12

Initial value problems for ODEs

12.1 Introduction

Ordinary differential equations frequently occur in mathematical models that arise in many branches of science, engineering and economics. Unfortunately it is seldom that these equations have solutions which 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

$$y' = f(x, y) \tag{12.1}$$

for the real-valued function y of the real variable x, where $y' \equiv \frac{dy}{dx}$ and f is a given real-valued function of two real variables. In order to select a particular integral from the infinite family of solution curves that constitute the general solution to (12.1), 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 (12.1) for $x > x_0$ such that

$$y(x_0) = y_0\,. \tag{12.2}$$

The differential equation (12.1) together with the initial condition (12.2) is called an **initial value problem**.

If you believe that any initial value problem of the form (12.1), (12.2) possesses a unique solution, take a look at the following example.

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

Example 12.1 *Consider the differential equation $y' = |y|^\alpha$, subject to the initial condition $y(0) = 0$, where α is a fixed real number, $\alpha \in (0,1)$.*

It is a simple matter to verify that, for any nonnegative real number c,

$$y_c(x) = \begin{cases} (1-\alpha)^{\frac{1}{1-\alpha}}(x-c)^{\frac{1}{1-\alpha}}, & c \le x < \infty, \\ 0, & 0 \le x \le c, \end{cases}$$

is a solution to the initial value problem on the interval $[0,\infty)$. Consequently the existence of the solution is ensured, but not its uniqueness; in fact, the initial value problem has an infinite family of solutions $\{y_c\}$, parametrised by $c \ge 0$.

We note in passing that in contrast with the case of $\alpha \in (0,1)$, when $\alpha \ge 1$, the initial value problem $y' = |y|^\alpha$, $y(0) = 0$ has the unique solution $y(x) \equiv 0$. ◇

Example 12.1 indicates that the function f has to obey a certain growth condition with respect to its second argument so as to ensure that (12.1), (12.2) has a unique solution. The precise hypotheses on f guaranteeing the existence of a unique solution to the initial value problem (12.1), (12.2) are stated in the next theorem.

Theorem 12.1 (Picard's Theorem[1]) *Suppose that the real-valued function $(x,y) \mapsto f(x,y)$ is continuous in the rectangular region D defined by $x_0 \le x \le X_M$, $y_0 - C \le y \le y_0 + C$; that $|f(x,y_0)| \le K$ when $x_0 \le x \le X_M$; and that f satisfies the Lipschitz condition: there exists $L > 0$ such that*

$$|f(x,u) - f(x,v)| \le L|u-v| \quad \text{for all } (x,u) \in D, \ (x,v) \in D.$$

Assume further that

$$C \ge \frac{K}{L}\left(\mathrm{e}^{L(X_M - x_0)} - 1\right). \tag{12.3}$$

Then, there exists a unique function $y \in \mathrm{C}^1[x_0, X_M]$ such that $y(x_0) = y_0$ and $y' = f(x,y)$ for $x \in [x_0, X_M]$; moreover,

$$|y(x) - y_0| \le C, \quad x_0 \le x \le X_M.$$

[1] Charles Emile Picard (24 July 1856, Paris, France – 11 December 1941, Paris, France). Although as a child he was a brilliant pupil, Picard disliked mathematics and only became interested in the subject during the vacation following his secondary studies. He was appointed to the chair of differential calculus at the Sorbonne in Paris at the age of 29 but could only take up his position a year later, as university regulations prevented anyone below the age of 30 holding a chair. Picard made important contributions to mathematical analysis and the theory of differential equations.

Proof We define a sequence of functions $(y_n)_{n=0}^{\infty}$ by

$$\begin{aligned} y_0(x) &\equiv y_0\,, \\ y_n(x) &= y_0 + \int_{x_0}^{x} f(s, y_{n-1}(s))\mathrm{d}s\,, \quad n = 1, 2, \ldots\,. \end{aligned} \tag{12.4}$$

Since f is continuous on D, it is clear that each function y_n is continuous on $[x_0, X_M]$. Further, since

$$y_{n+1}(x) = y_0 + \int_{x_0}^{x} f(s, y_n(s))\mathrm{d}s\,,$$

it follows by subtraction that

$$y_{n+1}(x) - y_n(x) = \int_{x_0}^{x} [f(s, y_n(s)) - f(s, y_{n-1}(s))]\,\mathrm{d}s\,. \tag{12.5}$$

We now proceed by induction, and assume that, for some positive value of n,

$$|y_n(x) - y_{n-1}(x)| \le \frac{K}{L}\frac{[L(x - x_0)]^n}{n!}\,, \quad x_0 \le x \le X_M\,, \tag{12.6}$$

and that

$$\begin{aligned} |y_k(x) - y_0| &\le \frac{K}{L}\sum_{j=1}^{k}\frac{[L(x - x_0)]^j}{j!}\,, \\ &\qquad x_0 \le x \le X_M\,, \quad k = 1, \ldots, n\,. \end{aligned} \tag{12.7}$$

Trivially, the hypotheses of the theorem and (12.4) imply that (12.6) and (12.7) hold for $n = 1$.

Now, (12.7) and (12.3) yield that

$$\begin{aligned} |y_k(x) - y_0| &\le \frac{K}{L}\left(\mathrm{e}^{L(X_M - x_0)} - 1\right) \le C\,, \\ &\qquad x_0 \le x \le X_M\,, \quad k = 1, \ldots, n\,. \end{aligned}$$

Therefore $(x, y_{n-1}(x)) \in D$ and $(x, y_n(x)) \in D$ for all $x \in [x_0, X_M]$. Hence, using (12.5), the Lipschitz condition and (12.6),

$$\begin{aligned} |y_{n+1}(x) - y_n(x)| &\le L\int_{x_0}^{x}\frac{K}{L}\frac{[L(s - x_0)]^n}{n!}\mathrm{d}s \\ &= \frac{K}{L}\frac{[L(x - x_0)]^{n+1}}{(n+1)!}\,, \end{aligned} \tag{12.8}$$

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

for all $x \in [x_0, X_M]$. Moreover, using (12.8) and (12.7),

$$\begin{aligned} |y_{n+1}(x) - y_0| &\leq |y_{n+1}(x) - y_n(x)| + |y_n(x) - y_0| \\ &\leq \frac{K}{L}\frac{[L(x-x_0)]^{n+1}}{(n+1)!} + \frac{K}{L}\sum_{j=1}^{n}\frac{[L(x-x_0)]^j}{j!} \\ &= \frac{K}{L}\sum_{j=1}^{n+1}\frac{[L(x-x_0)]^{j+1}}{(j+1)!}\,, \end{aligned} \tag{12.9}$$

for all $x \in [x_0, X_M]$. Thus, (12.6) and (12.7) hold with n replaced by $n+1$, and hence, by induction, they hold for all positive integers n.

Since the infinite series $\sum_{j=1}^{\infty}(c^j/j!)$ converges (to $\mathrm{e}^c - 1$) for any value of $c \in \mathbb{R}$, and for $c = L(X_M - x_0)$ in particular, it follows from (12.6) that the infinite series

$$\sum_{j=1}^{\infty}[y_j(x) - y_{j-1}(x)]$$

converges absolutely and uniformly for $x \in [x_0, X_M]$. However,

$$y_0 + \sum_{j=1}^{n}[y_j(x) - y_{j-1}(x)] = y_n(x)\,,$$

showing that the sequence of continuous functions (y_n) converges to a limit, uniformly on $[x_0, X_M]$, and hence that the limit itself is a continuous function. Calling this limit y, we see from (12.4) that

$$\begin{aligned} y(x) &= \lim_{n\to\infty} y_{n+1}(x) \\ &= y_0 + \lim_{n\to\infty}\int_{x_0}^{x} f(s, y_n(s))\mathrm{d}s\,, \\ &= y_0 + \int_{x_0}^{x}\lim_{n\to\infty} f(s, y_n(s))\mathrm{d}s\,, \\ &= y_0 + \int_{x_0}^{x} f(s, y(s))\mathrm{d}s\,, \end{aligned} \tag{12.10}$$

where we used the uniform convergence of the sequence of functions (y_n) in the transition from line two to line three to interchange the order of the limit process and integration, and the continuity of the function f in the transition from line three to line four. As $s \mapsto f(s, y(s))$ is a continuous function of s on the interval $[x_0, X_M]$, its integral over the interval $[x_0, x]$ is a continuously differentiable function of x. Hence, by

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

(12.10), y is a continuously differentiable function of x on $[x_0, X_M]$; *i.e.*, $y \in \mathrm{C}^1[x_0, X_M]$. On differentiating (12.10) we deduce that

$$y' = f(x, y)\,,$$

as required; also $y(x_0) = y_0$. We have already seen that $(x, y_n(x)) \in D$ when $x_0 \leq x \leq X_M$; as D is a closed set in $\mathbb{R}^2$, on letting $n \to \infty$ it then follows that also $(x, y(x)) \in D$ when $x_0 \leq x \leq X_M$.

To show that the solution of the initial value problem is unique, suppose, if possible, that there are two different solutions y and z. Then, by subtraction,

$$y(x) - z(x) = \int_{x_0}^{x} (f(s, y(s)) - f(s, z(s)))\,\mathrm{d}s\,, \qquad x \in [x_0, X_M]\,,$$

from which it follows that

$$|y(x) - z(x)| \leq L \int_{x_0}^{x} |y(s) - z(s)|\mathrm{d}s \tag{12.11}$$

for all $x \in [x_0, X_M]$. Suppose that m is the maximum value of the expression $|y(x) - z(x)|$ for $x_0 \leq x \leq X_M$, and that $m > 0$. Then,

$$|y(x) - z(x)| \leq mL(x - x_0)\,, \qquad x_0 \leq x \leq X_M\,.$$

Substituting this inequality into the right-hand side of (12.11) we find

$$|y(x) - z(x)| \leq L^2 m \int_{x_0}^{x} (s - x_0)\,\mathrm{d}s = m\frac{[L(x - x_0)]^2}{2!}\,.$$

Proceeding in a similar manner, it is easy to show by induction that

$$|y(x) - z(x)| \leq m\frac{[L(x - x_0)]^k}{k!}\,, \qquad k = 1, 2, \ldots,$$

for all $x \in [x_0, X_M]$. However, the right-hand side in the last inequality is bounded above by $m[L(X_M - x_0)]^k/k!$ for all $x \in [x_0, X_M]$, which can be made arbitrarily small by choosing k sufficiently large. Therefore, $|y(x) - z(x)|$ must be zero for all $x \in [x_0, X_M]$. Hence the solutions y and z are identical. □

In an application of this theorem it is necessary to choose a value of the constant C in Picard's Theorem so that the various hypotheses are satisfied, in particular (12.3); it is not difficult to see that if $\partial f/\partial y$ is continuous in a neighbourhood of (x_0, y_0) the conditions will be satisfied if $X_M - x_0$ is sufficiently small.

As a very simple example, consider the linear equation

$$y' = py + q\,, \tag{12.12}$$

where p and q are constants. Then, $L = |p|$, independently of C, and $K = |py_0| + |q|$. Hence, for any interval $[x_0, X_M]$, the conditions are satisfied by choosing C sufficiently large; therefore, the initial value problem has a unique continuously differentiable solution, defined for all $x \in [x_0, \infty)$.

Now, consider another example

$$y' = y^2\,, \quad y(0) = 1\,.$$

Here for any interval $[0, X_M]$ we have $K = 1$. Choosing any positive value of C we find that

$$|u^2 - v^2| = |u + v|\,|u - v| \le L|u - v| \qquad \forall u, v \in \mathbb{R}\,,$$

where $L = 2(1 + C)$. We therefore now require the condition

$$C \ge \frac{1}{2(1+C)} \left(\mathrm{e}^{2(1+C)X_M} - 1 \right) .$$

This is satisfied if

$$X_M \le F(C) \equiv \frac{1}{2(1+C)} \ln(1 + 2C + 2C^2)\,,$$

where ln means $\log_{\mathrm{e}}$. A sketch of the graph of the function F against C shows that F takes its maximum value near $C = 1.714$, and this gives the condition $X_M \le 0.43$ (see Figure 12.1).

Thus, we are *unable* to prove the existence of the solution over the infinite interval $[0, \infty)$. This is correct, of course, as the unique solution of the initial value problem is

$$y(x) = \frac{1}{1-x}\,, \quad 0 \le x < 1\,,$$

and this is not continuous, let alone continuously differentiable, on any interval $[0, X_M]$ with $X_M \ge 1$. The conditions of Picard's Theorem, which are sufficient but not necessary for the existence and the uniqueness of the solution, have given a rather more restrictive bound on the size of the interval over which the solution exists.

