Numerical Solution of Ordinary Differential Equations

E. Süli

October 16, 2001

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). 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 the mathematical analysis of their behaviour, covering the material taught in the M.Sc. in Mathematical Modelling and Scientific Computation in the eight-lecture course *Numerical Solution of Ordinary Differential Equations.*

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 predictor corrector methods, and a brief excursion into numerical methods for stiff systems of ODEs. The final sections are devoted to an overview of classical algorithms for the numerical solution of two-point boundary value problems.

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. Predictor-corrector methods.

Stiffness, stability regions, Gear's methods and their implementation. Nonlinear stability.

Boundary value problems: shooting methods, matrix methods and collocation.

Reading List:

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

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

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] K.W. Morton, *Numerical Solution of Ordinary Differential Equations.* Oxford University Computing Laboratory, 1987.

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

1 Picard's theorem

Ordinary differential equations frequently occur as mathematical models in many branches of science, engineering and economy. 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. In this section we shall be concerned with the construction and the analysis of numerical methods for first-order differential equations of the form

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

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**.

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 the **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, or the

[1] Consider, for example, the initial value problem $y' = y^{2/3}$, $y(0) = 0$; this has two solutions: $y(x) \equiv 0$ and $y(x) = x^3/27$.

[2] Emile Picard (1856–1941)

[3] Rudolf Lipschitz (1832–1903)

lecture notes of P. J. Collins, *Differential and Integral Equations, Part I,* Mathematical Institute Oxford, 1988 (reprinted 1990). 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*:

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

and show, using the conditions of the theorem, that $\{y_n\}_{n=0}^{\infty}$ converges uniformly on the interval $[x_0, X_M]$ to a 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 $\mathbf{y}_0 \in \mathbf{R}^m$ and $\mathbf{f} : [x_0, X_M] \times \mathbf{R}^m \to \mathbf{R}^m$. On introducing the Euclidean norm $\|\cdot\|$ on $\mathbf{R}^m$ by

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

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 $L > 0$ 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 $\mathbf{R}^m$. Indeed, when (7) holds, the Mean Value Theorem implies that (6) is also valid. The converse of this statement is not true; for the function $\mathbf{f}(\mathbf{y}) = (|y_1|, \ldots, |y_m|)^T$, 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 $\epsilon > 0$ there exists $\delta > 0$ 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 $\|\mathbf{v}(x) - \mathbf{w}(x)\| < \epsilon$ 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

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

that is

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

Now substituting (11) into (9) gives

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

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

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

Thus, given $\epsilon > 0$ as in Definition 1, we choose $\delta = \epsilon\exp(-L(X_M - x_0))$ 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 $\Re\lambda > 0$; the solution is stable in the sense of Lyapunov for $\Re\lambda \le 0$ and is asymptotically stable for $\Re\lambda < 0$.

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].

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 $y(x_0) = y_0$

[4] Leonard Euler (1707–1783)

is known, let us suppose that we have already calculated y_n, up to some n, $0 \le n \le N-1$; we define

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

Thus, taking in succession $n = 0, 1, \dots, N-1$, 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

this is called the **Trapezium Rule Method**.

The θ-method is an explicit method for $\theta = 0$ and is an implicit method for $0 < \theta \leq 1$, given that 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 calculate an accurate approximation to the analytical solution on the interval $[0, 0.4]$ and call this the "exact solution". Thus, we consider*

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

Hence,

$$\begin{aligned} y_0(x) &\equiv 0\,, \\ y_1(x) &= \frac{1}{2}x^2\,, \\ y_2(x) &= \frac{1}{2}x^2 - \frac{1}{20}x^5\,, \end{aligned}$$

[5] Using MAPLE V, 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	$y(x_k)$
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

$$y_3(x) = \frac{1}{2}x^2 - \frac{1}{20}x^5 + \frac{1}{160}x^8 - \frac{1}{4400}x^{11} .$$

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 "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 the MAPLE V package by typing in 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, $n = 0, \dots, N$, 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 $n = 0, \dots, N$. 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

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

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

So let us consider Euler's explicit method:

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

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 **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 $\xi_n \in (x_n, x_{n+1})$ 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

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

on 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 $|y_n - y_0| \le Y_M$ from the Lipschitz condition (3) we get

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

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

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

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

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

This estimate, together with the bound

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

which follows from (19), yields

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

To conclude, we note that pursuing an analogous argument it is possible to prove that, in the general case of the θ-method,

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

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 $|e_n| = O(h^2)$ for $\theta = 1/2$, while for $\theta = 0$ and $\theta = 1$, and indeed for any $\theta \neq 1/2$, $|e_n| = O(h)$ 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 truncation 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 then the forward Euler method, it is less convenient from the computational point of view given that 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

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

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 explicit one-step methods.

2.3 General explicit one-step method

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

$$y_{n+1} = y_n + h\Phi(x_n, y_n; h) , \quad n = 0, \ldots, N-1 , \quad y_0 = y(x_0)\,[= \text{specified by (2)}] , \tag{22}$$

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 (22), we define the **global error**, e_n, by

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

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

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - \Phi(x_n, y(x_n); h) \ . \tag{23}$$

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

Theorem 4 *Consider the general one-step method (22) 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,*

$$|\Phi(x, y; h) - \Phi(x, z; h)| \le L_\Phi |y - z|, \qquad \textit{for } (x, y),\ (x, z) \textit{ in } \mathsf{R}\ . \tag{24}$$

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

$$|e_n| \le \mathrm{e}^{L_\Phi (x_n - x_0)} |e_0| + \left[\frac{\mathrm{e}^{L_\Phi (x_n - x_0)} - 1}{L_\Phi} \right] T\ , \quad n = 0, \ldots, N\ , \tag{25}$$

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

PROOF: Subtracting (22) from (23) 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 (24) implies that

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

That is,

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

Hence

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

Observing that $1 + hL_\Phi \le \exp(hL_\Phi)$, we obtain (25). ◇

Let us note that the error bound (20) for Euler's explicit method is a special case of (25). We highlight the practical relevance of the error bound (25) 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 (25) 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*

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

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:

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

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

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

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

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

Thus, given a 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*

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

For such h *we shall have* $|y(x_n) - y_n| = |e_n| \le TOL$ *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 due to 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 (22), we consider the choice of the function Φ. Theorem 4 suggests that if the truncation 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 2 *The numerical method (22) is* ***consistent*** *with the differential equation (1) if the truncation error defined by (23) is such that for any* $\epsilon > 0$ *there exists a positive* $h(\epsilon)$ *for which* $|T_n| < \epsilon$ *for* $0 < h < h(\epsilon)$ *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 (22) 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 (23),

$$\lim_{h\to 0} T_n = y'(x_n) - \Phi(x_n, y(x_n); 0) .$$

This implies that the one-step method (22) is consistent if and only if

$$\Phi(x, y; 0) \equiv f(x, y) . \tag{26}$$

Now we are ready to state a convergence theorem for the general one-step method (22).

Theorem 5 *Suppose that the solution of the initial value problem (1–2) lies in* R *as does its approximation generated from (22) when* $h \le h_0$. *Suppose also that the function* $\Phi(\cdot,\cdot;\cdot)$ *is uniformly continuous on* $\mathsf{R} \times [0, h_0]$ *and satisfies the consistency condition (26) and the Lipschitz condition*

$$|\Phi(x, y; h) - \Phi(x, z; h)| \le L_\Phi |y - z| \qquad \text{on } \mathsf{R} \times [0, h_0] . \tag{27}$$

Then, if successive approximation sequences (y_n), *generated for* $x_n = x_0 + nh$, $n = 1, 2, \ldots, N$, *are obtained from (22) 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*

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

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 $h \le h_0$. Since $y(x_0) = y_0$ and therefore $e_0 = 0$, Theorem 4 implies that

$$|y(x_n) - y_n| \le \left[\frac{\mathrm{e}^{L_\phi(X_M - x_0)} - 1}{L_\phi}\right] \max_{0\le m\le n-1} |T_m| , \qquad n = 1, \ldots, N . \tag{28}$$

From the consistency condition (26) 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 $\epsilon > 0$ there exists $h_1(\epsilon)$ such that

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

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

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

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

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

Inserting this into (28) we deduce that $|y(x_n) - y_n| \to 0$ as $h \to 0$. 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 $h \to 0$ 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 truncation 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 truncation error as the step size h converges to zero, we introduce the following definition.

Definition 3 *The numerical method (22) is said to have* ***order of accuracy*** *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, we now focus on a specific family: explicit Runge–Kutta methods.

2.4 Runge–Kutta methods

In the sense of Definition 3 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 **R-stage Runge–Kutta family** is defined by

$$\begin{aligned}
y_{n+1} &= y_n + h\Phi(x_n, y_n; h)\ , \\
\Phi(x, y; h) &= \sum_{r=1}^{R} c_r k_r\ , \\
k_1 &= f(x, y)\ , \\
k_r &= f\left(x + ha_r, y + h\sum_{s=1}^{r-1} b_{rs} k_s\right)\ , \qquad r = 2, \ldots, R\ , \qquad (29)\\
a_r &= \sum_{s=1}^{r-1} b_{rs}\ , \quad r = 2, \ldots, R\ .
\end{aligned}$$

$$
\begin{array}{c|c}
a = Be & B \\
\hline
 & c^{\mathrm{T}}
\end{array}
\qquad \text{where } e = (1, \ldots, 1)^{\mathrm{T}}.
$$

Figure 1: Butcher table of a Runge–Kutta method

In compressed form, this information is usually displayed in the so-called Butcher table displayed in Figure 1.

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

$$y_{n+1} = y_n + hf(x_n, y_n) \ . \tag{30}$$

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

$$y_{n+1} = y_n + h(c_1 k_1 + c_2 k_2) \ , \tag{31}$$

where

$$
\begin{aligned}
k_1 &= f(x_n, y_n) \ , && (32)\\
k_2 &= f(x_n + a_2 h, y_n + b_{21} h k_1) \ , && (33)
\end{aligned}
$$

and where the parameters c_1, c_2, a_2 and b_{21} are to be determined.[6] Clearly (31–33) can be rewritten in the form (22) and therefore it is a family of one step methods. By the condition (26), 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 truncation error of (31–33) in powers of h, after some algebra we obtain

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

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 $T_n = O(h^2)$ 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 Runge–Kutta methods of the form (31–33):

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

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 truncation error is of the form, respectively,

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

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 (31–33) is referred to as the class of explicit two-stage Runge–Kutta methods.

Exercise 1 *Let α be a non-zero 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 truncation error, $T_n(h, \alpha)$, can be expressed as

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

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 $T_n(h, \alpha) = O(h^2)$ for every non-zero real number α. Show also that if $p \geq 2$ then there exists a non-zero real number α_0 such that $T_n(h, \alpha_0) = O(h^3)$.

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 truncation 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,

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

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

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

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

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

for all (non-zero) 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

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

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

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

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 Runge–Kutta methods. Let us now suppose that $R = 3$ to illustrate the general idea. Thus, we consider the family of methods:

$$y_{n+1} = y_n + h\left[c_1 k_1 + c_2 k_2 + c_3 k_3\right] ,$$

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:

$$\begin{aligned}
k_2 &= f + ha_2(f_x + k_1 f_y) + \frac{1}{2}h^2 a_2^2 (f_{xx} + 2k_1 f_{xy} + k_1^2 f_{yy}) + O(h^3) \\
&= f + ha_2(f_x + f f_y) + \frac{1}{2}h^2 a_2^2 (f_{xx} + 2f f_{xy} + f^2 f_{yy}) + O(h^3) \\
&= f + ha_2 F_1 + \frac{1}{2}h^2 a_2^2 F_2 + O(h^3) ,
\end{aligned}$$

where

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

and

$$\begin{aligned}
k_3 &= f + h\left\{a_3 f_x + \left[(a_3 - b_{32})k_1 + b_{32}k_2\right] f_y\right\} \\
&\quad + \frac{1}{2}h^2 \left\{a_3^2 f_{xx} + 2a_3\left[(a_3 - b_{32})k_1 + b_{32}k_2\right] f_{xy}\right. \\
&\quad \left. + \left[(a_3 - b_{32})k_1 + b_{32}k_2\right]^2 f_{yy}\right\} + O(h^3) \\
&= f + ha_3 F_1 + h^2\left(a_2 b_{32} F_1 f_y + \frac{1}{2}a_3^2 F_2\right) + O(h^3) .
\end{aligned}$$

Substituting these expressions for k_2 and k_3 into (29) with $R = 3$ we find that

$$\begin{aligned}
\Phi(x, y, h) &= (c_1 + c_2 + c_3)f + h(c_2 a_2 + c_3 a_3)F_1 \\
&\quad + \frac{1}{2}h^2\left[2c_3 a_2 b_{32} F_1 f_y + \left(c_2 a_2^2 + c_3 a_3^2\right) F_2\right] + O(h^3) . \qquad (34)
\end{aligned}$$

We match this with the Taylor series expansion:

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

This yields:

$$\begin{aligned} c_1+c_2+c_3 &= 1\ ,\\ c_2a_2+c_3a_3 &= \frac{1}{2}\ ,\\ c_2a_2^2+c_3a_3^2 &= \frac{1}{3}\ ,\\ c_3a_2b_{32} &= \frac{1}{6}\ .\end{aligned}$$

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 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

$$\begin{aligned} y_{n+1} &= y_n+\frac{1}{4}h\left(k_1+3k_3\right)\ ,\\ k_1 &= f(x_n,y_n)\ ,\\ k_2 &= f\left(x_n+\frac{1}{3}h, y_n+\frac{1}{3}hk_1\right)\ ,\\ k_3 &= f\left(x_n+\frac{2}{3}h, y_n+\frac{2}{3}hk_2\right)\ .\end{aligned}$$

(ii) **Standard third-order 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

$$\begin{aligned} y_{n+1} &= y_n+\frac{1}{6}h\left(k_1+4k_2+k_3\right)\ ,\\ k_1 &= f(x_n,y_n)\ ,\\ k_2 &= f\left(x_n+\frac{1}{2}h, y_n+\frac{1}{2}hk_1\right)\ ,\\ k_3 &= f\left(x_n+h, y_n-hk_1+2hk_2\right)\ .\end{aligned}$$

Four-stage 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

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

Here k_2 and k_3 represent approximations to the derivative $y'(\cdot)$ at points on the solution curve, intermediate between $(x_n, y(x_n))$ and $(x_{n+1}, y(x_{n+1}))$, and $\Phi(x_n, y_n; h)$ is a weighted average of the k_i, $i = 1, \ldots, 4$, the weights corresponding to those of Simpson's rule method (to which the fourth-order Runge–Kutta method reduces when $\frac{\partial f}{\partial y} \equiv 0$).

