, X_M]$, $X_M > 0$, whose solution is $y(x) = x$. The difference equation (38) now becomes

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \cdots + \alpha_0 y_n = h(\beta_k + \beta_{k-1} + \cdots + \beta_0), \tag{63}$$

where $X_M - x_0 = X_M - 0 = Nh$ and $1 \le n \le N - k$. For a convergent method every solution of (63) satisfying

$$\lim_{h\to 0} \eta_s(h) = 0, \qquad s = 0, 1, \ldots k-1, \tag{64}$$

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

$$\lim_{\substack{h\to 0 \\ nh=x}} y_n = x. \tag{65}$$

Since according to the previous theorem zero-stability is necessary for convergence, we may take it for granted that the first characteristic polynomial $\rho(z)$ of the method does not have a multiple root on the unit circle $|z| = 1$; therefore

$$\rho'(1) = k\alpha_k + \cdots + 2\alpha_2 + \alpha_1 \neq 0.$$

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

$$K = \frac{\beta_k + \cdots + \beta_1 + \beta_0}{k\alpha_k + \cdots + 2\alpha_2 + \alpha_1}; \tag{66}$$

this sequence clearly satisfies (64) and is the solution of (63). Furthermore, (65) implies that

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

and therefore $K = 1$. Hence, from (66),

$$C_1 = (k\alpha_k + \cdots + 2\alpha_2 + \alpha_1) - (\beta_k + \cdots + \beta_1 + \beta_0) = 0;$$

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

3.4.2 Sufficient conditions for convergence

We begin by establishing some preliminary results.

Start of optional material

Lemma 1 *Suppose that all roots of the polynomial $\rho(z) = \alpha_k z^k + \cdots + \alpha_1 z + \alpha_0$ lie in the closed unit disk $|z| \le 1$ and those which lie on the unit circle $|z| = 1$ are simple. Assume further that the numbers γ_l, $l = 0, 1, 2, \ldots$, are defined by*

$$\frac{1}{\alpha_k + \cdots + \alpha_1 z^{k-1} + \alpha_0 z^k} = \gamma_0 + \gamma_1 z + \gamma_2 z^2 + \cdots \ .$$

Then, $\Gamma \equiv \sup_{l\ge 0} |\gamma_l| < \infty$.

PROOF: Let us define $\hat{\rho}(z) = z^k \rho(1/z)$ and note that, by virtue of our assumptions about the roots of $\rho(z)$, the function $1/\hat{\rho}(z)$ is holomorphic in the open unit disc $|z| < 1$. As the roots $z_1, z_2, \ldots, z_m$ of $\rho(z)$ on $|z| = 1$ are simple, the same is true of the poles of $1/\hat{\rho}(z)$, and there exist constants $A_1, \ldots, A_m$ such that the function

$$f(z) = \frac{1}{\hat{\rho}(z)} - \frac{A_1}{z - \frac{1}{z_1}} - \cdots - \frac{A_m}{z - \frac{1}{z_m}} \tag{67}$$

is holomorphic for $|z| < 1$ and $|f(z)| \leq M$ for all $|z| \leq 1$. Thus by Cauchy's estimate[8] the coefficients of the Taylor expansion of f at $z = 0$ also form a bounded sequence. As

$$-\frac{A_i}{z - \frac{1}{z_i}} = A_i \sum_{l=0}^{\infty} z_i^l z^l, \qquad i = 1, \ldots, m,$$

and $|z_i| \leq 1$, we deduce from (67) that the coefficients in the Taylor series expansion of $1/\hat{\rho}(z)$ form a bounded sequence, which completes the proof. $\diamond$

Now we shall apply Lemma 1 to estimate the solution of the linear difference equation

$$\alpha_k e_{m+k} + \alpha_{k-1} e_{m+k-1} + \cdots + \alpha_0 e_0 = h(\beta_{k,m} e_{m+k} + \beta_{k-1,m} e_{m+k-1} + \cdots + \beta_{0,m} e_m) + \lambda_m. \tag{68}$$

The result is stated in the next Lemma.

Lemma 2 *Suppose that all roots of the polynomial $\rho(z) = \alpha_k z^k + \cdots + \alpha_1 z + \alpha_0$ lie in the closed unit disk $|z| \leq 1$ and those which lie on the unit circle $|z| = 1$ are simple. Let B^* and Λ denote nonnegative constants and β a positive constant such that*

$$|\beta_{k,n}| + |\beta_{k-1,n}| + \cdots + |\beta_{0,n}| \leq B^*,$$

$$|\beta_{k,n}| \leq \beta, \qquad |\lambda_n| \leq \Lambda, \qquad n = 0, 1, \ldots, N,$$

and let $0 \leq h < |\alpha_k| \beta^{-1}$. Then every solution of (68) for which

$$|e_s| \leq E, \qquad s = 0, 1, \ldots, k-1,$$

satisfies

$$|e_n| \leq K^* \exp(nhL^*), \qquad n = 0, 1, \ldots, N,$$

where

$$L^* = \Gamma^* B^*, \quad K^* = \Gamma^*(N\Lambda + AEk), \quad \Gamma^* = \Gamma/(1 - h|\alpha_k|^{-1}\beta),$$

Γ is as in Lemma 1, and

$$A = |\alpha_k| + |\alpha_{k-1}| + \cdots + |\alpha_0|.$$

PROOF: For a fixed k we consider the numbers γ_l, $l = 0, 1, \ldots, n-k$, defined in Lemma 1. After multiplying both sides of the equation (68) for $m = n - k - l$ by γ_l, $l = 0, 1, \ldots, n-k$ and summing the resulting equations, on denoting by S_n the sum, we find by manipulating the left-hand side in the sum that

$$\begin{aligned} S_n &= (\alpha_k e_n + \alpha_{k-1} e_{n-1} + \cdots + \alpha_0 e_{n-k})\gamma_0 \\ &\quad + (\alpha_k e_{n-1} + \alpha_{k-1} e_{n-2} + \cdots + \alpha_0 e_{n-k-1})\gamma_1 + \cdots \\ &\quad + (\alpha_k e_k + \alpha_{k-1} e_{k-1} + \cdots + \alpha_0 e_0)\gamma_{n-k} \\ &= \alpha_k \gamma_0 e_n + (\alpha_k \gamma_1 + \alpha_{k-1}\gamma_0) e_{n-1} + \cdots \\ &\quad + (\alpha_k \gamma_{n-k} + \alpha_{k-1}\gamma_{n-k-1} + \cdots + \alpha_0 \gamma_{n-2k}) e_k \\ &\quad + (\alpha_{k-1}\gamma_{n-k} + \cdots + \alpha_0 \gamma_{n-2k+1}) e_{k-1} + \cdots \\ &\quad + \alpha_0 \gamma_{n-k} e_0. \end{aligned}$$

[8] **Theorem (Cauchy's Estimate)** If f is a holomorphic function in the open disc $D(a, R)$, centre a and radius R, and $|f(z)| \leq M$ for all $z \in D(a, R)$, then $|f^{(n)}(a)| \leq M(n!/R^n)$, $n = 0, 1, 2, \ldots$. [For a proof of this result see, for example, *Walter Rudin: Real and Complex Analysis. 3rd edition. McGraw-Hill, New York, 1986.*]

Defining $\gamma_l = 0$ for $l < 0$ and noting that

$$\alpha_k\gamma_l + \alpha_{k-1}\gamma_{l-1} + \cdots + \alpha_0\gamma_{l-k} = \begin{cases} 1 & \text{for } l = 0, \\ 0 & \text{for } l > 0, \end{cases} \tag{69}$$

we have that

$$S_n = e_n + (\alpha_{k-1}\gamma_{n-k} + \cdots + \alpha_0\gamma_{n-2k+1})e_{k-1} + \cdots + \alpha_0\gamma_{n-k}e_0.$$

By manipulating similarly the right-hand side in the sum, we find that

$$\begin{aligned} &e_n + (\alpha_{k-1}\gamma_{n-k} + \cdots + \alpha_0\gamma_{n-2k+1})e_{k-1} + \cdots + \alpha_0\gamma_{n-k}e_0 \\ &\quad = h\,[\beta_{k,n-k}\gamma_0 e_n + (\beta_{k-1,n-k}\gamma_0 + \beta_{k,n-k-1}\gamma_1)e_{n-1} + \cdots \\ &\qquad + (\beta_{0,n-k}\gamma_0 + \cdots + \beta_{k,n-2k}\gamma_k)e_{n-k} + \cdots + \beta_{0,0}\gamma_{n-k}e_0] \\ &\qquad + (\lambda_{n-k}\gamma_0 + \lambda_{n-k-1}\gamma_1 + \cdots + \lambda_0\gamma_{n-k}). \end{aligned}$$

Thus, by our assumptions and noting that by (69) $\gamma_0 = \alpha_k^{-1}$, we have that

$$|e_n| \le h\beta|\alpha_k^{-1}|\,|e_n| + h\Gamma B^* \sum_{m=0}^{n-1} |e_m| + N\Gamma\Lambda + A\Gamma Ek.$$

Hence,

$$(1 - h\beta|\alpha_k^{-1}|)|e_n| \le h\Gamma B^* \sum_{m=0}^{n-1} |e_m| + N\Gamma\Lambda + A\Gamma Ek.$$

Recalling the definitions of L^* and K^* we can rewrite the last inequality as follows:

$$|e_n| \le K^* + hL^* \sum_{m=0}^{n-1} |e_m|, \qquad n = 0, 1, \ldots, N. \tag{70}$$

The final estimate is deduced from (70) by induction. First, we note that by virtue of (69), $A\Gamma \ge 1$. Consequently, $K^* \ge \Gamma AEk \ge Ek \ge E$. Now, letting $n = 1$ in (70),

$$|e_1| \le K^* + hL^*|e_0| \le K^* + hL^*E \le K^*(1 + hL^*).$$

Repeating this argument, we find that

$$|e_m| \le K^*(1 + hL^*)^m, \qquad m = 0, \ldots, k-1.$$

Now suppose that this inequality has already been shown to hold for $m = 0, 1, \ldots, n-1$, where $n \ge k$; we shall prove that it then also holds for $m = n$, which will complete the induction. Indeed, we have from (70) that

$$|e_n| \le K^* + hL^*K^*\frac{(1 + hL^*)^n - 1}{hL^*} = K^*(1 + hL^*)^n. \tag{71}$$

Further, as $1 + hL^* \le \mathrm{e}^{hL^*}$ we have from (71) that

$$|e_n| \le K^*\mathrm{e}^{hL^*n}, \qquad n = 0, 1, \ldots, N. \tag{72}$$

That completes the proof of the lemma. We remark that the implication (70) $\Rightarrow$ (72) is usually referred to as the **Discrete Gronwall Lemma**. $\diamond$

Using Lemma 2 we can now show that zero-stability and consistency, which have been shown to be necessary are also sufficient conditions for convergence.

Theorem 9 *For a linear multi-step method that is consistent with the ordinary differential equation (1) where f is assumed to satisfy a Lipschitz condition, and starting with consistent starting conditions, zero-stability is sufficient for convergence.*

PROOF: Let us define

$$\delta = \delta(h) = \max_{0\le s\le k-1} |\eta_s(h) - y(a+sh)|;$$

because the starting conditions $y_s = \eta_s(h)$, $s = 0, \ldots, k-1$, are assumed to be consistent, we have that $\lim_{h\to 0} \delta(h) = 0$. We have to prove that

$$\lim_{\substack{n\to\infty \\ nh = x - x_0}} y_n = y(x)$$

for all x in the interval $[x_0, X_M]$. We begin the proof by estimating the consistency error of (38):

$$T_n = \frac{1}{h\sigma(1)} \left[\sum_{j=0}^{k} \alpha_j y(x_{n+j}) - h\beta_j y'(x_{n+j}) \right]. \tag{73}$$

As $y' \in C[x_0, X_M]$, it makes sense to define, for $\varepsilon \ge 0$, the function

$$\chi(\varepsilon) = \max_{\substack{|x^*-x|\le\varepsilon \\ x,\, x^*\in[x_0,X_M]}} |y'(x^*) - y'(x)|.$$

For $s = 0, 1, \ldots, k-1$, we can then write

$$y'(x_{n+s}) = y'(x_n) + \theta_s \chi(sh),$$

where $|\theta_s| \le 1$. Further, by the Mean-Value Theorem, there exists a $\xi_s \in (x_n, x_{n+s})$ such that

$$y(x_{n+s}) = y(x_n) + shy'(\xi_s).$$

Thus,

$$y(x_{m+s}) = y(x_m) + sh\left[y'(x_m) + \theta'_s \chi(sh)\right],$$

where $|\theta'_s| \le 1$.

Now we can write

$$\begin{aligned} |\sigma(1)T_n| \;\le\; & \big|h^{-1}(\alpha_1 + \alpha_2 + \cdots + \alpha_k)y(x_n) + (\alpha_1 + 2\alpha_2 + \cdots + k\alpha_k)y'(x_n) \\ & - (\beta_0 + \beta_1 + \cdots + \beta_k)y'(x_n)\big| \\ & + (|\alpha_1| + 2|\alpha_2| + \cdots + k|\alpha_k|)|\chi(kh)| + (|\beta_0| + |\beta_1| + \cdots + |\beta_k|)|\chi(kh)|. \end{aligned}$$

Since the method has been assumed consistent, the first, second, and third terms on the right-had side cancel, giving

$$|\sigma(1)T_n| \le (|\alpha_1| + 2|\alpha_2| + \cdots + k|\alpha_k|)|\chi(kh)| + (|\beta_0| + |\beta_1| + \cdots + |\beta_k|)|\chi(kh)|.$$

Thus,

$$|\sigma(1)T_n| \le K\chi(kh), \tag{74}$$

where

$$K = |\alpha_1| + 2|\alpha_2| + \cdots + k|\alpha_k| + |\beta_0| + |\beta_1| + \cdots + |\beta_k|.$$

Comparing (38) with (73), we deduce that the global error $e_m = y(x_m) - y_m$ satisfies

$$\alpha_k e_{m+k} + \cdots + \alpha_0 e_0 = h\left(\beta_k g_{m+k} e_{m+k} + \cdots + \beta_0 g_m e_m\right) + \sigma(1)T_n h,$$

where

$$g_m = \begin{cases} [f(x_m, y(x_m)) - f(x_m, y_m)]/e_m, & e_m \neq 0, \\ 0, & e_m = 0. \end{cases}$$

By virtue of (74), we then have that

$$\alpha_k e_{m+k} + \cdots + \alpha_0 e_0 = h\left(\beta_k g_{m+k} e_{m+k} + \cdots + \beta_0 g_m e_m\right) + \theta K \chi(kh) h.$$

As f is assumed to satisfy the Lipschitz condition (3) we have that $|g_m| \leq L$, $m = 0, 1, \ldots$. By applying Lemma 2 with $E = \delta(h)$, $\Lambda = K\chi(kh)h$, $N = (X_M - x_0)/h$, $B^* = BL$, where $B = |\beta_0| + |\beta_1| + \cdots + |\beta_k|$, we find that

$$|e_n| \leq \Gamma^* \left[Ak\delta(h) + (X_M - x_0)K\chi(kh)\right] \exp[(x_n - x_0)L\Gamma^* B], \tag{75}$$

where

$$A = |\alpha_0| + |\alpha_1| + \cdots + |\alpha_k|, \qquad \Gamma^* = \frac{\Gamma}{1 - h|\alpha_k^{-1}\beta_k|L}.$$

Now, y' is a continuous function on the closed interval $[x_0, X_M]$; therefore it is uniformly continuous on $[x_0, X_M]$. Thus, $\chi(kh) \to 0$ as $h \to 0$; also, by virtue of the assumed consistency of the starting values, $\delta(h) \to 0$ as $h \to 0$. Passing to the limit $h \to 0$ in (75), we deduce that

$$\lim_{\substack{n\to\infty \\ x - x_0 = nh}} |e_n| = 0;$$

equivalently,

$$\lim_{\substack{n\to\infty \\ x - x_0 = nh}} |y(x) - y_n| = 0$$

so the method is convergent. $\diamond$

End of optional material

By combining Theorems 7, 8 and 9, we arrive at the following important result.