The method of proof of Picard's Theorem also suggests a possible technique for constructing approximations to the solution, by determining the functions y_n from (12.4). In practice it may be impossible, or very difficult, to evaluate the necessary integrals in closed form. We

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

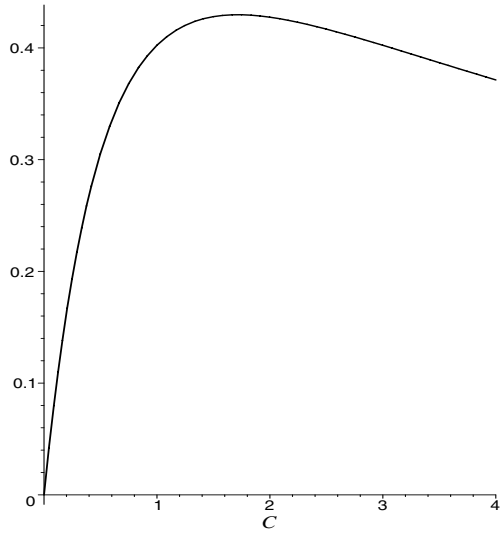


Fig. 12.1. Graph of the function $C \mapsto F(C)$ on the interval $[0, 4]$; F achieves its maximum value near $C = 1.714$ and $F(C) \leq 0.43$ for all $C \geq 0$.

leave it as an exercise (see Exercise 3) to show that for the simple linear equation (12.12), with initial condition $y(0) = 1$, the function y_n is the same as the approximation obtained from the exact solution by expanding the exponential function as a power series and retaining the terms up to the one involving x^n.

In the rest of this chapter we shall consider step-by-step numerical methods for the approximate solution of the initial value problem (12.1), (12.2). We shall suppose throughout that the function f satisfies the conditions of Picard's Theorem. Suppose that the initial value problem (12.1), (12.2) is to be solved on the interval $[x_0, X_M]$. We divide this interval by the **mesh points** $x_n = x_0 + nh$, $n = 0, 1, \ldots, N$, where $h = (X_M - x_0)/N$ and N is a positive integer. The positive real number h is called the **step size** or **mesh size**. 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; these values y_n are calculated in succession, for $n = 1, 2, \ldots, N$.

12.2 One-step methods

A one-step method expresses y_{n+1} in terms of the previous value y_n; later on we shall consider k-step methods, where y_{n+1} is expressed in terms of the k previous values $y_{n-k+1}, \ldots, y_n$, where $k \geq 2$. The simplest example of a one-step method for the numerical solution of the initial value problem (12.1), (12.2) is Euler's method.

Euler's method. Given that $y(x_0) = y_0$, let us suppose that we have already calculated y_n, up to some n, $0 \leq n \leq N-1$, $N \geq 1$; we define

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

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

In order to motivate the definition of Euler's method, let us observe that on expanding $y(x_{n+1}) = y(x_n + h)$ into a Taylor series about x_n, retaining only the first two terms, and writing $y'(x_n) = f(x_n, y(x_n))$, we have that

$$y(x_n + h) = y(x_n) + hf(x_n, y(x_n)) + \mathcal{O}(h^2)\,.$$

After replacing $y(x_n)$ and $y(x_n + h)$ by their numerical approximations, denoted by y_n and y_{n+1}, respectively, and discarding the $\mathcal{O}(h^2)$ term, we arrive at Euler's method.

More generally, a one-step method may be written in the form

$$y_{n+1} = y_n + h\Phi(x_n, y_n; h)\,, \qquad n = 0, 1, \ldots, N-1\,, \quad y(x_0) = y_0\,, \tag{12.13}$$

where $\Phi(\,\cdot\,,\,\cdot\,;\,\cdot\,)$ is a continuous function of its variables. For example, in the case of Euler's method, $\Phi(x_n, y_n; h) = f(x_n, y_n)$. More intricate examples of one-step methods will be discussed below.

In order to assess the accuracy of the numerical method (12.13), we define the **global error**, e_n, by

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

We also need the concept of **truncation error**, T_n, defined by

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - \Phi(x_n, y(x_n); h)\,. \tag{12.14}$$

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

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

Theorem 12.2 *Consider the general one-step method (12.13) 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, that is, there exists a positive constant L_Φ such that, for $0 \le h \le h_0$ and for all (x,u) and (x,v) in the rectangle*

$$D = \{(x,y)\colon\ x_0 \le x \le X_M,\ |y - y_0| \le C\}\,,$$

we have that

$$|\Phi(x,u;h) - \Phi(x,v;h)| \le L_\Phi |u - v|\,. \tag{12.15}$$

Then, assuming that $|y_n - y_0| \le C,\ n = 1,2,\ldots,N$, it follows that

$$|e_n| \le \frac{T}{L_\Phi}\left(\mathrm{e}^{L_\Phi(x_n - x_0)} - 1\right)\,, \quad n = 0,1,\ldots,N\,, \tag{12.16}$$

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

Proof Rewriting (12.14) as

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

and subtracting (12.13) from this, 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 D, the Lipschitz condition (12.15) implies that

$$|e_{n+1}| \le |e_n| + hL_\Phi|e_n| + h|T_n|\,, \qquad n = 0,1,\ldots,N-1\,. \tag{12.17}$$

That is,

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

It easily follows by induction that

$$|e_n| \le \frac{T}{L_\Phi}[(1 + hL_\Phi)^n - 1]\,, \qquad n = 0,1,\ldots,N,$$

since $e_0 = 0$. Observing that $1 + hL_\Phi \le \exp(hL_\Phi)$ gives (12.16). □

Let us apply this general result in order to obtain a bound on the global error in Euler's method. The truncation error for Euler's method is given by

$$\begin{aligned} T_n &= \frac{y(x_{n+1}) - y(x_n)}{h} - f(x_n, y(x_n)) \\ &= \frac{y(x_{n+1}) - y(x_n)}{h} - y'(x_n)\,. \end{aligned} \tag{12.18}$$

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

Assuming that $y \in \mathrm{C}^2[x_0, X_M]$, *i.e.*, that y is a twice continuously differentiable function of x on $[x_0, X_M]$, and expanding $y(x_{n+1})$ about the point x_n into a Taylor series with remainder (see Theorem A.4), we have that

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2!}y''(\xi_n)\,, \qquad x_n < \xi_n < x_{n+1}\,.$$

Substituting this expansion into (12.18) gives

$$T_n = \frac{1}{2}hy''(\xi_n)\,.$$

Let $M_2 = \max_{\zeta\in[x_0,X_M]} |y''(\zeta)|$. Then, $|T_n| \le T$, $n = 0, 1, \ldots, N-1$, where $T = \frac{1}{2}hM_2$. Inserting this into (12.16) and noting that for Euler's method $\Phi(x_n, y_n; h) \equiv f(x_n, y_n)$ and therefore $L_\Phi = L$ where L is the Lipschitz constant for f, we have that

$$|e_n| \le \frac{1}{2}M_2 \left[\frac{\mathrm{e}^{L(x_n-x_0)} - 1}{L}\right] h\,, \quad n = 0, 1, \ldots, N\,. \tag{12.19}$$

Let us highlight the practical relevance of our error analysis by focusing on a particular example.

Example 12.2 *Let us consider the initial value problem* $y' = \tan^{-1} y$, $y(0) = y_0$, *where* y_0 *is a given real number. In order to find an upper bound on the global error* $e_n = y(x_n) - y_n$, *where* y_n *is the Euler approximation to* $y(x_n)$, *we need to determine the constants* L *and* M_2 *in the inequality (12.19).*

Here $f(x, y) = \tan^{-1} y$; so, by the Mean Value Theorem (Theorem A.3),

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

where η lies between u and v. In our case

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

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 the variable x:

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

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

Therefore $|y''(x)| \leq M_2 = \frac{1}{2}\pi$. Inserting the values of L and M_2 into (12.19) and noting that $x_0 = 0$, we have

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

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

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

For such h we shall have $|y(x_n) - y_n| = |e_n| \leq \mathtt{TOL}$, for $n = 0, 1, \ldots, N$, as required. Thus, at least in principle, we can calculate the numerical solution to arbitrarily high accuracy by choosing a sufficiently small step size h.

A numerical experiment shows that this error estimate is rather pessimistic. Taking, for example, $y_0 = 1$ and $X_M = 1$, our bound implies that the tolerance $\mathtt{TOL} = 0.01$ will be achieved with $h \leq 0.0074$; hence, it would appear that we need $N \geq 135$. In fact, using $N = 27$ gives a result from Euler's method which is just within this tolerance, so the error estimate has predicted the use of a step size which is five times smaller than is actually required. ◇

Example 12.3 *As a more typical practical example, consider the problem*

$$y' = y^2 + g(x)\,, \quad y(0) = 2\,, \tag{12.20}$$

where

$$g(x) = \frac{x^4 - 6x^3 + 12x^2 - 14x + 9}{(1+x)^2}\,,$$

is so chosen that the solution is known, and is

$$y(x) = \frac{(1-x)(2-x)}{1+x}\,.$$

The results of some numerical calculations on the interval $x \in [0, 1.6]$ are shown in Figure 12.2. They use step sizes 0.2, 0.1 and 0.05, and show how halving the step size gives a reduction of the error also by a factor of roughly 2, in agreement with the error bound (12.19). ◇

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

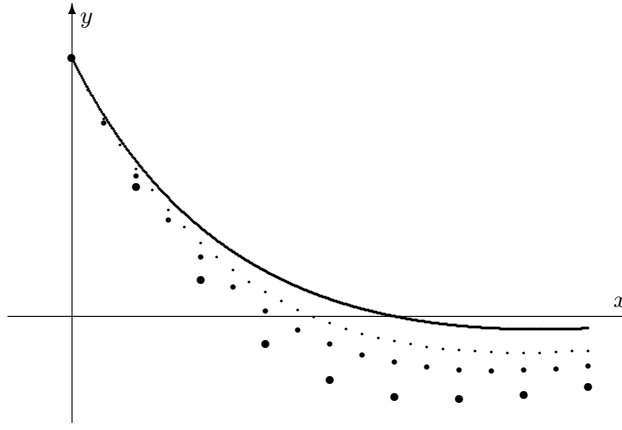


Fig. 12.2. Euler's method for the solution of (12.20). The exact solution (solid curve) and three sets of results are shown (large, medium and small dots), using respectively 8 steps of size 0.2, 16 steps of size 0.1 and 32 steps of size 0.05 on the interval $[0, 1.6]$.

12.3 Consistency and convergence

Returning to the general one-step method (12.13), we consider the choice of the function Φ. Theorem 12.2 suggests that if the truncation error 'approaches zero' as $h \to 0$, then the global error 'converges to zero' also. This observation motivates the following definition.

Definition 12.1 *The numerical method (12.13) is* **consistent** *with the differential equation (12.1) if the truncation error, defined by (12.14), is such that for any $\varepsilon > 0$ there exists a positive $h(\varepsilon)$ for which $|T_n| < \varepsilon$ for $0 < h < h(\varepsilon)$ and any pair of points $(x_n, y(x_n))$, $(x_{n+1}, y(x_{n+1}))$ on any solution curve in D.*

For the general one-step method (12.13) we have assumed that the function $\Phi(\,\cdot\,,\cdot\,;\cdot\,)$ is continuous; since y' is also a continuous function on $[x_0, X_M]$ it follows from (12.14) that, in the limit of

$$h \to 0 \text{ and } n \to \infty, \text{ with } \lim_{n\to\infty} x_n = x \in [x_0, X_M]\,,$$

we have

$$\lim_{n\to\infty} T_n = y'(x) - \Phi(x, y(x); 0)\,.$$

In this limit h tends to zero and n tends to infinity in such a way that x_n tends to a limit point x which lies in the interval $[x_0, X_M]$. This implies

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

that the one-step method (12.13) is consistent if, and only if,

$$\Phi(x, y; 0) \equiv f(x, y)\,. \tag{12.21}$$

This condition is sometimes taken as the definition of consistency. We shall henceforth always assume that (12.21) holds.

Now, we are ready to state a convergence theorem for the general one-step method (12.13).