In this section, we have constructed R-stage Runge–Kutta methods of order of accuracy $O(h^R)$, $R = 1, 2, 3, 4$. Is 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 $\leq R - 2$.

2.5 Absolute stability of Runge–Kutta methods

It is instructive to consider the model problem

$$y' = \lambda y , \quad y(0) = y_0 \ (\neq 0), \tag{35}$$

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 a Runge–Kutta method over an interval $[x_0, X_M]$ with $X_M >> x_0$. 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 first-order accurate Runge–Kutta method is Euler's explicit method. Applying (30–35) 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 Euler's explicit 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 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 (35) 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 Runge–Kutta method to the model problem (35) 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 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 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

$$y_{n+1} = y_{n-1} + \frac{1}{3}h\left[f(x_{n-1}, y_{n-1}) + 4f(x_n, y_n) + f(x_{n+1}, y_{n+1})\right] . \tag{36}$$

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 (36) is not a one-step method.

In this section we consider a class of methods of the type (36) 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**

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(x_{n+j}, y_{n+j}) , \tag{37}$$

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 $f(x_{n+k}, y_{n+k})$; due to this implicit dependence on y_{n+k} the method is then called **implicit**. The numerical method (37) 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 (36); 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

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,

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

and

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

Now suppose that u is a real-valued function defined on $\mathbf{R}$ 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

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

[7] J. C. Adams (1819–1892)

and hence

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

Thus, formally,

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

and therefore, again by Taylor series expansion,

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

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

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

On successive truncation 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

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

Letting, again, $u(x) = y(x)$ where y is the solution of the initial-value problem (1–2) and noting that $u'(x) = y'(x) = f(x, y(x))$, on successive truncation of the infinite series on the right results 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 mid-point 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,

$$\begin{aligned}(E-I)D^{-1} &= \Delta_+ D^{-1} = E\Delta_- D^{-1} = hE\Delta_-(hD)^{-1} \\ &= -hE\Delta_-\left[\ln(I-\Delta_-)\right]^{-1}.\end{aligned} \tag{38}$$

Furthermore,

$$\begin{aligned}(E-I)D^{-1} &= E\Delta_- D^{-1} = \Delta_- ED^{-1} = \Delta_-(DE^{-1})^{-1} = h\Delta_-(hDE^{-1})^{-1} \\ &= -h\Delta_-\left[(I-\Delta_-)\ln(I-\Delta_-)\right]^{-1}\ .\end{aligned} \tag{39}$$

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 (38) and (39) we deduce that

$$\begin{aligned}y(x_{n+1}) - y(x_n) &= \int_{x_n}^{x_{n+1}} y'(\xi)\,\mathrm{d}\xi = (E-I)D^{-1}y'(x_n) = (E-I)D^{-1}f(x_n, y(x_n)) \\ &= \begin{cases} -hE\Delta_-\left[\ln(I-\Delta_-)\right]^{-1} f(x_n, y(x_n)) \\ -h\Delta_-\left[(I-\Delta_-)\ln(I-\Delta_-)\right]^{-1} f(x_n, y(x_n))\ . \end{cases}\end{aligned} \tag{40}$$

On expanding $\ln(I-\Delta_-)$ into a Taylor series on the right-hand side of (40) we find that

$$y(x_{n+1}) - y(x_n) \approx h\left[I - \frac{1}{2}\Delta_- - \frac{1}{12}\Delta_-^2 - \frac{1}{24}\Delta_-^3 - \frac{19}{720}\Delta_-^4 - \dots\right] f(x_n, y(x_n)) \tag{41}$$

and

$$y(x_{n+1}) - y(x_n) \approx h\left[I + \frac{1}{2}\Delta_- + \frac{5}{12}\Delta_-^2 + \frac{3}{8}\Delta_-^3 + \frac{251}{720}\Delta_-^4 + \dots\right] f(x_n, y(x_n))\ . \tag{42}$$

Successive truncation of (41) yields the family of Adams–Moulton methods, while similar successive truncation of (42) 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.

3.2 Zero-stability

As is clear from (37) we need k starting values, $y_0, \dots, y_{k-1}$, 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, $y_1, \dots, y_{k-1}$, 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 (37). Thus, we wish to consider the 'stability' of the numerical method with respect to 'small perturbations' in the starting conditions.

Definition 4 *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 different initial data $y_0, y_1, \dots, y_{k-1}$ and $\hat{y}_0, \hat{y}_1, \dots, \hat{y}_{k-1}$, respectively, we have*

$$|y_n - \hat{y}_n| \leq K \max\{|y_0 - \hat{y}_0|, |y_1 - \hat{y}_1|, \dots, |y_{k-1} - \hat{y}_{k-1}|\} \tag{43}$$

for $x_n \leq X_M$, and as h tends to 0.

We shall prove later on 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 4 is called *zero* stability. While Definition 4 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 4 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 (37) 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 $y' = 0$:

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \ldots + \alpha_1 y_{n+1} + \alpha_0 y_n = 0 \ . \tag{44}$$

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

$$y_n = \sum_s p_s(n) z_s^n \ , \tag{45}$$

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 $\hat{y}_0 = \hat{y}_1 = \ldots = \hat{y}_{k-1} = 0$ for which the corresponding solution of (44) is $(\hat{y}_n)$ with $\hat{y}_n = 0$ for all n, we see that (43) 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, K.W. Morton, *Numerical Solution of Ordinary Differential Equations*, Oxford University Computing Laboratory, 1987, or P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962. ◇

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 (37). For this purpose, as in the case of one-step methods, we introduce the notion of truncation error. Thus, suppose that $y(x)$ is a solution of the ordinary differential equation (1). Then the truncation error of (37) is defined as follows:

$$T_n = \frac{\sum_{j=0}^{k} [\alpha_j y(x_{n+j}) - h\beta_j y'(x_{n+j})]}{h \sum_{j=0}^{k} \beta_j} . \tag{46}$$

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 truncation error can be thought of as the residual that is obtained by inserting the solution of the differential equation into the formula (37) 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 5 *The numerical scheme (37) is said to be* ***consistent*** *with the differential equation (1) if the truncation error defined by (46) is such that for any $\epsilon > 0$ there exists $h(\epsilon)$ for which*

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

and 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 (46) to obtain

$$T_n = \frac{1}{h\sigma(1)}[C_0 y(x_n) + C_1 h y'(x_n) + C_2 h^2 y''(x_n) + \dots] \tag{47}$$

where

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

For consistency we need that $T_n \to 0$ as $h \to 0$ and this requires that $C_0 = 0$ and $C_1 = 0$; in terms of the characteristic polynomials 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 6 *The numerical method (37) is said to have* **order of accuracy** 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* $(x_n, y(x_n)), \dots, (x_{n+k}, y(x_{n+k}))$ *on the solution curve.*

Thus we deduce from (47) that the method is of order of accuracy p if and only if

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

In this case,

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

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, containing one free parameter. Determine the order 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 (47) are all equal to zero, given that we wish to maximise the order 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

$$y_{n+2} - (1 + a) y_{n+1} + a y_n = \frac{1}{12} h \left[(5 + a) f_{n+2} + 8(1 - a) f_{n+1} - (1 + 5a) f_n \right] . \tag{48}$$

Further,

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

If $a \neq -1$ then $C_4 \neq 0$, and the method (48) is third order accurate. If, on the other hand, $a = -1$, then $C_4 = 0$ and $C_5 \neq 0$ and the method (48) becomes Simpson's rule method – a fourth-order accurate two-step method. The error constant is:

$$C_4 = -\frac{1}{4!}(1 + a) , \qquad a \neq -1 , \tag{49}$$

$$C_5 = -\frac{4}{3 * 5!} , \qquad a = -1 . \tag{50}$$

◇

Exercise 3 *Determine all values of the real parameter b 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 the method is 4. Is the method convergent for this value of b? Show further that if the method is zero-stable than its order cannot exceed 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 truncation error can be written in the form

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

If $b = 6$, then $T_n = O(h^4)$ and so the method is of order 4. 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 $T_n = O(h^2)$. $\diamond$

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, $n = 0, \dots, N$, 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.

Definition 7 *The linear multistep method (37) is said to be* ***convergent*** *if, for all initial value problems (1–2) subject to the hypotheses of Theorem 1, we have that*

$$\lim_{\substack{h \to 0 \\ nh = x - x_0}} y_n = y(x_n) \tag{51}$$

holds for all $x \in [x_0, X_M]$ and for all solutions $\{y_n\}_{n=0}^N$ of the difference equation (37) with ***consistent starting conditions****, i.e. with starting conditions $y_s = \eta_s(h)$, $s = 0, 1, \dots, k-1$, for which $\lim_{h\to 0} \eta_s(h) = y_0$, $s = 0, 1, \dots, k-1$.*

We emphasize here that Definition 7 requires that (51) hold *not only* for those sequences $\{y_n\}_{n=0}^N$ which have been generated from (37) 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 (37) is that it be zero-stable.*

PROOF: Let us suppose that the linear multi-step method (37) 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 (37) to this problem yields the difference equation

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \ldots + \alpha_0 y_n = 0 . \tag{52}$$

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

$$\lim_{\substack{h \to 0 \\ nh = x}} y_n = 0 , \tag{53}$$

for all solutions of (52) satisfying $y_s = \eta_s(h)$, $s = 0, \ldots, k-1$, where

$$\lim_{h \to 0} \eta_s(h) = 0 , \qquad s = 0, 1, \ldots k-1 . \tag{54}$$

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 (52) satisfying (54). 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 (37) lies in the closed unit disc.

Next we prove that any root of the first characteristic polynomial of (37) that lies on the unit circle must be *simple*. Assume, instead, that $z = re^{i\phi}$, is a *multiple* root of $\rho(z)$; $|r| = 1$, $0 \leq \phi < 2\pi$. We shall prove below that this contradicts our assumption that the method (52) is convergent. It is easy to check that the numbers

$$y_n = h^{1/2} n r^n \cos(n\phi) \tag{55}$$

define a solution to (52) which satisfies (54), for

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

If $\phi = 0$ or $\phi = \pi$, it follows from (55) with $h = x/n$ that

$$|y_n| = x^{1/2} n^{1/2} r^n \ . \tag{56}$$

Since, by assumption, $|r| = 1$, we deduce from (56) that $\lim_{n\to\infty} |y_n| = \infty$, which contradicts (53).

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

$$\frac{z_n^2 - z_{n+1} z_{n-1}}{\sin^2 \phi} = r^{2n} \ , \tag{57}$$

where $z_n = n^{-1} h^{-1/2} y_n = h^{1/2} x^{-1} y_n$. Since, by (53), $\lim_{n\to\infty} z_n = 0$, it follows that the left-hand side of (57) converges to 0 as $n \to \infty$. But then the same must be true of the right-hand side of (57); however, the right-hand side of (57) of cannot converge to 0 as $n \to \infty$, since $|z| = 1$. 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 (37) 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. ◇

Theorem 8 *A necessary condition for the convergence of the linear multi-step method (37) is that it be consistent.*

Proof: Let us suppose that the linear multi-step method (37) 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 (37) to this problem yields the difference equation

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \ldots + \alpha_0 y_n = 0 \ . \tag{58}$$

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

$$\lim_{\substack{h\to 0 \\ nh = x}} y_n = 1 \ . \tag{59}$$

Since in the present case y_n is independent of the choice of h, (59) is equivalent to saying that

$$\lim_{n\to\infty} y_n = 1 \ . \tag{60}$$

Passing to the limit $n \to \infty$ in (58), we deduce that

$$\alpha_k + \alpha_{k-1} + \ldots + \alpha_0 = 0 \ . \tag{61}$$

Recalling the definition of C_0, (61) 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 (37) now becomes

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \ldots + \alpha_0 y_n = h(\beta_k + \beta_{k-1} + \ldots + \beta_0) \ , \tag{62}$$

where $X_M - x_0 = X_M - 0 = Nh$ and $1 \le n \le N - k$. For a convergent method every solution of (62) satisfying

$$\lim_{h\to 0} \eta_s(h) = 0 \ , \qquad s = 0, 1, \ldots k-1 \ , \tag{63}$$

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

$$\lim_{\substack{h\to 0\\ nh=x}} y_n = x \ . \tag{64}$$

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

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

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

$$K = \frac{\beta_k + \ldots + \beta_2 + \beta_1}{k\alpha_k + \ldots + 2\alpha_2 + \alpha_1} \ ; \tag{65}$$

this sequence clearly satisfies (63) and is the solution of (62). Furthermore, (64) 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 (65),

$$C_1 = (k\alpha_k + \ldots + 2\alpha_2 + \alpha_1) - (\beta_k + \ldots + \beta_2 + \beta_1) = 0 \ ;$$

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

3.4.2 Sufficient conditions for convergence

We begin by establishing some preliminary results.