Theorem 10 (**Dahlquist**) *For a linear multi-step method that is consistent with the ordinary differential equation (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 $y(x)$ has continuous derivative 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 linear multi-step method 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 linear multi-step method which will grow by an arbitrarily large factor in a fixed interval of x, however accurate the starting conditions are. This result 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

Start of optional material

Let us suppose that we have already chosen the coefficients α_j, $j = 0, \ldots, k$, of the k-step method (38). The question we shall be concerned with in this section is how to choose the coefficients β_j, $j = 0, \ldots, k$, so that the order of accuracy of the resulting method (38) is as high as possible.

In view of Theorem 10 we shall only be interested in consistent methods, so it is natural to assume that the first and second characteristic polynomials $\rho(z)$ and $\sigma(z)$ associated with (38) satisfy $\rho(1) = 0$, $\rho'(1) - \sigma(1) = 0$, with $\sigma(1) \neq 0$ (the last condition being required for the sake of correctness of the definition of the consistency error (47)).

Consider the function ϕ of the complex variable z, defined by

$$\phi(z) = \frac{\rho(z)}{\log z} - \sigma(z); \tag{76}$$

the function $\log z$ appearing in the denominator is made single-valued by cutting the complex plane along the half-line $\mathrm{Re}\, z \leq 0$. We begin our analysis with the following fundamental lemma.

Lemma 3 *Suppose that p is a positive integer. The linear multistep method (38), with stability polynomials $\rho(z)$ and $\sigma(z)$, is of order of accuracy p if, and only if, the function $\phi(z)$ defined by (76) has a zero of multiplicity p at $z = 1$.*

Proof: Let us suppose that the k-step method (38) for the numerical approximation of the initial-value problem (1), (2) is of order p. Then, for any sufficiently smooth function $f(x, y)$, the resulting solution to (1), (2) yields a consistency error of the form:

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

as $h \to 0$, $C_{p+1} \neq 0$, $x_n = x_0 + nh$. In particular, for the initial-value problem

$$y' = y, \qquad y(0) = 1,$$

we get

$$T_n \equiv \frac{\mathrm{e}^{nh}}{h\sigma(1)} \Big[\rho(\mathrm{e}^h) - h\sigma(\mathrm{e}^h)\Big] = \mathrm{e}^{nh} \frac{C_{p+1}}{\sigma(1)} h^p + \mathcal{O}(h^{p+1}), \tag{77}$$

as $h \to 0$, $C_{p+1} \neq 0$. Thus, the function

$$f(h) = \frac{1}{h} \Big[\rho(\mathrm{e}^h) - h\sigma(\mathrm{e}^h)\Big]$$

is holomorphic in a neighbourhood of $h = 0$ and has a zero of order p at $h = 0$. The function $z = \mathrm{e}^h$ is a bijective mapping of a neighbourhood of $h = 0$ onto a neighbourhood of $z = 1$. Therefore $\phi(z)$ is holomorphic in a neighbourhood of $z = 1$ and has a zero of multiplicity p at $z = 1$.

Conversely, suppose that $\phi(z)$ has a zero of multiplicity p at $z = 1$. Then $f(h) = \phi(\mathrm{e}^h)$ is a holomorphic function in the vicinity of $h = 0$ and has a zero of multiplicity p at $h = 0$. Therefore,

$$g(h) = \sum_{j=0}^{k} (\alpha_j - h\beta_j)\mathrm{e}^{jh}$$

has a zero of multiplicity $(p+1)$ at $h = 0$, implying that $g(0) = g'(0) = \cdots = g^{(p)}(0) = 0$, but $g^{(p+1)}(0) \neq 0$. First,

$$g(0) = 0 = \sum_{j=0}^{k} \alpha_j = C_0.$$

Now, by successive differentiation of g with respect to h,

$$g'(0) = 0 = \sum_{j=0}^{k} (j\alpha_j - \beta_j) = C_1,$$
$$g''(0) = 0 = \sum_{j=0}^{k} (j^2\alpha_j - 2j\beta_j) = 2C_2,$$
$$g'''(0) = 0 = \sum_{j=0}^{k} (j^3\alpha_j - 3j^2\beta_j) = 6C_3,$$
etc.
$$g^{(p)}(0) = 0 = \sum_{j=0}^{k} (j^p\alpha_j - pj^{p-1}\beta_j) = p!\, C_p.$$

We deduce that $C_0 = C_1 = C_2 = \cdots = C_p = 0$; since $g^{(p+1)}(0) \neq 0$ we have that $C_{p+1} \neq 0$. Consequently (38) is of order of accuracy p. $\diamond$

We shall use this lemma in the next theorem to supply a lower bound for the maximum order of a linear multistep method with prescribed first stability polynomial $\rho(z)$.

Theorem 11 *Suppose that $\rho(z)$ is a polynomial of degree k such that $\rho(1) = 0$ and $\rho'(1) \neq 0$, and let $\hat{k}$ be an integer such that $0 \leq \hat{k} \leq k$. Then, there exists a unique polynomial $\sigma(z)$ of degree $\hat{k}$ such that $\rho'(1) - \sigma(1) = 0$ and the order of the linear multi-step method associated with $\rho(z)$ and $\sigma(z)$ is $\geq \hat{k} + 1$.*

PROOF: Since the function $\rho(z)/\log(z)$ is holomorphic in the neighbourhood of $z = 1$ it can be expanded into a convergent Taylor series:

$$\frac{\rho(z)}{\log z} = c_0 + c_1(z-1) + c_2(z-1)^2 + \cdots .$$

By multiplying both sides by $\log z$ and differentiating we deduce that $c_0 = \rho'(1)\,(\neq 0)$. Let us define

$$\sigma(z) = c_0 + c_1(z-1) + \cdots + c_{\hat{k}}(z-1)^{\hat{k}}.$$

Clearly $\sigma(1) = c_0 = \rho'(1)\,(\neq 0)$. With this definition,

$$\phi(z) = \frac{\rho(z)}{\log z} - \sigma(z) = c_{\hat{k}+1}(z-1)^{\hat{k}+1} + \cdots ,$$

and therefore $\phi(z)$ has a zero at $z = 1$ of multiplicity not less than $\hat{k} + 1$. By Lemma 3 the linear k-step method associated with $\rho(z)$ and $\sigma(z)$ is of order $\geq \hat{k} + 1$.

The uniqueness of $\sigma(z)$ possessing the desired properties follows from the uniqueness of the Taylor series expansion of $\phi(z)$ about the point $z = 1$. $\diamond$

We note in connection with this theorem that for most methods of practical interest either $\hat{k} = k - 1$ resulting in an explicit method or $\hat{k} = k$ corresponding to an implicit method. In the next example we shall encounter the latter situation.

Example 5 *Consider a linear two-step method with $\rho(z) = (z-1)(z-\lambda)$. The method will be zero-stable as long as $\lambda \in [-1, 1)$. Consider the Taylor series expansion of the function $\rho(z)/\log(z)$ about the point*

$z = 1$:

$$\begin{aligned}
\frac{\rho(z)}{\log z} &= \frac{(z-1)(1-\lambda+(z-1))}{\log[1+(z-1)]} \\
&= [1-\lambda+(z-1)] \times \left\{1 - \frac{z-1}{2} + \frac{(z-1)^2}{3} - \frac{(z-1)^3}{4} + \mathcal{O}((z-1)^4)\right\}^{-1} \\
&= [1-\lambda+(z-1)] \times \left\{1 + \frac{z-1}{2} - \frac{(z-1)^2}{12} + \frac{(z-1)^3}{24} + \mathcal{O}((z-1)^4)\right\} \\
&= 1-\lambda+\frac{3-\lambda}{2}(z-1) + \frac{5+\lambda}{12}(z-1)^2 - \frac{1+\lambda}{24}(z-1)^3 + \mathcal{O}((z-1)^4).
\end{aligned}$$

A two-step method of maximum order is obtained by selecting

$$\begin{aligned}
\sigma(z) &= 1-\lambda+\frac{3-\lambda}{2}(z-1) + \frac{5+\lambda}{12}(z-1)^2 \\
&= -\frac{1+5\lambda}{12} + \frac{2-2\lambda}{3}z + \frac{5+\lambda}{12}z^2.
\end{aligned}$$

If $\lambda \neq -1$, the resulting method is of third order, with error constant

$$C_4 = -\frac{1+\lambda}{24},$$

whereas if $\lambda = -1$ the method is of fourth order.

In the former case the method is

$$y_{n+2} - (1+\lambda)y_{n+1} + \lambda y_n = h\left(\frac{5+\lambda}{12}f_{n+2} + \frac{2-2\lambda}{3}f_{n+1} - \frac{1+5\lambda}{12}f_n\right)$$

with λ a parameter contained in the interval $(-1, 1)$. In the latter case, the method has the form

$$y_{n+2} - y_n = \frac{h}{3}(f_{n+2} + 4f_{n+1} + f_n),$$

and is referred to as the Simpson rule method.

By inspection, the linear k-step method (38) has $2k+2$ coefficients: α_j, β_j, $j = 0, \ldots, k$, 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 (48), if the method is required to have order p, $p+1$ linear relationships $C_0 = 0, \ldots, C_p = 0$ involving α_j, β_j, $j = 0, \ldots, k$, must be satisfied. Thus, in the case of the implicit method, we can impose $p+1 = 2k+1$ linear constraints $C_0 = 0, \ldots, C_{2k+1} = 0$ 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 which is of order $> (k+2)$. Therefore the maximum orders $2k$ and $2k-1$ cannot be attained without violating the condition of zero-stability. We formalise these facts in the next theorem.

Theorem 12 *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.*

PROOF: Consider a linear k-step method (38) with associated first and second stability polynomials ρ and σ. Further, consider the transformation

$$\zeta \in \mathbf{C} \mapsto \frac{\zeta - 1}{\zeta + 1} \in \mathbf{C},$$

which maps the open unit disc $|\zeta| < 1$ of the ζ-plane onto the left open complex half-plane $\operatorname{Re} z < 0$ of the z-plane; the circle $|\zeta| = 1$ is mapped onto the imaginary axis $\operatorname{Re} z = 0$, the point $\zeta = 1$ onto $z = 0$, and the point $\zeta = -1$ onto $\zeta = \infty$.

It is clear that the functions r and s defined by

$$r(z) = \left(\frac{1-z}{2}\right)^k \rho\left(\frac{1+z}{1-z}\right), \qquad s(z) = \left(\frac{1-z}{2}\right)^k \rho\left(\frac{1+z}{1-z}\right),$$

are in fact polynomials, $\deg(r) \le k$ and $\deg(s) \le k$.

If $\rho(\zeta)$ has a root of multiplicity p, $0 \le p \le k$, at $\zeta = \zeta_0 \ne -1$, then $r(z)$ has a root of the same multiplicity at $z = (\zeta_0 - 1)/(\zeta_0 + 1)$; if $\rho(\zeta)$ has a root of multiplicity $p \ge 1$, $0 \le p \le k$, at $\zeta = -1$, then $r(z)$ is of degree $k - p$.

Since, by assumption, the method is zero-stable, $\zeta = 1$ is a simple root of $\rho(\zeta)$; consequently, $z = 0$ is a simple root of $r(z)$. Thus,

$$r(z) = a_1 z + a_2 z^2 + \cdots + a_k z^k, \qquad a_1 \ne 0, \qquad a_j \in \mathbf{R}.$$

It can be assumed, without loss of generality, that $a_1 > 0$. Since by zero stability all roots of $\rho(\zeta)$ are contained in the closed unit disc, we deduce that all roots of $r(z)$ have real parts ≤ 0. Therefore, all coefficients a_j, $j = 1, \ldots, k$, of $r(z)$ are nonnegative.

Now let us consider the function

$$q(z) = \left(\frac{1-z}{2}\right)^k \phi\left(\frac{1+z}{1-z}\right) = \frac{1}{\log \frac{1+z}{1-z}}\, r(z) - s(z).$$

The function $q(z)$ has a zero of multiplicity p at $z = 0$ if, and only if, $\phi(\zeta)$ defined by (76) has a zero of multiplicity p at $\zeta = 1$; according to Lemma 3 this is equivalent to the linear k-step method associated with $\rho(\zeta)$ and $\sigma(\zeta)$ having order p. Thus if the linear k-step method associated with $\rho(z)$ and $\sigma(z)$ has order p then

$$s(z) = b_0 + b_1 z + b_2 z^2 + \cdots + b_{p-1} z^{p-1},$$

where

$$\frac{z}{\log \frac{1+z}{1-z}}\, \frac{r(z)}{z} = b_0 + b_1 z + b_2 z^2 + \cdots .$$

As the degree of $s(z)$ is $\le k$, the existence of a consistent zero-stable k-step linear multistep method of order $p > k + 1$ (or $p > k + 2$) now hinges on the possibility that

$$b_{k+1} = \cdots = b_{p-1} = 0, \qquad (\text{or } b_{k+2} = \cdots = b_{p-1} = 0).$$

Let us consider whether this is possible.

We denote by c_0, c_1, $c_2, \ldots$, the coefficients in the Taylor series expansion of the function

$$\frac{z}{\log \frac{1+z}{1-z}},$$

namely,

$$\frac{z}{\log \frac{1+z}{1-z}} = c_0 + c_2 z^2 + c_4 z^4 + \cdots .$$

Then, adopting the notational convention that $a_\nu = 0$ for $\nu > k$, we have that

$$\begin{aligned}
b_0 &= c_0 a_0, \\
b_1 &= c_0 a_2, \\
&\text{etc.} \\
b_{2\nu} &= c_0 a_{2\nu+1} + c_2 a_{2\nu-1} + \cdots + c_{2\nu} a_1, \\
b_{2\nu+1} &= c_0 a_{2\nu+2} + c_2 a_{2\nu} + \cdots + c_{2\nu} a_2, \qquad \nu = 1, 2, \ldots .
\end{aligned}$$

It is a straightforward matter to prove that $c_{2\nu} < 0$, $\nu = 1, 2, \ldots$ (see also Lemma 5.4 on page p.233 of Henrici's book).

(i) If k is an odd number, then, since $a_\nu = 0$ for $\nu > k$, we have that

$$b_{k+1} = c_2 a_k + c_4 a_{k-2} + \cdots + c_{k+1} a_1.$$

Since $a_1 > 0$ and no a_ν is negative, it follows that $b_{k+1} < 0$.

(ii) If k is an even number, then

$$b_{k+1} = c_2 a_k + c_4 a_{k-2} + \cdots + c_k a_2.$$

Since $c_{2\nu} < 0$, $\nu = 1, 2, \ldots$, and $a_\mu \geq 0$, $\mu = 2, 3, \ldots, k$, we deduce that $b_{k+1} = 0$ if, and only if, $a_2 = a_4 = \cdots = a_k = 0$, i.e. when $r(z)$ is an odd function of z. This, together with the fact that all roots of $r(z)$ have real part ≤ 0, implies that all roots of $r(z)$ mush have real part equal to zero. Consequently, all roots of $\rho(\zeta)$ lie on $|\zeta| = 1$. Since $a_k = 0$, the degree of $r(z)$ is $k-1$, and therefore -1 is a (simple) root of $\rho(\zeta)$.

As $c_{2\nu} < 0$, $a_\mu \geq 0$ and $a_1 > 0$, it follows that

$$b_{k+2} = c_4 a_{k-1} + c_6 a_{k-3} + \cdots + c_{k+2} a_1 < 0,$$

showing that $b_{k+2} \neq 0$.

Thus if a k-step method is zero-stable and k is odd then $b_{k+1} \neq 0$, whereas if k is even then b_{k+2}. This proves that there is no zero-stable k-step method whose order exceeds $k+1$ if k is odd or $k+2$ if k is even. $\diamond$

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

According to the proof of the previous theorem, all roots of the first characteristic polynomial ρ associated with an optimal linear multistep method have modulus 1.

Example 6 *As $k + 2 = 2k$ if an only if $k = 2$ and the Simpson rule method is the zero-stable linear 2-step method of maximum order, we deduce that the Simpson rule method is the only zero-stable linear multistep method 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.

