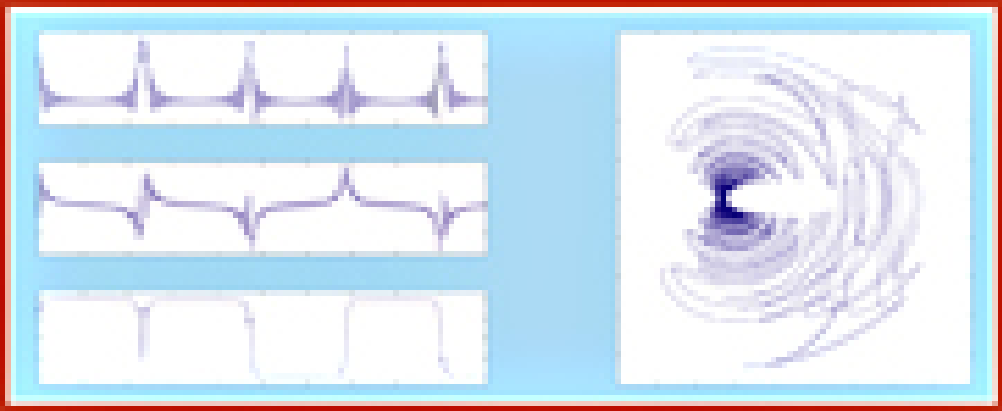
CAMBRIDGE TEXTS IN APPLIED MATHEMATICS

A First Course in the Numerical Analysis of Differential Equations



SECOND EDITION

ARIEH ISERLES

CAMBRIDGE www.cambridge.org/9780521734905

4

Stiff equations

4.1 What are stiff ODEs?

Let us try to solve the seemingly innocent linear ODE

$$\boldsymbol{y}' = \Lambda \boldsymbol{y}, \qquad \boldsymbol{y}(0) = \boldsymbol{y}_0, \qquad \text{where} \qquad \Lambda = \begin{bmatrix} -100 & 1 \\ 0 & -\frac{1}{10} \end{bmatrix}, \tag{4.1}$$

by Euler's method (1.4). We obtain

$$\boldsymbol{y}_1 = \boldsymbol{y}_0 + h\Lambda \boldsymbol{y}_0 = (I + h\Lambda)\boldsymbol{y}_0, \qquad \boldsymbol{y}_2 = \boldsymbol{y}_1 + h\Lambda \boldsymbol{y}_1 = (I + h\Lambda)\boldsymbol{y}_1 = (I + h\Lambda)^2 \boldsymbol{y}_0$$

(where I is the identity matrix) and, in general, it is easy to prove by elementary induction that

$$\boldsymbol{y}_n = (I + h\Lambda)^n \boldsymbol{y}_0, \qquad n = 0, 1, 2, \ldots \tag{4.2}$$

Since the *spectral factorization* (A.1.5.4) of Λ is

$$\Lambda = VDV^{-1}, \qquad \text{where} \qquad V = \begin{bmatrix} 1 & 1 \\ 0 & \frac{999}{10} \end{bmatrix} \qquad \text{and} \qquad D = \begin{bmatrix} -100 & 0 \\ 0 & -\frac{1}{10} \end{bmatrix},$$

we deduce that the exact solution of (4.1) is

$$\boldsymbol{y}(t) = \mathrm{e}^{t\Lambda} = V\mathrm{e}^{tD}V^{-1}\boldsymbol{y}_0, \qquad t \geq 0, \qquad \text{where} \qquad \mathrm{e}^{tD} = \begin{bmatrix} \mathrm{e}^{-100t} & 0 \\ 0 & \mathrm{e}^{-t/10} \end{bmatrix}.$$

In other words, there exist two vectors, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, say, dependent on $\boldsymbol{y}_0$ but not on t, such that

$$\boldsymbol{y}(t) = \mathrm{e}^{-100t}\boldsymbol{x}_1 + \mathrm{e}^{-t/10}\boldsymbol{x}_2, \qquad t \geq 0. \tag{4.3}$$

The function $g(t) = \mathrm{e}^{-100t}$ decays exceedingly fast: $g\left(\frac{1}{10}\right) \approx 4.54 \times 10^{-5}$ and $g(1) \approx 3.72 \times 10^{-44}$, while the decay of $\mathrm{e}^{-t/10}$ is a thousandfold more sedate. Thus, even for small $t > 0$ the contribution of $\boldsymbol{x}_1$ is nil to all intents and purposes and $\boldsymbol{y}(t) \approx \mathrm{e}^{-t/10}\boldsymbol{x}_2$. What about the Euler solution $\{\boldsymbol{y}_n\}_{n=0}^{\infty}$, though? It follows from (4.2) that

$$\boldsymbol{y}_n = V(I + hD)^n V^{-1}\boldsymbol{y}_0, \qquad n = 0, 1, \ldots$$

and, since

$$(I + hD)^n = \begin{bmatrix} (1 - 100h)^n & 0 \\ 0 & (1 - \frac{1}{10}h)^n \end{bmatrix},$$