Theorem 12.3 *Suppose that the initial value problem (12.1), (12.2) satisfies the conditions of Picard's Theorem, and also that its approximation generated from (12.13) when $h \le h_0$ lies in the region D. Assume further that the function $\Phi(\,\cdot\,,\cdot\,;\cdot\,)$ is continuous on $D \times [0, h_0]$, and satisfies the consistency condition (12.21) and the Lipschitz condition*

$$|\Phi(x, u; h) - \Phi(x, v; h)| \le L_\Phi |u - v| \qquad \text{on } D \times [0, h_0]\,. \tag{12.22}$$

Then, if successive approximation sequences (y_n), generated by using the mesh points $x_n = x_0 + nh$, $n = 1, 2, \ldots, N$, are obtained from (12.13) with successively smaller values of h, each h less than h_0, we have convergence of the numerical solution to the solution of the initial value problem in the sense that

$$\lim_{n\to\infty} y_n = y(x) \qquad \text{as} \quad x_n \to x \in [x_0, X_M] \text{ when } h \to 0 \text{ and } n \to \infty\,.$$

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 12.2 implies that

$$|y(x_n) - y_n| \le \left(\frac{\mathrm{e}^{L_\Phi (X_M - x_0)} - 1}{L_\Phi}\right) \max_{0\le m\le n-1} |T_m|\,, \qquad n = 1, 2, \ldots, N\,. \tag{12.23}$$

From the consistency condition (12.21) we have

$$\begin{aligned} T_n &= \left(\frac{y(x_{n+1}) - y(x_n)}{h} - f(x_n, y(x_n))\right) \\ &\quad + (\Phi(x_n, y(x_n); 0) - \Phi(x_n, y(x_n); h))\,. \end{aligned} \tag{12.24}$$

According to the Mean Value Theorem, Theorem A.3, the expression in the first bracket is equal to $y'(\xi_n) - y'(x_n)$, where $\xi_n \in [x_n, x_{n+1}]$. By Picard's Theorem, y' is continuous on the closed interval $[x_0, X_M]$; therefore, it is uniformly continuous on this interval. Hence, for each $\varepsilon > 0$ there exists $h_1(\varepsilon)$ such that

$$|y'(\xi_n) - y'(x_n)| \le \tfrac{1}{2}\varepsilon \qquad \text{for } h < h_1(\varepsilon)\,, \quad n = 0, 1, \ldots, N - 1\,.$$

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

Also, since $\Phi(\,\cdot\,,\cdot\,;\cdot\,)$ is a continuous function on the closed set $D\times[0,h_0]$ and is, therefore, uniformly continuous on $D\times[0,h_0]$, there exists $h_2(\varepsilon)$ such that

$$|\Phi(x_n,y(x_n);0)-\Phi(x_n,y(x_n);h)|\le\tfrac{1}{2}\varepsilon$$

for $h<h_2(\varepsilon)$, $n=0,1,\ldots,N-1$. On defining $h(\varepsilon)=\min\{h_1(\varepsilon),h_2(\varepsilon)\}$, we then have that

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

Inserting this into (12.23) we deduce that

$$\begin{aligned}|y(x)-y_n| &\le |y(x)-y(x_n)|+|y(x_n)-y_n|\\ &\le |y(x)-y(x_n)|+\varepsilon\,\frac{\mathrm{e}^{L_\Phi(X_M-x_0)}-1}{L_\Phi}\,.\end{aligned}\tag{12.25}$$

Now, in the limit of $h\to0$, $n\to\infty$ with $x_n\to x\in[x_0,X_M]$, we have $\lim_{n\to\infty}y(x_n)=y(x)$, since y is a continuous function on $[x_0,X_M]$. Further, the second term on the right-hand side of (12.25) can be made arbitrarily small, independently of h and n, by letting $\varepsilon\to0$. Therefore, in the limit of $h\to0$, $n\to\infty$ with $x_n\to x\in[x_0,X_M]$, we have that $\lim_{n\to\infty}y_n=y(x)$, as stated. □

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

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

where K is a positive constant, independent of h. However, there are other one-step methods (a class of which, called Runge–Kutta[1] methods, will be considered below) for which we can do better. Thus, in order to quantify the asymptotic rate of decay of the truncation error as the step size h converges to 0, we introduce the following definition.

Definition 12.2 *The numerical method (12.13) 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 D of the initial value problem (12.1), (12.2), there exist constants K and h_0 such that*

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

[1] After Carle David Tolmé Runge (30 August 1856, Bremen, Germany – 3 January 1927, Göttingen, Germany) and Martin Wilhelm Kutta (3 November 1867, Pitschen, Upper Silesia, Prussia, North Germany (now Byczyna, Poland) – 25 December 1944, Fürstenfeldbruck, Germany).

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

for any pair of points $(x_n, y(x_n))$, $(x_{n+1}, y(x_{n+1}))$ *on the solution curve.*

12.4 An implicit one-step method

A one-step method with second-order accuracy is the **trapezium rule method**

$$y_{n+1} = y_n + \tfrac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1})]\,. \tag{12.26}$$

This method is easily motivated by writing

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

and approximating the integral by the trapezium rule. Since the right-hand side involves the integral of the function $x \mapsto y'(x) = f(x, y(x))$ we see at once from (7.6) that the truncation error

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - \tfrac{1}{2}\left[f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))\right]$$

of the trapezium rule method satisfies the bound

$$|T_n| \le \tfrac{1}{12}h^2 M_3\,, \quad \text{where } M_3 = \max_{x\in[x_0,X_M]} |y'''(x)|\,. \tag{12.27}$$

The important difference between this method and Euler's method is that the value y_{n+1} appears on both sides of (12.26). To calculate y_{n+1} from the known y_n therefore requires the solution of an equation, which will usually be nonlinear. This additional complication means an increase in the amount of computation required, but not usually a very large increase. The equation (12.26) is easily solved for y_{n+1} by Newton's method, assuming that the derivative $\partial f/\partial y$ can be calculated quickly; as a starting point for the Newton iteration the obvious estimate

$$y_n + hf(x_n, y_n)\,,$$

will usually be close, and a couple of iterations will then suffice.

Methods of this type, which require the solution of an equation to determine the new value y_{n+1}, are known as **implicit methods**.

Writing the trapezium rule method in the standard form (12.13) we see that

$$\begin{aligned} h\Phi(x_n, y_n; h) &= \tfrac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1})] \\ &= \tfrac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_n + h\Phi(x_n, y_n; h)]\,. \end{aligned} \tag{12.28}$$

Hence, the function Φ is also defined in an implicit form.

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

In order to employ Theorem 12.2 to estimate the error in the trapezium rule method we need a value for the Lipschitz constant L_Φ. From (12.28) we find that

$$|\Phi(x_n,u;h)-\Phi(x_n,v;h)| = \tfrac{1}{2}|f(x_n,u)-f(x_n+h,u+h\Phi(x_n,u;h)) \\ -f(x_n,v)-f(x_n+h,v+h\Phi(x_n,v;h))|\,.$$

Hence,

$$\begin{aligned}
&|\Phi(x_n,u;h)-\Phi(x_n,v;h)| \\
&\quad\le \tfrac{1}{2}|f(x_n,u)-f(x_n,v)| \\
&\qquad +\tfrac{1}{2}|f(x_n+h,u+h\Phi(x_n,u;h))-f(x_n+h,v+h\Phi(x_n,v;h))| \\
&\quad\le \tfrac{1}{2}L_f|u-v| \\
&\qquad +\tfrac{1}{2}L_f|u+h\Phi(x_n,u;h)-v-h\Phi(x_n,v;h)| \\
&\quad\le \tfrac{1}{2}L_f|u-v|+\tfrac{1}{2}L_f|u-v|+\tfrac{1}{2}L_fh|\Phi(x_n,u;h)-\Phi(x_n,v;h)|\,.
\end{aligned}$$

This shows that

$$\left(1-\tfrac{1}{2}hL_f\right)|\Phi(x_n,u;h)-\Phi(x_n,v;h)| \le L_f|u-v|\,,$$

and, therefore,

$$L_\Phi \le \frac{L_f}{1-\frac{1}{2}hL_f}\,, \qquad \text{provided that } \tfrac{1}{2}hL_f<1\,.$$

Consequently, (12.16) and (12.27) imply that the global error in the trapezium rule method is $\mathcal{O}(h^2)$, as h tends to 0.

Figure 12.3 depicts the results of some numerical calculations on the interval $x\in[0,1.6]$ for the same problem as in Figure 12.2. The step sizes are 0.4 and 0.2, larger than for Euler's method; nevertheless we see a much reduced error in comparison with Euler's method, and also how the reduction in the step size h by a factor of 2 gives a reduction in the error by a factor of about 4, as predicted by our error analysis.

12.5 Runge–Kutta methods

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

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

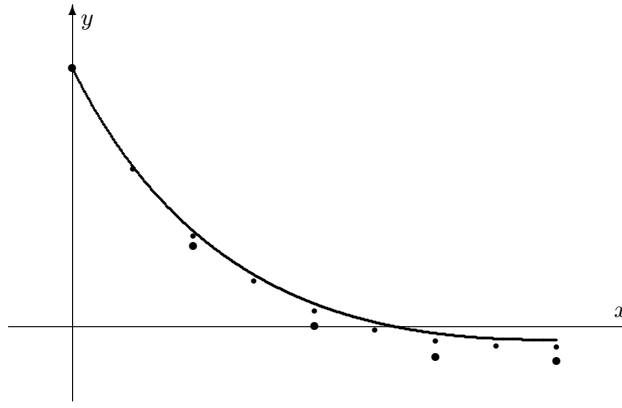


Fig. 12.3. Trapezium rule method for the solution of (12.20). The exact solution (solid curve) and two sets of results are shown (large and small dots), using respectively 4 steps of size 0.4, and 8 steps of size 0.2 on $[0, 1.6]$.

$(x_n, y(x_n))$ and $(x_{n+1}, y(x_{n+1}))$. Consider, for example, the following family of methods:

$$y_{n+1} = y_n + h(ak_1 + bk_2)\,, \tag{12.29}$$

where

$$k_1 = f(x_n, y_n)\,, \tag{12.30}$$

$$k_2 = f(x_n + \alpha h, y_n + \beta h k_1)\,, \tag{12.31}$$

and where the parameters a, b, α and β are to be determined.

Note that Euler's method is a member of this family of methods, corresponding to $a = 1$ and $b = 0$. However, we are now seeking methods that are at least second-order accurate. Clearly (12.29)–(12.31) can be written in the form (12.13) with

$$\Phi(x_n, y_n; h) = af(x_n, y_n) + bf(x_n + \alpha h, y_n + \beta h f(x_n, y_n))\,.$$

By the condition (12.21), a method from this family will be consistent if, and only if, $a + b = 1$. Further conditions on the parameters are found by attempting to maximise the order of accuracy of the method.

To determine the truncation error of the method from (12.14) we need the higher derivatives of $y(x)$, which are obtained by differentiating the function f:

$$y'(x_n) = f\,,$$

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

$$\begin{aligned} y''(x_n) &= f_x + f_y y' = f_x + f_y f\,, \\ y'''(x_n) &= f_{xx} + f_{xy} f + (f_{xy} + f_{yy} f) f + f_y (f_x + f_y f)\,, \end{aligned}$$

and so on; in these expressions the subscripts x and y denote partial derivatives, and all functions appearing on the right-hand sides are to be evaluated at $(x_n, y(x_n))$. We also need to expand $\Phi(x_n, y(x_n); h)$ in powers of h, giving (with the same notational conventions as before)

$$\begin{aligned} \Phi(x_n, y(x_n); h) &= af + b\big(f + \alpha h f_x + \beta h f f_y + \tfrac{1}{2}(\alpha h)^2 f_{xx} \\ &\qquad + \alpha\beta h^2 f f_{xy} + \tfrac{1}{2}(\beta h)^2 f^2 f_{yy} + \mathcal{O}(h^3)\big)\,. \end{aligned}$$

Thus, we obtain the truncation error in the form

$$\begin{aligned} T_n &= \frac{y(x_n + h) - y(x_n)}{h} - \Phi(x_n, y(x_n); h) \\ &= f + \tfrac{1}{2} h (f_x + f f_y) \\ &\quad + \tfrac{1}{6} h^2 [f_{xx} + 2 f_{xy} f + f_{yy} f^2 + f_y (f_x + f_y f)] \\ &\quad - \big\{ af + b[f + \alpha h f_x + \beta h f f_y + \tfrac{1}{2}(\alpha h)^2 f_{xx} \\ &\qquad + \alpha\beta h^2 f f_{xy} + \tfrac{1}{2}(\beta h)^2 f^2 f_{yy}] \big\} + \mathcal{O}(h^3)\,. \end{aligned}$$

As $1 - a - b = 0$, the term $(1 - a - b) f$ is equal to 0. The coefficient of the term in h is

$$\tfrac{1}{2}(f_x + f f_y) - b\alpha f_x - b\beta f f_y$$

which vanishes for all functions f provided that

$$b\alpha = b\beta = \tfrac{1}{2}\,.$$

The method is therefore second-order accurate if

$$\beta = \alpha\,, \quad a = 1 - \tfrac{1}{2\alpha}\,, \quad b = \tfrac{1}{2\alpha}\,, \qquad \alpha \neq 0\,,$$

showing that there is a one-parameter family of second-order methods of this form, parametrised by $\alpha \neq 0$. The truncation error of the method then becomes

$$\begin{aligned} T_n &= h^2 \{ (\tfrac{1}{6} - \tfrac{\alpha}{4})(f_{xx} + f_{yy} f^2) + (\tfrac{1}{3} - \tfrac{\alpha}{2}) f f_{xy} \\ &\qquad + \tfrac{1}{6}(f_x f_y + f f_y^2) \} + \mathcal{O}(h^3)\,. \end{aligned} \tag{12.32}$$

Evidently there is no choice of the free parameter α which will make this method third-order accurate for all functions f; this can be seen, for example, by considering the initial value problem $y' = y$, $y(0) = 1$, and noting that in this case (12.32), with $f(x, y) = y$, yields

$$T_n = \tfrac{1}{6} h^2 y(x_n) + \mathcal{O}(h^3) = \tfrac{1}{6} h^2 \mathrm{e}^{x_n} + \mathcal{O}(h^3)\,.$$

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

Two examples of second-order Runge–Kutta methods of the form (12.29)–(12.31) are the modified Euler method and the improved Euler method.