End of optional material

3.6 Absolute stability of linear multistep methods

Lecture 8

Up to now we have been concerned with the stability and accuracy properties of linear multistep methods 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 initial-value problem whose solution decays to zero as $x \to \infty$, the linear multistep method 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

$$y' = \lambda y, \quad x > 0, \qquad y(0) = y_0 \ (\neq 0), \tag{78}$$

where $\operatorname{Re}\lambda < 0$. Indeed,

$$y(x) = y_0 \mathrm{e}^{\imath x \operatorname{Im}\lambda} \mathrm{e}^{x \operatorname{Re}\lambda},$$

and therefore,

$$|y(x)| \leq |y_0| \, \exp(-x|\operatorname{Re}\lambda|), \qquad x \geq 0,$$

yielding $\lim_{x\to\infty} y(x) = 0$. Thus, using the terminology introduced in the last paragraph of Section 1, the solution is asymptotically stable.

In the rest of the section we shall assume, for simplicity, that λ is a negative real number, but everything we shall say extends straightforwardly to the general case of λ complex, with $\operatorname{Re}\lambda < 0$.

Now consider the linear k-step method (38) and apply it to the model problem (78) with λ real and negative. This yields the linear difference equation

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

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

$$\pi(z;\bar{h}) = \rho(z) - \bar{h}\sigma(z), \qquad (\bar{h} = h\lambda), \tag{79}$$

it follows that y_n will converge to zero for $h > 0$ fixed and $n \to \infty$ if, and only if, all roots of $\pi(z;\bar{h})$ have modulus < 1. The kth degree polynomial $\pi(z;\bar{h})$ defined by (79) is called the **stability polynomial** of the linear k-step method with first and second characteristic polynomials $\rho(z)$ and $\sigma(z)$, respectively. This motivates the following definition.

Definition 10 *The linear multistep method (38) is called* ***absolutely stable*** *for a given $\bar{h}$ if, and only if, for that $\bar{h}$ all the roots $r_s = r_s(\bar{h})$ of the stability polynomial $\pi(z,\bar{h})$ defined by (79) satisfy $|r_s| < 1$, $s = 1, \ldots, k$. Otherwise, the method is said to be* ***absolutely unstable****. An interval (α, β) of the real line is called the* ***interval of absolute stability*** *if 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 (78) exhibits exponential growth, it is reasonable to expect that a consistent and zero-stable (and, therefore, convergent) linear multistep method will have a similar behaviour for $h > 0$ sufficiently small, and will be therefore absolutely unstable for small $\bar{h} = \lambda h$. According to the next theorem, this is indeed the case.

Theorem 13 *Every consistent and zero-stable linear multistep method is absolutely unstable for small positive $\bar{h}$.*

PROOF: Because the method is consistent, there exists an integer $p \geq 1$ such that $C_0 = C_1 = \cdots = C_p = 0$

and $C_{p+1} \neq 0$. Let us consider

$$\begin{aligned}
\pi(\mathrm{e}^{\bar{h}};\bar{h}) &= \rho(\mathrm{e}^{\bar{h}}) - \bar{h}\sigma(\mathrm{e}^{\bar{h}}) = \sum_{j=0}^{k}\left[\alpha_j \mathrm{e}^{\bar{h}j} - \bar{h}\beta_j \mathrm{e}^{\bar{h}j}\right] \\
&= \sum_{j=0}^{k}\left[\alpha_j \sum_{q=0}^{\infty}\frac{(\bar{h}j)^q}{q!} - \beta_j \sum_{q=0}^{\infty}\frac{\bar{h}^{q+1}j^q}{q!}\right] \\
&= \sum_{j=0}^{k}\left[\alpha_j \sum_{q=0}^{\infty}\frac{(\bar{h}j)^q}{q!} - \beta_j \sum_{q=1}^{\infty}\frac{\bar{h}^{q}j^{q-1}}{(q-1)!}\right] \\
&= \sum_{j=0}^{k}\alpha_j + \sum_{j=0}^{k}\left[\alpha_j \sum_{q=1}^{\infty}\frac{(\bar{h}j)^q}{q!} - \beta_j \sum_{q=1}^{\infty}\frac{\bar{h}^{q}j^{q-1}}{(q-1)!}\right] \\
&= \sum_{j=0}^{k}\alpha_j + \sum_{q=1}^{\infty}\bar{h}^q\left[\sum_{j=0}^{k}\alpha_j\frac{j^q}{q!} - \sum_{j=0}^{k}\beta_j\frac{j^{q-1}}{(q-1)!}\right] \\
&= C_0 + \sum_{q=1}^{\infty}\bar{h}^q C_q \\
&= \sum_{q=p+1}^{\infty} C_q \bar{h}^q = \mathcal{O}(\bar{h}^{p+1}). \qquad (80)
\end{aligned}$$

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

$$\pi(z,\bar{h}) = (\alpha_k - \bar{h}\beta_k)(z - r_1)\cdots(z - r_k)$$

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

$$\pi(\mathrm{e}^{\bar{h}};\bar{h}) = (\alpha_k - \bar{h}\beta_k)(\mathrm{e}^{\bar{h}} - r_1(\bar{h}))\cdots(\mathrm{e}^{\bar{h}} - r_k(\bar{h})). \qquad (81)$$

As $\bar{h} \to 0$, $\alpha_k - \bar{h}\beta_k \to \alpha_k \neq 0$ and $r_s(\bar{h}) \to \zeta_s$, $s = 1,\ldots,k$, where ζ_s, $s = 1,\ldots,k$, are the roots of the first stability polynomial $\rho(z)$. Since, by assumption, the method is consistent, 1 is a root of $\rho(z)$; furthermore, by zero-stability 1 is a simple root of $\rho(z)$. 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 (81) 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 (80), $\pi(\mathrm{e}^{\bar{h}};\bar{h}) = \mathcal{O}(\bar{h}^{p+1})$, so it follows that

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

Thus we have shown that

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

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 $k+2$. We have also seen in the proof of Theorem 12 that all roots of the first characteristic polynomial of an optimal k-step method lie on the unit circle. By refining the proof of Theorem 13 it can be shown that an optimal linear multistep method has no interval of absolute stability.

It also follows from Theorem 13 that whenever a consistent zero-stable linear multistep method is used for the numerical solution of the initial-value problem (1), (2) where $\frac{\partial f}{\partial y} > 0$, the error of the method will increase as the computation proceeds.

3.7 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.

3.7.1 The Schur criterion

Consider the polynomial

$$\phi(r) = c_k r^k + \cdots + c_1 r + c_0, \qquad c_k \neq 0, \quad c_0 \neq 0,$$

with complex coefficients. The polynomial ϕ is said to be a **Schur polynomial** if each of its roots r_s satisfies $|r_s| < 1$, $s = 1, \ldots, k$.

Let us consider the polynomial

$$\hat{\phi}(r) = \bar{c}_0 r^k + \bar{c}_1 r^{k-1} + \cdots + \bar{c}_{k-1} r + \bar{c}_k,$$

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

$$\phi_1(r) = \frac{1}{r}\left[\hat{\phi}(0)\phi(r) - \phi(0)\hat{\phi}(r)\right].$$

Clearly ϕ_1 has degree $\leq k-1$.

The following key result is stated without proof.

Theorem 14 (**Schur's Criterion**) *The polynomial ϕ is a Schur polynomial if, and only if, $|\hat{\phi}(0)| > |\phi(0)|$ and ϕ_1 is a Schur polynomial.*

We illustrate Schur's criterion by a simple example.

Exercise 4 *Use Schur's criterion to determine the interval of absolute stability of the linear multistep method*

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

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

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

Let us restrict ourselves to the case when $\lambda \in \mathbb{R}$ with $\lambda < 0$, and therefore $\bar{h} := h\lambda$ is also a (negative) real number. Now,

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

Clearly, $|\hat{\pi}(0;\bar{h})| > |\pi(0,\bar{h})|$ if, and only if, $\bar{h} \in (-\frac{4}{3}, 0)$. As

$$\pi_1(r,\hat{h}) = -\frac{1}{2}\bar{h}\left(2 + \frac{3}{2}\bar{h}\right)(3r+1)$$

has the unique root $-\frac{1}{3}$ and is, therefore, a Schur polynomial, we deduce from Schur's criterion that $\pi(r;\bar{h})$ is a Schur polynomial if, and only if, $\bar{h} \in (-\frac{4}{3}, 0)$. Therefore the interval of absolute stability is $(-\frac{4}{3}, 0)$. $\diamond$

3.7.2 The Routh–Hurwitz criterion

Consider the mapping

$$z = \frac{r-1}{r+1}$$

of the open unit disc $|r| < 1$ of the complex r-plane to the left open complex half-plane $\operatorname{Re} z < 0$ of the complex z-plane. The inverse of this mapping is given by

$$r = \frac{1+z}{1-z}.$$

Under this transformation the function

$$\pi(r,\bar{h}) = \rho(r) - \bar{h}\sigma(r)$$

becomes

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

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

$$a_0 z^k + a_1 z^{k-1} + \cdots + a_k. \tag{82}$$

The roots of the stability polynomial $\pi(r,\bar{h})$ lie inside the open unit disk $|r| < 1$ if, and only if, the roots of the polynomial (82) lie in the left open complex half-plane $\operatorname{Re} z < 0$.

Theorem 15 (**Routh–Hurwitz Criterion**) *The roots of (82) lie in the left open complex half-plane if, and only if, all the leading principal minors of the $k \times k$ matrix*

$$Q = \begin{bmatrix} a_1 & a_3 & a_5 & \cdots & a_{2k-1} \\ a_0 & a_2 & a_4 & \cdots & a_{2k-2} \\ 0 & a_1 & a_3 & \cdots & a_{2k-3} \\ 0 & a_0 & a_2 & \cdots & a_{2k-4} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & a_k \end{bmatrix}$$

are positive and $a_0 > 0$; we assume that $a_j = 0$ if $j > k$. In particular:

a) for $k = 2$: $a_0 > 0$, $a_1 > 0$, $a_2 > 0$;

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

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

represent the necessary and sufficient conditions for ensuring that all roots of (82) lie in the left open complex half-plane.

We illustrate this result by a simple exercise.

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

SOLUTION: By applying the substitution

$$r = \frac{1+z}{1-z}$$

in the stability polynomial

$$\pi(r, \bar{h}) = r^2 - \frac{1}{2}\bar{h}r - \left(1 + \frac{3}{2}\bar{h}\right)$$

and multiplying the resulting function by $(1-z)^2$, we get

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

with

$$a_0 = -\bar{h}, \qquad a_1 = 4 + 3\bar{h}, \qquad a_2 = -2\bar{h}.$$

Applying part a) of Theorem 15 we deduce that the method is zero-stable if, and only if, $\bar{h} \in (-\frac{4}{3}, 0)$; hence 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 of the stated properties is left to the reader as exercise.

k-step Adams–Bashforth (explicit) methods:

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

$$y_1 - y_0 = h f_0;$$

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

$$y_2 - y_1 = \frac{h}{2}(3f_1 - f_0);$$

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

$$y_3 - y_2 = \frac{h}{12}(23f_2 - 16f_1 + 5f_0);$$

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

$$y_4 - y_3 = \frac{h}{24}(55f_3 - 59f_2 + 37f_1 - 9f_0).$$

k-step Adams–Moulton (implicit) methods:

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

$$y_1 - y_0 = \frac{h}{2}(f_1 + f_0);$$

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

$$y_2 - y_1 = \frac{h}{12}(5f_2 + 8f_1 - f_0);$$

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

$$y_3 - y_2 = \frac{h}{24}(9f_3 + 19f_2 - 5f_1 + f_0);$$

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

$$y_4 - y_3 = \frac{h}{720}(251f_4 + 646f_3 - 264f_2 + 106f_1 - 19f_0).$$

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

3.8 Predictor-corrector methods

Let us suppose that we wish to use the implicit linear k-step method

Start of optional material

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j}, \qquad \alpha_k,\ \beta_k \neq 0.$$

Then, at each step we have to solve for y_{n+k} the equation

$$\alpha_k y_{n+k} - h\beta_k f(x_{n+k}, y_{n+k}) = \sum_{j=0}^{k-1} \left(h\beta_j f_{n+j} - \alpha_j y_{n+j}\right).$$

If $h < |\alpha_k|/(L|\beta_k|)$, where L is the Lipschitz constant of f with respect to y (as in Picard's Theorem 1), then this equation has a unique solution, y_{n+k}; moreover, y_{n+k} can be computed by the fixed-point iteration

$$\alpha_k y_{n+k}^{[s+1]} + \sum_{j+0}^{k-1} \alpha_j y_{n+j} = h\beta_k f(x_{n+k}, y_{n+k}^{[s]}) + h \sum_{j=0}^{k-1} \beta_j f_{n+j}, \qquad s = 0, 1, 2, \ldots,$$

with $y_{n+k}^{[0]}$ a suitably chosen starting value.

Theoretically, we would iterate until the iterates $y_{n+k}^{[s]}$ have converged (in practice, until some stopping criterion such as $|y_{n+k}^{[s+1]} - y_{n+k}^{[s]}| < \varepsilon$ is satisfied, where ε is some preassigned tolerance). We would then regard the converged value as an acceptable approximation y_{n+k} to the unknown analytical solution-value $y(x_{n+k})$. This procedure is usually referred to as **correcting to convergence**.

Unfortunately, in practice, such an approach is usually unacceptable because of the amount of work involved: each step of the iteration involves an evaluation of $f(x_{n+k}, y_{n+k}^{[s]})$, which may be quite time-consuming. In order to keep to a minimum the number of times $f(x_{n+k}, y_{n+k}^{[s]})$ is evaluated, the initial guess $y_{n+k}^{[0]}$ must be chosen accurately. This is achieved by evaluating $y_{n+k}^{[0]}$ by a separate *explicit* method called the **predictor**, and taking this as the initial guess for the iteration based on the implicit method. The implicit method is called the **corrector**.

For the sake of simplicity we shall suppose that the predictor and the corrector have the same number of steps, say k, but in the case of the corrector we shall allow that both α_0 and β_0 vanish. Let the linear multistep method used as predictor have the characteristic polynomials

$$\rho^*(z) = \sum_{j=0}^{k} \alpha_j^* z^j, \quad \alpha_k^* = 1, \qquad \sigma^*(z) = \sum_{j=0}^{k-1} \beta_j^* z^j, \tag{83}$$

and suppose that the corrector has characteristic polynomials

$$\rho(z) = \sum_{j=0}^{k} \alpha_j z^j, \quad \alpha_k = 1, \qquad \sigma(z) = \sum_{j=0}^{k} \beta_j z^j. \tag{84}$$

Suppose that m is a positive integer: it will denote the number of times the corrector is applied; in practice $m \leq 2$. If P indicates the application of the predictor, C a single application of the corrector, and E an evaluation of f in terms of the known values of its arguments, then $P(EC)^m E$ and $P(EC)^m$ denote the following procedures.

a) $P(EC)^m E$

$$\begin{aligned}
y_{n+k}^{[0]} + \sum_{j=0}^{k-1} \alpha_j^* y_{n+j}^{[m]} &= h \sum_{j=0}^{k-1} \beta_j^* f_{n+j}^{[m]}, \\
f_{n+k}^{[s]} &= f(x_{n+k}, y_{n+k}^{[s]}), \\
y_{n+k}^{[s+1]} + \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} &= h\beta_k f_{n+k}^{[s]} + h \sum_{j=0}^{k-1} \beta_j f_{n+j}^{[m]}, \qquad s = 0, \ldots, m-1, \\
f_{n+k}^{[m]} &= f(x_{n+k}, y_{n+k}^{[m]}),
\end{aligned}$$

for $n = 0, 1, 2, \ldots$.

b) $P(EC)^m$

$$\begin{aligned}
y_{n+k}^{[0]} + \sum_{j=0}^{k-1} \alpha_j^* y_{n+j}^{[m]} &= h \sum_{j=0}^{k-1} \beta_j^* f_{n+j}^{[m-1]}, \\
f_{n+k}^{[s]} &= f(x_{n+k}, y_{n+k}^{[s]}), \\
y_{n+k}^{[s+1]} + \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} &= h\beta_k f_{n+k}^{[s]} + h \sum_{j=0}^{k-1} \beta_j f_{n+j}^{[m-1]}, \qquad s = 0, \ldots, m-1,
\end{aligned}$$

for $n = 0, 1, 2, \ldots$.