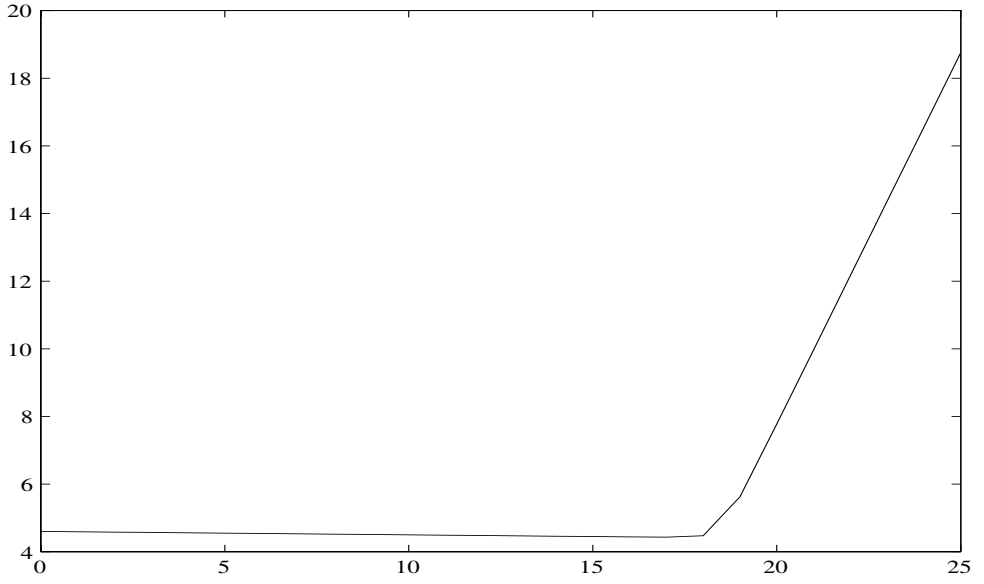


Figure 4.1 The logarithm of the Euclidean norm $\|\boldsymbol{y}_n\|$ of the Euler steps, as applied to the equation (4.1) with $h = \frac{1}{10}$ and an initial condition identical to the second (i.e., the 'stable') eigenvector. The divergence is thus entirely due to roundoff error!

it follows that

$$\boldsymbol{y}_n = (1 - 100h)^n \boldsymbol{x}_1 + (1 - \tfrac{1}{10}h)^n \boldsymbol{x}_2, \qquad n = 0, 1, \ldots \tag{4.4}$$

(it is left to the reader to prove in Exercise 4.1 that the constant vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are the same in (4.3) and (4.4)). Suppose that $h > \frac{1}{50}$. Then $|1 - 100h| > 1$ and it is a consequence of (4.4) that, for sufficiently large n, *the Euler iterates grow geometrically in magnitude,* in contrast with the asymptotic behaviour of the true solution.

Suppose that we choose an initial condition identical to an eigenvector corresponding to the eigenvalue -0.1, for example

$$\boldsymbol{y}_0 = \begin{bmatrix} 1 \\ \frac{999}{10} \end{bmatrix}.$$

Then, in exact arithmetic, $\boldsymbol{x}_1 = \boldsymbol{0}$, $\boldsymbol{x}_2 = \boldsymbol{y}_0$ and $\boldsymbol{y}_n = \left(1 - \frac{1}{10}h\right)^n \boldsymbol{y}_0$, $n = 0, 1, \ldots$; the latter converges to $\boldsymbol{0}$ as $n \to \infty$ for all reasonable values of $h > 0$ (specifically, for $h < 20$). Hence, we might hope that all will be well with the Euler method. Not so! Real computers produce roundoff errors and, unless $h < \frac{1}{50}$, sooner or later these are bound to attribute a nonzero contribution to an eigenvector corresponding to the eigenvalue -100. As soon as this occurs, the unstable component grows geometrically, as $(1 - 100h)^n$, and rapidly overwhelms the true solution.

Figure 4.1 displays $\ln \|\boldsymbol{y}_n\|$, $n = 0, 1, \ldots, 25$, with the above initial condition and the time step $h = \frac{1}{10}$. The calculation was performed on a computer equipped with

the ubiquitous IEEE arithmetic,[1] which is correct (in a single algebraic operation) to about 15 decimal digits. The norm of the first 17 steps decreases at the right pace, dictated by $\left(1-\frac{1}{10}h\right)^n = \left(\frac{99}{100}\right)^n$. However, everything then breaks down and, after just two steps, the norm increases geometrically, as $|1-100h|^n = 9^n$. The reader is welcome to check that the slope of the curve in Fig. 4.1 is indeed $\ln\frac{99}{100} \approx -0.0101$ initially but becomes $\ln 9 \approx 2.1972$ in the second, unstable, regime.