(a) **The modified Euler method.** In this case we take $\alpha = \frac{1}{2}$ 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 $\alpha = 1$ 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 using (12.32) that the truncation error is of the form, respectively,

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

A similar but more complicated analysis is used to construct Runge–Kutta methods of higher order. One of the most frequently used methods of the Runge–Kutta family is often known as the **classical fourth-order method**:

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

where

$$\left.\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f\left(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hk_1\right), \\ k_3 &= f\left(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hk_2\right), \\ k_4 &= f(x_n + h, y_n + hk_3). \end{aligned}\right\} \tag{12.33}$$

Here k_2 and k_3 represent approximations to the derivative y' 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, $i = 1, 2, 3, 4$, the weights corresponding to those of Simpson's rule (to which the classical fourth-order Runge–Kutta method reduces when $\frac{\partial f}{\partial y} \equiv 0$).

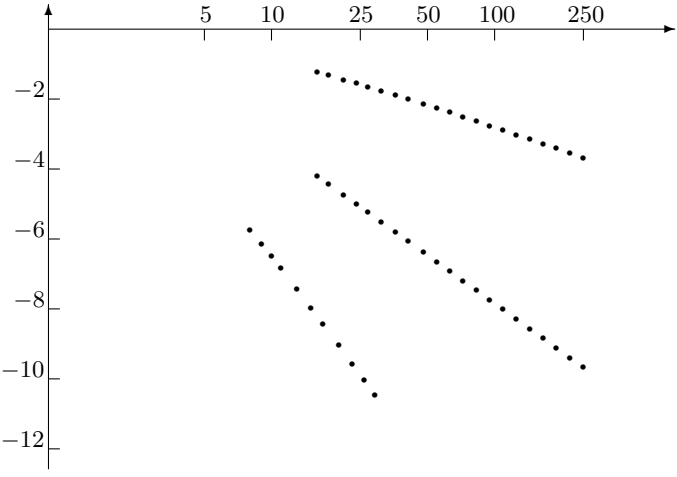


Fig. 12.4. The errors in three methods for the solution of (12.20) on the interval $[0, 1.6]$. Reading from the top, the lines (whose slopes indicate first-, second- and fourth-order convergence) represent the errors of Euler's method, the trapezium rule method, and the classical Runge–Kutta method respectively. The horizontal axis indicates the number $N = 1.6/h$, on a logarithmic scale, and the vertical axis shows $\ln|e_N| = \ln|y(1.6) - y_N|$.

To illustrate the behaviour of the one-step methods which we have discussed, Figure 12.4 shows the errors in the calculation of $y(1.6)$, where $y(x)$ is the solution to the problem (12.20) on the interval $[0, 1.6]$. The horizontal axis indicates N, the number of equally spaced mesh points used in the interval $(0, 1.6]$, on a logarithmic scale, and the vertical axis shows $\ln|e_N| = \ln|y(1.6) - y_N|$. The three methods employed are Euler's method, the trapezium rule method, and the classical Runge–Kutta method (12.33). The three lines show clearly the improved accuracy of the higher-order methods, and the rate at which the accuracy improves as N increases.

12.6 Linear multistep methods

While Runge–Kutta methods give 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},

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

yields

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

and applying Simpson's rule to approximate the integral on the right-hand side then leads to the method

$$y_{n+1} = y_{n-1} + \frac{1}{3}h\left[f(x_{n-1}, y_{n-1}) + 4f(x_n, y_n) + f(x_{n+1}, y_{n+1})\right], \tag{12.34}$$

requiring only three function evaluations per step. 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 (12.34) is *not* a one-step method.

In this section we consider a class of methods of the type (12.34) for the numerical solution of the initial value problem (12.1), (12.2), called **linear multistep methods**.

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

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(x_{n+j}, y_{n+j})\,, \tag{12.35}$$

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 0. 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 method (12.35) is called *linear* because it involves only linear combinations of the y_{n+j} and the $f(x_{n+j}, y_{n+j})$, $j = 0, 1, \ldots, k$; for the sake of notational simplicity, henceforth we shall often write f_n instead of $f(x_n, y_n)$.

Example 12.4 *We have already seen an example of a linear two-step method in (12.34); here we present further examples of linear multistep methods.*

(a) Euler's method is a trivial case: it is an explicit linear one-step

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

method. The **implicit Euler method**

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) \tag{12.36}$$

is an implicit linear one-step method. Another trivial example is the **trapezium rule method**, given by

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

it, too, is an implicit linear one-step method.

(b) The **Adams**[1]**–Bashforth**[2] **method**

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

is an example of an explicit linear four-step method, while the **Adams–Moulton**[3] **method**

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

is an implicit linear three-step method. ◇

There are systematic ways of generating linear multistep methods, but these constructions will not be discussed here. Instead, we turn our attention to the analysis of linear multistep methods and introduce the concepts of *(zero-) stability*, *consistency* and *convergence*. The significance of these properties cannot be overemphasised: the failure of any of the three will render the linear multistep method practically useless.

12.7 Zero-stability

As is clear from (12.35) we need k starting values, $y_0, \ldots, y_{k-1}$, before we can apply a linear k-step method to the initial value problem (12.1), (12.2): of these, y_0 is given by the initial condition (12.2), but the others,

[1] John Couch Adams (5 June 1819, Laneast, Cornwall, England – 21 January 1892, Cambridge, Cambridgeshire, England) was educated at St John's College in Cambridge. In 1841 while he was still an undergraduate, he began to study the irregularities of the motion of Uranus to discover whether these can be attributed to the action of an undiscovered planet. Four years later he gave accurate information about the position of the new planet (Neptune) to the director of the Cambridge Observatory. Adams made several other contributions to astronomy.

[2] F. Bashforth: *An Attempt to Test the Theories of Capillary Action by Comparing the Theoretical and Measured Forms of Drops of Fluid. With an Explanation of the Method of Integration in Constructing Tables Which Give the Theoretical Form of Such Drops, by J.C. Adams*, Cambridge University Press, 1883.

[3] F.R. Moulton: *New Methods in Exterior Ballistics*, University of Chicago Press, 1926.

$y_1, \ldots, y_{k-1}$, have to be computed by other means: say, by using a suitable one-step method (*e.g.* a 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 (12.35). Thus, we wish to consider the 'stability' of the numerical method with respect to 'small perturbations' in the starting conditions.

Definition 12.3 *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 (z_n) that have been generated by the same formulae but different starting values $y_0, y_1, \ldots, y_{k-1}$ and $z_0, z_1, \ldots, z_{k-1}$, respectively, we have*

$$|y_n - z_n| \leq K \max\{|y_0 - z_0|, |y_1 - z_1|, \ldots, |y_{k-1} - z_{k-1}|\} \qquad (12.37)$$

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 (12.1) with $f(x, y) \equiv 0$; it is for this reason that the concept of stability formulated in Definition 12.3 is referred to as *zero*-stability. While Definition 12.3 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 multistep method using Definition 12.3 alone. Thus, we shall next formulate an algebraic equivalent of zero-stability, known as the **Root Condition**, which will simplify this task. Before doing so, however, we introduce some notation.

Given the linear k-step method (12.35) we consider its **first** and **second characteristic polynomials**, 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

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

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

Before stating the main theorem of this section, we recall a classical result from the theory of kth-order linear recurrence relations.

Lemma 12.1 *Consider the kth-order homogeneous linear recurrence relation*

$$\alpha_k y_{n+k} + \cdots + \alpha_1 y_{n+1} + \alpha_0 y_n = 0\,, \qquad n = 0, 1, 2, \ldots\,, \tag{12.38}$$

with $\alpha_k \neq 0$, $\alpha_0 \neq 0$, $\alpha_j \in \mathbb{R}$, $j = 0, 1, \ldots, k$, and the corresponding **characteristic polynomial**

$$\rho(z) = \alpha_k z^k + \cdots + \alpha_1 z + \alpha_0\,.$$

Let z_r, $1 \le r \le \ell$, $\ell \le k$, be the distinct roots of the polynomial ρ, and let $m_r \ge 1$ denote the multiplicity of z_r, with $m_1 + \cdots + m_\ell = k$. If a sequence (y_n) of complex numbers satisfies (12.38), then

$$y_n = \sum_{r=1}^{\ell} p_r(n) z_r^n\,, \qquad \textit{for all } n \ge 0\,, \tag{12.39}$$

where $p_r(\,\cdot\,)$ is a polynomial in n of degree $m_r - 1$, $1 \le r \le \ell$. In particular, if all roots are simple, that is $m_r = 1$, $1 \le r \le k$, then the p_r, $r = 1, \ldots, k$, are constants.

Proof We give a sketch of the proof.[1] Let us first consider the case when all of the (distinct) roots $z_1, z_2, \ldots, z_k$ are simple. As, by assumption, $\alpha_0 \neq 0$, none of the roots is equal to 0. It is then easy to verify by direct substitution that, since $\rho(z_r) = 0$, $r = 1, 2, \ldots, k$, each of the sequences $(y_n) = (z_r^n)$, $r = 1, 2, \ldots, k$, satisfies (12.38).

In order to prove that any solution (y_n) of (12.38) can be expressed as a linear combination of the sequences $(z_1^n), (z_2^n), \ldots, (z_k^n)$, it suffices to show that these k sequences are linearly independent. To do so, let us suppose that

$$C_1 z_1^n + C_2 z_2^n + \cdots + C_k z_k^n = 0\,, \qquad \text{for all } n = 0, 1, 2, \ldots.$$

Then, in particular,

$$\begin{array}{lllll} C_1 & + C_2 & + \cdots + C_k & = & 0\,, \\ C_1 z_1 & + C_2 z_2 & + \cdots + C_k z_k & = & 0\,, \\ \multicolumn{3}{c}{\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots} & & \\ C_1 z_1^{k-1} & + C_2 z_2^{k-1} & + \cdots + C_k z_k^{k-1} & = & 0\,. \end{array}$$

[1] For details, see, for example, pp. 213–214 of P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962.

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

The matrix of this system of k simultaneous linear equations for the k unknowns $C_1, C_2, \ldots, C_k$ has the determinant

$$\mathcal{D} = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ z_1 & z_2 & \ldots & z_k \\ \ldots & \ldots & \ldots & \ldots \\ z_1^{k-1} & z_2^{k-1} & \ldots & z_k^{k-1} \end{vmatrix},$$

known as the Vandermonde determinant, and $\mathcal{D} = \prod_{r<s}(z_s - z_r)$. Since the roots are distinct, $\mathcal{D} \neq 0$, so the matrix of the system is nonsingular. Therefore $C_1 = C_2 = \cdots = C_k = 0$ is the unique solution, which then means that the sequences $(z_1^n), (z_2^n), \ldots, (z_k^n)$ are linearly independent.