3.8.1 Absolute stability of predictor-corrector methods

Let us apply the predictor-corrector method $P(EC)^m E$ to the model problem

$$y' = \lambda y, \qquad y(0) = y_0 \ (\neq 0), \tag{85}$$

where $\lambda < 0$, whose solution is, trivially, the decaying exponential $y(x) = y_0 \exp(\lambda x)$, $x \geq 0$. Our aim is to identify the values of the step size h for which the numerical solution computed by the $P(EC)^m E$ method exhibits a similar exponential decay. The resulting recursion is

$$\begin{aligned}
y_{n+k}^{[0]} + \sum_{j=0}^{k-1} \alpha_j^* y_{n+j}^{[m]} &= \bar{h} \sum_{j=0}^{k-1} \beta_j^* y_{n+j}^{[m]}, \\
y_{n+k}^{[s+1]} + \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} &= \bar{h}\beta_k y_{n+k}^{[s]} + \bar{h} \sum_{j=0}^{k-1} \beta_j y_{n+j}^{[m]}, \qquad s = 0, \ldots, m-1,
\end{aligned}$$

for $n = 0, 1, 2, \ldots$, where $\bar{h} = \lambda h$. In order to rewrite this set of equations as a single difference equation involving $y_n^{[m]}, y_{n+1}^{[m]}, \ldots y_{n+k}^{[m]}$ only, we have to eliminate the intermediate stages involving $y_{n+k}^{[0]}, \ldots, y_{n+k}^{[m-1]}$ from the above recursion.

Let us first take $s = 0$ and eliminate $y_{n+k}^{[0]}$ form the resulting pair of equations to obtain

$$y_{n+k}^{[1]} + \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} = \bar{h}\beta_k \left(\bar{h} \sum_{j=0}^{k-1} \beta_j^* y_{n+j}^{[m]} - \sum_{j=0}^{k-1} \alpha_j^* y_{n+j}^{[m]} \right) + \bar{h} \sum_{j=0}^{k-1} \beta_j y_{n+j}^{[m]}.$$

Now take $s = 1$ and use the last equation to eliminate $y_{n+k}^{[1]}$; this gives,

$$\begin{aligned} y_{n+k}^{[2]} + \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} &= \bar{h}\beta_k \left[\bar{h}\beta_k \left(\bar{h} \sum_{j=0}^{k-1} \beta_j^* y_{n+j}^{[m]} - \sum_{j=0}^{k-1} \alpha_j^* y_{n+j}^{[m]} \right) \right. \\ &\quad \left. + \bar{h} \sum_{j=0}^{k-1} \beta_j y_{n+j}^{[m]} - \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} \right] + \bar{h} \sum_{j=0}^{k-1} \beta_j y_{n+j}^{[m]}. \end{aligned}$$

Equivalently,

$$\begin{aligned} & y_{n+k}^{[2]} + \left(1 + \bar{h}\beta_k\right) \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} \\ & = (\bar{h}\beta_k)^2 \left(\bar{h} \sum_{j=0}^{k-1} \beta_j^* y_{n+j}^{[m]} - \sum_{j=0}^{k-1} \alpha_j^* y_{n+j}^{[m]} \right) + (1 + \bar{h}\beta_k)\bar{h} \sum_{j=0}^{k-1} \beta_j y_{n+j}^{[m]}. \end{aligned}$$

By induction,

$$\begin{aligned} & y_{n+k}^{[m]} + \left(1 + \bar{h}\beta_k + \cdots + (\bar{h}\beta_k)^{m-1}\right) \sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} \\ & = (\bar{h}\beta_k)^m \left(\bar{h} \sum_{j=0}^{k-1} \beta_j^* y_{n+j}^{[m]} - \sum_{j=0}^{k-1} \alpha_j^* y_{n+j}^{[m]} \right) + \left(1 + \bar{h}\beta_k + \cdots + (\bar{h}\beta_k)^{m-1}\right) \bar{h} \sum_{j=0}^{k-1} \beta_j y_{n+j}^{[m]}. \end{aligned}$$

For m fixed, this is a kth order difference equation involving $y_n^{[m]}, \ldots, y_{n+k}^{[m]}$. In order to ensure that the solution to this exhibits exponential decay as $n \to \infty$, we have to assume that all roots to the associated characteristic equation

$$\begin{aligned} & z^k + \left(1 + \bar{h}\beta_k + \cdots + (\bar{h}\beta_k)^{m-1}\right) \sum_{j=0}^{k-1} \alpha_j z^j \\ & = (\bar{h}\beta_k)^m \left(\bar{h} \sum_{j=0}^{k-1} \beta_j^* z^j - \sum_{j=0}^{k-1} \alpha_j^* z^j \right) + \left(1 + \bar{h}\beta_k + \cdots + (\bar{h}\beta_k)^{m-1}\right) \bar{h} \sum_{j=0}^{k-1} \beta_j z^j \end{aligned}$$

have modulus < 1. This can be rewritten in the equivalent form

$$Az^k + \left(1 + \bar{h}\beta_k + \cdots + (\bar{h}\beta_k)^{m-1}\right) (\rho(z) - \bar{h}\sigma(z)) + (\bar{h}\beta_k)^m \left(\rho^*(z) - \bar{h}\sigma^*(z)\right) = 0,$$

where

$$A = 1 + \left(1 + \bar{h}\beta_k + \cdots + (\bar{h}\beta_k)^{m-1}\right) (\bar{h}\beta_k - \alpha_k) + (\bar{h}\beta_k)^m (\bar{h}\beta_k^* - \alpha_k^*),$$

Now, since $\alpha_k = \alpha_k^* = 1$ and $\beta_k^* = 0$, we deduce that $A = 0$, and therefore the characteristic equation of the $P(EC)^m E$ method is

$$\pi_{P(EC)^m E}(z;\bar{h}) \equiv \rho(z) - \bar{h}\sigma(z) + M_m(\bar{h})(\rho^*(z) - \bar{h}\sigma^*(z)) = 0,$$

where

$$M_m(\bar{h}) = \frac{(\bar{h}\beta_k)^m}{1 + \bar{h}\beta_k + \cdots + (\bar{h}\beta_k)^{m-1}}, \qquad m \geq 1.$$

Here, $\pi_{P(EC)^m E}(z;\bar{h})$ is referred to as the stability polynomial of the predictor-corrector method $P(EC)^m E$.

By pursuing a similar argument we can also deduce that the characteristic equation of the predictor corrector method $P(EC)^m$ is

$$\pi_{P(EC)^m}(z;\bar{h}) \equiv \rho(z) - \bar{h}\sigma(z) + \frac{M_m(\bar{h})}{\bar{h}\beta_k}\left(\rho^*(z)\sigma(z) - \rho(z)\sigma^*(z)\right) = 0.$$

Here, $\pi_{P(EC)^m}(z;\bar{h})$ is referred to as the stability polynomial of the predictor-corrector method $P(EC)^m$.

With the predictor and corrector specified, one can now check using the Schur criterion or the Routh–Hurwitz criterion, just as in the case of a single multi-step method, whether the roots of $\pi_{P(EC)^m E}(z;\bar{h})$ and $\pi_{P(EC)^m}(z;\bar{h})$ all lie in the open unit disk $|z| < 1$ thereby ensuring the absolute stability of the $P(EC)^m E$ and $P(EC)^m$ method, respectively.

Let us suppose, for example, that $|\bar{h}\beta_k| < 1$, i.e. that $0 < h < 1/|\lambda\beta_k|$; then, $\lim_{m\to\infty} M_m(\bar{h}) = 0$, and consequently,

$$\lim_{m\to\infty} \pi_{P(EC)^m E}(z;\bar{h}) = \pi(z,\bar{h}), \qquad \lim_{m\to\infty} \pi_{P(EC)^m}(z;\bar{h}) = \pi(z,\bar{h}),$$

where $\pi(z;\bar{h}) = \rho(z) - \bar{h}\sigma(z)$ is the stability polynomial of the corrector. This implies that in the mode of correcting to convergence the absolute stability properties of the predictor-corrector method are those of the corrector alone, provided that $|\bar{h}\beta_k| < 1$.

3.8.2 The accuracy of predictor-corrector methods

Let us suppose that the predictor P has order of accuracy p^* and the corrector has order of accuracy p. The question we would like to investigate here is: *What is the overall accuracy of the predictor-corrector method?*

Let us consider the $P(EC)^m E$ method applied to the model problem (85) with $m \geq 1$. We have that

$$\begin{aligned}
\pi_{P(EC)^m E}(\mathrm{e}^{\bar{h}};\bar{h}) &= \rho(\mathrm{e}^{\bar{h}}) - \bar{h}\sigma(\mathrm{e}^{\bar{h}}) + M_m(\bar{h})(\rho^*(\mathrm{e}^{\bar{h}}) - \bar{h}\sigma^*(\mathrm{e}^{\bar{h}})) \\
&= \mathcal{O}(\bar{h}^{p+1}) + M_m(\bar{h})\mathcal{O}(\bar{h}^{p^*+1}) \\
&= \mathcal{O}(\bar{h}^{p+1} + \bar{h}^{p^*+m+1}) \\
&= \begin{cases} \mathcal{O}(\bar{h}^{p+1} + \bar{h}^{p+2}) & \text{if } p^* \geq p \\ \mathcal{O}(\bar{h}^{p+1}) & \text{if } p^* = p-q,\ 0 < q \leq p \text{ and } m \geq q \\ \mathcal{O}(\bar{h}^{p+1} + \bar{h}^{p-q+m+1}) & \text{if } p^* = p-q,\ 0 < q \leq p \text{ and } m \leq q-1. \end{cases}
\end{aligned}$$

Consequently, denoting by $T_n^{P(EC)^m E}$ the consistency error of the method $P(EC)^m E$, we have that

$$T_n^{P(EC)^m E} = \begin{cases} \mathcal{O}(\bar{h}^p) & \text{if } p^* \geq p \\ \mathcal{O}(\bar{h}^p) & \text{if } p^* = p-q,\ 0 < q \leq p \text{ and } m \geq q \\ \mathcal{O}(\bar{h}^{p-q+m}) & \text{if } p^* = p-q,\ 0 < q \leq p \text{ and } m \leq q-1. \end{cases}$$

This implies that from the point of view of overall accuracy there is no particular advantage in using a predictor of order $p^* \geq p$. Indeed, as long as $p^* + m \geq p$, the predictor-corrector method $P(EC)^m E$ will have order of accuracy p.

Similar statements can be made about $P(EC)^m$ type predictor-corrector methods with $m \geq 1$. On writing

$$\rho^*(z)\sigma(z) - \sigma^*(z)\rho(z) = (\rho^*(z) - \bar{h}\sigma^*(z))\sigma(z) - \sigma^*(z)(\rho(z) - \bar{h}\sigma(z)),$$

we deduce that

$$\pi_{P(EC)^m}(\mathrm{e}^{\bar{h}}; \bar{h}) = \mathcal{O}(\bar{h}^{p+1} + \bar{h}^{p^*+m} + \bar{h}^{p+m}).$$

Consequently, as long as $p^* + m \geq p + 1$ the predictor-corrector method $P(EC)^m$ will have order of accuracy p.

End of optional material

4 Stiff problems

Let us consider an initial-value problem for a *system* of m ordinary differential equations of the form: **Lecture 9**

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \qquad \mathbf{y}(a) = \mathbf{y}_0, \tag{86}$$

where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)^{\mathrm{T}}$. A linear k-step method for the numerical solution of (86) has the form

$$\sum_{j=0}^{k} \alpha_j \mathbf{y}_{n+j} = h \sum_{j=0}^{k} \beta_j \mathbf{f}_{n+j}, \tag{87}$$

where $\mathbf{f}_{n+j} = \mathbf{f}(x_{n+j}, y_{n+j})$. Let us suppose, for simplicity, that $\mathbf{f}(x, \mathbf{y}) = A\mathbf{y} + \mathbf{b}$ where A is a constant matrix of size $m \times m$ and $\mathbf{b}$ is a constant (column) vector of size m; then (87) becomes

$$\sum_{j=0}^{k} (\alpha_j I - h\beta_j A)\mathbf{y}_{n+j} = h\sigma(1)\mathbf{b}, \tag{88}$$

where $\sigma(1) = \sum_{j=0}^{k} \beta_j \, (\neq 0)$ and I is the $m \times m$ identity matrix. Let us suppose that the eigenvalues λ_i, $i = 1, \ldots, m$, of the matrix A are distinct. Then, there exists a nonsingular matrix H such that

$$HAH^{-1} = \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix}. \tag{89}$$

Let us define $\mathbf{z} = H\mathbf{y}$ and $\mathbf{c} = h\sigma(1)H\mathbf{b}$. Thus, (88) can be rewritten as

$$\sum_{j=0}^{k} (\alpha_j I - h\beta_j \Lambda)\mathbf{z}_{n+j} = \mathbf{c}, \tag{90}$$

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, $i = 1, \ldots, m$, 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 linear multistep methods for a single differential equation. However, there is a new feature here: because the numbers λ_i, $i = 1, \ldots, m$, are eigenvalues of the matrix A, they need not be real numbers. As a consequence the parameter $\bar{h} = h\lambda$, where λ is any of the m eigenvalues, can be a complex number. 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 $\mathcal{R}_A$ of the complex plane if, for all $\bar{h} \in \mathcal{R}_A$, all roots r_s, $s = 1, \ldots, k$, of the stability polynomial $\pi(r, \bar{h})$ associated with the method, and defined by (79), satisfy $|r_s| < 1$. The set $\mathcal{R}_A$ is called the* ***region of absolute stability*** *of the method.*

Clearly, the interval of absolute stability of a linear multistep method is a subset of its region of absolute stability.

Exercise 6 *a) Find the region of absolute stability of Euler's explicit method when applied to $y' = \lambda y$, $y(x_0) = y_0$, $\lambda \in \mathbb{C}$, $\mathrm{Re}\ \lambda < 0$.*

b) Suppose that Euler's explicit method is applied to the second-order differential equation

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

rewritten as a first-order system in the vector $(u, v)^{\mathrm{T}}$, with $u = y$ and $v = y'$, $\lambda \in \mathbb{C}$, $\mathrm{Re}\ \lambda < 0$, and let $|\lambda| \gg 1$. What choice of the step size $h \in (0, 1)$ will guarantee absolute stability in the sense of the last definition?

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

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

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

$$\mathcal{R}_A = \{\bar{h} \in \mathbf{C} : |1 + \bar{h}| < 1\},$$

which is an open unit disc centred at -1.

b) Now writing $u = y$ and $v = y'$ and $\mathbf{y} = (u, v)^{\mathrm{T}}$, the initial-value problem for the given second-order differential equation can be recast as

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

where

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

The eigenvalues of A are the roots of the characteristic polynomial of A,

$$\det(A - zI) = z^2 + (1 - \lambda)z - \lambda,$$

whose roots are -1 and λ, and we deduce that the method is absolutely stable provided that $|1 + h\lambda| < 1$. Indeed, Euler's explicit method for this system has the form $\mathbf{y}_{n+1} = (I + hA)\mathbf{y}_n$, $n = 0, 1, 2, \ldots$, with $\mathbf{y}_0$ given, where I denotes the (in this particular case 2×2) identity matrix. By diagonalising the matrix $I + hA$, we have that $(I + hA) = S^{-1}DS$, where D is a diagonal matrix containing the eigenvalues of the matrix $I + hA$ on its diagonal, and S is a nonsingular (in this particular case 2×2) matrix. Hence, $S\mathbf{y}_{n+1} = D(S\mathbf{y}_n)$ for $n = 0, 1, \ldots$, and therefore $S\mathbf{y}_n = D^n(S\mathbf{y}_0)$. Now, $\lim_{n\to\infty} \mathbf{y}_n = \mathbf{0}$ if, and only if, $\lim_{n\to\infty} S\mathbf{y}_n = \mathbf{0}$; on the other hand, $\lim_{n\to\infty} S\mathbf{y}_n = \mathbf{0}$ if, and only if, the sequence of matrices $(D^n)_{n=1}^{\infty}$ converges to the zero matrix. Since D is diagonal, the same is true of D^n for each $n = 1, 2, \ldots$, and the diagonal entries of D^n are $(1 - h)^n$ and $(1 + h\lambda)^n$, which will converge to 0 as $n \to \infty$ if, and only if, $|1 - h| < 1$ and $|1 + h\lambda| < 1$. Since we are interested in the case when $\mathrm{Re}\ \lambda < 0$ and $|\lambda| \gg 1$, the first of these two requirements automatically follows from the second requirement. Hence we deduce the stated requirement for absolute stability, that $|1 + h\lambda| < 1$.