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

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

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

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

$$f(z) = \frac{1}{\hat{\rho}(z)} - \frac{A_1}{z - \frac{1}{z_1}} - \ldots - \frac{A_m}{z - \frac{1}{z_m}} \tag{66}$$

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 (66) 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

$$\alpha_k e_{m+k} + \alpha_{k-1} e_{m+k-1} + \ldots + \alpha_0 e_0 = h(\beta_{k,m} e_{m+k} + \beta_{k-1,m} e_{m+k-1} + \ldots + \beta_{0,m} e_m) + \lambda_m \ . \tag{67}$$

The result is stated in the next Lemma.

Lemma 2 *Suppose that all roots of the polynomial $\rho(z) = \alpha_k z^k + \ldots + \alpha_1 z + \alpha_0$ 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*

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

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

and let $0 \leq h < |\alpha_k| \beta^{-1}$. Then every solution of (67) for which

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

[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.*]

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*

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

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 (67) 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

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

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

$$\alpha_k \gamma_l + \alpha_{k-1}\gamma_{l-1} + \ldots + \alpha_0 \gamma_{l-k} = \begin{cases} 1 & \text{for } l = 0 \\ 0 & \text{for } l > 0 \end{cases} \tag{68}$$

we have that

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

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

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

Thus, by our assumptions and noting that by (68) $\gamma_0 = \alpha_k^{-1}$,

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

Hence,

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

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

$$|e_n| \leq K^* + hL^* \sum_{m=0}^{n-1} |e_m| \,, \qquad n = 0, 1, \ldots, N \,. \tag{69}$$

The final estimate is deduced from (69) by induction. First, we note that by virtue of (68), $A\Gamma \geq 1$. Consequently, $K^* \geq \Gamma AEk \geq Ek \geq E$. Now, letting $n = 1$ in (69),

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

Repeating this argument, we find that

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

Now suppose that this inequality has already been shown to hold for $m = 0, 1, \ldots, n-1$, where $n \geq k$; we shall prove that it then also holds for $m = n$, which will complete the induction. Indeed, we have from (69) that

$$|e_n| \leq K^* + hL^*K^* \frac{(1 + hL^*)^n - 1}{hL^*} = K^*(1 + hL^*)^n \,. \tag{70}$$

Further, as $1 + hL^* \leq \mathrm{e}^{hL^*}$ we have from (70) that

$$|e_n| \leq K^* e^{hL^* n} \,, \qquad n = 0, 1, \ldots, N \,. \tag{71}$$

That completes the proof of the lemma. We remark that the implication (69) $\Rightarrow$ (71) 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

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

given that the starting conditions $y_s = \eta_s(h)$, $s = 0, \ldots, k$, 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 truncation error of (37):

$$T_n = \frac{1}{h\sigma(1)} \left[\sum_{j=0}^{k} \alpha_j y(x_{n+j}) - h\beta_j y'(x_{n+j}) \right] \,. \tag{72}$$

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

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

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

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

where $|\theta_s| \leq 1$. Further, by the Mean Value Theorem, there exists $\xi_s \in (x_n, x_{n+s})$ 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 $|\theta'_s| \leq 1$.

Now we can write

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

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

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

Thus,

$$|\sigma(1)T_n| \leq K\chi(kh) \ . \tag{73}$$

where

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

Comparing (37) with (72), we conclude that the global error $e_m = y(x_m) - y_m$ satisfies

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

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 (73), we then have that

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

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

$$|e_n| \leq \Gamma^*\left[Ak\delta(h) + (X_M - x_0)K\chi(kh)\right] \exp[(x_n - x_0)L\Gamma^* B] \ , \tag{74}$$

where

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

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 (74), 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. ◇

On 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 truncation error $O(h^p)$, then the global error $e_n = y(x_n) - y_n$ is also $O(h^p)$.*

By virtue of 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 a zero-stable linear multi-step method

Let us suppose that we have already chosen the coefficients α_j, $j = 0, \ldots, k$, of the k-step method (37). 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 the resulting method (37) 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 (37) 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 truncation error (46)).

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

$$\phi(z) = \frac{\rho(z)}{\log z} - \sigma(z) ; \tag{75}$$

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

Lemma 3 *Suppose that p is a positive integer. The linear multistep method (37), with stability polynomials $\rho(z)$ and $\sigma(z)$, is of order of accuracy p if and only if the function $\phi(z)$ defined by (75) has a zero of multiplicity p at $z = 1$.*

PROOF: Let us suppose that the k-step method (37) 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 truncation error of the form:

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

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

$$T_n \equiv \frac{\mathrm{e}^{nh}}{h\sigma(1)} \left[\rho(\mathrm{e}^h) - h\sigma(\mathrm{e}^h)\right] = \mathrm{e}^{nh} \frac{C_{p+1}}{\sigma(1)} h^p + O(h^{p+1}) , \tag{76}$$

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

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

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 $g(0) = g'(0) = \ldots = g^{(p)}(0) = 0$, 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,

$$g'(0) = 0 = \sum_{j=0}^{k} (j\alpha_j - \beta_j) = C_1 ,$$

$$g''(0) = 0 = \sum_{j=0}^{k} (j^2\alpha_j - 2j\beta_j) = 2C_2 ,$$

$$g'''(0) = 0 = \sum_{j=0}^{k} (j^3\alpha_j - 3j^2\beta_j) = 2C_3 ,$$

etc.

$$g^{(p)}(0) = 0 = \sum_{j=0}^{k} (j^p\alpha_j - pj^{p-1}\beta_j) = p!\, C_p .$$

We deduce that $C_0 = C_1 = C_2 = \ldots = C_p = 0$; since $g^{(p+1)}(0) \neq 0$ we have that $C_{p+1} \neq 0$. Consequently (37) 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:

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

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

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

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

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

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 $z = 1$:*

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

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

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

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 Simpson's method.

By inspection, the linear k-step method (37) 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 implicit (given that in the latter case β_k is fixed to have value 0). According to (47), 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 (37) 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 open half-plane $\Re z < 0$ of the z-plane; the circle $|\zeta| = 1$ is mapped onto the imaginary axis $\Re z = 0$, 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, $\deg(r) \leq k$ and $\deg(s) \leq k$.

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,

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

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 (75) 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

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

where

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

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

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

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,

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

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

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

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

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

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

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

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 $a_2 = a_4 = \ldots = a_k = 0$, 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

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

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 8 *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 Simpson's rule method is the zero-stable linear 2-step method of maximum order, we deduce that Simpson's 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.

3.6 Absolute stability of linear multistep methods

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

$$y' = \lambda y \ , \quad x > 0 \ , \qquad y(0) = y_0 \ (\neq 0) \ , \tag{77}$$

where $\Re\lambda < 0$. Indeed,

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

and therefore,

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

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, $\Re\lambda < 0$.

Now consider the linear k-step method (37) and apply it to the model problem (77) with λ real and negative. This yields the linear difference equation

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

Given that 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

$$\pi(z; \bar{h}) = \rho(z) - \bar{h}\sigma(z) \ , \qquad (\bar{h} = h\lambda) \ , \tag{78}$$

it follows that y_n will converge to zero for $h > 0$ fixed and $n \to \infty$ if and only if all roots of $\pi(z; h)$ have modulus < 1. The kth degree polynomial $\pi(z; \bar{h})$ defined by (78) 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 9 *The linear multistep method (37) 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 (78) 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 (77) 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: Given that the method is consistent, there exists an integer $p \geq 1$ such that $C_0 = C_1 = \ldots = C_p = 0$ and $C_{p+1} \neq 0$. Let us consider

$$\begin{aligned}
\pi(\mathrm{e}^{\bar{h}};\bar{h}) &= \rho(\mathrm{e}^{\bar{h}}) - \bar{h}\sigma(\mathrm{e}^{\bar{h}}) \\
&= \sum_{j=0}^{k}\left[\alpha_j \mathrm{e}^{\bar{h}j} - \bar{h}\beta_j e^{\bar{h}j}\right] \\
&= \sum_{j=0}^{k}\left[\alpha_j \sum_{q=0}^{\infty}\frac{(\bar{h}j)^q}{q!} - \beta_j \sum_{q=0}^{\infty}\frac{\bar{h}^{q+1}j^q}{q!}\right] \\
&= \sum_{j=0}^{k}\left[\alpha_j \sum_{q=0}^{\infty}\frac{(\bar{h}j)^q}{q!} - \beta_j \sum_{q=1}^{\infty}\frac{\bar{h}^{q}j^{q-1}}{(q-1)!}\right] \\
&= \sum_{j=0}^{k}\alpha_j + \sum_{j=0}^{k}\left[\alpha_j \sum_{q=1}^{\infty}\frac{(\bar{h}j)^q}{q!} - \beta_j \sum_{q=1}^{\infty}\frac{\bar{h}^{q}j^{q-1}}{(q-1)!}\right] \\
&= \sum_{j=0}^{k}\alpha_j + \sum_{q=1}^{\infty}\bar{h}^q\left[\sum_{j=0}^{k}\alpha_j\frac{j^q}{q!} - \sum_{j=0}^{k}\beta_j\frac{j^{q-1}}{(q-1)!}\right] \\
&= C_0 + \sum_{q=1}^{\infty}\bar{h}^q C_q \\
&= \sum_{q=p+1}^{\infty} C_q\bar{h}^q = O(\bar{h}^{p+1})\ .
\end{aligned} \tag{79}$$

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

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

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

$$\pi(\mathrm{e}^{\bar{h}};\bar{h}) = (\alpha_k - \bar{h}\beta_k)(\mathrm{e}^{\bar{h}} - r_1(\bar{h}))\ldots(\mathrm{e}^{\bar{h}} - r_k(\bar{h}))\ . \tag{80}$$

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 $\zeta(z)$. Since, by assumption, the method is consistent, 1 is a root of $\zeta(z)$; furthermore, by zero-stability 1 is a simple root of $\zeta(z)$. 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 conclude from (80) 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 (79), $\pi(\mathrm{e}^{\bar{h}};\bar{h}) = O(\bar{h}^{p+1})$, so it follows that

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

Thus we have shown that

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

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

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

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

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

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

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

Now,

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

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

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

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 open left half-plane $\Re z < 0$ 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

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

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

$$a_0 z^k + a_1 z^{k-1} + \ldots + a_k \, . \tag{81}$$

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 (81) lie in the open left half-plane $\Re z < 0$.

Theorem 15 (**Routh–Hurwitz Criterion**) *The roots of (81) lie in the open left half-plane if and only if all the leading principal minors of the $k \times k$ matrix*

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

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 (81) lie in the open left 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: On 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

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

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 $(-\frac{4}{3}, 0)$. ⋄

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

$$y_1 - y_0 = hf_0\,;$$

(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

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

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 means of the fixed-point iteration

$$\alpha_k y_{n+k}^{[s+1]}+\sum_{j+0}^{k-1}\alpha_j y_{n+j}=h\beta_k f(x_{n+k},y_{n+k}^{[s]})+h\sum_{j=0}^{k-1}\beta_j f_{n+j}\ ,\qquad s=1,2,3,\dots\ ,$$

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

Theoretically, we would iterate until the iterates $y^{[s]}_{n+k}$ have converged (in practice, until some stopping criterion such as $|y^{[s+1]}_{n+k} - y^{[s]}_{n+k}| < \epsilon$ 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 due to the amount of work involved: each step of the iteration involves an evaluation of $f(x_{n+k}, y^{[s]}_{n+k})$ which may be quite time-consuming. In order to keep to a minimum the number of times $f(x_{n+k}, y^{[s]}_{n+k})$ is evaluated, the initial guess $y^{[0]}_{n+k}$ must be chosen accurately. This is achieved by evaluating $y^{[0]}_{n+k}$ 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

$$\rho^*(z) = \sum_{j=0}^{k} \alpha^*_j z^j \ , \quad \alpha^*_k = 1 \ , \qquad \sigma^*(z) = \sum_{j=0}^{k-1} \beta^*_j z^j \ , \tag{82}$$

and suppose that the corrector has characteristic polynomials

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

Suppose that m is a positive integer: it will denote the number of times the corrector is applied; in practice $m \le 2$. 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) $P(EC)^m E$

$$\begin{aligned}
y^{[0]}_{n+k} + \sum_{j=0}^{k-1} \alpha^*_j y^{[m]}_{n+j} &= h \sum_{j=0}^{k-1} \beta^*_j f^{[m]}_{n+j} \ , \\
f^{[s]}_{n+k} &= f(x_{n+k}, y^{[s]}_{n+k}) \ , \\
y^{[s+1]}_{n+k} + \sum_{j=0}^{k-1} \alpha_j y^{[m]}_{n+j} &= h\beta_k f^{[s]}_{n+k} + h \sum_{j=0}^{k-1} \beta_j f^{[m]}_{n+j} \ , \qquad s = 0, \dots, m-1, \\
f^{[m]}_{n+k} &= f(x_{n+k}, y^{[m]}_{n+k}) \ ,
\end{aligned}$$

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

b) $P(EC)^m$

$$y^{[0]}_{n+k} + \sum_{j=0}^{k-1} \alpha^*_j y^{[m]}_{n+j} = h \sum_{j=0}^{k-1} \beta^*_j f^{[m-1]}_{n+j} \ ,$$