Now, suppose that (y_n) is any solution of (12.38); as $\mathcal{D} \neq 0$, there exists a unique set of k constants, $C_1, C_2, \ldots, C_k$, such that

$$y_m = C_1 z_1^m + C_2 z_2^m + \cdots + C_k z_k^m, \qquad m = 0, 1, \ldots, k-1. \quad (12.40)$$

Substituting these equalities into (12.38) for $n = 0$, we conclude that

$$\begin{aligned} 0 &= \alpha_k y_k + \alpha_{k-1}(C_1 z_1^{k-1} + \cdots + C_k z_k^{k-1}) + \cdots \\ &\quad + \alpha_0(C_1 z_1^0 + \cdots + C_k z_k^0) \\ &= \alpha_k y_k + C_1(\rho(z_1) - \alpha_k z_1^k) + \cdots + C_k(\rho(z_k) - \alpha_k z_k^k) \\ &= \alpha_k(y_k - (C_1 z_1^k + \cdots + C_k z_k^k)). \end{aligned}$$

As $\alpha_k \neq 0$, it follows that

$$y_k = C_1 z_1^k + \cdots + C_k z_k^k,$$

which, together with (12.40), proves (12.39) for $0 \leq n \leq k$ in the case of simple roots. Next, we select $n = 1$ in (12.38) and proceed in the same manner as in the case of $n = 0$ discussed above to show that (12.39) holds for $0 \leq n \leq k+1$. Continuing in the same way, we deduce by induction that (12.39) holds for all $n \geq 0$.

In the case when $\rho(z)$ has repeated roots, the proof is similar, except that instead of (z_r^n), $r = 1, 2, \ldots, n$, the following k sequences are used:

$$\left.\begin{array}{l} (z_r^n), \\ (n z_r^n), \\ \ldots\ldots\ldots \\ (n(n-1)\ldots(n-m_r+2) z_r^n), \quad r = 1, 2, \ldots, \ell. \end{array}\right\} \quad (12.41)$$

These can be shown to satisfy (12.38) by direct substitution on noting that $\rho(z_r) = \rho'(z_r) = \cdots = \rho^{(m_r-1)}(z_r) = 0$, given that z_r is a root of

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

$\rho(z)$ of multiplicity m_r, $r = 1, 2, \ldots, \ell$. The linear independence of the sequences (12.41) follows as before, except instead of $\prod_{r<s}(z_s - z_r)$, the value of the corresponding determinant is now

$$\mathcal{D}_1 = \prod_{1 \le r < s \le \ell} (z_r - z_s)^{m_r + m_s} \prod_{r=1}^{\ell} (m_r - 1)!!$$

where $0!! = 1$, $m!! = m!\,(m-1)!\ldots 1!$ for $m = 1, 2, \ldots$. As the roots $z_1, z_2, \ldots, z_\ell$ are distinct, we have that $\mathcal{D}_1 \neq 0$, and therefore the sequences (12.41) are linearly independent. The rest of the argument is identical as in the case of simple roots.[1] □

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

Theorem 12.4 (Root Condition) *A linear multistep method is zero-stable for any initial value problem of the form (12.1), (12.2), where f satisfies the hypotheses of Picard's Theorem, if, and only if, all roots of the first characteristic polynomial of the method are inside the closed unit disc in the complex plane, 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 of theorem Necessity. Consider the method (12.35), applied to $y' = 0$:

$$\alpha_k y_{n+k} + \cdots + \alpha_1 y_{n+1} + \alpha_0 y_n = 0\,. \tag{12.42}$$

According to Lemma 12.1, every solution of this kth-order linear recurrence relation has the form

$$y_n = \sum_{r=1}^{\ell} p_r(n) z_r^n\,, \tag{12.43}$$

where z_r is a root, of multiplicity $m_r \ge 1$, of the first characteristic polynomial ρ of the method, and the polynomial p_r has degree $m_r - 1$, $1 \le r \le \ell$, $\ell \le k$. Clearly, if $|z_r| > 1$ for some r, then there are starting values $y_0, y_1, \ldots, y_{k-1}$ for which the corresponding solution grows like

[1] We warn the reader that in certain mathematical texts the notation $m!!$ is, instead, used to mean $m \cdot (m-2) \ldots 5 \cdot 3 \cdot 1$ for m odd and $m \cdot (m-2) \ldots 6 \cdot 4 \cdot 2$ for m even.

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

$|z_r|^n$, and if $|z_r| = 1$ and the multiplicity is $m_r > 1$, then there is a solution growing like n^{m_r-1}. In either case there are solutions that grow unboundedly as $n \to \infty$, *i.e.*, as $h \to 0$ with nh fixed. Considering starting values $y_0, y_1, \ldots, y_{k-1}$ which give rise to such an unbounded solution (y_n), and starting values $z_0 = z_1 = \cdots = z_{k-1} = 0$ for which the corresponding solution of (12.42) is (z_n) with $z_n = 0$ for all n, we see that (12.37) 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, the interested reader is referred to Theorem 3.1 on page 353 of W. Gautschi, *Numerical Analysis: an Introduction*, Birkhäuser, Boston, MA, 1997. □

Example 12.5 *We shall explore the zero-stability of the methods from Example 12.4 using the Root Condition.*

(a) The Euler method and the implicit Euler method 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 rule method.

(b) The Adams–Bashforth and Adams–Moulton methods considered in Example 12.4 have first characteristic polynomials, respectively, $\rho(z) = z^3(z-1)$ and $\rho(z) = z^2(z-1)$. These have multiple root $z = 0$ and simple root $z = 1$, and therefore both methods are zero-stable.

(c) The three-step method

$$\begin{aligned} 11y_{n+3} + 27y_{n+2} - 27y_{n+1} - 11y_n \\ = 3h\,(f_{n+3} + 9f_{n+2} + 9f_{n+1} + f_n) \end{aligned} \qquad (12.44)$$

is *not* zero-stable. Indeed, the corresponding first characteristic polynomial $\rho(z) = 11z^3 + 27z^2 - 27z - 11$ has roots at $z_1 = 1$, $z_2 \approx -0.32$, $z_3 = -3.14$, so $|z_3| > 1$.

(d) The first characteristic polynomial of the three-step method

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

is $\rho(z) = z^3 + z^2 - z - 1 = (z+1)(z^2-1)$, which has roots $z_{1/2} = -1$, $z_3 = 1$. The first of these is a double root lying on the unit circle; therefore, the method is *not* zero-stable. ◇

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

12.8 Consistency

In this section we consider the accuracy of the linear k-step method (12.35). For this purpose, as in the case of one-step methods, we introduce the notion of truncation error. Thus, suppose that y is a solution to the ordinary differential equation (12.1). The truncation error of (12.35) is then defined as follows:

$$T_n = \frac{\sum_{j=0}^{k} [\alpha_j y(x_{n+j}) - h\beta_j f(x_{n+j}, y(x_{n+j}))]}{h \sum_{j=0}^{k} \beta_j} . \tag{12.45}$$

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 (12.35) 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 12.4 *The numerical method (12.35) is said to be* **consistent** *with the differential equation (12.1) if the truncation error defined by (12.45) is such that for any $\varepsilon > 0$ there exists an $h(\varepsilon)$ for which*

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

and any $k+1$ points $(x_n, y(x_n)), \ldots, (x_{n+k}, y(x_{n+k}))$ on any solution curve in D of the initial value problem (12.1), (12.2).

Now, let us suppose that the solution to the differential equation is sufficiently smooth, and let us expand the expressions $y(x_{n+j})$ and $f(x_{n+j}, y(x_{n+j})) = y'(x_{n+j})$ into Taylor series about the point x_n. On substituting these expansions into the numerator in (12.45) we obtain

$$T_n = \frac{1}{h\sigma(1)} \left[C_0 y(x_n) + C_1 h y'(x_n) + C_2 h^2 y''(x_n) + \cdots \right] \tag{12.46}$$

where

$$\left.\begin{aligned} C_0 &= \sum_{j=0}^{k} \alpha_j , \\ C_1 &= \sum_{j=1}^{k} j\alpha_j - \sum_{j=0}^{k} \beta_j , \\ C_2 &= \sum_{j=1}^{k} \tfrac{j^2}{2!} \alpha_j - \sum_{j=1}^{k} j\beta_j , \\ &\cdots \\ C_q &= \sum_{j=1}^{k} \tfrac{j^q}{q!} \alpha_j - \sum_{j=1}^{k} \tfrac{j^{q-1}}{(q-1)!} \beta_j . \end{aligned}\right\} \tag{12.47}$$

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

For consistency we need that, as $h \to 0$ and $n \to \infty$ with $x_n \to x \in [x_0, X_M]$, the truncation error T_n tends to 0. This requires that $C_0 = 0$ and $C_1 = 0$ in (12.46). 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 multistep 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 root.

Definition 12.5 *The numerical method (12.35) is said to have* **order of accuracy** *p, if p is the largest positive integer such that, for any sufficiently smooth solution curve in D of the initial value problem (12.1), (12.2), there exist constants K and h_0 such that*

$$|T_n| \le K h^p \quad \textit{for } 0 < h \le h_0\,,$$

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

Thus, we deduce from (12.46) that the method is of order of accuracy p if, and only if,

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

In this case,

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

The number $C_{p+1}/\sigma(1)$ is called the **error constant** of the method.

Example 12.6 *Let us determine all values of the real parameter b, $b \neq 0$, for which the linear multistep method*

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

is zero-stable. We shall show that there exists a value of b for which the order of the method is 4, and that if the method is zero-stable for some value of b, then its order cannot exceed 2.

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

According to the Root Condition, this linear multistep 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:

$$\rho(z) = (z-1)\rho_1(z)\,, \qquad \text{where} \quad \rho_1(z) = z^2 - 2(1-b)z + 1\,.$$

Thus, the method is zero-stable if, and only if, all roots of the polynomial $\rho_1(z)$ 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. It then 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 $z = -1$. Further, since the product of the two roots of $\rho_1(z)$ is equal to 1, both have modulus less than or equal to 1, and neither of them is equal to ± 1, it follows that they must both be strictly complex; hence the discriminant of the quadratic polynomial $\rho_1(z)$ must be negative. That is, $4(1-b)^2 - 4 < 0$. 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

$$\begin{aligned} T_n &= \frac{1}{\sigma(1)}\left[\left(1 - \frac{b}{6}\right)h^2 y'''(x_n) + \frac{1}{4}(6-b)h^3 y^{\mathrm{iv}}(x_n)\right. \\ &\qquad \left. + \frac{1}{120}(150 - 23b)h^4 y^{\mathrm{v}}(x_n) + \mathcal{O}(h^5)\right], \end{aligned}$$

where $\sigma(1) = 2b \neq 0$. If $b = 6$, then $T_n = \mathcal{O}(h^4)$ 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 $T_n = \mathcal{O}(h^2)$. ◇

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

12.9 Dahlquist's theorems

An important result connecting the concepts of zero-stability, consistency and convergence of a linear multistep method was proved by the Swedish mathematician Germund Dahlquist.

Theorem 12.5 (Dahlquist's Equivalence Theorem) *For a linear k-step method that is consistent with the ordinary differential equation (12.1) where f is assumed to satisfy a Lipschitz condition, and with consistent starting values,*[1] *zero-stability is necessary and sufficient for convergence. Moreover if the solution y has continuous derivative of order $p+1$ and truncation error $\mathcal{O}(h^p)$, then the global error of the method, $e_n = y(x_n) - y_n$, is also $\mathcal{O}(h^p)$.*

The proof of this result is long and technical; for details of the argument, see Theorem 6.3.4 on page 357 of W. Gautschi, *Numerical Analysis: an Introduction*, Birkhäuser, Boston, MA, 1997, or Theorem 5.10 on page 244 of P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962.

By virtue of Dahlquist's theorem, if a linear multistep 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 multistep 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.

A second theorem by Dahlquist imposes a restriction on the order of accuracy of a zero-stable linear multistep method.

Theorem 12.6 (Dahlquist's Barrier Theorem) *The order of accuracy of a zero-stable k-step method cannot exceed $k+1$ if k is odd, or $k+2$ if k is even.*

A proof of this result will be found in Section 4.2 of Gautschi's book or in Section 5.2-8 of Henrici's book, cited above.

Theorem 12.6 makes it very difficult to choose a 'best' multistep method of a given order. Suppose, for example, that we consider five-step methods. The general five-step method involves 12 parameters, of

[1] That is, with starting values $y_j = \eta_j \equiv \eta_j(h)$, $j = 0, \ldots, k-1$, which all converge to the exact initial value y_0, as $h \to 0$.

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

which 11 are independent: the method is obviously unaffected by multiplying all the parameters by a nonzero constant. Now it would be possible to construct a five-step method of order 10, by solving the 11 equations of the form $C_q = 0,\ q = 0, 1, \ldots, 10$, where C_q is given in (12.47). But the Barrier Theorem states that this method would not be zero-stable, and the order of a zero-stable five-step method cannot exceed 6. There is a family of stable five-step methods of order 6, involving 4 free parameters, and there is no obvious way of deciding whether any one of these methods is better than the others.