We note in passing that it is an easy matter to show that

$$u(x) = 2\mathrm{e}^{-x} - \mathrm{e}^{\lambda x}, \qquad v(x) = -2\mathrm{e}^{-x} + \lambda\mathrm{e}^{\lambda x}.$$

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

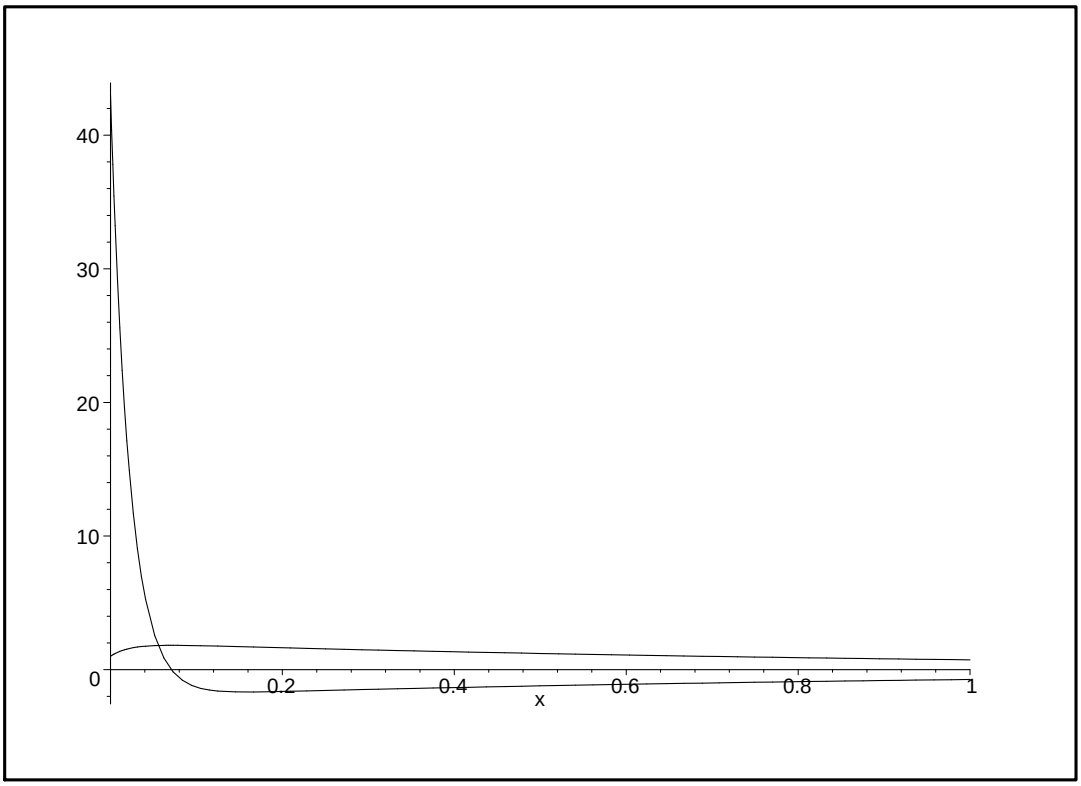

Figure 2: The functions u and v plotted against x for $x \in [0, 1]$.

In the example the v component of the solution exhibited two vastly different time scales; in addition, the fast transition (which occurs between $x = 0$ and $x \approx 1/(-\lambda)$ 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, $h < -2\text{Re}\ \lambda/|\lambda|^2$, smaller than an accurate approximation of the solution for $x \gg 1/|\lambda|$ would necessitate. Systems of differential equations 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.[9] A historic and pragmatic 'definition' by Curtis and Hirschfelder[10] (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 Euler's implicit method for the initial-value problem

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

where λ is a complex number. The stability polynomial of the method is $\pi(z, \bar{h}) = \rho(z) - \bar{h}\sigma(z)$ where $\bar{h} = h\lambda$, $\rho(z) = z - 1$ and $\sigma(z) = z$. Since the only root of the stability polynomial is $z = 1/(1 - \bar{h})$, we

[9] 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.

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

deduce that the method has the region of absolute stability

$$\mathcal{R}_A = \{\bar{h} \in \mathbb{C} : |1 - \bar{h}| > 1\}.$$

In particular $\mathcal{R}_A$ includes the whole of the left open complex half-plane. The next definition is due to Dahlquist (1963).

Definition 12 *A linear multistep method is said to be A-stable if its region of absolute stability, $\mathcal{R}_A$, contains the whole of the left open complex half-plane* $\mathrm{Re}(h\lambda) < 0$.

Thus, 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 16

(i) No explicit *linear multistep method is A-stable.*

(ii) The order of accuracy an A-stable implicit *linear multistep method cannot exceed* 2.

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

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

Definition 13 *A linear multistep method is said to be $A(\alpha)$-**stable**, $\alpha \in (0, \pi/2)$, if its region of absolute stability $\mathcal{R}_A$ contains the infinite open wedge in the complex plane*

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

*A linear multistep method is said to be $A(0)$-**stable** if it is $A(\alpha)$-stable for some $\alpha \in (0, \pi/2)$. A linear multistep method is 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 $\mathrm{Re}\,\lambda < 0$ for a given λ then $\bar{h} = h\lambda$ 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 (89)) happen to lie in some wedge W_α then an $A(\alpha)$-stable method can be used for the numerical solution of the initial-value problem (86) 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 (stated here without proof) can be regarded the analogue of Theorem 16 for the case of $A(\alpha)$ and $A(0)$ stability.

Theorem 17

(i) No explicit *linear multistep method is $A(0)$-stable.*

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

(iii) For each $\alpha \in [0, \pi/2)$ 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 $\mathrm{Re}\,\bar{h} = -a$, $a > 0$, in the complex plane, while those that are responsible for the slow transients are clustered around zero.

Definition 14 *A linear multistep method is said to be* **stiffly stable** *if there exist positive real numbers a and c such that $\mathcal{R}_A \supset \mathcal{R}_1 \cup \mathcal{R}_2$ where*

$$\mathcal{R}_1 = \{\bar{h} \in \mathbf{C} : \operatorname{Re}\bar{h} < -a\} \qquad \text{and} \qquad \mathcal{R}_2 = \{\bar{h} \in \mathbf{C} : -a \le \operatorname{Re}\bar{h} < 0, \ -c \le \operatorname{Im}\hat{h} \le c\}.$$

It is clear that stiff stability implies $A(\alpha)$-stability with $\alpha = \arctan(c/a)$. 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 two sections we shall consider linear multistep methods which are particularly well suited for the numerical solution of stiff systems of ordinary differential equations.

4.2 Backward differentiation methods for stiff systems

Start of optional material

Consider a linear multistep method 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 the closed unit disk when $\bar{h}$ is real and $\bar{h} \to -\infty$. Hence,

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

in other words, the roots of $\pi(\cdot, \bar{h})$ approach those of $\sigma(\cdot)$. Thus it is natural to choose $\sigma(\cdot)$ 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, so-called, **backward differentiation methods** 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 and 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 $p = 1, \ldots, 6$. For $p \ge 7$ backward differentiation methods of order p of the kind considered here are *not* zero-stable and are therefore irrelevant from the practical point of view.

4.3 Gear's method

Since backward differentiation methods are implicit, they have to be used in conjunction with a predictor. Instead of iterating the corrector to convergence via a fixed point iteration, Newton's method can be used to accelerate the iterative convergence of the corrector. Rewriting the resulting predictor-corrector multistep pair as a one step method gives rise to **Gear's method** which allows the local alteration of the order of the method as well as of the mesh size. We elaborate on this below.

As we have seen in Section 4.1, in the numerical solution of stiff systems of ordinary differential equations, the stability considerations highlighted in parts (i) of Theorems 16 and 17 necessitate the use of implicit methods. Indeed, if a predictor-corrector method is used with a backward differentiation formula as corrector, a system of nonlinear equations of the form

$$\mathbf{y}_{n+k} - h\beta_k \mathbf{f}(x_{n+k}, \mathbf{y}_{n+k}) = \mathbf{g}_{n+k}$$

will have to be solved at each step, where

$$\mathbf{g}_{n+k} = -\sum_{j=0}^{k-1} \alpha_j \mathbf{y}_{n+j}$$

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

is a term that involves information which has already been computed at previous steps and can be considered known. If this equation is solved by a fixed-point iteration, the Contraction Mapping Theorem will require that

$$Lh|\beta_k| < 1 \tag{91}$$

in order to ensure convergence of the iteration; here L is the Lipschitz constant of the function $\mathbf{f}(x,\cdot)$. In fact, since the function $\mathbf{f}(x,\cdot)$ is assumed to be continuously differentiable,

$$L = \max_{(x,\mathbf{y})\in\mathsf{R}} \left\| \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(x,\mathbf{y}) \right\|.$$

For a stiff system L is typically very large, thus the restriction on the steplength h expressed by (91) is just as severe as the condition on h that one encounters when using an explicit method with a bounded region of absolute stability. In order to overcome this difficulty, Gear proposed to use Newton's method:

$$\mathbf{y}_{n+k}^{[s+1]} = \mathbf{y}_{n+k}^{[s]} - \left[I - h\beta_k \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(x_{n+k}, \mathbf{y}_{n+k}^{[s]}) \right]^{-1} \left[\mathbf{y}_{n+k}^{[s]} - h\beta_k \mathbf{f}(x_{n+k}, \mathbf{y}_{n+k}^{[s]}) - \mathbf{g}_{n+k} \right], \tag{92}$$

for $s = 0, 1, \ldots$, with a suitable initial guess $y_{n+k}^{[0]}$. Even when applied to a stiff system, convergence of the Newton iteration (92) can be attained without further restrictions on the mesh size h provided that we can supply a sufficiently accurate initial guess $y_{n+k}^{[0]}$ (by using an appropriately accurate predictor, for example).

On the other hand, the use of Newton's method in this context has the disadvantage that the Jacobi matrix $\partial \mathbf{f}/\partial \mathbf{y}$ has to be reevaluated and the matrix $I - h\beta_k \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(x_{n+k}, y_{n+k}^{[s]})$ inverted at each step of the iteration and at each mesh point x_{n+k}.

One aspect of Gear's method is that the matrix $I - h\beta_k \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(x_{n+k}, y_{n+k}^{[s]})$ involved in the Newton iteration described above is only calculated occasionally (i.e. at the start of the iteration, for $s = 0$, and thereafter

only if the Newton iteration exhibits slow convergence); the inversion of this matrix is performed by an LU decomposition.

A further aspect of Gear's method is a strategy for varying the order of the backward differentiation formula and the step size according to the intermediate results in the computation. This is achieved by rewriting the multistep predictor-corrector pair as a one-step method (in the so-called Nordsieck form). For further details, we refer to Chapter III.6 in the book of Hairer, Norsett and Wanner. We shall therefore confine ourselves here to a brief discussion of adaptive one-step methods for stiff systems.

End of optional material

5 Adaptivity for stiff problems

Ideally, we would like to compute an approximate solution of the following initial-value problem for a system of first-order ordinary differential equations: **Lecture 10**

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \qquad \mathbf{y}(x_0) = \mathbf{y}_0, \tag{93}$$

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.

First attempt: A simple strategy could be to:

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

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 $h = (X_M - x_0)/N$;
3. check whether $\|\tilde{\mathbf{y}}_{\tilde{N}} - \mathbf{y}_N\| < \texttt{TOL}$.

If $\|\tilde{\mathbf{y}}_{\tilde{N}} - \mathbf{y}_N\|$ is smaller than the target absolute tolerance `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 $\tilde{h} = (X_M - x_0)/\tilde{N}$. 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 $\alpha := \tilde{h}/h = N/\tilde{N} < 1$. For h sufficiently small,

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

and thus,

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

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, and a new numerical approximation has to be computed on such a globally refined mesh).

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

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

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

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

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

$$\mathrm{ERR}(x_n; h) \coloneqq \|\tilde{\boldsymbol{\Psi}}(x_n, \mathbf{y}_n; h) - \boldsymbol{\Psi}(x_n, \mathbf{y}_n; h)\|. \tag{94}$$

The idea behind using (94) 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,

$$\begin{aligned} hT_n &= \mathbf{y}(x_{n+1}) - \boldsymbol{\Psi}(x_n, \mathbf{y}(x_n); h) \\ &= \mathbf{y}(x_{n+1}) - \tilde{\boldsymbol{\Psi}}(x_n, \mathbf{y}(x_n); h) + \tilde{\boldsymbol{\Psi}}(x_n, \mathbf{y}(x_n); h) - \boldsymbol{\Psi}(x_n, \mathbf{y}(x_n); h) \\ &\approx \mathbf{y}(x_{n+1}) - \tilde{\boldsymbol{\Psi}}(x_n, \mathbf{y}(x_n); h) + \tilde{\boldsymbol{\Psi}}(x_n, \mathbf{y}_n; h) - \boldsymbol{\Psi}(x_n, \mathbf{y}_n; h) \\ &\approx Ch^{\tilde{p}+1} + \tilde{\boldsymbol{\Psi}}(x_n, \mathbf{y}_n; h) - \boldsymbol{\Psi}(x_n, \mathbf{y}_n; h). \end{aligned} \tag{95}$$

Since the left-hand side of (95) is of the order $\mathcal{O}(h \times h^p) = \mathcal{O}(h^{p+1})$ 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" $\tilde{\boldsymbol{\Psi}}(x_n, \mathbf{y}_n; h) - \boldsymbol{\Psi}(x_n, \mathbf{y}_n; h)$. Hence, $hT_n \approx \tilde{\boldsymbol{\Psi}}(x_n, \mathbf{y}_n; h) - \boldsymbol{\Psi}(x_n, \mathbf{y}_n; h)$.

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 this is smaller than a target tolerance, we set $\mathbf{y}_{n+1} = \boldsymbol{\Psi}(x_n, \mathbf{y}_n; h_n)$;
   otherwise, we choose a smaller h_n and return 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 3 *Let* `TOL` *be a target absolute error tolerance and let* $\mathrm{ERR}(x_n; h_n) <$ `TOL`*. Then, the "optimal"* β *is*

$$\beta = \beta_n = \sqrt[(p+1)]{\texttt{TOL}/\mathrm{ERR}(x_n; h_n)}. \tag{96}$$

Indeed, if $\mathrm{ERR}(x_n; h_n) <$ `TOL`*, we could have chosen a larger* h_n *and still satisfied the tolerance criterion. Let* β_n *be such that* $\mathrm{ERR}(x_n, \beta_n h_n) =$ `TOL`*, so that* $\beta_n h_n$ *is the ideal step size. Then, we deduce* (96)*, 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 Runge–Kutta methods, which limit the number of function evaluations.

Definition 15 *Two Runge–Kutta methods are* embedded *if they use the same stages. The Butcher tableau of two embedded Runge–Kutta 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 tableaux of the two Runge–Kutta methods, respectively.

Example 7 *The Heun–Euler method has the Butcher tableau:*

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

are the Butcher tableaux 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 8 MATLAB *integrators for ODEs (such as the functions* `ode45`*,* `ode23`*, etc.) are based on embedded Runge–Kutta methods.*[11]

6 Structure-preserving integrators

Lecture 11

Many physical phenomena are modeled by initial-value problems and, by analyzing these, one can show that certain relevant physical quantities, such as energy, mass, volume, etc., are preserved during the course of evolution, that is, they are constant in time. The goal of this section is to study numerical methods that preserve some of these quantities also at the discrete level.[12] To begin with, we clarify the concept of solution to an ODE to allow generic initial values. For simplicity, we restrict ourselves to the autonomous ODE

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}), \quad \text{where} \quad \mathbf{f} : D \to \mathbb{R}^d, \tag{97}$$

(where now $\mathbf{y}$ is considered to be a function of $t \in [0, \infty)$, and $\mathbf{y}' := \mathrm{d}\mathbf{y}/\mathrm{d}t$), subject to the initial condition

$$\mathbf{y}(0) = \mathbf{x},$$

where $\mathbf{x} \in D$, and D is a nonempty open subset of $\mathbb{R}^d$.