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

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

3.8.1 Absolute stability of predictor-corrector methods

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

$$
y' = \lambda y , \qquad y(0) = y_0 \ (\neq 0) , \tag{84}
$$

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

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

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

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

Equivalently,

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

By induction,

$$
\begin{aligned}
&y^{[m]}_{n+k} + \left(1 + \bar{h}\beta_k + \ldots + (\bar{h}\beta_k)^{m-1}\right) \sum_{j=0}^{k-1} \alpha_j y^{[m]}_{n+j} \\
&= (\bar{h}\beta_k)^m \left(\bar{h} \sum_{j=0}^{k-1} \beta^*_j y^{[m]}_{n+j} - \sum_{j=0}^{k-1} \alpha^*_j y^{[m]}_{n+j}\right) + \left(1 + \bar{h}\beta_k + \ldots + (\bar{h}\beta_k)^{m-1}\right) \bar{h} \sum_{j=0}^{k-1} \beta_j y^{[m]}_{n+j} \,.
\end{aligned}
$$

For m fixed, this is a kth order difference equation involving $y^{[m]}_n, \ldots, y^{[m]}_{n+k}$. 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

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

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

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

where

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

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

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

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

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

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 (84) with $m \geq 1$. We have that

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

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

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

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

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

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

4 Stiff problems

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

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}) \ , \qquad \mathbf{y}(a) = \mathbf{y}_0 \ , \tag{85}$$

where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)^T$. A linear k-step method for the numerical solution of (85) has the form

$$\sum_{j=0}^{k} \alpha_j \mathbf{y}_{n+j} = h \sum_{j=0}^{k} \beta_j \mathbf{f}_{n+j} \ , \tag{86}$$

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 (86) becomes

$$\sum_{j=0}^{k} (\alpha_j I - h\beta_j A)\mathbf{y}_{n+j} = h\sigma(1)\mathbf{b} \ , \tag{87}$$

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

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

Let us define $\mathbf{z} = H\mathbf{y}$ and $\mathbf{c} = h\sigma(1)H\mathbf{b}$. Thus, (87) can be rewritten as

$$\sum_{j=0}^{k} (\alpha_j I - h\beta_j \Lambda)\mathbf{z}_{n+j} = \mathbf{c} \ , \tag{88}$$

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: given that 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 the definition of absolute stability.

Definition 10 *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 (78), 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 the ordinary differential equation $y' = \lambda y$, $y(x_0) = y_0$.

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

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

rewritten as a first-order system in the vector $(u, v)^T$, with $u = y$ and $v = y'$. Suppose that $\lambda >> 1$. What choice of the step size h will guarantee absolute stability in the sense of Definition 10?

SOLUTION: a) For Euler's explicit method $\rho(z) = z - 1$ and $\sigma(z) = 1$, so that $\pi(z; \bar{h}) = \rho(z) - \bar{h}\sigma(z) = (z - 1) - \bar{h} = z - (1 + \bar{h})$. 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 circle centred at -1.

b) Now writing $u = y$ and $v = y'$ and $\mathbf{y} = (u, v)^T$, 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,

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

Hence, $r_1 = -1$, $r_2 = -\lambda$; we deduce that the method is absolutely stable provided that $h < 2/\lambda$. 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 shown in the figure below for $\lambda = 45$. Note that v exhibits a rapidly varying behaviour (fast time scale) 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 most of the interval $[0, 1]$ both u and v are slowly varying when $\lambda = 45$, we are forced to use a step size of $h < 2/45$ in order to ensure that the method is absolutely stable. $\diamond$

In the example the v component of the solution exhibited two vastly different time scales; in addition, the fast time scale (which occurs between $x = 0$ and $x \approx 1/\lambda$) 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, $h < 2/\lambda$, 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.

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

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

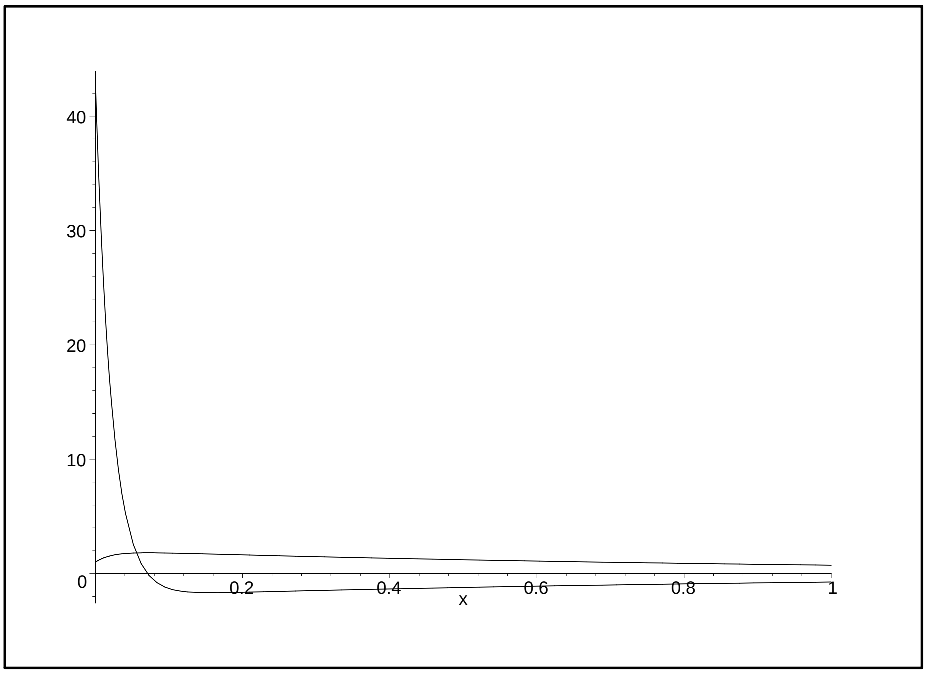

Figure 2: The functions u and v plotted against x for $x \in [0,1]$.

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 deduce that the method has the region of stability

$$\mathcal{R} = \{\bar{h} \, : \, |1 - \bar{h}| > 1\} \, .$$

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

Definition 11 *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-hand complex half-plane $\Re(h\lambda) < 0$.*

As the next theorem by Dahlquist (1963) shows, Definition 11 is far too restrictive.

Theorem 16

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

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

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

Thus we adopt the following definition due to Widlund (1967).

Definition 12 *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 wedge*

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

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 $\Re\lambda < 0$ 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 happen to lie in some wedge W_β then an $A(\beta)$-stable method can be used for the numerical solution of the initial value problem (85) without any restrictions on the step size h. In particular, if all eigenvalues of A are real and nonnegative, 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 $\Re\bar{h} = -a$, $a > 0$, in the complex plane, while those that are responsible for the slow transients are clustered around zero.

Definition 13 *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*

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

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.

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

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

4.2 Backward differentiation methods for stiff systems

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:

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

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 $p \geq 7$ 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 multi-step 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

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

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

$$Lh|\beta_k| < 1 \tag{89}$$

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 (89) 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:

$$\mathbf{y}_{n+k}^{[s+1]} = \mathbf{y}_{n+k}^{[s]} - \left[I - h\beta_k \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(x_{n+k}, \mathbf{y}_{n+k}^{[s]})\right]^{-1} \left[\mathbf{y}_{n+k}^{[s]} - h\beta_k \mathbf{f}(x_{n+k}, \mathbf{y}_{n+k}^{[s]}) - \mathbf{g}_{n+k}\right] , \tag{90}$$

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 (90) 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.

5 Nonlinear Stability

All notions of stability which were considered so far in these notes rest on the assumption that $f(x, y) = \lambda y$ or $\mathbf{f}(x, \mathbf{y}) = A\mathbf{y} + b$. The purpose of this section is to develop an appropriate theoretical framework which is directly applicable to the stability analysis of numerical methods for nonlinear ODEs.

Consider the linear system of differential equations $\mathbf{y}' = A\mathbf{y}$, $\mathbf{y}(x_0) = \mathbf{y}_0$, where all eigenvalues of the $m \times m$ matrix A have negative real part. Then $\|\mathbf{y}(x)\|$ decreases as x increases; also, neighbouring solution curves get closer as x increases: if $\mathbf{u}(x)$ and $\mathbf{v}(x)$ are two solutions to $\mathbf{y}' = A\mathbf{y}$ subject to $\mathbf{u}(x_0) = \mathbf{u}_0$ and $\mathbf{v}(x_0) = \mathbf{v}_0$, respectively, then

$$\|\mathbf{u}(x_2) - \mathbf{v}(x_2)\| \leq \mathrm{e}^{(x_2 - x_1)\max_i \Re\lambda_i(A)}\|\mathbf{u}(x_1) - \mathbf{v}(x_1)\| , \qquad x_2 \geq x_1 \geq x_0 , \tag{91}$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbf{R}^m$. Clearly,

$$0 < \mathrm{e}^{(x_2 - x_1)\max_i \Re\lambda_i(A)} \leq 1$$

for $x_2 \geq x_1 \geq x_0$ and therefore,

$$\|\mathbf{u}(x_2) - \mathbf{v}(x_2)\| \leq \|\mathbf{u}(x_1) - \mathbf{v}(x_1)\| , \qquad x_2 \geq x_1 \geq x_0 .$$

It is the property (91) that has a natural extension to nonlinear differential equations and leads to the following definition.

Definition 14 *Suppose that $\mathbf{u}$ and $\mathbf{v}$ are two solutions of the differential equation $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$ subject to respective initial conditions $\mathbf{u}(x_0) = \mathbf{u}_0$, $\mathbf{v}(x_0) = \mathbf{v}_0$. If*

$$\|\mathbf{u}(x_2) - \mathbf{v}(x_1)\| \leq \|\mathbf{u}(x_1) - \mathbf{v}(x_1)\|$$

*for all real numbers x_2 and x_1 such that $x_2 \geq x_1 \geq x_0$, where $\|\cdot\|$ denotes the Euclidean norm on $\mathbf{C}^m$, then the solutions $\mathbf{u}$ and $\mathbf{v}$ are said to be **contractive** in the norm $\|\cdot\|$.*

To see what assumptions of $\mathbf{f}$ we must make to ensure that two solutions $\mathbf{u}$ and $\mathbf{v}$ to the differential equations $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$, with respective initial conditions $\mathbf{u}(x_0) = \mathbf{u}_0$, $\mathbf{v}(x_0) = \mathbf{v}_0$, are contractive, let $\langle \cdot, \cdot \rangle$ denote the inner product of $\mathbf{R}^m$ and consider the inner product of $\mathbf{u}' - \mathbf{v}' = \mathbf{f}(x, \mathbf{u}) - \mathbf{f}(x, \mathbf{v})$ with $\mathbf{u} - \mathbf{v}$. This yields,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}\|\mathbf{u}(x) - \mathbf{v}(x)\|^2 = \langle \mathbf{f}(x, \mathbf{u}(x)) - \mathbf{f}(x, \mathbf{v}(x)), \mathbf{u}(x) - \mathbf{v}(x) \rangle .$$

Thus, if

$$\langle \mathbf{f}(x, \mathbf{u}(x)) - \mathbf{f}(x, \mathbf{v}(x)), \mathbf{u}(x) - \mathbf{v}(x) \rangle \leq 0 \tag{92}$$

for all $x \geq x_0$ then

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}\|\mathbf{u}(x) - \mathbf{v}(x)\|^2 \leq 0$$

for all $x \geq x_0$, and therefore the solutions $\mathbf{u}$ and $\mathbf{v}$ corresponding to initial conditions $\mathbf{u}_0$ and $\mathbf{v}_0$, respectively, are contractive in the norm $\|\cdot\|$. The inequality (92) motivates the following definition.

Definition 15 *The system of differential equations $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$ is said to be **dissipative** in the interval $[x_0, \infty)$ if*

$$\langle \mathbf{f}(x, \mathbf{u}) - \mathbf{f}(x, \mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \leq 0 \tag{93}$$

for all $x \geq x_0$ and all $\mathbf{u}$ and $\mathbf{v}$ in $\mathbf{R}^m$.

Thus we have proved that if the system of differential equations is dissipative then any solutions $\mathbf{u}$ and $\mathbf{v}$ corresponding to respective initial values $\mathbf{u}_0$ and $\mathbf{v}_0$ are contractive. A slight generalisation of (92) results in the following definition.

Definition 16 *The function $\mathbf{f}(x, \cdot)$ is said to satisfy a **one-sided Lipschitz condition** on the interval $[x_0, \infty)$ if there exists a function $\nu(x)$ such that*

$$\langle \mathbf{f}(x, \mathbf{u}) - \mathbf{f}(x, \mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \leq \nu(x)\|\mathbf{u} - \mathbf{v}\|^2 \tag{94}$$

for all $x \in [x_0, \infty)$.

In particular, if $\mathbf{f}(x, \cdot)$ satisfies a one-sided Lipschitz condition on $[x_0, \infty)$ and $\nu(x) \leq 0$ for all $x \in [x_0, \infty)$, then the differential equation $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$ is dissipative on this interval, and therefore any pair of solutions $\mathbf{u}$ and $\mathbf{v}$ to this equations, corresponding to respective initial values $\mathbf{u}_0$ and $\mathbf{v}_0$ are also contractive.