Example 12.7 (i) *The Barrier Theorem says that when $k = 1$ the order of accuracy of a zero-stable method cannot exceed 2. The trapezium rule method has order 2, and is zero-stable.*

(ii) *The two-step method*

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

is zero-stable, as the roots of the first characteristic polynomial, $\rho(z) = z^2 - 1$, are 1 and -1. A simple calculation shows that its order of accuracy is 4; by the Barrier Theorem, this is the highest order which could be achieved by a two-step method.

(iii) *The three-step method*

$$\begin{aligned} 11y_{n+3} + 27y_{n+2} - 27y_{n+1} - 11y_n \\ = 3h\,(f_{n+3} + 9f_{n+2} + 9f_{n+1} + f_n) \end{aligned}$$

has order 6. The Barrier Theorem therefore implies that this method is not zero-stable. We have already shown this in Example 12.5(c) using the Root Condition.

It is found that all the zero-stable k-step methods of highest possible order are *implicit*, with β_k nonzero.

12.10 Systems of equations

In this section we discuss the application of numerical methods to simultaneous systems of differential equations, which we shall write in the form

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}x} = \boldsymbol{f}(x, \boldsymbol{y})\,.$$

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

Here $\boldsymbol{y}$ is an m-component vector function of x, and $\boldsymbol{f}$ is an m-component vector function of the independent variable x and the vector variable $\boldsymbol{y}$. In component form the system becomes

$$\frac{\mathrm{d}y_j}{\mathrm{d}x} = f_j(x, y_1, \ldots, y_m)\,, \qquad j = 1, 2, \ldots, m\,.$$

The system comprises m simultaneous differential equations. To single out a unique solution we need m side conditions, and we shall suppose that all these conditions are given at the same value of x, and have the form

$$\boldsymbol{y}(x_0) = \boldsymbol{y}_0\,,$$

or, in component form,

$$y_j(x_0) = y_{j,0}\,, \qquad j = 1, 2, \ldots, m\,,$$

where the values of $y_{j,0}$ are given. This is called an initial value problem for a system of ordinary differential equations; we may also require a solution of the system on an interval $[a, b]$, with r conditions given at one end of the interval and $m - r$ conditions at the other end. This constitutes a boundary value problem, and requires different numerical methods which are considered in the next chapter.

All the numerical methods which we have discussed apply without change to systems of differential equations; it is only necessary to realise that we are dealing with vectors. For example, the first stage of the classical Runge–Kutta method (12.33) becomes

$$\boldsymbol{k}_1 = \boldsymbol{f}(x_n, \boldsymbol{y}_n)\,;$$

we must evaluate all the elements of the vector $\boldsymbol{k}_1$ before proceeding to the next stage to calculate $\boldsymbol{k}_2$, and so on.

The most important difference which arises in dealing with a system of differential equations is in the practical use of an *implicit* multistep method. As we have seen, this almost always requires an iterative method for the solution of an equation to determine y_{n+1}. Applying such a method to a system of differential equations now involves the solution of a system of equations, which will usually be nonlinear, to determine the elements of the vector $\boldsymbol{y}_{n+1}$. In real-life problems it is quite common to deal with systems of several hundred differential equations, and it then becomes very important to be sure that the improved efficiency of the implicit method justifies the very considerable extra work in each step of the process.

We shall not discuss the extension of our earlier analysis to deal with

systems of differential equations; in almost all cases we simply need to introduce vector notation, and replace the absolute value of a number by the norm of a vector. For example, in the proof of Theorem 12.2, (12.17) becomes

$$\|\boldsymbol{e}_{n+1}\| \le \|\boldsymbol{e}_n\| + hL_\Phi\|\boldsymbol{e}_n\| + h\|\mathbf{T}_n\|\,, \qquad n = 0, 1, \ldots, N-1\,,$$

where $\|\cdot\|$ is any norm on $\mathbb{R}^m$, with obvious definitions of the global error $\boldsymbol{e}_n$ and the truncation error $\mathbf{T}_n$. Similarly, Picard's Theorem and its proof, discussed at the beginning of the chapter in the case of a single ordinary differential equation, can be easily extended to an m-component system of differential equations by replacing the absolute value sign with a vector norm on $\mathbb{R}^m$ throughout.

12.11 Stiff systems

The phenomenon of stiffness usually appears only in a system of differential equations, but we begin by discussing an almost trivial example of a single equation,

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

where λ is a constant. The solution of this equation is evidently $y(x) = y_0 \exp(\lambda x)$. When $\lambda < 0$ the absolute value of the solution is exponentially decreasing, so it is sensible to require that the absolute value of our numerical solution also decreases. It is very easy to give expressions for the result of a numerical solution using Euler's method and the implicit Euler method (12.36). They are, respectively,

$$y_n^{\mathrm{E}} = (1 + h\lambda)^n y_0\,, \qquad y_n^{\mathrm{I}} = (1 - h\lambda)^{-n} y_0\,.$$

When $\lambda < 0$ and $h > 0$, we have $(1 - h\lambda) > 1$; therefore, the sequence $(|y_n^{\mathrm{I}}|)$ decreases monotonically with increasing n. On the other hand, for $\lambda < 0$ and $h > 0$,

$$|1 + h\lambda| < 1 \qquad \text{if, and only if,} \qquad 0 < h|\lambda| < 2\,.$$

This gives the restriction $h|\lambda| < 2$ on the size of h for which the sequence $(|y_n^{\mathrm{E}}|)$ decreases monotonically; if h exceeds $2/|\lambda|$, the numerical solution obtained by Euler's method will oscillate with increasing magnitude with increasing n and fixed $h > 0$, instead of converging to zero as $n \to \infty$.

We now consider the same two methods applied to the initial value problem for a system of differential equations of the form

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

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

where A is a square matrix of order m, each of whose elements is a constant. For simplicity we assume that the eigenvalues of A are distinct, so there exists a matrix M such that $MAM^{-1} = \Lambda$ is a diagonal matrix. The system of differential equations is therefore equivalent to

$$\boldsymbol{z}' = \Lambda \boldsymbol{z}\,, \qquad \boldsymbol{z}(0) = \boldsymbol{z}_0 = M\boldsymbol{y}_0\,,$$

with $\boldsymbol{z} = M\boldsymbol{y}$. In this form the system reduces to a set of m independent equations, whose solutions are

$$z_j = z_j(0)\exp(\lambda_j x)\,, \qquad j = 1, 2, \ldots, m\,,$$

where the numbers λ_j, $j = 1, 2, \ldots, m$, are the diagonal elements of the matrix Λ, and are therefore the eigenvalues of A. In particular, if all the λ_j, $j = 1, 2, \ldots, m$, are real and negative, then $\lim_{x\to+\infty} \|\boldsymbol{z}(x)\| = 0$ and since

$$\|\boldsymbol{y}(x)\| = \|M^{-1}\,\boldsymbol{z}(x)\| \le \|M^{-1}\|\,\|\boldsymbol{z}(x)\|\,,$$

also

$$\lim_{x\to+\infty} \|\boldsymbol{y}(x)\| = 0\,.$$

Here $\|\cdot\|$ is any norm on $\mathbb{R}^m$, and the norm on M^{-1} is the associated subordinate matrix norm defined in Chapter 2.

In just the same way, Euler's method applied to the system gives

$$\boldsymbol{y}_{n+1} = (I + hA)\boldsymbol{y}_n\,,$$

which leads to

$$\begin{aligned} \boldsymbol{z}_{n+1} &= M\boldsymbol{y}_{n+1} = M(I + hA)\boldsymbol{y}_n \\ &= M(I + hA)M^{-1}\boldsymbol{z}_n = (I + h\Lambda)\boldsymbol{z}_n\,. \end{aligned}$$

Thus, the result $\boldsymbol{y}_{n+1}$ of Euler's method applied to the initial value problem $\boldsymbol{y}' = A\boldsymbol{y}$, $\boldsymbol{y}(0) = \boldsymbol{y}_0$, is exactly the same as $M^{-1}\boldsymbol{z}_{n+1}$, where $\boldsymbol{z}_{n+1}$ is the result of applying Euler's method to the transformed problem $\boldsymbol{z}' = \Lambda\boldsymbol{z}$, $\boldsymbol{z}(0) = M\boldsymbol{y}_0$; an analogous remark applies to the use of the implicit Euler method.

Suppose that all the eigenvalues λ_j, $j = 1, 2, \ldots, m$, are real and negative. Then, in order to ensure that, for a fixed positive value of h,

$$\lim_{n\to\infty} \|\boldsymbol{y}_n\| = 0\,,$$

we must require that, for Euler's method, $h|\lambda_j| < 2$, $j = 1, 2, \ldots, m$; for the implicit Euler method no such condition is required. The importance of this fact is highlighted by a numerical example.

We consider the system where A is the 2×2 matrix

$$A = \begin{pmatrix} -8003 & 1999 \\ 23988 & -6004 \end{pmatrix},$$

and the initial condition is

$$\boldsymbol{y}(0) = \begin{pmatrix} 1 \\ 4 \end{pmatrix}.$$

The eigenvalues of A are $\lambda_1 = -7$ and $\lambda_2 = -14000$; the solution of the problem is

$$\boldsymbol{y}(x) = \begin{pmatrix} \mathrm{e}^{-7x} \\ 4\mathrm{e}^{-7x} \end{pmatrix}.$$

Clearly, $\lim_{x \to +\infty} \|\boldsymbol{y}(x)\| = 0$.

The numerical solution uses 12 steps of size $h = 0.004$; the results are shown in Table 12.1. The second column gives the first component of the solution, $y_1(x) = \mathrm{e}^{-7x}$, the third column shows the result from the implicit Euler method, and the last gives the result of the standard Euler method. The last column is a dramatic example of what happens when the step size h is too large; in this case $h|\lambda_2| = 56$. The numerical values given by the implicit Euler method have an error of a few units in the third decimal digit; to get the same accuracy from the Euler method would require a step size about 30 times smaller, and about 30 times as much work.

It is clear that the difficulty in the numerical example is caused by the size of the eigenvalue -14000, but what is important is its size relative to the other eigenvalue. The special constant-coefficient system $y' = Ay$ is said to be **stiff** if all the eigenvalues of A have negative real parts, and if the ratio of the largest of the real parts to the smallest of the real parts is large. Most practical problems are nonlinear, and for such problems it is quite difficult to define precisely what is meant by stiffness.[1] To begin with we may replace the system by a linearised approximation, the first terms of an expansion

$$\boldsymbol{y}'(x) = \boldsymbol{y}'(x_n) + \frac{\partial \boldsymbol{f}}{\partial x}(x_n, y(x_n))(x - x_n) + J(x_n)(\boldsymbol{y}(x) - \boldsymbol{y}(x_n)) + \cdots$$

[1] Indeed, even in the case of variable-coefficient linear systems of differential equations, stiffness can be defined in several (nonequivalent) ways; for a discussion of the pros and cons of the various definitions, we refer to Section 6.2 of J.D. Lambert, *Numerical Methods for Ordinary Differential Systems*, Wiley, Chichester, 1991.

Table 12.1. *The use of Euler's method and the implicit Euler method to solve a stiff system.*

x	$y_1(x)$	Implicit Euler	Euler
0.000	1.000	1.000	1.000
0.004	0.972	0.973	0.972
0.008	0.946	0.946	0.945
0.012	0.919	0.920	0.918
0.016	0.894	0.895	0.893
0.020	0.869	0.871	0.868
0.024	0.845	0.847	0.843
0.028	0.822	0.824	0.820
0.032	0.799	0.802	0.794
0.036	0.777	0.780	0.941
0.040	0.756	0.759	−8.430
0.044	0.735	0.738	505.769
0.048	0.715	0.718	−27776.357

where J is the Jacobian matrix of the function $\boldsymbol{f}$, whose (i, j)-entry is

$$(J(x_n))_{ij} = \frac{\partial f_i}{\partial y_j}(x_n, \boldsymbol{y}(x_n)) .$$

We can then think of the system as being stiff if the eigenvalues of the matrix $J(x_n)$ have negative real parts and if the ratio of the largest of the real parts to the smallest is large. Although this gives some indication of the sort of problems which may cause difficulty, the behaviour of nonlinear systems is much more complicated than this. It is not difficult to construct examples in which all the eigenvalues of the Jacobian matrix have negative real parts, yet the norm of the solution of the differential equation is exponentially increasing as $x \to +\infty$.