Definition 16 *For* $t \geq 0$*, let* $\mathbf{\Phi}^t : D \to \mathbb{R}^d$ *denote the function that maps an initial datum* $\mathbf{x} \in D$ *into* $\mathbf{y}(t) \in \mathbb{R}^d$*, where* $\mathbf{y}(t)$ *is the solution at time* t *to* $\mathbf{y}' = \mathbf{f}(\mathbf{y})$*,* $\mathbf{y}(0) = \mathbf{x}$ *(tacitly assuming that the solution* $t \in [0, \infty) \mapsto \mathbf{y}(t) \in \mathbb{R}^d$ *to this initial-value problem, for each* $\mathbf{x} \in D$*, exists and that it is unique). The family* $\{\mathbf{\Phi}^t\}_{t \geq 0}$ *is called the* flow *of* (97) *(defined on* $D \subset \mathbb{R}^d$*).*

Remark 4 *The function* $t \mapsto \mathbf{\Phi}^t(\mathbf{x})$ *is the solution to* $\mathbf{y}' = \mathbf{f}(\mathbf{y})$*,* $\mathbf{y}(0) = \mathbf{x}$*.*

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

[12] We could have named this section *Geometric numerical integration*, a term introduced by J.M. Sanz-Serna; see his article *Geometric integration* in the proceedings *The State of the Art in Numerical Analysis*, I.S. Duff and G.A. Watson, eds., Clarendon Press, Oxford, 1997, pp. 121–143.

Using the concept of flow, we can clarify what is a "preserved quantity".

Definition 17 *Suppose that $\mathbf{\Phi}^t(D) \subseteq D$ for every admissible $t \geq 0$. A* first integral *of* (97) *is a function $I : D \to \mathbb{R}$ that satisfies $I(\mathbf{\Phi}^t(\mathbf{x})) = I(\mathbf{x})$ for every $\mathbf{x} \in D$ and every admissible $t \geq 0$.*

Lemma 4 *Suppose that $\mathbf{\Phi}^t(D) \subseteq D$ for every admissible $t \geq 0$. I is a first integral of* (97) *if, and only if, $\frac{\mathrm{d}}{\mathrm{d}t} I(\mathbf{\Phi}^t(\mathbf{x})) = 0$ for every $\mathbf{x} \in D$ and every admissible $t \geq 0$. This is equivalent to: $\mathbf{D}I(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) = 0$ for every $\mathbf{x} \in D$, where $\mathbf{D}I := \mathbf{grad}\, I$.*

Proof. The first part of the lemma is trivial. Indeed, if I is a first integral then if follows from Definition 17 that $\frac{\mathrm{d}}{\mathrm{d}t} I(\mathbf{\Phi}^t(\mathbf{x})) = \frac{\mathrm{d}}{\mathrm{d}t} I(\mathbf{x}) = 0$. Conversely, if $\frac{\mathrm{d}}{\mathrm{d}t} I(\mathbf{\Phi}^t(\mathbf{x})) = 0$, then by integration with respect to t, we have $I(\mathbf{\Phi}^t(\mathbf{x})) = C$ for some constant C for every admissible $t \geq 0$. By taking $t = 0$ in particular, we deduce that $C = I(\mathbf{\Phi}^0(\mathbf{x})) = I(\mathbf{x})$, and therefore, by Definition 17, I is a first integral.

The second part follows by applying the chain rule on the left-hand side of the equality $\frac{\mathrm{d}}{\mathrm{d}t} I(\mathbf{\Phi}^t(\mathbf{x})) = 0$. Indeed,

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} I(\mathbf{\Phi}^t(\mathbf{x})) = \frac{\mathrm{d}}{\mathrm{d}t} I(\mathbf{y}(t)) = \mathbf{D}I(\mathbf{y}(t)) \cdot \mathbf{y}'(t) = \mathbf{D}I(\mathbf{y}(t)) \cdot \mathbf{f}(\mathbf{y}(t))$$

for every admissible $t \geq 0$, where $\mathbf{y}$ is the solution of the initial-value problem $\mathbf{y}'(t) = \mathbf{f}(\mathbf{y}(t))$, $\mathbf{y}(0) = \mathbf{x} \in D$. Thus in particular $0 = \mathbf{D}I(\mathbf{y}(0)) \cdot \mathbf{f}(\mathbf{y}(0))$, and the assertion follows, because $\mathbf{y}(0) = \mathbf{x} \in D$; i.e., $\mathbf{D}I(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) = 0$ for every $\mathbf{x} \in D$. Conversely, if $\mathbf{D}I(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) = 0$ for every $\mathbf{x} \in D$, then in particular because $\mathbf{\Phi}^t(D) \subseteq D$ for every admissible $t \geq 0$, we have that $\mathbf{D}I(\mathbf{\Phi}^t(\mathbf{x})) \cdot \mathbf{f}(\mathbf{\Phi}^t(\mathbf{x})) = 0$ for every $\mathbf{x} \in D$ and every admissible $t \geq 0$; in other words, $\mathbf{D}I(\mathbf{y}(t)) \cdot \mathbf{f}(\mathbf{y}(t)) = 0$ for every admissible $t \geq 0$. Hence, by reversing the chain of equalities displayed above,

$$0 = \mathbf{D}I(\mathbf{y}(t)) \cdot \mathbf{f}(\mathbf{y}(t)) = \mathbf{D}I(\mathbf{y}(t)) \cdot \mathbf{y}'(t) = \frac{\mathrm{d}}{\mathrm{d}t} I(\mathbf{y}(t)) = \frac{\mathrm{d}}{\mathrm{d}t} I(\mathbf{\Phi}^t(\mathbf{x}))$$

for every $\mathbf{x} \in D$ and every admissible $t \geq 0$. That completes the proof. ⋄

For a systematic investigation, we consider first integrals that can be expressed as polynomials. Note that, according to Lemma 4, for such first integrals the solution $\mathbf{y}(t)$ remains on the zero level-surface of the first integral (which could be a plane, or a sphere, or an ellipsoid, etc.) for all $t > 0$ provided that the initial datum $\mathbf{x} = \mathbf{y}(0)$ belongs to the zero level-surface of the first integral.

Definition 18 *We shall say that a first integral I of an autonomous system is a polynomial of degree $n \in \mathbb{N}$ if*

$$I(\mathbf{x}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}_0^d,\, |\boldsymbol{\alpha}| \leq n} \beta_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}, \tag{98}$$

where $\beta_{\boldsymbol{\alpha}} \in \mathbb{R}$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$, $|\boldsymbol{\alpha}| = \sum_{i=1}^d \alpha_i$, and $\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$; in other words, I is a multivariate polynomial of degree n in $\mathbf{x} \in \mathbb{R}^d$. Here, $\mathbb{N}_0$ denotes the set of all nonnegative integers, and $\mathbb{N}_0^d$ signifies the d-fold Cartesian product $\mathbb{N}_0 \times \cdots \times \mathbb{N}_0$.

Example 9 *Linear first integrals are of the form $I(\mathbf{x}) = \mathbf{b}^{\mathrm{T}}\mathbf{x} + c$ (with $\mathbf{b} \in \mathbb{R}^d$, and $c \in \mathbb{R}$).*

Example 10 *Quadratic first integrals are of the form $I(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x} + c$ (with $\mathbf{M} = \mathbf{M}^{\mathrm{T}} \in \mathbb{R}^{d \times d}$, $\mathbf{b} \in \mathbb{R}^d$, and $c \in \mathbb{R}$).*

The following theorems summarize a few key facts about structure-preserving Runge–Kutta methods.

Theorem 18 *Every Runge–Kutta method preserves linear first integrals.*

Theorem 19 *Gauss-collocation methods (i.e. Runge–Kutta methods based on function-evaluations at points of Gaussian quadrature rules) preserve quadratic first integrals.*

PROOF: (of Theorem 18) Clearly, using the same notation as in Example 9, the equality $I(\mathbf{x}) = \mathbf{b}^{\mathrm{T}}\mathbf{x} + c$ implies that $\mathbf{D}I(\mathbf{x}) \equiv \mathbf{b}$; thus, by applying Lemma 4, we have that $0 = \mathbf{D}I(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) = \mathbf{b} \cdot \mathbf{f}(\mathbf{x})$ for all $\mathbf{x} \in D$. Hence, for all $n \geq 0$, and $\mathbf{y}_{n+1}$ computed from $\mathbf{y}_n$ using an R-stage Runge–Kutta method $\mathbf{y}_{n+1} = \mathbf{y}_n + h(c_1\mathbf{k}_1 + \cdots + c_R\mathbf{k}_R)$, we have that

$$\begin{aligned} I(\mathbf{y}_{n+1}) - I(\mathbf{y}_n) &= \mathbf{b}^{\mathrm{T}}(\mathbf{y}_{n+1} - \mathbf{y}_n) \\ &= h\mathbf{b}^{\mathrm{T}}(c_1\mathbf{k}_1 + \cdots + c_R\mathbf{k}_R) = 0 + \cdots + 0 = 0, \end{aligned}$$

because each of the functions $\mathbf{k}_i$, $i = 1, \ldots, R$, is defined by evaluating $\mathbf{f}$ at a certain point $\mathbf{x} \in D$, whereby $\mathbf{b}^{\mathrm{T}}\mathbf{k}_i = \mathbf{b} \cdot \mathbf{k}_i = 0$ for $i = 1, \ldots, R$. Therefore $I(\mathbf{y}_n) = I(\mathbf{y}_0) = I(\mathbf{x})$ for all all $n \geq 0$, all $\mathbf{x} \in D$, and $\mathbf{y}_0 = \mathbf{x}$. That completes the proof. $\diamond$

Simple examples of Gauss-collocation methods are the Gauss–Legendre–Runge–Kutta methods, based on function-evaluations at points of Gauss–Legendre quadrature rules. The Gauss–Legendre method of order two is the **implicit midpoint rule**,

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\,\mathbf{f}\left(\frac{1}{2}\mathbf{y}_n + \frac{1}{2}\mathbf{y}_{n+1}\right),$$

which has Butcher tableau

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array} \quad .$$

The **Gauss–Legendre method of order four**:

$$\begin{aligned} \mathbf{y}_{n+1} &= \mathbf{y}_n + h\left(\frac{1}{2}\mathbf{k}_1 + \frac{1}{2}\mathbf{k}_2\right), \qquad \text{where} \\ \mathbf{k}_1 &= \mathbf{f}\left(t + \left(\frac{1}{2} - \frac{1}{6}\sqrt{3}\right)h, \mathbf{y}_n + \frac{1}{4}\mathbf{k}_1 + \left(\frac{1}{4} - \frac{1}{6}\sqrt{3}\right)\mathbf{k}_2\right), \\ \mathbf{k}_2 &= \mathbf{f}\left(t + \left(\frac{1}{2} + \frac{1}{6}\sqrt{3}\right)h, \mathbf{y}_n + \left(\frac{1}{4} + \frac{1}{6}\sqrt{3}\right)\mathbf{k}_1 + \frac{1}{4}\mathbf{k}_2\right). \end{aligned}$$

has Butcher tableau

$$\begin{array}{c|cc} \frac{1}{2} - \frac{1}{6}\sqrt{3} & \frac{1}{4} & \frac{1}{4} - \frac{1}{6}\sqrt{3} \\ \frac{1}{2} + \frac{1}{6}\sqrt{3} & \frac{1}{4} + \frac{1}{6}\sqrt{3} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \quad .$$

Unfortunately, there is no consistent Runge–Kutta method that preserves polynomial first integrals of degree higher than 2; more precisely, the following negative result holds.

Theorem 20 *If $n \geq 3$, then there is no consistent Runge–Kutta method that preserves every polynomial first integral of degree n for every autonomous ODE.*

We conclude this section with a few results concerning the conservation of a structure that is at the heart of classical mechanics: conservation of the symplectic product. First, we recall the notion of Hamiltonian differential equation from classical mechanics.

Definition 19 *A **Hamiltonian differential equation** is an ODE of the form*

$$\mathbf{p}' = -\mathbf{D}_{\mathbf{q}}H(\mathbf{p}, \mathbf{q}), \quad \mathbf{q}' = \mathbf{D}_{\mathbf{p}}H(\mathbf{p}, \mathbf{q}), \tag{99}$$

where $\mathbf{p}(t) := (p_1(t), \ldots, p_d(t))^{\mathrm{T}}$, $\mathbf{q}(t) := (q_1(t), \ldots, q_d(t))^{\mathrm{T}}$, $\mathbf{D}_{\mathbf{p}} := (\frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_d})^{\mathrm{T}}$, $\mathbf{D}_{\mathbf{q}} := (\frac{\partial}{\partial q_1}, \ldots, \frac{\partial}{\partial q_d})^{\mathrm{T}}$. The function $H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called the Hamiltonian.

Theorem 21 *The Hamiltonian H is a first integral of* (99).

PROOF. By applying the chain rule and noting (99), we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H(\mathbf{p}(t),\mathbf{q}(t)) &= \mathbf{D_p}H(\mathbf{p}(t),\mathbf{q}(t))\cdot\mathbf{p}'(t)+\mathbf{D_q}H(\mathbf{p}(t),\mathbf{q}(t))\cdot\mathbf{q}'\\ &= -\mathbf{D_p}H(\mathbf{p}(t),\mathbf{q}(t))\cdot\mathbf{D_q}H(\mathbf{p}(t),\mathbf{q}(t))+\mathbf{D_q}H(\mathbf{p}(t),\mathbf{q}(t))\cdot\mathbf{D_p}H(\mathbf{p}(t),\mathbf{q}(t))=0,\end{aligned}$$

which implies the assertion by recalling the definition of *first integral*. $\diamond$

Lemma 5 *The ODE* (99) *is equivalent to*

$$\mathbf{y}'=\mathbf{J}^{-1}\mathbf{D}H(\mathbf{y}),\quad \textit{where}\quad \mathbf{J}=\begin{pmatrix}\mathbf{0}&\mathbf{I}\\-\mathbf{I}&\mathbf{0}\end{pmatrix}\in\mathbb{R}^{2d\times 2d}\quad \textit{and}\quad \mathbf{y}\coloneqq\begin{pmatrix}\mathbf{p}\\\mathbf{q}\end{pmatrix}\in\mathbb{R}^{2d}. \tag{100}$$

PROOF. The ODE system (99) is equivalent to $\mathbf{J}\mathbf{y}'=\mathbf{grad}\,H(\mathbf{y})$. Indeed,

$$\mathbf{J}\mathbf{y}'=\begin{pmatrix}\mathbf{0}&\mathbf{I}\\-\mathbf{I}&\mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{p}'\\\mathbf{q}'\end{pmatrix}=\begin{pmatrix}\mathbf{q}'\\-\mathbf{p}'\end{pmatrix}=\begin{pmatrix}\mathbf{D_p}H(\mathbf{p},\mathbf{q})\\\mathbf{D_q}H(\mathbf{p},\mathbf{q})\end{pmatrix}=\mathbf{D}H(\mathbf{p},\mathbf{q}),$$

which directly implies the assertion of the lemma. $\diamond$

The next definition is inspired by the previous lemma.

Definition 20 *The bilinear map*

$$\omega:\mathbb{R}^{2d}\times\mathbb{R}^{2d}\to\mathbb{R},\quad (\mathbf{a},\mathbf{b})\mapsto\omega(\mathbf{a},\mathbf{b})\coloneqq\mathbf{a}^{\mathrm{T}}\mathbf{J}\,\mathbf{b}$$

is called the symplectic product *of* $\mathbf{a}$ *and* $\mathbf{b}$.

Definition 21 *A continuously differentiable map $\mathbf{\Phi}:D\subset\mathbb{R}^{2d}\to\mathbb{R}^{2d}$ is called* symplectic *if*

$$\omega(\mathbf{D\Phi}(\mathbf{x})\mathbf{a},\mathbf{D\Phi}(\mathbf{x})\mathbf{b})=\omega(\mathbf{a},\mathbf{b})$$

for every $\mathbf{x}\in D$ and every pair $(\mathbf{a},\mathbf{b})\in\mathbb{R}^{2d}\times\mathbb{R}^{2d}$. Here $\mathbf{D\Phi}(\mathbf{x})$ denotes the Jacobian matrix of $\mathbf{\Phi}$ evaluated at $\mathbf{x}\in D$, i.e. the $2d\times 2d$ matrix whose (i,j) entry is $\partial\Phi_i/\partial x_j$, $i,j=1,\dots,2d$.