Now we shall consider numerical methods for the solution of an initial value problem for a dissipative differential equation $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$ and develop conditions under which numerical solutions are also contractive in a suitable norm. In order to keep the presentation simple we shall suppose that $\mathbf{f}$ is independent of x and, instead of a linear k-step method,

$$\sum_{j=0}^{k} \alpha_j \mathbf{y}_{n+j} = h \sum_{j=0}^{k} \beta_j \mathbf{f}(\mathbf{y}_{n+j}) \tag{95}$$

for the numerical solution of $\mathbf{y}' = \mathbf{f}(\mathbf{y})$, $\mathbf{y}(x_0) = \mathbf{y}_0$, we shall consider its **one-leg twin**

$$\sum_{j=0}^{k} \alpha_j \mathbf{y}_{n+j} = h\mathbf{f}\left(\sum_{j=0}^{k} \beta_j \mathbf{y}_{n+j}\right) . \tag{96}$$

For example, the one-leg twin of the trapezium rule method

$$\mathbf{y}_{n+1} - \mathbf{y}_n = \frac{1}{2}h\left[\mathbf{f}(\mathbf{y}_{n+1}) + \mathbf{f}(\mathbf{y}_n)\right]$$

is the implicit midpoint rule method

$$\mathbf{y}_{n+1} - \mathbf{y}_n = h\mathbf{f}\left(\frac{1}{2}(\mathbf{y}_{n+1} + \mathbf{y}_n)\right) .$$

Let us recall the notation from Section 3.1 to simplify writing: putting $y_{n+1} = Ey_n$, we can write the linear k-step method (95) as

$$\rho(E)\mathbf{y}_n = h\sigma(E)\mathbf{f}(\mathbf{y}_n) ,$$

where $\rho(z) = \sum_{j=0}^{k} \alpha_j z^j$ and $\sigma(z) = \sum_{j=0}^{k} \beta_j z^j$ are the first and second characteristic polynomial of the method; the associated one-leg twin (96) is then

$$\rho(E)\mathbf{y}_n = h\mathbf{f}(\sigma(E)\mathbf{y}_n) .$$

There is a close relationship between the linear multistep method (95) and its one-leg twin (96). Let $\mathbf{z}_n = \sigma(E)\mathbf{y}_n$; then

$$\rho(E)\mathbf{z}_n = \rho(E)\sigma(E)\mathbf{y}_n = \sigma(E)\rho(E)\mathbf{y}_n = h\sigma(E)\mathbf{f}(\sigma(E)\mathbf{y}_n) = h\sigma(E)\mathbf{f}(\mathbf{z}_n) .$$

In other words, if $\{\mathbf{y}_n\}_{n\geq 0}$ is a solution to (96) then $\{\mathbf{z}_n\}_{n\geq 0}$, with $\mathbf{z}_n = \sigma(E)\mathbf{y}_n$, is the solution of the linear multistep method (95) whose one-leg twin (96) is. This connection allows results for the one-leg twin to be translated into results for the linear multistep method.

Now we shall state a definition of nonlinear stability due to Dahlquist (1975). Before we do so, we introduce the concept of **G-norm**. Consider a vector

$$\mathbf{Z}_n = (\mathbf{z}_{n+k-1}^T, \ldots, \mathbf{z}_n^T)^T$$

in $\mathbf{R}^{mk}$ where $\mathbf{z}_{n+j} \in \mathbf{R}^m$, $j = 0, 1, \ldots, k-1$. Given that $G = (g_{ij})$ is a $k \times k$ symmetric positive definite matrix, the G-norm $\|\cdot\|_G$ is defined by

$$\|\mathbf{Z}_n\|_G = \left(\sum_{i=1}^{k} \sum_{j=1}^{k} g_{ij} \langle \mathbf{z}_{n+k-i}, \mathbf{z}_{n+k-j} \rangle \right)^{1/2} .$$

Definition 17 *The k-step method (96) is said to be **G-stable** if there exists a symmetric positive definite matrix G such that*

$$\|\mathbf{Z}_{n+1}\|_G^2 - \|\mathbf{Z}_n\|_G^2 \leq 2\langle \rho(E)\mathbf{z}_n, \sigma(E)\mathbf{z}_n \rangle / \sigma^2(1) .$$

Let $\{\mathbf{u}_n\}$ and $\{\mathbf{v}_n\}$ be two solutions of the differential equation $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ given by (96) with different starting values, and suppose that (96) is G-stable. Define the vectors $\mathbf{U}_n$ and $\mathbf{V}_n$ in $\mathbf{R}^{mk}$ by

$$\mathbf{U}_n = (\mathbf{u}_{n+k-1}^T, \ldots, \mathbf{u}_n^T)^T , \qquad \mathbf{V}_n = (\mathbf{v}_{n+k-1}^T, \ldots, \mathbf{v}_n^T)^T .$$

Since we are assuming that the differential equation $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ is dissipative, it follows that

$$\begin{aligned}
\|\mathbf{U}_{n+1} - \mathbf{V}_{n+1}\|_G^2 - \|\mathbf{U}_n - \mathbf{V}_n\|_G^2 &\leq 2\langle \rho(E)(\mathbf{u}_n - \mathbf{v}_n), \sigma(E)(\mathbf{u}_n - \mathbf{v}_n) \rangle / \sigma^2(1) \\
&\leq 2\langle h\mathbf{f}(\sigma(E)\mathbf{u}_n) - h\mathbf{f}(\sigma(E)\mathbf{v}_n), \sigma(E)(\mathbf{u}_n - \mathbf{v}_n) \rangle / \sigma^2(1) \\
&= \frac{2h}{\sigma(1)^2} \langle \mathbf{f}(\sigma(E)\mathbf{u}_n) - h\mathbf{f}(\sigma(E)\mathbf{v}_n), \sigma(E)\mathbf{u}_n - \sigma(E)\mathbf{v}_n \rangle \leq 0 .
\end{aligned}$$

Hence,

$$\|\mathbf{U}_{n+1} - \mathbf{V}_{n+1}\|_G^2 \le \|\mathbf{U}_n - \mathbf{V}_n\|_G^2 \qquad \text{for } n \ge 0 \ .$$

Thus we deduce that a G-stable approximation of a dissipative ordinary differential equation $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ is contractive in the G-norm; this is a discrete counterpart of the property we established at the beginning of this section that analytical solutions to a dissipative equation of this kind are contractive in the Euclidean norm. For further developments of these ideas, we refer to the books of J.D. Lambert, and A.M. Stuart and A.R. Humphries.

6 Boundary value problems

In many applications a system of m simultaneous first-order ordinary differential equations in m unknowns $y_1(x), y_2(x), \ldots, y_m(x)$ has to be solved. If each of these variables satisfies a given condition at the same value a of x then we have an initial value problem for a system of first-order ordinary differential equations. If the y_i, $i = 1, \ldots, m$, satisfy given conditions at different values a, b, c, $\ldots$ of the independent variable x then we have a multi-point boundary value problem. In particular, if conditions on the y_i, $i = 1, \ldots, m$, are imposed at two different values a and b then we have a two-point boundary value problem.

Example 7 *Here is an example of a multipoint (in this case, three-point) boundary value problem:*

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

The exact solution is $y(x) = e^x$.

Example 8 *This is an example of a two-point boundary value problem:*

$$y'' - 2y^3 = 0 \ , \qquad y(1) = 1 \ , \quad y'(2) + [y(2)]^2 = 0 \ .$$

The exact solution is $y(x) = 1/x$.

In this section we shall consider three classes of methods for the numerical solution of two-point boundary value problems: shooting methods, matrix methods and collocation methods.

6.1 Shooting methods

Let us consider the two-point boundary value problem

$$y'' = f(x, y, y') \ , \qquad y(a) = A \ , \quad y(b) = B \ , \tag{97}$$

with $a < b$ and $x \in [a, b]$. We shall suppose that (97) has a unique solution. The motivation behind shooting methods is to convert the two-point boundary value problem into solving a sequence of initial value problems whose solutions converge to that of the boundary value problem, so that one can use existing software developed for the numerical solution of initial value problems: observe that an attempt to solve the boundary value problem (97) directly will lead to a coupled system of nonlinear equations whose solution may be a hard problem.

Let us make an initial guess s for $y'(a)$ and denote by $y(x;s)$ the solution of the initial value problem

$$y'' = f(x, y, y') , \qquad y(a) = A , \quad y'(a) = s . \tag{98}$$

Introducing the notation $u(x;s) = y(x;s)$, $v(x;s) = \frac{\partial}{\partial x} y(x;s)$, we can rewrite (98) as a system of first-order ordinary differential equations:

$$\begin{aligned} \frac{\partial}{\partial x} u(x;s) &= v(x;s) , & u(a;s) &= A , \\ \frac{\partial}{\partial x} v(x;s) &= f(x, u(x;s), v(x;s)) , & v(a;s) &= s . \end{aligned} \tag{99}$$

The solution $u(x;s)$ of the initial value problem (99) will coincide with with the solution $y(x)$ of the boundary value problem (97) provided that that we can find a value of s such that

$$\phi(s) \equiv u(b;s) - B = 0 . \tag{100}$$

The essence of the shooting method for the numerical solution of the boundary value problem (97) is to find a root to the equation (100). Any standard root-finding technique can be used; here we shall consider two: bisection of the interval which is known to contain the root and Newton's method.

6.1.1 The method of bisection

Let us suppose that that two numbers s_1 and s_2 are known such that

$$\phi(s_1) < 0 \qquad \text{and} \qquad \phi(s_2) > 0 .$$

We assume, for the sake of definiteness, that $s_1 < s_2$. Given that the solution of the initial value problem (99) depends continuously on the initial data, there must exist at least one value of s in the interval (s_1, s_2) such that $\phi(s) = 0$. Thus the interval $[s_1, s_2]$ contains a root of the equation (100).

The root of (100) can now be calculated approximately using the method of bisection. We take the midpoint s_3 of the interval $[s_1, s_2]$, compute $u(b, s_3)$ and consider whether $\phi(s_3) = u(b; s_3) - B$ is positive or negative. If $\phi(s_3) > 0$ then it is the interval $[s_1, s_3]$ that contains a root of ϕ, whereas if $\phi(s_3) < 0$ then the interval in question is $[s_3, s_2]$. By repeating this process, one can construct a sequence of numbers $\{s_n\}_{n=1}^{\infty}$ converging the s. In practice the process of bisection is terminated after a finite number of steps when the length of the interval containing s has become sufficiently small.

6.1.2 The Newton–Raphson method

An alternative to the method of bisection is to compute a sequence $\{s_n\}_{n=1}^{\infty}$ generated by the Newton–Raphson method:

$$s_{n+1} = s_n - \phi(s_n)/\phi'(s_n) , \tag{101}$$

with the starting value s_0 chosen arbitrarily in a sufficiently small interval surrounding the root. For example, a suitable s_0 may be found by performing a few steps of the method of bisection. If s_0 is a sufficiently good approximation to the required root of (100) the theory of the Newton–Raphson method ensures that, in general, we have quadratic convergence of the sequence $\{s_n\}_{n=0}^{\infty}$ to the root s.

From the point of view of implementing (101) the first question that we need to clarify is how one can compute compute $\phi'(s_n)$. To do so, we introduce the new dependent variables

$$\xi(x;s) = \frac{\partial u(x;s)}{\partial s}\,, \qquad \eta(x;s) = \frac{\partial v(x;s)}{\partial s}$$

and differentiate the initial value problem (99) with respect to s to obtain a second initial value problem:

$$\begin{aligned} \frac{\partial \xi(x;s)}{\partial x} &= \eta(x;s)\,, & \xi(a;s) = 0\,, \\ \frac{\partial \eta(x;s)}{\partial x} &= p(x;s)\xi(x;s) + q(x;s)\eta(x;s)\,, & \eta(a;s) = 1\,, \end{aligned} \tag{102}$$

where

$$\begin{aligned} p(x;s) &= \frac{\partial f(x,u(x;s),v(x;s))}{\partial u}\,, \\ q(x;s) &= \frac{\partial f(x,u(x;s),v(x;s))}{\partial v}\,. \end{aligned} \tag{103}$$

If we assign the value s_n to s, $n \geq 0$, then the initial value problem (99), (102) can be solved by a predictor-corrector method or a Runge–Kutta method on the interval $[a,b]$. Thus we obtain $u(b;s_n)$ or, more precisely, an approximation to $u(b;s_n)$ from which we can calculate $\phi(s_n) = u(b;s_n) - B$; in addition, we obtain an approximation to $\xi(b;s_n) = \phi'(s_n)$. Having calculated $\phi(s_n)$ and $\phi'(s_n)$, we obtain the next Newton–Raphson iterate s_{n+1} from (101). The process is then repeated until the iterates s_n settle to a fixed number of digits.

Two remarks are in order:

1) According to (103), the initial value problems (99) and (102) are coupled and therefore they must be solved simultaneously over the interval $[a,b]$, with s set to s_n, $n = 0, 1, 2, \ldots$;

2) The coupled initial value problem (99), (102) may be very sensitive to perturbations of the initial guess s_0; a bad initial guess of s_0 may result in a sequence of Newton–Raphson iterates $\{s_n\}_{n=0}^{\infty}$ which does not converge to the root s.