Even though any classification of nonlinear systems of differential equations into stiff and nonstiff, based only on monitoring the eigenvalues of $J(x_n)$, is somewhat simplistic, it does highlight some of the key difficulties. Stiff systems of differential equations arise in many application areas, a typical one being chemical engineering. For example, in parts of an oil refinery there may be a large number of substances undergoing chemical reactions with widely different reaction rates. These reaction rates correspond to the eigenvalues of the Jacobian matrix, and it is not unusual to find the ratio of the largest of the real parts to the smallest to be in excess of 10^{10}. For such problems it is essential to find a numerical method which imposes no restriction on the step size;

Euler's method, which might require the restriction $10^{10}h < 2$, would evidently be quite useless.

Application of the linear multistep method

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

to the equation $y' = \lambda y$ leads to the kth-order linear recurrence relation

$$\sum_{j=0}^{k} (\alpha_j - \lambda h \beta_j) y_{n+j} = 0 \,. \tag{12.48}$$

The characteristic polynomial of the linear recurrence relation (12.48) is

$$\pi(z; \lambda h) = \sum_{j=0}^{k} (\alpha_j - \lambda h \beta_j) z^j \,.$$

Alternatively, we can write this in terms of the first and second characteristic polynomials of the linear multistep method as

$$\pi(z; \lambda h) = \rho(z) - \lambda h \sigma(z) \,.$$

In the present context, the polynomial $\pi(\,\cdot\,; \lambda h)$ is usually referred to as the **stability polynomial** of the linear multistep method. According to Lemma 12.1, the general solution of the recurrence relation (12.48) can be expressed in terms of the distinct roots z_r, $1 \le r \le \ell$, $\ell \le k$, of $\pi(\,\cdot\,; \lambda h)$. Letting m_r denote the multiplicity of the root z_r, $1 \le r \le \ell$, $m_1 + \cdots + m_\ell = k$, we have that

$$y_n = \sum_{r=1}^{\ell} p_r(n) z_r^n \,, \tag{12.49}$$

where the polynomial $p_r(\,\cdot\,)$ has degree $m_r - 1$, $1 \le r \le \ell$.

Clearly, the roots z_r are functions of λh. For $\lambda \in \mathbb{C}$, with $\mathrm{Re}(\lambda) < 0$, the solution of the model problem

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

converges in $\mathbb{C}$ to 0 as $x \to \infty$. Thus, we would like to ensure that, when a linear multistep method is applied to this problem, the step size h can be chosen so that the resulting sequence of numerical approximations (y_n) exhibits an analogous behaviour as $n \to \infty$, that is, $\lim_{n\to\infty} y_n = 0$. By virtue of (12.49), this can be guaranteed by demanding that each root $z_r = z_r(\lambda h)$ has modulus less then 1.

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

Definition 12.6 *A linear multistep method is said to be* **absolutely stable** *for a given value of λh if each root $z_r = z_r(\lambda h)$ of the associated stability polynomial $\pi(\,\cdot\,;\lambda h)$ satisfies $|z_r(\lambda h)| < 1$.*

Our aim is, therefore, to single out those values of λh for which the linear multistep method is absolutely stable.

Definition 12.7 *The* **region of absolute stability** *of a linear multistep method is the set of all points λh in the complex plane for which the method is absolutely stable.*

Ideally, the region of absolute stability of the method should admit all values of λ, $\mathrm{Re}(\lambda) < 0$, so as to ensure that there is no limitation on the size of h, however large $|\lambda|$ may be. This leads us to the next definition.

Definition 12.8 *A linear multistep method is said to be* **A-stable** *if its region of absolute stability contains the negative (left) complex half-plane.*

Unfortunately, the condition of A-stability is extremely demanding. Dahlquist[1] has shown the following results which are collectively known as his **Second Barrier Theorem**:

(i) No *explicit* linear multistep method is A-stable;
(ii) No A-stable linear multistep method can have order greater than 2.
(iii) The second-order A-stable linear multistep method with the smallest error constant is the trapezium rule method.

The trapezium rule method is a one-step method, so the associated stability polynomial has only one root, given by

$$z = \frac{1 + \frac{1}{2}\lambda h}{1 - \frac{1}{2}\lambda h}\,.$$

Evidently $|z| < 1$ if $\mathrm{Re}(h\lambda) = h\,\mathrm{Re}(\lambda) < 0$, so the trapezium rule method is indeed A-stable.

To construct useful methods of higher order we need to relax the condition of A-stability by requiring that the region of absolute stability should include a large part of the negative half-plane, and certainly that it contains the whole of the negative real axis.

[1] G. Dahlquist, A special stability problem for linear multistep methods, *BIT* **3**, 27–43, 1963.

The most efficient methods of this kind in current use are the **Backward Differentiation Formulae**, or BDF methods. These are the linear multistep methods (12.35) in which $\beta_j = 0$, $0 \le j \le k-1$, $k \ge 1$, and $\beta_k \ne 0$. Thus,

$$\alpha_k y_{n+k} + \cdots + \alpha_0 y_n = h\beta_k f_{n+k}\,.$$

The coefficients are obtained by requiring that the order of accuracy of the method is as high as possible, *i.e.*, by making the coefficients C_j zero in (12.47) for $j = 0, 1, \ldots, k$. For $k = 1$ this yields the implicit Euler method (BDF1), whose order of accuracy is, of course, 1; the method is A-stable. The choice of $k = 6$ results in the sixth-order, six-step BDF method (BDF6):

$$147y_{n+6} - 360y_{n+5} + 450y_{n+4} - 400y_{n+3} + 225y_{n+2} - 72y_{n+1} + 10y_n = 60hf_{n+6}\,. \qquad (12.50)$$

Although the method (12.50) is not A-stable, its region of absolute stability includes the whole of the negative real axis (see Figure 12.5). For the intermediate values, $k = 2, 3, 4, 5$, we have the following kth-order, k-step BDF methods, respectively:

$$\begin{aligned} 3y_{n+2} - 4y_{n+1} + y_n &= 2hf_{n+2}\,,\\ 11y_{n+3} - 18y_{n+2} + 9y_{n+1} - 2y_n &= 6hf_{n+3}\,,\\ 25y_{n+4} - 48y_{n+3} + 36y_{n+2} - 16y_{n+1} + 3y_n &= 12hf_{n+4}\,,\\ 137y_{n+5} - 300y_{n+4} + 300y_{n+3} - 200y_{n+2} + 75y_{n+1} - 12y_n &= 60hf_{n+5}\,, \end{aligned}$$

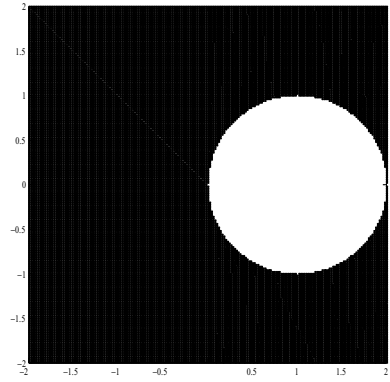
referred to as BDF2, BDF3, BDF4 and BDF5. Their regions of absolute stability are also shown in Figure 12.5. In each case the region of absolute stability includes the negative real axis. Higher-order methods of this type cannot be used, as all BDF methods, with $k > 6$, are zero-unstable.

12.12 Implicit Runge–Kutta methods

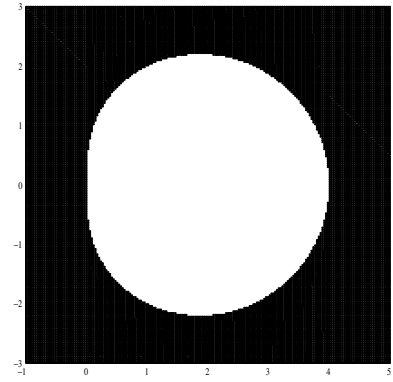
For Runge–Kutta methods absolute stability is defined in much the same way as for linear multistep methods; *i.e.*, by applying the method in question to the model problem $y' = \lambda y$, $y(0) = y_0$, $\lambda \in \mathbb{C}$, $\mathrm{Re}(\lambda) < 0$, and demanding that the resulting sequence (y_n) converges to 0 as $n \to \infty$, with $h\lambda$ held fixed. The set of all values of $h\lambda$ in the complex plane for which the method is absolutely stable is called the region of absolute stability of the Runge–Kutta method.

Classical Runge–Kutta methods are explicit, and are unsuitable for

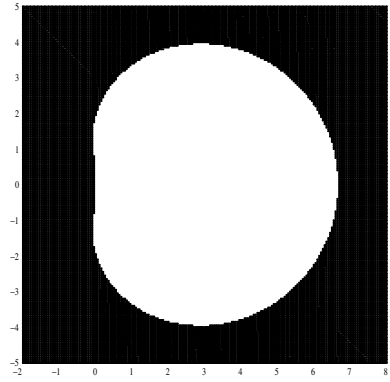
https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press



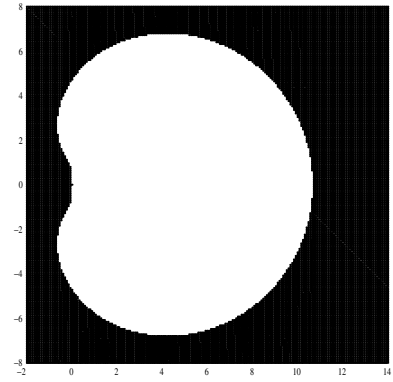
(a) BDF1



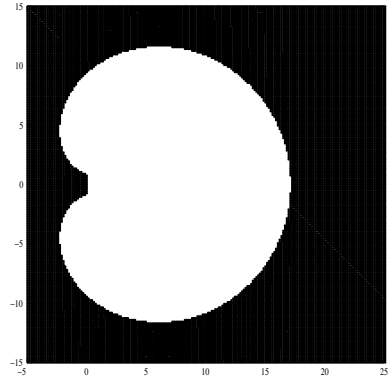
(b) BDF2



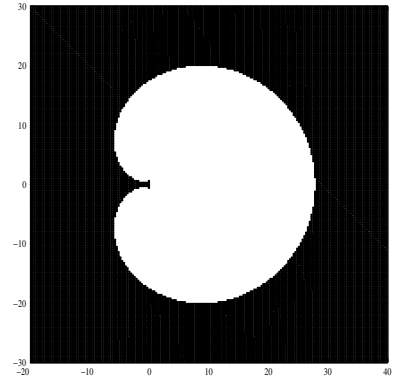
(c) BDF3



(d) BDF4



(e) BDF5



(f) BDF6

Fig. 12.5. Absolute stability regions in the complex plane for k-step Backward Differentiation Formulae, $k = 1, 2, \ldots, 6$. In each case the region of absolute stability is the set of points in the complex plane outside the white region. In each case, the region of absolute stability contains the whole of the negative real axis.

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

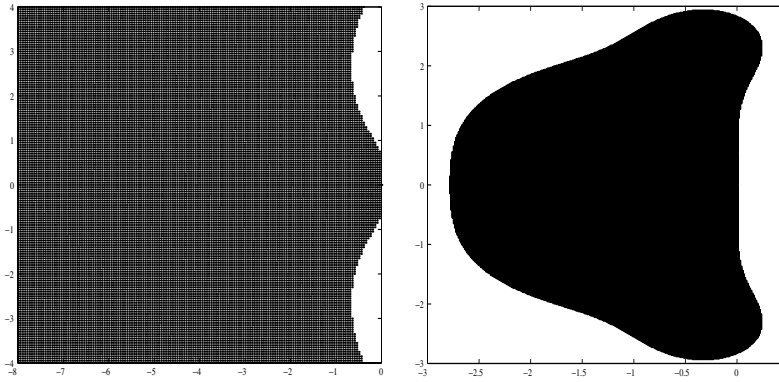


Fig. 12.6. The dark chequered region in the figure on the left indicates part of the absolute stability region in the complex plane for the four-step, fourth-order Backward Differentiation Formula, BDF4 (zoom into Figure 12.5(d)); here we only show the section of the region of absolute stability for BDF4 which lies in the rectangle $-8 < \mathrm{Re}(\lambda h) < 0$ and $-4 < \mathrm{Im}(\lambda h) < 4$, with $\mathrm{Re}(\lambda) < 0$, $h > 0$. The dark region in the figure on the right shows the region of absolute stability for the classical explicit fourth-order Runge–Kutta method, RK4. For BDF4, the region of absolute stability includes the whole of the negative real axis; clearly, this is not the case for RK4.

stiff systems because of their small region of absolute stability. Figure 12.6 depicts the region of absolute stability of the classical fourth-order Runge–Kutta method, together with that of the fourth-order Backward Differentiation Formula, BDF4. The contrast is striking: while the region of absolute stability of BDF4 includes most of the negative half-plane and, in particular, all of the negative real axis, for RK4 the region of absolute stability is bounded[1] (for example, along the negative real axis it does not extend to the left of, approximately, -2.8).