Remark 5 *A map is symplectic if its Jacobian matrix $\mathbf{D\Phi}(\mathbf{x})$ (evaluated at a generic point $\mathbf{x}$) preserves the simplectic product. This concept is similar to the property of orthogonal matrices that they preserve the Euclidean inner product, i.e. if $\mathbf{O}\in\mathbb{R}^{d\times d}$ is an orthogonal matrix then $\langle\mathbf{Oa},\mathbf{Ob}\rangle=\langle\mathbf{a},\mathbf{b}\rangle$ for every pair $(\mathbf{a},\mathbf{b})\in\mathbb{R}^d\times\mathbb{R}^d$, where $\langle\cdot,\cdot\rangle$ is the Euclidean inner product in $\mathbb{R}^d$.*

The following result, due to Poincaré, asserts that a Hamiltonian flow is a symplectic map, which explains why the concept of symplectic map is so relevant.

Theorem 22 (Poincaré) *Suppose that H is a twice continuously differentiable Hamiltonian. Then, the flow $\mathbf{\Phi}^t$ of* (100) *satisfies the following property: for each $\mathbf{x}\in D$ there exists a $\delta>0$ such that*

$$\omega(\mathbf{D\Phi}^t(\mathbf{x})\mathbf{a},\mathbf{D\Phi}^t(\mathbf{x})\mathbf{b})=\omega(\mathbf{a},\mathbf{b})\quad \textit{for every } (\mathbf{a},\mathbf{b})\in\mathbb{R}^{2d}\times\mathbb{R}^{2d}\textit{ and all } t\in[0,\delta),$$

where $\mathbf{D\Phi}^t(\mathbf{x})$ denotes the Jacobian matrix of $\mathbf{\Phi}^t$ evaluated at $\mathbf{x}\in D$.

PROOF. Let $\mathbf{y}$ be the solution of the initial-value problem $\mathbf{y}'(t) = \mathbf{J}^{-1}\mathbf{D}H(\mathbf{y}(t))$, $\mathbf{y}(0) = \mathbf{x} \in D$ for $t \in [0, \delta)$, and let $\mathbf{\Phi}^t(\mathbf{x}) := \mathbf{y}(t)$, with $\mathbf{y}(0) = \mathbf{x} \in D$. Thanks to Lemma 5 and the chain rule, we have that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{D}\mathbf{\Phi}^t(\mathbf{x}) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{D}\mathbf{y}(t) = \mathbf{D}\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{D}\mathbf{J}^{-1}\,\mathbf{grad}\,H(\mathbf{y}(t)) = \mathbf{J}^{-1}\mathbf{D}(\mathbf{grad}\,H(\mathbf{y}(t)))\\ = \mathbf{J}^{-1}\,\mathbf{D}^2H(\mathbf{y}(t))\,\mathbf{D}\mathbf{y}(t) = \mathbf{J}^{-1}\,(\mathbf{D}^2H(\mathbf{\Phi}^t(\mathbf{x})))\,(\mathbf{D}\mathbf{\Phi}^t(\mathbf{x})),\end{aligned}$$

where $\mathbf{D}^2H$ denotes the Hessian of H, i.e. the $2d \times 2d$ matrix whose (i,j) entry is the second partial derivative of H with respect to its ith and jth arguments, for $i, j = 1, \ldots, 2d$. Thanks to the assumption that H is twice continuously differentiable, the matrix $\mathbf{D}^2H$ is symmetric, i.e. $(\mathbf{D}^2H)^{\mathrm{T}} = \mathbf{D}^2H$. Therefore, by the product rule and because $\mathbf{J}$ is a constant matrix, for any $\mathbf{x} \in D$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}((\mathbf{D}\mathbf{\Phi}^t(\mathbf{x}))^{\mathrm{T}}\mathbf{J}(\mathbf{D}\mathbf{\Phi}^t(\mathbf{x}))) &= \left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{D}\mathbf{\Phi}^t(\mathbf{x})\right)^{\mathrm{T}}\mathbf{J}(\mathbf{D}\mathbf{\Phi}^t(\mathbf{x})) + (\mathbf{D}\mathbf{\Phi}^t(\mathbf{x}))^{\mathrm{T}}\mathbf{J}\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{D}\mathbf{\Phi}^t(\mathbf{x})\right)\\ &= (\mathbf{D}\mathbf{\Phi}^t(\mathbf{x}))^{\mathrm{T}}\,(\mathbf{D}^2H(\mathbf{\Phi}^t(\mathbf{x})))\,\mathbf{J}^{-\mathrm{T}}\mathbf{J}\,(\mathbf{D}\mathbf{\Phi}^t(\mathbf{x})) + (\mathbf{D}\mathbf{\Phi}^t(\mathbf{x}))^{\mathrm{T}}\mathbf{J}\mathbf{J}^{-1}\,(\mathbf{D}^2H(\mathbf{\Phi}^t(\mathbf{x})))\,(\mathbf{D}\mathbf{\Phi}^t(\mathbf{x}))\\ &= -(\mathbf{D}\mathbf{\Phi}^t(\mathbf{x}))^{\mathrm{T}}\,(\mathbf{D}^2H(\mathbf{\Phi}^t(\mathbf{x})))\,(\mathbf{D}\mathbf{\Phi}^t(\mathbf{x})) + (\mathbf{D}\mathbf{\Phi}^t(\mathbf{x}))^{\mathrm{T}}\,(\mathbf{D}^2H(\mathbf{\Phi}^t(\mathbf{x})))\,(\mathbf{D}\mathbf{\Phi}^t(\mathbf{x})) = \mathbf{0},\end{aligned}$$

where we have used that $\mathbf{J}^{-\mathrm{T}}\mathbf{J} = -\mathbf{I}$ and $\mathbf{J}\mathbf{J}^{-1} = \mathbf{I}$. This implies that for any $\mathbf{x} \in D$ and any $t \in [0, \delta)$,

$$(\mathbf{D}\mathbf{\Phi}^t(\mathbf{x}))^{\mathrm{T}}\mathbf{J}(\mathbf{D}\mathbf{\Phi}^t(\mathbf{x})) = (\mathbf{D}\mathbf{\Phi}^0(\mathbf{x}))^{\mathrm{T}}\mathbf{J}(\mathbf{D}\mathbf{\Phi}^0(\mathbf{x})) = \mathbf{J}$$

as an equality between $2d \times 2d$ matrices, where in the last equality we have used that $\mathbf{\Phi}^0\mathbf{x} = \mathbf{x}$ (and therefore $\mathbf{D}\mathbf{\Phi}^0(\mathbf{x}) = \mathbf{I}$). Hence, for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{2d}$ we have that

$$\omega(\mathbf{D}\mathbf{\Phi}^t(\mathbf{x})\mathbf{a}, \mathbf{D}\mathbf{\Phi}^t(\mathbf{x})\mathbf{b}) = \mathbf{a}^{\mathrm{T}}(\mathbf{D}\mathbf{\Phi}^t(\mathbf{x}))^{\mathrm{T}}\mathbf{J}(\mathbf{D}\mathbf{\Phi}^t(\mathbf{x}))\mathbf{b} = \mathbf{a}^{\mathrm{T}}\mathbf{J}\mathbf{b} = \omega(\mathbf{a}, \mathbf{b}),$$

and that completes the proof. $\diamond$

Since Hamiltonian flows are symplectic, we are interested in symplectic one-step methods, in the sense of the following definition.

Definition 22 *Consider* (100) *subject to the initial condition* $\mathbf{y}(0) = \mathbf{x}$*, for* $\mathbf{x} \in D$*, and let* $\mathbf{x} \mapsto \mathbf{\Psi}(0, \mathbf{x}; h)$ *be a one-step method for* (100)*, which maps the initial datum* $\mathbf{x} \in D$ *into a numerical approximation* $\mathbf{\Psi}(0, \mathbf{x}; h) \in \mathbb{R}^d$ *of* $\mathbf{y}(h) \in \mathbb{R}^d$ *over a single time step of length* $h > 0$*. The one-step method* $\mathbf{x} \mapsto \mathbf{\Psi}(0, \mathbf{x}; h)$ *is said to be* symplectic *if* $\mathbf{x} \mapsto \mathbf{\Psi}(0, \mathbf{x}; h)$ *defines a symplectic map on every compact subset* $K \subset D$*, whenever* H *is twice continuously differentiable and* $h > 0$ *is sufficiently small.*

The following theorem provides a convenient sufficient condition for a Runge–Kutta being symplectic, although for an arbitrary one-step method one would still need to appeal to Definition 22 to verify that the method in question is symplectic.

Theorem 23 *Every Runge–Kutta method that preserves quadratic first integrals is symplectic.*

We conclude by mentioning that there exist also explicit numerical methods that are symplectic. The most important example for an ODEs with a "separated" right-hand side is the Störmer–Verlet method. For furter details, we refer to E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration*, Springer Series in Computational Mathematics, (2006).

7 Finite difference approximation of parabolic equations

The final section of these lecture notes is concerned with the construction and mathematical analysis of finite difference methods for the numerical solution of parabolic equations. As a simple yet representative model problem we shall focus on the unsteady diffusion equation (heat equation) in one space dimension: **Lecture 12**

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \tag{101}$$

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

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

where u_0 is a given function.

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

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

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

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

By Fourier-transforming the partial differential equation (101) we obtain

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

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

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

whereby

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

and therefore

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

The inverse Fourier transform of a function is defined by

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

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

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

where the function w, defined by

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

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

$$u(x,t) = \frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} \mathrm{e}^{-(x-y)^2/(4t)}u_0(y)\, \mathrm{d}y. \tag{102}$$

This formula gives an explicit expression of the solution of the heat equation (101) in terms of the initial datum u_0. Because $w(x,t) > 0$ for all $x \in (-\infty, \infty)$ and all $t > 0$, and

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

we deduce from (102) that if u_0 is a bounded continuous function, then

$$\sup_{x\in(-\infty,+\infty)}|u(x,t)| \le \sup_{x\in(-\infty,\infty)}|u_0(x)|, \qquad t > 0. \tag{103}$$

In other words, the 'largest' and 'smallest' values of $u(\cdot,t)$ at $t > 0$ cannot exceed those of $u_0(\cdot)$. Similar bounds on the 'magnitude' of the solution at future times in terms of the 'magnitude' of the initial datum can be obtained in other norms as well, and we shall focus here on the L^2 norm in particular. We will show, using Parseval's identity, that the L^2 norm of the solution, at any time $t > 0$, is bounded by the L^2 norm of the initial datum. We shall then try to mimic this property when using various numerical approximations of the initial-value problem for the heat equation.

Lemma 6 (Parseval's identity) *Let $L^2(-\infty,\infty)$ denote the set of all complex-valued square-integrable functions defined on the real line. Suppose that $u \in L^2(-\infty,\infty)$. Then, $\hat{u} \in L^2(-\infty,\infty)$, and the following equality holds:*

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

where

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

Proof. We begin by observing that

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

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

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

and substitute this into the identity above to complete the proof. $\diamond$

Returning to the equation (101), we thus have by Parseval's identity that

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

and therefore

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

Thus we have shown that

$$\|u(\cdot,t)\|_{L^2(-\infty,\infty)} \leq \|u_0\|_{L^2(-\infty,\infty)} \qquad \text{for all } t > 0. \tag{104}$$

This is a useful result as it can be used to deduce stability of the solution of the equation (101) with respect to perturbations of the initial datum in a sense which we shall now explain. Suppose that u_0 and $\tilde{u}_0$ are two functions contained in $L^2(-\infty,\infty)$ and denote by u and $\tilde{u}$ the solutions to (101) resulting from the initial functions u_0 and $\tilde{u}_0$, respectively. Then $u - \tilde{u}$ solves the heat equation with initial datum $u_0 - \tilde{u}_0$, and therefore, by (104), we have that

$$\|u(\cdot,t) - \tilde{u}(\cdot,t)\|_{L^2(-\infty,\infty)} \leq \|u_0 - \tilde{u}_0\|_{L^2(-\infty,\infty)} \qquad \text{for all } t > 0. \tag{105}$$

This inequality implies continuous dependence of the solution on the initial function: small perturbations in u_0 in the $L^2(-\infty,\infty)$ norm will result in small perturbations in the associated analytical solution $u(\cdot,t)$ in the $L^2(-\infty,\infty)$ norm for all $t > 0$.

The inequality (104) is therefore a relevant property, which we shall try to mimic with our numerical approximations of the equation (101).

7.1 Finite difference approximation of the heat equation

We take our computational domain to be **Lecture 13**

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

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

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

and

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

This then motivates us to approximate the heat equation (101) at the point (x_j,t_m) by the following numerical method, called the **explicit Euler scheme**:

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

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

Equivalently, we can write this as

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

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

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

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

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

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

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

$$\begin{aligned} \frac{U_j^{m+1} - U_j^m}{\Delta t} &= (1-\theta)\,\frac{U_{j+1}^m - 2U_j^m + U_{j-1}^m}{(\Delta x)^2} + \theta\,\frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{(\Delta x)^2}, \\ U_j^0 &= u_0(x_j), \qquad j = 0, \pm 1, \pm 2, \ldots, \end{aligned}$$

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

7.1.1 Accuracy of the θ-method

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

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

where

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

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

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

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

Therefore,

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

Similarly,

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

Combining these, we deduce that

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

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

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

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

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

while the Crank–Nicolson scheme has consistency error

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

7.2 Stability of finite difference schemes

In order to be able to replicate the stability property (104) at the discrete level, we require an appropriate notion of stability. We shall say that a finite difference scheme for the unsteady heat equation is **(practically) stable in the ℓ_2 norm**, if **Lecture 14**

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

where

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

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

Definition 23 *The semidiscrete Fourier transform of a function U defined on the infinite mesh $x_j = j\Delta x$, $j = 0, \pm 1, \pm 2, \ldots$, is:*

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

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

Definition 24 *Let $\hat{U}$ be defined on the interval $[-\pi/\Delta x, \pi/\Delta x]$. The inverse semidiscrete Fourier transform of $\hat{U}$ is defined by*

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

We then have the following result.

Lemma 7 (Discrete Parseval's identity) *Let*

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

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

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

The proof of this result is very similar to the proof of Lemma 6, and we shall therefore leave it to the reader as an exercise.

7.2.1 Stability analysis of the explicit Euler scheme

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

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

into the Euler scheme we deduce that

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

Therefore, we have that

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

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

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

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

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

where

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

is the **amplification factor** and

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

is called the CFL number (after Richard Courant, Kurt Friedrichs, and Hans Levy, who first performed an analysis of this kind).[13] By the discrete Parseval identity stated in Lemma 7 we have that

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

[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 (104) we would like to ensure that

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

Thus we demand that

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

i.e., that

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

Using Euler's formula

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

and the trigonometric identity

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

we can restate this as follows:

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

Equivalently, we need to ensure that

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

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

Theorem 24 *Suppose that U_j^m is the solution of the explicit Euler scheme*

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

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

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

$$\|U^m\|_{\ell_2} \leq \|U^0\|_{\ell_2}, \qquad m = 1, 2, \ldots, M. \tag{106}$$

In other words the explicit Euler scheme is **conditionally practically stable**, the condition for stability being that $\mu = \Delta t/\Delta x^2 \leq 1/2$. One can also show that if $\mu > 1/2$, then (106) will fail. In other words, once Δx has been chosen, one must choose Δt so that $\Delta t/\Delta x^2 \leq 1/2$ in order to ensure that the bound (106) holds.

7.2.2 Stability analysis of the implicit Euler scheme

We shall now perform a similar analysis for the **implicit Euler scheme** for the heat equation (101), which is defined as follows:

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

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

Equivalently,

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

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

where, again,

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

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

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

Clearly,

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

for all values of

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

Thus we have the following result.