The latter difficulty may be overcome by using the **multiple shooting method** which we describe below. First, however, we show how the simple shooting method may be extended to the nonlinear two-point boundary value problem

$$\mathbf{y}' = \mathbf{f}(x,\mathbf{y})\,, \qquad a < x < b\,, \qquad \mathbf{g}(\mathbf{y}(a),\mathbf{y}(b)) = \mathbf{0}\,, \tag{104}$$

where $\mathbf{y}$, $\mathbf{f}$ and $\mathbf{g}$ are m-component vector functions of their respective arguments. In the simple shooting method the boundary value problem (104) is transformed into the initial value problem

$$\mathbf{u}'(x;\mathbf{s}) = \mathbf{f}(x,\mathbf{u}(x;\mathbf{s}))\,, \qquad a < x < b\,, \qquad \mathbf{u}(a;\mathbf{s}) = \mathbf{s}\,, \tag{105}$$

whose solution is required to satisfy the condition

$$\mathbf{G}(\mathbf{s}) \equiv \mathbf{g}(\mathbf{s}, \mathbf{u}(b; \mathbf{s})) = \mathbf{0} \tag{106}$$

for an unknown value of s. Thus the problem has been transformed into one of finding a solution to the system of nonlinear equations $\mathbf{G}(\mathbf{s}) = \mathbf{0}$. In order to evaluate the function $\mathbf{G}$ for a specific value of $\mathbf{s}$ a numerical method for the solution of the initial value problem (105) has to be employed on the interval $[a, b]$ to compute (an approximation to) $\mathbf{u}(b; \mathbf{s})$. In fact, as we have already seen earlier, a root-finding procedure such as Newton's method will require the computation of the Jacobi matrix

$$J(\mathbf{s}) = \frac{\partial \mathbf{G}}{\partial \mathbf{s}} \ .$$

This in turn requires the solution of a coupled initial value problem for m^2 linear ordinary differential equations to obtain $\partial \mathbf{u}/\partial \mathbf{s}$ for $a \le x \le b$.

The shooting method is said to converge if the root-finding algorithm results in a sequence $\{\mathbf{s}_n\}_{n=0}^{\infty}$ which converges to a root $\mathbf{s}$ of $\mathbf{G}$; then $\mathbf{s} = \mathbf{y}(a)$ where $\mathbf{y}(x)$ is the desired solution of the boundary value problem (104). The convergence of this sequence and therefore the success of the shooting method may be hampered by two effects:

1) A well-conditioned boundary value problem of the form (104) may easily lead to an ill-posed initial value problem (104);

2) A bounded solution to the initial value problem (104) may exist only for $\mathbf{s}$ in a small neighbourhood of $\mathbf{y}(s)$ (which, of course, in unknown).

The idea behind the **multiple shooting method** is to divide the interval $[a, b]$ into smaller subintervals

$$a = x_0 < x_1 < \ldots < x_{K-1} < x_K = b \ ; \tag{107}$$

the problem then is to find a vector $\mathbf{S}^T = (\mathbf{s}_0^T, \ldots, \mathbf{s}_{K-1}^T)$ such that the solutions $\mathbf{u}_k(x; \mathbf{s}_k)$ of the initial value problems

$$\begin{aligned} \mathbf{u}_k'(x; \mathbf{s}_k) &= \mathbf{f}(x, \mathbf{u}_k(x; \mathbf{s}_k)) \ , \qquad x_k < x \le x_{k+1} \ , \\ \mathbf{u}_k(x_k; \mathbf{s}_k) &= \mathbf{s}_k \ , \qquad\qquad\qquad\quad k = 0, \ldots, K-1 \ , \end{aligned} \tag{108}$$

satisfy the conditions

$$\mathbf{u}_k(x_{k+1}; \mathbf{s}_k) - \mathbf{s}_{k+1} = \mathbf{0} \ , \quad k = 0, \ldots, K-2 \ , \qquad \mathbf{g}(\mathbf{s}_0, \mathbf{u}_{K-1}(b; \mathbf{s}_{K-1})) = \mathbf{0} \ . \tag{109}$$

The equations (109) can be written in the compact form $\mathbf{G}(\mathbf{S}) = \mathbf{0}$. A clear advantage of the multiple shooting method over simple shooting is that the growth of the solutions to the initial value problems (108) and the related linear initial value problems for the $\partial \mathbf{u}_k(x; \mathbf{s}_k)/\partial \mathbf{s}_k$, $k = 0, \ldots, K-1$, can be approximated accurately by selecting a sufficiently fine subdivision (107) of the interval $[a, b]$.

Indeed, it is possible to prove that, under reasonable assumptions, the multiple shooting method leads to an exponential increase of the size of the domain of initial values for which the first iteration of the root-finding procedure is defined. This is consistent with the practical observation that multiple shooting is less sensitive to the choice of the starting values than simple shooting.

6.2 Matrix methods

In this section, rather than attempting to convert the boundary value problem to an initial value problem, we approximate it directly by using a finite difference method. This results in a system of equations for the unknown values of the numerical solution at the mesh points. We begin by considering a linear boundary value problem. In this case the calculation of the numerical solution amounts to solving a system of linear equations with a sparse matrix. We then consider a nonlinear boundary value problem; upon the application of Newton's method, this again involves, in each Newton iteration, the solution of a system of linear equations with a sparse matrix. Methods of this kind are usually referred to in the ODE literature as **matrix methods**.

6.2.1 Linear boundary value problem

Let us consider the two-point boundary value problem

$$\begin{aligned} y'' + p(x)y' + q(x)y &= f(x)\,, \qquad x \in (a,b)\,, \\ a_0 y(a) + b_0 y'(a) &= c_0\,, \\ a_1 y(b) + b_1 y'(b) &= c_1\,. \end{aligned} \tag{110}$$

We discretise (110) using a finite difference method on the uniform mesh

$$\{x_i \mid x_i = a + ih\,,\ \ i = 0, \ldots, N\}$$

of step size $h = (b-a)/N$, $N \geq 2$. The essence of the method is to approximate the derivatives in the differential equation and the boundary conditions by divided differences.

Assuming that y is sufficiently smooth, it is a simple matter to show using Taylor series expansions that

$$y''(x_i) = \frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} + O(h^2)\,, \tag{111}$$

$$y'(x_i) = \frac{y(x_{i+1}) - y(x_{i-1})}{2h} + O(h^2)\,, \tag{112}$$

$$y'(x_0) = \frac{-3y(x_0) + 4y(x_1) - y(x_2)}{2h} + O(h^2)\,, \tag{113}$$

$$y'(x_N) = \frac{y(x_{N-2}) - 4y(x_{N-1}) + 3y(x_N)}{2h} + O(h^2)\,. \tag{114}$$

Now, using (111–114) we can construct a finite difference method for the numerical solution of (110); we denote by y_i the numerical approximation to $y(x_i)$ for $i = 0, \ldots, N$:

$$\begin{aligned} \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + p(x_i)\frac{y_{i+1} - y_{i-1}}{2h} + q(x_i)y_i &= f(x_i)\,, \qquad i = 1, \ldots, N-1\,, \\ a_0 y_0 + b_0 \frac{-3y_0 + 4y_1 - y_2}{2h} &= c_0\,, \\ a_1 y_N + b_1 \frac{y_{N-2} - 4y_{N-1} + 3y_N}{2h} &= c_1\,. \end{aligned}$$

After rearranging these, we obtain

$$\left(\frac{1}{h^2} - \frac{p(x_i)}{2h}\right) y_{i-1} - \left(\frac{2}{h^2} - q(x_i)\right) y_i + \left(\frac{1}{h^2} + \frac{p(x_i)}{2h}\right) y_{i+1} = f(x_i), \quad 1 \leq i \leq N-1\,,$$

$$\begin{aligned} \left(a_0 - \frac{3b_0}{2h}\right) y_0 + \frac{4b_0}{2h} y_1 - \frac{b_0}{2h} y_2 &= c_0 \ , \\ \frac{b_1}{2h} y_{N-2} - \frac{4b_1}{2h} y_{N-1} + \left(a_1 + \frac{3b_1}{2h}\right) y_{N-2} &= c_1 \ . \end{aligned}$$

This is a system of linear equations of the form

$$\begin{aligned} A_0 y_0 + C_0 y_1 + B_0 y_2 &= F_0 \ , \\ A_i y_{i-1} + C_i y_i + B_i y_{i+1} &= F_i \ , \qquad i = 1, \dots, N-1 \ , \\ A_N y_{N-2} + C_N y_{N-1} + B_N y_N &= F_N \ . \end{aligned}$$

The matrix of the system is

$$M = \begin{bmatrix} A_0 & C_0 & B_0 & 0 & 0 & \dots & 0 \\ A_1 & C_1 & B_1 & 0 & 0 & \dots & 0 \\ 0 & A_2 & C_2 & B_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & A_{N-1} & C_{N-1} & B_{N-1} \\ 0 & 0 & 0 & 0 & A_N & C_N & B_N \end{bmatrix} .$$

Let us consider the following cases:

1) If $B_0 = 0$ and $A_N = 0$ then M is a tridiagonal matrix.
2) If $B_0 \neq 0$ and $B_1 = 0$ (and/or $A_N \neq 0$ and $A_{N-1} = 0$) then we can interchange the first two rows of the matrix (and/or the last two rows) and, again, we obtain a tridiagonal matrix.
3) If $B_0 \neq 0$ and $B_1 \neq 0$ (and/or $A_N \neq 0$ and $A_{N-1} \neq 0$) then we can eliminate B_0 from the first row (and/or A_N from the last row) to obtain a tridiagonal matrix.

In any case the matrix M can be transformed into a tridiagonal matrix. Thus, from now on, without any restrictions on generality, we shall suppose that M is tridiagonal. To summarise the situation, we wish to solve the system of linear equations

$$M\mathbf{y} = \mathbf{F}$$

where M is a tridiaginal matrix,

$$\mathbf{y} = (y_0, y_1, \dots, y_N)^T, \qquad \mathbf{F} = (F_0, F_1, \dots, F_N)^T \ .$$

The algorithm we present below for the solution of this linear system is usually referred to as the **Thomas algorithm**. It is a special case of Gaussian elimination or LU decomposition; in fact, it is this latter interpretation that we adopt here.

We wish to express M as a product LU of a lower-triangular matrix

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ l_1 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & l_2 & 1 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & l_{N-1} & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & l_N & 1 \end{bmatrix}$$

and an upper-triangular matrix

$$U = \begin{bmatrix} u_0 & v_0 & 0 & 0 & 0 & \dots & 0 \\ 0 & u_1 & v_1 & 0 & 0 & \dots & 0 \\ 0 & 0 & u_2 & v_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & u_{N-1} & v_{N-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & u_N \end{bmatrix} .$$

On multiplying L and U and equating the resulting matrix with M, we find that

$$u_0 = C_0 \ , \quad v_0 = B_0 \ , \quad \left.\begin{aligned} l_i u_{i-1} &= A_i \\ l_i v_{i-1} + u_i &= C_i \\ v_i &= B_i \end{aligned}\right\} \quad i = 1, \dots, N \ .$$

Hence

$$v_i = B_i \ , \quad i = 0, \dots, N \ , \quad u_0 = C_0 \ , \quad \left.\begin{aligned} u_i &= C_i - (A_i B_{i-1})/u_{i-1} \\ l_i &= A_i / u_{i-1} \end{aligned}\right\} \quad i = 1, \dots, N \ .$$

Given that the entries of the matrix M are known, the components of L and U can be computed from these formulae, and the set of linear equations $M\mathbf{y} = \mathbf{F}$ can be written in the following equivalent form:

$$\left\{ \begin{aligned} L\mathbf{z} &= \mathbf{F} \\ U\mathbf{y} &= \mathbf{z} \end{aligned} \right. .$$

Thus, instead of solving $M\mathbf{y} = \mathbf{F}$ directly, we solve two linear systems in succession, each with a triangular matrix: $L\mathbf{z} = \mathbf{F}$ is solved for $\mathbf{z}$, followed by solving $U\mathbf{y} = \mathbf{z}$ for $\mathbf{y}$. Writing this out in detail yields

$$\left\{ \begin{aligned} & z_1 = F_1 \ , \\ & z_i = F_i - l_{i-1} z_{i-1} \ , \qquad i = 1, \dots, N \ , \end{aligned} \right.$$

and

$$\left\{ \begin{aligned} & y_N = z_N / u_N \ , \\ & y_i = (z_i - v_i y_{i+1})/u_i \ , \qquad i = N-1, \dots, 0 \ . \end{aligned} \right.$$

Expressing in these formulae the values of u_i and v_i in terms of A_i, B_i, C_i, we find that

$$\begin{aligned} y_i &= \alpha_{i+1} y_{i+1} + \beta_{i+1} \ , \qquad i = N-1, \dots, 0 \ , \\ y_N &= \beta_{N+1} \ , \end{aligned}$$

where

$$\begin{aligned} \alpha_{i+1} &= -\frac{B_i}{C_i + \alpha_i A_i} \ , \qquad i = 1, 2, \dots, N-1 \ , \qquad \alpha_1 = -\frac{B_0}{C_0} \ , \\ \beta_{i+1} &= \frac{F_i - \beta_i A_i}{C_i + \alpha_i A_i} \ , \qquad i = 1, 2, \dots, N \ , \qquad \beta_1 = \frac{F_0}{C_0} \ . \end{aligned}$$

The last set of formulae are usually referred to as the **Thomas algorithm**.