Motivated by the fact that BDF methods are implicit, we now go on to introduce *implicit* Runge–Kutta methods, which can also have a large region of absolute stability.

The general s-stage Runge–Kutta method is written

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i \,,$$

[1] This is not a peculiarity of RK4. It can be shown that every explicit Runge–Kutta method has bounded region of absolute stability; see, for example, Section 5.12, in J.D. Lambert's book, cited in the previous section.

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

where

$$k_i = f(x_n + hc_i, y_n + h\sum_{j=1}^{s} a_{ij}k_j)\,, \qquad 1 \le i \le s\,. \tag{12.51}$$

It is convenient to display the coefficients in a **Butcher tableau**

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \dots & \dots & \dots & \dots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \dots & b_s \end{array}$$

The method is then defined by the matrix $A = (a_{ij}) \in \mathbb{R}^{s\times s}$, of order s, and the two vectors $\boldsymbol{b} = (b_1, \dots, b_s)^{\mathrm{T}} \in \mathbb{R}^s$ and $\boldsymbol{c} = (c_1, \dots, c_s)^{\mathrm{T}} \in \mathbb{R}^s$. For example, the classical four-stage Runge–Kutta method is defined by the tableau

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{2}{3} & \frac{1}{6} \end{array}$$

The 4×4 array representing the matrix A for this method, displayed in the upper right quadrant of the tableau, follows the usual notational convention that zero elements after the last nonzero element in each row of the matrix A are omitted.

This is an explicit method, shown by the fact that the matrix A is *strictly lower triangular*, with $a_{ij} = 0$ when $1 \le i \le j \le 4$. Each value k_i can therefore be calculated in sequence, all the quantities on the right-hand side of (12.51) being known.

It is not difficult to construct s-stage implicit methods which are A-stable. For example, this can be done by choosing the coefficients c_i and b_i to be the quadrature points and weights respectively in the Gauss quadrature formula for the evaluation of

$$\int_0^1 g(x)\mathrm{d}x \approx \sum_{i=1}^{s} b_i g(c_i)\,.$$

The numbers a_{ij} can then be chosen so that the method has order $2s$, and is A-stable.

For example, the array

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

$$\begin{array}{c|cc} \frac{1}{6}(3-\sqrt{3}) & \frac{1}{4} & \frac{1}{12}(3-2\sqrt{3}) \\ \frac{1}{6}(3+\sqrt{3}) & \frac{1}{12}(3+2\sqrt{3}) & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

defines a 2-stage A-stable method of order 4.

However, there is a heavy price to pay for using implicit methods of this kind, as we now have to calculate all the numbers k_i, $i = 1, 2, \ldots, s$, simultaneously, not in succession. For a system of m differential equations an implicit linear multistep method requires the solution of m simultaneous equations at each step; an s-stage implicit Runge–Kutta method requires the solution of sm simultaneous equations. This is a considerable increase in cost, and the general implicit Runge–Kutta methods cannot compete in efficiency with the Backward Differentiation Formulae such as (12.50); their use is almost exclusively limited to stiff systems of ODEs.

The overall computational effort can be somewhat reduced by using **diagonally implicit Runge–Kutta** (or DIRK) methods, in which the matrix A is lower triangular, so that $a_{ij} = 0$ if $j > i$. A further improvement in efficiency is possible by requiring in addition that all the diagonal elements a_{ii} are the same; unfortunately it has proved difficult to construct such methods with order greater than 4.

12.13 Notes

In this chapter we have only been able to introduce some of the basic ideas in what has become a vast area of numerical analysis. In particular we have not discussed the practical implementation of the various methods. The questions of how to choose the step size h to obtain efficiently a prescribed accuracy, and when and how to adjust h during the course of the calculation, are dealt with in the following books.

- E. Hairer, S.P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems,* Second Edition, Springer Series in Computational Mathematics, 8, Springer, Berlin, 1993.
- A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, Cambridge, 1996.
- J.D. Lambert, *Numerical Methods for Ordinary Differential Systems,* John Wiley & Sons, Chichester, 1991.

For a study of dynamical systems and their numerical analysis, with focus on long-time behaviour, we refer to

- A.M. Stuart and A.R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, 1999.

The numerical solution of stiff initial value problems for systems of ordinary differential equations is discussed in

- E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, Springer Series in Computational Mathematics, 14, Springer, Berlin, 1991.

An extensive survey of the theory of Runge–Kutta and linear multistep methods is found in

- J.C. Butcher, *The Numerical Analysis of Ordinary Differential Equations. Runge–Kutta and General Linear Methods*, Wiley-Interscience, John Wiley & Sons, Chichester, 1987.

Satisfactory theoretical treatment of nonlinear systems of differential equations from the point of view of stiffness requires the development of a genuinely nonlinear stability theory which does not involve the rather dubious idea of defining stiffness through linearisation based on the 'frozen Jacobian matrix'. We close by mentioning just one concept in this direction – that of *algebraic stability*. Given a Runge–Kutta method with Butcher tableau

$$\begin{array}{c|c} \boldsymbol{c} & A \\ \hline & \boldsymbol{b}^T \end{array}$$

we define the matrices

$$B = \operatorname{diag}(b_1, b_2, \ldots, b_s) \quad \text{and} \quad M = BA + A^{\mathrm{T}}B - bb^{\mathrm{T}}\,.$$

The method is said to be **algebraically stable** if the matrices B and M are both positive semidefinite, *i.e.*, $\boldsymbol{x}^{\mathrm{T}}B\boldsymbol{x} \geq 0$ and $\boldsymbol{x}^{\mathrm{T}}M\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^s$. Algebraic stability can be seen to ensure that approximations to solutions of nonlinear systems of differential equations exhibit acceptable numerical behaviour. For example, the Gauss–Runge–Kutta methods discussed in the last section are algebraically stable. For further details, see, for example,

- K. Dekker and J.G. Verver, *Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations,* North-Holland, Amsterdam, 1984.

Exercises

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

(a) $f(x,y) = 2yx^{-4}\,, \qquad x \in [1,\infty)\,;$
(b) $f(x,y) = \mathrm{e}^{-x^2}\tan^{-1} y\,, \qquad x \in [1,\infty)\,;$
(c) $f(x,y) = 2y(1+y^2)^{-1}(1+\mathrm{e}^{-|x|})\,, \qquad x \in (-\infty,\infty)\,.$

12.2 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?

12.3 Write down the solution y of the initial value problem

$$y' = py + q\,, \quad y(0) = 1\,,$$

where p and q are constants. Suppose that the method in the proof of Picard's Theorem is used to generate the sequence of approximations $y_n(x)$, $n = 0, 1, 2, \ldots$; show that $y_n(x)$ is a polynomial of degree n, and consists of the first $n+1$ terms in the series expansion of $y(x)$ in powers of x.

12.4 Show that Euler's 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 approximating the same problem with the implicit Euler method. Show that there is a solution of the form $y_n = (C_n h)^{5/4}$, $n \geq 0$, with $C_0 = 0$ and $C_1 = 1$ and $C_n > 1$ for all $n \geq 2$.

12.5 Write down Euler's method for the solution of the problem

$$y' = x\mathrm{e}^{-5x} - 5y\,, \qquad y(0) = 0$$

on the interval $[0,1]$ with step size $h = 1/N$. Denoting by y_N the resulting approximation to $y(1)$, show that $y_N \to y(1)$ as $N \to \infty$.

12.6 Consider the initial value problem

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

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

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

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

using the mesh points $x_n = nh$, $n = 0, 1, \ldots, N$, with spacing $h = 1/N$.

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

(ii) Verify that

$$|y(x_{n+1}) - y_{n+1}| \leq (1 + hL)|y(x_n) - y_n| + h|T_n|$$

for $n = 0, 1, \ldots, N-1$, where $L = 1/(2\ln 4)$.

(iii) 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 $N \geq N_0$.

12.7 Define the truncation error T_n of the trapezium rule method

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

for the numerical solution of $y' = f(x, y)$ with $y(0) = y_0$ 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

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

for all real x, u, v, where L is a positive constant independent

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

of x, and that $|y'''(x)| \leq M$ for some positive constant M independent of x. Show that the global error $e_n = y(x_n) - y_n$ satisfies the inequality

$$|e_{n+1}| \leq |e_n| + \frac{1}{2}hL\,(|e_{n+1}| + |e_n|) + \frac{1}{12}h^3M\,.$$

For a constant step size $h > 0$ satisfying $hL < 2$, deduce that, if $y_0 = y(x_0)$, then

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

12.8 Show that the one-step method defined by

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

where

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

is consistent and has truncation error

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

12.9 When the classical fourth-order Runge–Kutta method is applied to the differential equation $y' = \lambda y$, where λ is a constant, show that

$$y_{n+1} = (1 + h\lambda + \tfrac{1}{2}h^2\lambda^2 + \tfrac{1}{6}h^3\lambda^3 + \tfrac{1}{24}h^4\lambda^4)y_n\,.$$

Compare this with the Taylor series expansion of $y(x_{n+1}) = y(x_n + h)$ about the point $x = x_n$.

12.10 Consider the one-step method

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

where α, β and γ are real parameters and $h > 0$. 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,

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

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

12.11 Find the values of α and β so that the three-step method

$$y_{n+3} + \alpha(y_{n+2} - y_{n+1}) - y_n = h\beta(f_{n+2} + f_{n+1})$$

has order of accuracy 4, and show that the resulting method is *not* zero-stable.

12.12 Consider approximating the initial value problem $y' = f(x, y)$, $y(0) = y_0$ by the linear multistep method

$$y_{n+1} + by_{n-1} + ay_{n-2} = hf(x_n, y_n)$$

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

(i) 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.

(ii) Although the method is consistent for the choice of a and b from part (i), 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.

12.13 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 $\{x_n\colon\ x_n = x_0 + nh, n = 1, 2, \ldots, N\}$ 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?

12.14 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?

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

For the methods under (a) and (c) explore absolute stability when applied to the differential equation $y' = \lambda y$ with $\lambda < 0$.

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

12.15 Determine the order of the linear multistep method

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

and investigate its zero-stability and absolute stability.

12.16 Assuming that $\sigma(z) = z^2$ is the second characteristic polynomial of a linear two-step method, find a quadratic polynomial $\rho(z)$ such that the order of the method is 2. Is this method convergent? By applying the method to $y' = \lambda y$, $y(0) = 1$, where λ is a negative real number, show that the method is absolutely stable for all $h > 0$.

12.17 Consider the θ-method

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

for $\theta \in [0,1]$. Show that the method is A-stable if, and only if, $\theta \geq 1/2$.

12.18 Write down an expression for the Lagrange interpolation polynomial of degree 2 for a function $x \mapsto y(x)$, using the interpolation points x_n, $x_{n+1} = x_n + h$ and $x_{n+2} = x_n + 2h$, $h > 0$. Differentiate this polynomial to show that

$$y'(x_{n+2}) = \frac{1}{2h}\left(3y(x_{n+2}) - 4y(x_{n+1}) + y(x_n)\right) + \mathcal{O}(h^2)\,,$$

provided that $y \in \mathrm{C}^3[x_n, x_{n+2}]$. Confirm this result by determining the truncation error of the BDF2 method

$$3y_{n+2} - 4y_{n+1} + y_n = 2hf_{n+2}\,.$$

12.19 When the general two-stage implicit Runge–Kutta method is applied to the single constant-coefficient differential equation $y' = \lambda y$, show that

$$\begin{aligned} k_1 &= [1 + \lambda h(a_{12} - a_{22})]\lambda y_n/\Delta\,, \\ k_2 &= [1 + \lambda h(a_{21} - a_{11})]\lambda y_n/\Delta\,, \end{aligned}$$

where Δ is the determinant of the matrix $I - \lambda hA$ with

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

For the method defined by the Butcher tableau

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press

$\frac{1}{6}(3-\sqrt{3})$	$\frac{1}{4}$	$\frac{1}{12}(3-2\sqrt{3})$
$\frac{1}{6}(3+\sqrt{3})$	$\frac{1}{12}(3+2\sqrt{3})$	$\frac{1}{4}$
	$\frac{1}{2}$	$\frac{1}{2}$

deduce that $y_{n+1} = R(\lambda h)y_n$, where

$$R(\lambda h) = \frac{1 + \frac{1}{2}\lambda h + \frac{1}{12}\lambda^2 h^2}{1 - \frac{1}{2}\lambda h + \frac{1}{12}\lambda^2 h^2}\,.$$

By writing $R(z)$ in the factorised form $(z+p)(z+q)/(z-p)(z-q)$, deduce that this Runge–Kutta method is A-stable.

https://doi.org/10.1017/CBO9780511801181.013 Published online by Cambridge University Press