Theorem 25 *Suppose that U_j^m is the solution of the implicit Euler scheme*

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

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

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

$$\|U^m\|_{\ell_2} \leq \|U^0\|_{\ell_2}, \qquad m = 1, 2, \ldots, M. \tag{107}$$

In other words, the implicit Euler scheme is **unconditionally practically stable**, meaning that the bound (107) holds without any restrictions on Δx and Δt.

7.3 Von Neumann stability

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

Start of optional material

Definition 25 *We shall say that a finite difference scheme for the unsteady heat equation on the time interval $[0, T]$ is **von Neumann stable** in the ℓ_2 norm, if there exists a positive constant $C = C(T)$ such that*

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

where

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

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

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

Lemma 8 *Suppose that the semidiscrete Fourier transform of the solution* $\{U_j^m\}_{j=1}^{\infty}$, $m = 0, 1, \ldots, \frac{T}{\Delta t}$, *of a finite difference scheme for the heat equation satisfies*

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

and

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

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

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

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

Hence,

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

Therefore,

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

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

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

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

End of optional material

7.4 Stability of the θ-scheme

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

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

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

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

To analyse the practical stability of the θ-scheme in the ℓ_2 norm, we shall use Lemma 8 with $C_0 = 0$. Suppose that

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

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

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

Therefore,

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

For practical stability, we demand that

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

which holds if, and only if,

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

Thus we have shown that:

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

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

7.5 Boundary-value problems for parabolic problems

Lecture 15

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

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

subject to the initial condition

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

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

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

Remark 6 *We note in passing that the Neumann initial-boundary-value problem for the heat equation is:*

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

subject to the initial condition

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

and the Neumann boundary conditions

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

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

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

subject to the initial condition

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

and the mixed Dirichlet–Neumann boundary conditions

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

and

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

Matlab code for the Crank–Nicolson scheme

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

7.5.2 The discrete maximum principle

We shall now try to prove a bound, analogous to (103), for the θ-scheme **Lecture 16**

Theorem 26 (Discrete maximum principle for the θ-scheme)
The θ-scheme for the Dirichlet initial-boundary-value problem for the heat equation, with $0 \le \theta \le 1$ and $\mu(1-\theta) \le \frac{1}{2}$, yields a sequence of numerical approximations $\{U_j^m\}_{j=0,\ldots,J;\ m=0,\ldots,M}$ satisfying

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

where

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

and

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

Proof: We rewrite the θ-scheme as

$$(1+2\theta\mu)\,U_j^{m+1} = \theta\mu\left(U_{j+1}^{m+1} + U_{j-1}^{m+1}\right) + (1-\theta)\mu\left(U_{j+1}^m + U_{j-1}^m\right) + [1-2(1-\theta)\mu]\,U_j^m, \qquad (108)$$

and recall that, by hypothesis,

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

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

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

Then,

$$(1+2\theta\mu)\,U_j^{m+1} \;\le\; 2\theta\mu U^\star + 2(1-\theta)\mu U^\star + [1-2(1-\theta)\mu]U^\star = (1+2\theta\mu)\,U^\star, \tag{109}$$

and therefore

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

However, also,

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

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

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

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

In summary then, for

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

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

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

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

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

We close our discussion of finite difference schemes for the heat equation (101) in one space-dimension with the convergence analysis of the θ-scheme for the Dirichlet initial-boundary-value problem. We begin by rewriting the scheme as follows:

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

The scheme is considered subject to the initial condition

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

and the boundary conditions

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

[14] To see that the maximum value $U_j^{m+1} = U^*$ is attained at *each* of points neighbouring (x_j, t_{m+1}) present in the scheme, first observe that if: (a) $\theta = 0$, then U_{j+1}^{m+1} and U_{j-1}^{m+1} are absent from the right-hand side of (108); (b) if $\theta = 1$ then U_{j+1}^{m} and U_{j-1}^{m} are absent from the right-hand side of (108); (c) if $2(1-\theta)\mu = 1$, then U_j^m is absent from the right-hand side of (108), 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 (108). There are therefore four different cases to be discussed: (a), (b), (c) and (d). Suppose that we are in case (d) (the cases (a), (b) and (c) being dealt with identically); if one of U_{j+1}^{m+1}, U_{j-1}^{m+1}, U_{j+1}^{m}, U_{j-1}^{m}, and U_j^m were strictly smaller than $U_j^{m+1} = U^*$, then, by returning to the transition from (108) to (109), we would deduce (109) from (108), but now with the $\le$ symbol in (109) replaced by $<$, which would then imply that $U_j^{m+1} < U^*$. This would, however, contradict the equality $U_j^{m+1} = U^*$ we have already proved. Thus the value $U^{m+1} = U^*$ is attained at *each* of the five point neighbouring (x_j, t_{m+1}).

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

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

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

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

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

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

It then follows that

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

and

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

We define,

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

As, by hypothesis,

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

we have that

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

Hence,

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

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

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

which then implies that

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

Recall that the consistency error of the θ-scheme is

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

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

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

one has the following bound on the global error:

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

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

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

one has

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

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

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

8 Finite difference approximation in two space-dimensions

Start of optional material

Consider the heat equation

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

subject to the initial condition

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

and the Dirichlet boundary condition

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

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

8.1 The explicit Euler scheme

Let

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

and

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

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

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

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

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

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

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

and the boundary condition

$$U_{ij}^m = B(x_i, y_j, t_m), \;\; \text{at the boundary mesh points, for } m = 1, \ldots, M.$$

8.2 The implicit Euler scheme

The implicit Euler scheme is defined analogously. Let $\Delta x := (b-a)/J_x$, $\Delta y := (d-c)/J_y$, $\Delta t := T/M$, and define

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

The implicit Euler finite difference scheme for the problem under consideration is then

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

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

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

and the boundary condition

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

8.3 The θ-scheme

By taking the convex combination of the explicit and implicit Euler schemes, with a parameter $\theta \in [0, 1]$, with $\theta = 0$ corresponding to the explicit Euler scheme and $\theta = 1$ to the implicit Euler scheme, we obtain a one-parameter family of schemes, called the θ-scheme. It is defined as follows.

Let $\Delta x := (b-a)/J_x$, $\Delta y := (d-c)/J_y$, $\Delta t := T/M$, and, for $\theta \in [0, 1]$, consider the finite difference scheme

$$\frac{U_{ij}^{m+1} - U_{ij}^m}{\Delta t} = (1-\theta)\left(\frac{\delta_x^2 U_{ij}^m}{(\Delta x)^2} + \frac{\delta_y^2 U_{ij}^m}{(\Delta y)^2}\right) + \theta\left(\frac{\delta_x^2 U_{ij}^{m+1}}{(\Delta x)^2} + \frac{\delta_y^2 U_{ij}^{m+1}}{(\Delta y)^2}\right),$$

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

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

and the boundary condition

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

The practical stability of the θ-scheme (in the absence of boundary conditions now, i.e. for the pure initial-value problem rather than the initial-boundary-value problem) in the ℓ^2 norm is easily assessed by inserting the Fourier mode

$$U_{ij}^m = [\lambda(k_x, k_y)]^m \, \mathrm{e}^{\imath(k_x x_i + k_y y_j)}$$

into the scheme. This gives

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

where

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

Hence,

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

For practical stability in the ℓ_2 norm, we require that

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

Thus, we demand that

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

which can be restated in the following equivalent form:

$$2(1 - 2\theta)(\mu_x + \mu_y) \le 1.$$

For example, the implicit Euler scheme ($\theta = 1$) and the Crank–Nicolson scheme ($\theta = 1/2$) are unconditionally stable, while the explicit Euler scheme ($\theta = 0$) is only conditionally stable, the stability condition being that Δx, Δy, and Δt satisfy the following inequality:

$$\mu_x + \mu_y \equiv \Delta t \left(\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2}\right) \le \frac{1}{2}.$$

Under a suitable condition the θ-scheme for the initial-boundary-value problem also satisfies a discrete maximum principle. To see this, we rewrite the θ-scheme as

$$\begin{aligned}(1 + 2\theta(\mu_x + \mu_y))U_{ij}^{m+1} = {} & (1 - 2(1-\theta)(\mu_x + \mu_y))U_{ij}^m \\ & + (1-\theta)\mu_x(U_{i+1,j}^m + U_{i-1,j}^m) \\ & + (1-\theta)\mu_y(U_{i,j+1}^m + U_{i,j-1}^m) \\ & + \theta\mu_x(U_{i+1,j}^{m+1} + U_{i-1,j}^{m+1}) \\ & + \theta\mu_y(U_{i,j+1}^{m+1} + U_{i,j-1}^{m+1}),\end{aligned}$$

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

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

and the boundary condition

$$U_{ij}^m = B(x_i, y_j, t_m), \;\; \text{at the boundary mesh points, for } m = 1, \ldots, M.$$

Theorem 27 *Suppose that*

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

Then, the θ-scheme satisfies the following discrete maximum principle:

$$U_{\min} \leq U_{ij}^m \leq U_{\max},$$

where

$$U_{\min} = \min \left\{ \min\{U_{ij}^0\}_{i,j=0}^{J_x, J_y}, \; \min\{U_{ij}^m\}_{m=0}^{M}|_{(x_i, y_j) \in \partial\Omega} \right\}$$

and

$$U_{\max} = \max \left\{ \max\{U_{ij}^0\}_{i,j=0}^{J_x, J_y}, \; \max\{U_{ij}^m\}_{m=0}^{M}|_{(x_i, y_j) \in \partial\Omega} \right\}.$$

PROOF: The proof proceeds by an obvious modification of the proof of the discrete maximum principle for the θ-scheme in one space-dimension. $\square$

In summary, then, for

$$(\mu_x + \mu_y)(1 - \theta) \leq \frac{1}{2}$$

the θ-scheme satisfies the discrete maximum principle. This condition is more demanding than the one for the ℓ_2-stability of the scheme, which requires that

$$(\mu_x + \mu_y)(1 - 2\theta) \leq \frac{1}{2} \qquad \text{for} \quad 0 \leq \theta \leq \tfrac{1}{2}.$$

For example, the Crank–Nicolson scheme is unconditionally stable in the ℓ_2 norm, but for the discrete maximum principle to hold we had to assume that

$$\mu_x + \mu_y = \frac{\Delta t}{(\Delta x)^2} + \frac{\Delta t}{(\Delta y)^2} \leq 1.$$

We close our discussion of the θ-scheme with its error analysis. The starting point is to rewrite the scheme as follows:

$$\begin{aligned}(1 + 2\theta(\mu_x + \mu_y))U_{ij}^{m+1} = {} & (1 - 2(1 - \theta)(\mu_x + \mu_y))U_{ij}^m \\ & + (1 - \theta)\mu_x(U_{i+1,j}^m + U_{i-1,j}^m) \\ & + (1 - \theta)\mu_y(U_{i,j+1}^m + U_{i,j-1}^m) \\ & + \theta\mu_x(U_{i+1,j}^{m+1} + U_{i-1,j}^{m+1}) \\ & + \theta\mu_y(U_{i,j+1}^{m+1} + U_{i,j-1}^{m+1}),\end{aligned}$$

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

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

and the boundary condition

$$U_{ij}^m = B(x_i, y_j, t_m), \;\; \text{at the boundary mesh points, for } m = 1, \ldots, M.$$

Suppose further that

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

The consistency error of the θ-scheme is defined as follows:

$$T_{ij}^m := \frac{u_{ij}^{m+1} - u_{ij}^m}{\Delta t} - (1-\theta)\left(\frac{\delta_x^2 u_{ij}^m}{(\Delta x)^2} + \frac{\delta_y^2 u_{ij}^m}{(\Delta y)^2}\right) - \theta\left(\frac{\delta_x^2 u_{ij}^{m+1}}{(\Delta x)^2} + \frac{\delta_y^2 u_{ij}^{m+1}}{(\Delta y)^2}\right),$$

where

$$u_{ij}^m \equiv u(x_i, y_j, t_m).$$

By performing some elementary but tedious Taylor series expansions, one can deduce that

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

It follows from the definition of the consistency error T_{ij}^m for the θ-scheme that

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

for $i = 1, \ldots, J_x - 1$, $j = 1, \ldots, J_y - 1$, $m = 0, 1, \ldots, M-1$. We define the **global error** as

$$e_{ij}^m := u(x_i, y_j, t_m) - U_{ij}^m.$$

Then, $e_{ij}^0 = 0$ and $e_{ij}^m = 0$ for $(x_i, y_j) \in \partial\Omega$, and

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

We further define,

$$E^m := \max_{i,j} |e_{ij}^m| \quad \text{and} \quad T^m := \max_{i,j} |T_{ij}^m|.$$

As, by hypothesis,

$$1 - 2(1-\theta)(\mu_x + \mu_y) \ge 0,$$

we have

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

Hence,

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

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

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

and we have that

$$\max_{i,j} \max_{0\leq m\leq M} |u(x_i, y_j, t_m) - U_{ij}^m| \leq T \max_{i,j} \max_{0\leq m\leq M} |T_{ij}^m|.$$

The explicit and implicit Euler schemes therefore satisfy:

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

where in the case of the explicit Euler scheme we are assuming that $\mu_x + \mu_y \leq \frac{1}{2}$, while for the Crank–Nicolson scheme we have that

$$\max_{i,j} \max_{0\leq m\leq M} |u(x_i, y_j, t_m) - U_{ij}^m| \leq \text{Const.} \left((\Delta x)^2 + (\Delta y)^2 + (\Delta t)^2\right),$$

assuming that $\mu_x + \mu_y \leq 1$.

8.4 The alternating direction (ADI) method

Except for $\theta = 0$ corresponding to the explicit Euler scheme, for all other values of $\theta \in (0, 1]$ the θ-scheme is an implicit scheme, and its implementation therefore involves the solution of large systems of linear algebraic equations. This is true, in particular, for the Crank–Nicolson scheme corresponding to $\theta = \frac{1}{2}$. Our objective here is to propose a more economical scheme, which replaces the tedious task of solving such large systems of algebraic equations with the successive solution of smaller linear systems in the x and y co-ordinate directions respectively, alternating between solves in the x and y co-ordinate directions. The resulting finite difference scheme is called the alternating direction (or ADI) scheme. We describe its construction starting from the Crank–Nicolson scheme, which has the form:

$$\left(1 - \frac{1}{2}\mu_x\delta_x^2 - \mu_y\frac{1}{2}\delta_y^2\right) U_{ij}^{m+1} = \left(1 + \frac{1}{2}\mu_x\delta_x^2 + \mu_y\frac{1}{2}\delta_y^2\right) U_{ij}^m,$$

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

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

and the boundary condition

$$U_{ij}^m = B(x_i, y_j, t_m), \;\; \text{at the boundary mesh points, for } m = 1, \ldots, M.$$

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

$$\left(1 - \frac{1}{2}\mu_x\delta_x^2\right)\left(1 - \mu_y\frac{1}{2}\delta_y^2\right) U_{ij}^{m+1} = \left(1 + \frac{1}{2}\mu_x\delta_x^2\right)\left(1 + \mu_y\frac{1}{2}\delta_y^2\right) U_{ij}^m.$$

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

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

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

The equivalence is seen by applying

$$\left(1+\frac{1}{2}\mu_x\delta_x^2\right) \text{ to eq. (1)} \quad \text{and} \quad \left(1-\frac{1}{2}\mu_x\delta_x^2\right) \text{ to eq. (2).}$$

The stability in the ℓ^2 norm of the ADI scheme (for the pure initial-value problem now, i.e. with no boundary conditions assumed) is easily seen by substituting the Fourier mode

$$U_{ij}^m = [\lambda(k_x,k_y)]^m \mathrm{e}^{\imath(k_x x_i + k_y y_j)}$$

into the scheme. Hence,

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

Clearly,

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

Consequently, the ADI scheme is unconditionally stable in the ℓ_2 norm. The consistency error of the ADI scheme can be shown (again, after tedious Taylor series expansions) to be

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

The ADI scheme satisfies a discrete maximum principle for $\mu_x \leq 1$ and $\mu_y \leq 1$. The proof of this is similar to the case of the θ-scheme in one space-dimension (cf. the textbook by K.W. Morton and D.F. Mayers, Numerical Solution of Partial Differential Equations: An Introduction, 2nd Edition, CUP, 2005. ISBN: 978-0-521607-93-3. pp. 64–65).

End of optional material