6.2.2 Nonlinear boundary value problem

Now consider a nonlinear second-order differential equation subject to linear boundary conditions:

$$\begin{aligned}
y'' &= f(x, y, y') \ , \qquad x \in (a, b) \ , \\
a_0 y(a) + b_0 y'(a) &= c_0 \ , \\
a_1 y(b) + b_1 y'(b) &= c_1 \ .
\end{aligned}$$

On approximating the derivatives with divided differences, we obtain the following set of difference equations:

$$\begin{aligned}
\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} &= f\left(x_i, y_i, \frac{y_{i+1} - y_{i-1}}{2h}\right) \ , \qquad i = 1, \ldots, N-1 \ , \\
a_0 y_0 + b_0 \frac{-3y_0 + 4y_1 - y_2}{2h} &= c_0 \ , \\
a_1 y_N + b_1 \frac{y_{N-2} - 4y_{N-1} + 3y_N}{2h} &= c_1 \ .
\end{aligned}$$

After rearranging these, we obtain

$$\begin{aligned}
\frac{1}{h^2} y_{i-1} - \frac{2}{h^2} y_i + \frac{1}{h^2} y_{i+1} &= f\left(x_i, y_i, \frac{y_{i+1} - y_{i-1}}{2h}\right), \quad 1 \le i \le N-1 \ , \\
\left(a_0 - \frac{3b_0}{2h}\right) y_0 + \frac{4b_0}{2h} y_1 - \frac{b_0}{2h} y_2 &= c_0 \ , \\
\frac{b_1}{2h} y_{N-2} - \frac{4b_1}{2h} y_{N-1} + \left(a_1 + \frac{3b_1}{2h}\right) y_N &= c_1 \ .
\end{aligned}$$

This is a system of system of nonlinear equations of the form

$$\begin{aligned}
A_0 y_0 + C_0 y_1 + B_0 y_2 &= F_0 \ , \\
A_i y_{i-1} + C_i y_i + B_i y_{i+1} &= F_i \ , \qquad i = 1, \ldots, N-1 \ , \\
A_N y_{N-2} + C_N y_{N-1} + B_N y_N &= F_N \ ,
\end{aligned} \tag{115}$$

where

$$\begin{aligned}
A_0 &= a_0 - (3b_0)/(2h) \ , & C_0 &= 4b_0/(2h) \ , & B_0 &= -b_0/(2h) \ , \\
A_i &= 1/h^2 \ , & C_i &= -2/h^2 \ , & B_i &= 1/h^2 \ , \quad i = 1, \ldots, N-1 \ , \\
A_N &= b_1/(2h) \ , & C_N &= -4b_1/(2h) \ , & B_N &= a_1 + (3b_1)/(2h) \ ,
\end{aligned}$$

and

$$F_0 = c_0 \ , \qquad F_i = f(x_i, y_i, (y_{i+1} - y_{i-1})/(2h)) \ , \quad i = 1, \ldots, N-1 \ , \qquad F_N = c_1 \ .$$

Thus, (115) can be written in the compact form

$$M\mathbf{y} = \mathbf{F}(\mathbf{y})$$

where M is a tridiagonal matrix, $\mathbf{y} = (y_0, \ldots, y_N)^T$ and $\mathbf{F}(\mathbf{y}) = (F_0, \ldots, F_N)^T$. The nonlinear system of equations

$$\mathbf{G}(\mathbf{y}) \equiv M\mathbf{y} - \mathbf{F}(\mathbf{y}) = \mathbf{0}$$

can now be solved by Newton's method, for example, assuming that a solution exists. Given a starting value $\mathbf{y}^{(0)}$ for the Newton iteration, subsequent iterates are computed successively from

$$J(\mathbf{y}^{(n)})(\mathbf{y}^{(n+1)} - \mathbf{y}^{(n)}) = -\mathbf{G}(\mathbf{y}^{(n)}) , \qquad n = 0, 1, 2, \dots , \tag{116}$$

where

$$J(\mathbf{w}) = \frac{\partial \mathbf{G}}{\partial \mathbf{y}}(\mathbf{w}) = \left(\frac{\partial G_i}{\partial y_j}(\mathbf{w}) \right)_{0 \le i,j \le N}$$

is the Jacobi matrix of $\mathbf{G}$. As G_i is independent of y_j with $|i - j| > 1$, the matrix $J(\mathbf{y}^{(n)})$ is tridiagonal. Consequently, each step of the Newton iteration (116) involves the solution of a system of linear equations with a tridiagonal matrix; this may be accomplished by using the Thomas algorithm described above.

6.3 Collocation method

Consider the boundary value problem

$$\begin{aligned} \mathcal{L}y \equiv y'' + p(x)y' + q(x)y &= f(x) , \qquad x \in (a, b) , \\ a_0 y(a) + b_0 y'(a) &= c_0 , \\ a_1 y(b) + b_1 y'(b) &= c_1 , \end{aligned} \tag{117}$$

where a_0, a_1, b_0, b_1 are real numbers such that

$$(a_0 b_1 - a_1 b_0) + a_0 a_1 (b - a) \neq 0 . \tag{118}$$

It can be assumed, without loss of generality, that $c_0 = 0$ and $c_1 = 0$; for if this is not the case then we consider the function $\tilde{y}$ defined by

$$\tilde{y}(x) = y(x) - d(x) ,$$

with $d(x) = \alpha x + \beta$ where α and β are real numbers chosen so that $d(x)$ satisfies the (nonhomogeneous) boundary conditions at $x = a$ and $x = b$ stated in (117). It is a straightforward matter to see that provided (118) holds there are unique such α and β. The function $\tilde{y}$ then obeys the differential equation $\mathcal{L}\tilde{y} = \tilde{f}$, where $\tilde{f}(x) = f(x) - \mathcal{L}d(x)$, and satisfies the boundary conditions in (117) with $c_0 = c_1 = 0$.

Suppose that $\{\psi_i\}_{i=1}^{N}$ is a set of linearly independent functions defined on the interval $[a, b]$ satisfying the homogeneous counterparts of the boundary conditions in (117) (i.e. $c_0 = c_1 = 0$). We shall suppose that each function ψ_i is twice continuously differentiable on $[a, b]$.

The essence of the **collocation method** is to seek an approximate solution $y_N(x)$ to (110) in the form

$$y_N(x) = \sum_{i=1}^{N} \xi_i \psi_i(x) , \tag{119}$$

and demand that

$$\mathcal{L}y_N(x_j) = f(x_j) , \qquad j = 1, \dots, N , \tag{120}$$

at $(N+1)$ distinct points x_j, $j = 1, \dots, N$, referred to as the **collocation points**. We note that since each of the functions $\psi_i(x)$ satisfies the (homogeneous) boundary conditions at $x = a$ and $x = b$ the same is true of $y_N(x)$.

Now, (119–120) yield the system of linear equations

$$\sum_{i=1}^{N} \xi_i \mathcal{L}\psi_i(x_j) = f(x_j) , \qquad j = 1, \dots, N , \tag{121}$$

for the coefficients ξ_i, $i = 1, \dots, N$. The specific properties of the collocation method depend on the choice of the basis functions ψ_i and the collocation points x_j. In general the matrix $M = (\mathcal{L}\psi_i(x_j))_{1 \le i,j \le N}$ is full. However, if each of the basis functions ψ_i has compact support contained in (a, b) (for example, they are B-splines) then M is a band-matrix, given that both $\psi_i(x_j) = 0$ and $\mathcal{L}\psi_i(x_j) = 0$ for $|i - j| > M$ for some integer M, $1 < M < N$. If M is a fixed integer, independent of N, then the resulting system of linear equations can be solved in $O(N)$ arithmetic operations.

In spectral collocation methods, the functions $\psi_i(x)$ are chosen as trigonometric polynomials. Consider, for example, the simple boundary value problem

$$-y''(x) + q(x)y(x) = f(x) , \quad x \in (0, \pi) , \qquad y(0) = y(\pi) = 0 ,$$

where $q(x) \ge 0$ for all x in $[0, \pi]$. The functions $\psi_i(x) = \sin ix$, $i = 1, \dots, N$, satisfy the boundary conditions and are linearly independent on $[0, \pi]$. Thus we seek an approximate solution $y_N(x)$ in the form

$$y_N(x) = \sum_{i=1}^{N} \xi_i \sin ix .$$

Substitution of this expansion into the differential equation results in the system of linear equations

$$\sum_{i=1}^{N} \xi_i \left(i^2 + q(x_j)\right) \sin ix_j = f(x_j) , \qquad i = 1, \dots, N ,$$

for $\xi_1, \dots, \xi_N$. The collocation points are usually chosen as

$$x_j = \frac{j\pi}{N+1} , \qquad j = 1, \dots, N .$$

This choice is particularly convenient when q is a (nonnegative) constant, for then the linear system

$$\sum_{i=1}^{N} \xi_i \left(i^2 + q\right) \sin \frac{ij\pi}{N+1} = f(x_j) , \qquad i = 1, \dots, N ,$$

can be solved for $\xi_1, \dots, \xi_n$ in $O(N)$ arithmetic operations using a Fast Fourier Transform, despite the fact that the matrix of the system is full.

1. Verify that the following functions satisfy a Lipschitz condition on the respective intervals and find the associated Lipschitz constants:

   a) $f(x,y) = 2yx^{-4}$, $\quad x \in [1,\infty)$;

   b) $f(x,y) = \mathrm{e}^{-x^2}\tan^{-1} y$, $\quad x \in [1,\infty)$;

   c) $f(x,y) = 2y(1+y^2)^{-1}(1+\mathrm{e}^{-x})$, $\quad x \in (-\infty,\infty)$.

2. Suppose that m is a fixed positive integer. Show that the initial value problem

$$y' = y^{2m/(2m+1)} , \qquad y(0) = 0 ,$$

   has infinitely many continuously differentiable solutions. Why does this not contradict Picard's theorem?

3. Show that the explicit Euler method fails to approximate the solution $y(x) = (4x/5)^{5/4}$ of the initial value problem $y' = y^{1/5}$, $y(0) = 0$. Justify your answer. Consider the same problem with the implicit Euler method.

4. Write down the explicit Euler method for the numerical solution of the initial value problem $y' + 5y = x\mathrm{e}^{-5x}$, $y(0) = 0$, on the interval $[0,1]$ with step size $h = 1/N$, $N \geq 1$. Denoting by y_N the Euler approximation to $y(1)$ at $x = 1$, show that $\lim_{N\to\infty} y_N = y(1)$. Find an integer N_0 such that

$$|y(1) - y_N| \leq 10^{-5} , \qquad \text{for all } N \geq N_0 .$$

5. Consider the initial value problem

$$y' = \log\log(4+y^2) , \qquad x \in [0,1] , \qquad y(0) = 1 ,$$

   and the sequence $\{y_n\}_{n=0}^{N}$, $N \geq 1$, generated by the explicit Euler method

$$y_{n+1} = y_n + h\log\log(4+y_n^2) , \qquad n = 0,\ldots,N-1 , \qquad y_0 = 1 ,$$

   using the mesh points $x_n = nh$, $n = 0,\ldots,N$, with spacing $h = 1/N$. Here log denotes the logarithm with base e.

   a) Let T_n denote the truncation error of Euler's method at $x = x_n$ for this initial value problem. Show that $|T_n| \leq h/4$.

   b) Verify that

$$|y(x_{n+1}) - y_{n+1}| \leq (1+hL)|y(x_n) - y_n| + h|T_n| , \qquad n = 0,\ldots,N-1 ,$$

   where $L = 1/(2\log 4)$.

   c) Find a positive integer N_0, as small as possible, such that

$$\max_{0\leq n\leq N} |y(x_n) - y_n| \leq 10^{-4}$$

   whenever $N \geq N_0$.

6. Define the truncation error T_n of the trapezium rule method

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

for the numerical solution of the initial value problem $y' = f(x,y)$, $y(0)$ given, where $f_n = f(x_n, y_n)$ and $h = x_{n+1} - x_n$.

By integrating by parts the integral

$$\int_{x_n}^{x_{n+1}} (x - x_{n+1})(x - x_n)y'''(x)\mathrm{d}x \ ,$$

or otherwise, show that

$$T_n = -\frac{1}{12}h^2 y'''(\xi_n)$$

for some ξ_n in the interval (x_n, x_{n+1}), where y is the solution of the initial value problem.

Suppose that f satisfies the Lipschitz condition

$$|f(x,u) - f(x,v)| \le L|u - v|$$

for all real x, u, v, where L is a positive constant independent of x, and that $|y'''(x)| \le M$ for some positive constant M independent of x. Show that the global error $e_n = y(x_n) - y_n$ satisfies the inequality

$$|e_{n+1}| \le |e_n| + \frac{1}{2}hL\left(|e_{n+1}| + |e_n|\right) + \frac{1}{12}h^2 M \ .$$

For a uniform step h satisfying $hL < 2$ deduce that, if $y_0 = y(x_0)$, then

$$|e_n| \le \frac{h^2 M}{12L}\left[\left(\frac{1 + \frac{1}{2}hL}{1 - \frac{1}{2}hL}\right)^n - 1\right] \ .$$

7. Consider the following one-step method for the numerical solution of the initial value problem $y' = f(x,y)$, $y(x_0) = y_0$:

$$y_{n+1} = y_n + \frac{1}{2}h(k_1 + k_2) \ ,$$

where

$$k_1 = f(x_n, y_n), \qquad k_2 = f(x_n + h, y_n + hk_1) \ .$$

Show that the method is consistent and has truncation error

$$T_n = \frac{1}{6}h^2\left[f_y(f_x + f_y f) - \frac{1}{2}(f_{xx} + 2f_{xy}f + f_{yy}f^2)\right] + O(h^3) \ .$$

8. Show that for $R = 1, 2$ there is no R-stage Runge–Kutta method of order $R + 1$.

9. Consider the one-step method

$$y_{n+1} = y_n + h(a\,k_1 + b\,k_2)\ ,$$

where

$$\begin{aligned} k_1 &= f(x_n, y_n)\ , \\ k_2 &= f(x_n + \alpha h, y_n + \beta h k_1)\ , \end{aligned}$$

and where a, b, α, β are real parameters. Show that there is a choice of these parameters such that the order of the method is 2. Is there a choice of the parameters for which the order exceeds 2?

10. Consider the one-step method

$$y_{n+1} = y_n + \alpha h f(x_n, y_n) + \beta h f(x_n + \gamma h, y_n + \gamma h f(x_n, y_n))\ ,$$

where α, β and γ are real parameters. Show that the method is consistent if and only if $\alpha + \beta = 1$. Show also that the order of the method cannot exceed 2.

Suppose that a second-order method of the above form is applied to the initial value problem $y' = -\lambda y$, $y(0) = 1$, where λ is a positive real number. Show that the sequence $(y_n)_{n \geq 0}$ is bounded if and only if $h \leq \frac{2}{\lambda}$. Show further that, for such λ,

$$|y(x_n) - y_n| \leq \frac{1}{6}\lambda^3 h^2 x_n\ , \quad n \geq 0\ .$$

11. Derive the mid-point rule method and the Simpson rule method by integrating the differential equation $y' = f(x, y(x))$ over suitable intervals of the real line and applying appropriate numerical integration rules.

12. Write down the general form of a linear multistep method for the numerical solution of the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$. What does it mean to say that such a method is *zero-stable*? Explain the significance of zero-stability. What is the truncation error of such a linear multistep method?

Determine for what values of the real parameter b the linear multistep method defined by the formula

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

is zero-stable. Show that there exists a value of b for which the order of the method is 4. Show further that if this method is zero-stable then its order cannot exceed 2.

13. Consider the initial value problem $y' = f(x, y)$, $y(0) = y_0$. Consider attempting to solve this problem by the linear multistep method

$$a y_{n-2} + b y_{n-1} + y_{n+1} = h f(x_n, y_n)$$

on the regular mesh $x_n = nh$ where a and b are constants.

   a) For a certain (unique) choice of a and b, this method is consistent. Find these values of a and b and verify that the order of accuracy is 1.

b) Although the method is consistent for the choice of a and b from part a), the numerical solution it generates will not, in general, converge to the solution of the initial value problem as $h \to 0$, because the method is not zero-stable. Show that the method is not zero-stable for these a and b, and describe quantitatively what the unstable solutions will look like for small h.

14. Consider the linear two-step method

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

Show that the method is zero-stable; show further that it is third-order accurate, namely, $T_n = O(h^3)$.

15. Show that the linear three-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 sixth order accurate. Find the roots of the first characteristic polynomial and deduce that the method is not zero-stable.

16. Write down the general form of a linear multi-step method for the numerical solution of the initial value problem

$$y' = f(x, y) , \qquad y(x_0) = y_0 ,$$

on the closed real interval $[x_0, x_N]$, where f is a continuous function of its arguments and y_0 is a given real number. Define the *truncation error* of the method. What does it mean to say that the method has *order of accuracy* p?

Given that α is a positive real number, consider the linear two-step method

$$y_{n+2} - \alpha y_n = \frac{h}{3}\left[f(x_{n+2}, y_{n+2}) + 4f(x_{n+1}, y_{n+1}) + f(x_n, y_n)\right] ,$$

on the mesh $\{x_n : x_n = x_0 + nh, n = 1, \dots, N\}$ of spacing h, $h > 0$. Determine the set of all α such that the method is zero-stable. Find α such that the order of accuracy is as high as possible; is the method convergent for this value of α?

17. Compute a numerical solution to the initial value problem $y' + y = 0$, $y(0) = 1$, on the interval $[0, 1]$ with $h = 2^{-k}$ for $k = 1, 2, \dots, 10$, using the linear two-step method

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

where the missing starting value y_1 is computed by the explicit Euler method. Tabulate the global error at $x = 1$ for $k = 1, 2, \dots, 10$, and comment on its rate of decay.

18. Solve numerically the initial value problem

$$y' = x - y^2 , \qquad y(0) = 0 , \qquad x \in [0, 1] ,$$

with step size $h = 0.01$ by a fourth-order Milne-Simpson method. Use the classical fourth-order Runge–Kutta method to compute the necessary starting values.

19. Find which of the following linear multistep methods for the solution of the initial value problem $y' = f(x,y)$, $y(0)$ given, are zero-stable. For any which are zero-stable, find limits on the value of $h = x_{n+1} - x_n$ for which they are absolutely stable when applied to the equation $y' = \lambda y$, $\lambda < 0$.

   a) $y_{n+1} - y_n = hf_n$
   b) $y_{n+1} + y_n - 2y_{n-1} = h(f_{n+1} + f_n + f_{n-1})$
   c) $y_{n+1} - y_{n-1} = \frac{1}{3}h(f_{n+1} + 4f_n + f_{n-1})$
   d) $y_{n+1} - y_n = \frac{1}{2}h(3f_n - f_{n_1})$
   e) $y_{n+1} - y_n = \frac{1}{12}h(5f_{n+1} + 8f_n - f_{n-1})$

20. Determine the order of the linear multistep method

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

and investigate its zero-stability and absolute stability.

21. If $\sigma(z) = z^2$ is the second characteristic polynomial of a linear multistep method, find a quadratic polynomial $\rho(z)$ such that the order of the associated linear multistep method is 2. Is this method convergent? What is its interval of absolute stability?

22. Find the interval of absolute stability for the two-step Adams–Bashforth method

$$y_{n+2} - y_{n+1} = \frac{1}{2}h\,[3f_{n+1} - f_n]$$

using Schur's criterion and the Routh–Hurwitz criterion.

23. Given that α is a positive real number, consider the linear two-step method

$$y_{n+2} - \alpha y_n = \frac{h}{3}\,[f_{n+2} + 4f_{n+1} + f_n]$$

on the mesh $\{x_n \,:\, x_n = x_0 + nh,\ n = 0, \ldots, N\}$ of spacing h, $h > 0$. For values of α such that the method is zero-stable investigate its absolute stability using Schur's criterion.

24. A predictor P and a corrector C are defined by their characteristic polynomials:

$$\begin{array}{lll} P: & \rho^*(z) = z^4 - 1\,, & \sigma^*(z) = \frac{4}{3}(2z^3 - z^2 + 2z)\,, \\ C: & \rho(z) = z^2 - 1\,, & \sigma(z) = \frac{1}{3}(z^2 + 4z + 1)\,. \end{array}$$

   a) Write down algorithms which use P and C in the $P(EC)^mE$ and $P(EC)^m$ modes.
   b) Find the stability polynomials $\pi_{P(EC)^mE}(z;\bar{h})$ and $\pi_{P(EC)^m}(z;\bar{h})$ of these methods. Assuming that $m = 1$, use Schur's criterion to calculate the associated intervals of absolute stability.
   c) Express the truncation errors $T_n^{P(EC)^mE}$ and $T_n^{P(EC)^m}$ of these methods in the form $O(h^r)$ where $r = r(p^*, p, m)$, with p^* and p denoting the orders of accuracy of P and C, respectively.

25. Which of the following would you consider a stiff initial value problem?

a) $y' = -(10^5 e^{-10^4 x} + 1)(y - 1)$, $y(0) = 0$ on the interval $x \in [0, 1]$. Note that the solution can be found in closed form:

$$y(x) = e^{10(e^{-10^4 x} - 1)} e^{-x} + 1 .$$

b)

$$\begin{aligned} y_1' &= -0.5y_1 + 0.501y_2 , & y_1(0) &= 1.1 , \\ y_2' &= 0.501y_1 - 0.5y_2 , & y_2(0) &= -0.9 , \end{aligned}$$

on the interval $x \in [0, 10^3]$.

c) $\mathbf{y}' = A(x)\mathbf{y}$, $\mathbf{y}(0) = (1, 1, 1)^T$, where

$$A(x) = \begin{pmatrix} -1 + 100\cos 200x & 100(1 - \sin 200x) & 0 \\ -100(1 + \sin 200x) & -(1 + 100\cos 200x) & 0 \\ 1200(\cos 100x + \sin 100x) & 1200(\cos 100x - \sin 100x) & -501 \end{pmatrix} .$$

26. Consider the θ-method

$$y_{n+1} = y_n + h\left[(1 - \theta)f_n + \theta f_{n+1}\right]$$

for $\theta \in [0, 1]$.

a) Show that the method is A-stable for $\theta \geq 1/2$.

b) Show that the method is $A(\alpha)$-stable when it is A-stable.

27. Show that the second-order backward differentiation method is A-stable. Show that the third-order backward differentiation method is not A-stable, but that it is stiffly stable in the sense of Definition 13.

28. Find $X_M > 0$ as large as possible such that the system of differential equations

$$\begin{aligned} y_1' &= -y_1 + xy_2 \\ y_2' &= x^2(y_1 - y_2) \end{aligned}$$

is dissipative in the interval $[0, X_M]$. Deduce that any two solutions with respective initial conditions are then contractive on $[0, X_M]$ in the Euclidean norm.

29. Show that the trapezium rule method is G-stable with $G = 1$.

30. Show that the second-order backward differentiation method and its one-leg twin are G-stable with

$$G = \begin{bmatrix} 5/2 & -1 \\ -1 & 1/2 \end{bmatrix} .$$

31. Develop a shooting method for the numerical solution of the boundary value problem

$$-y'' + ye^y = 1 , \qquad y(0) = y(1) = 0 ,$$

on the interval $[0, 1]$ using:

a) The method of bisection;

b) The Newton-Raphson method.

Explain how your algorithm can be extended to the case of multiple shooting.

32. Construct a three-point finite difference scheme for the numerical solution of the boundary value problem

$$-y'' + x^2 y = 0\,, \qquad y(0) = 1\,, \qquad y'(1) = 0\,,$$

on the interval $[0, 1]$. Show that the resulting system of equations can be written so that its matrix is tridiagonal. Apply the Thomas algorithm to the linear system when the spacing between the mesh points is $h = 1/10$.

33. Suppose that real numbers a_i, b_i and c_i satisfy

$$a_i > 0\,, \qquad b_i > 0\,, \qquad c_i > a_i + b_i\,, \qquad i = 1, 2, \ldots, N-1\,,$$

and let

$$e_i = \frac{b_i}{c_i - a_i e_{i-1}}\,, \qquad i = 1, 2, \ldots, N-1\,,$$

with $e_0 = 0$. Show by induction that $0 < e_i < 1$ for $i = 1, 2, \ldots, N-1$, and that the conditions

$$c_i > 0\,, \qquad c_i > |a_i| + |b_i|\,, \qquad i = 1, 2, \ldots, N-1\,,$$

are sufficient for $e_0 = 0$ to imply that $|e_i| < 1$ for $i = 1, 2, \ldots, N-1$.

How is this method used to solve the system of equations

$$-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i\,, \qquad i = 1, 2, \ldots, N-1\,,$$

with

$$y_0 = 0\,, \qquad y_N = 0\,?$$

34. Assume that the boundary value problem

$$y'' + f(x, y) = 0\,, \quad 0 < x < 1\,, \qquad y(0) = y(1) = 0\,,$$

has a unique solution with a continuous fourth derivative on the interval $[0, 1]$. Suppose further that a unique approximate solution can be computed satisfying the finite difference scheme

$$h^{-2}\delta^2 y_n + f(x_n, y_n) = 0\,, \quad 1 \le n \le N-1\,, \qquad y_0 = y_N = 0\,,$$

where $\delta^2 y_n \equiv y_{n+1} - 2y_n + y_{n-1}$, $x_n = nh$ $(0 \le n \le N)$, and $Nh = 1$ for some integer $N \ge 2$.

a) Show that the truncation error of the finite difference scheme is given by

$$T_n = \frac{1}{12} h^2 y^{(4)}(\xi_n)$$

for some $\xi_n \in (x_{n-1}, x_{n+1})$.

Show further that the global error $e_n = y(x_n) - y_n$ satisfies

$$h^{-2}\delta^2 e_n + p_n e_n = T_n \ , \quad 1 \le n \le N-1 \ , \qquad e_0 = e_N = 0 \ ,$$

where $p_n = f_y(x_n, \eta_n)$ for some η_n between $y(x_n)$ and y_n, and it is assumed that $f_y(x,y)$ is a continuous function of x and y for $x \in [0,1]$ and $y \in \mathbf{R}$.

b) Suppose now that $f_y(x,y) \le 0$ for all $x \in [0,1]$ and $y \in \mathbf{R}$. Let $|T_n| \le M$, $1 \le n \le N-1$. Show that $w_n = \frac{1}{2}h^2 M n(N-n)$ satisfies $h^{-2}\delta^2 w_n = -M$, that

$$h^{-2}\delta^2 w_n + p_n w_n \le -M \ , \quad 1 \le n \le N-1 \ ,$$

and that, if $v_n = w_n + e_n$ or $v_n = w_n - e_n$, then v_n satisfies

$$h^{-2}\delta^2 v_n + p_n v_n \le 0 \ , \quad 1 \le n \le N-1 \ , \qquad v_0 = v_N = 0 \ .$$

c) Assume that v_n has a negative minimum for some value of n between 1 and $N-1$ and show that this leads to a contradiction. Deduce that

$$v_n \ge 0 \ , \qquad 1 \le n \le N-1 \ ,$$

that

$$|e_n| \le w_n \ , \qquad 1 \le n \le N-1 \ ,$$

and that

$$|e_n| \le \frac{1}{96} h^2 \max_{0 \le x \le 1} |y^{(4)}(x)| \ .$$

35. Consider the boundary value problem

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

Develop a collocation method for the numerical solution of this problem on the interval $[0,1]$ using the collocation points

$$x_j = \frac{j}{N+1} \ , \qquad j = 1, \ldots, N \ .$$

and the basis functions $\psi_i(x) = \sin(i\pi x)$, $i = 1, \ldots, N$. Solve the resulting system of linear equations for increasing values of N and compare the numerical solution with the exact solution to the problem.