The choice of $\boldsymbol{y}_0$ as a 'stable' eigenvector is not contrived. Faced with an equation like (4.1) (with an arbitrary initial condition) we are likely to employ a small step size in the initial *transient* regime, in which the contribution of the 'unstable' eigenvector is still significant. However, as soon as this has disappeared and the solution is completely described by the 'stable' eigenvector, it is tempting to increase h. This must be resisted: like a malign version of the Cheshire cat, the rogue eigenvector might seem to have disappeared, but its hideous grin stays and is bound to thwart our endeavours.

It is important to understand that this behaviour has nothing to do with the local error of the numerical method; the step size is depressed not by accuracy considerations (to which we should be always willing to pay heed) but by instability.

Not every numerical method displays a similar breakdown in stability. Thus, solving (4.1) with the trapezoidal rule (1.9), we obtain

$$\boldsymbol{y}_1 = \left(\frac{I+\frac{1}{2}h\Lambda}{I-\frac{1}{2}h\Lambda}\right)\boldsymbol{y}_0, \qquad \boldsymbol{y}_2 = \left(\frac{I+\frac{1}{2}h\Lambda}{I-\frac{1}{2}h\Lambda}\right)\boldsymbol{y}_1 = \left(\frac{I+\frac{1}{2}h\Lambda}{I-\frac{1}{2}h\Lambda}\right)^2\boldsymbol{y}_0,$$

noting that since $(I-\frac{1}{2}h\Lambda)^{-1}$ and $\left(I+\frac{1}{2}h\Lambda\right)$ commute the order of multiplication does not matter, and, in general,

$$\boldsymbol{y}_n = \left(\frac{I+\frac{1}{2}h\Lambda}{I-\frac{1}{2}h\Lambda}\right)^n\boldsymbol{y}_0, \qquad n=0,1,\ldots \tag{4.5}$$

Substituting for Λ from (4.1) and factorizing, we deduce, in the same way as for (4.4), that

$$\boldsymbol{y}_n = \left(\frac{1-50h}{1+50h}\right)^n\boldsymbol{x}_1 + \left(\frac{1-\frac{1}{20}h}{1+\frac{1}{20}h}\right)^n\boldsymbol{x}_2, \qquad n=0,1,\ldots$$

Thus, since

$$\left|\frac{1-50h}{1+50h}\right|, \left|\frac{1-\frac{1}{20}h}{1+\frac{1}{20}h}\right| < 1$$

for every $h>0$, we deduce that $\lim_{n\to\infty}\boldsymbol{y}_n = \boldsymbol{0}$. This recovers the correct asymptotic behaviour of the ODE (4.1) (cf. (4.3)) regardless of the size of h.

In other words, the trapezoidal rule does not require any restriction in the step size to avoid instability. We hasten to say that this does not mean, of course, that *any* h is suitable. It is necessary to choose $h>0$ small enough to ensure that the local error is within reasonable bounds and the exact solution is adequately approximated. However, there is no need to decrease h to a minuscule size to prevent rogue components of the solution growing out of control.

[1] The current standard of computer arithmetic on workstations and personal computers.

The equation (4.1) is an example of a *stiff ODE*. Several attempts at a rigorous definition of stiffness appear in the literature, but it is perhaps more informative to adopt an operative (and slightly vague) designation. Thus, we say that an ODE system

$$\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}), \quad t \geq t_0, \qquad \boldsymbol{y}(t_0) = \boldsymbol{y}_0, \tag{4.6}$$

is *stiff* if its numerical solution by some methods requires (perhaps in a portion of the solution interval) a significant depression of the step size to avoid instability. Needless to say this is not a proper mathematical definition, but then we are not aiming to prove theorems of the sort 'if a system is stiff then ...'. The main importance of the above concept is in helping us to choose and implement numerical methods – a procedure that, anyway, is far from an exact science!

We have already seen the most important mechanism generating stiffness, namely, that modes with vastly different scales and 'lifetimes' are present in the solution. It is sometimes the practice to designate the quotient of the largest and the smallest (in modulus) eigenvalues of a linear system (and, for a general system (4.6), the eigenvalues of the Jacobian matrix) as the *stiffness ratio*. The stiffness ratio of (4.1) is 10^3. This concept is helpful in elucidating the behaviour of many ODE systems and, in general, it is a safe bet that if (4.6) has a large stiffness ratio then it is stiff. Having said this, it is also valuable to stress the shortcomings of linear analysis and emphasize that the stiffness ratio might fail to elucidate the behaviour of a nonlinear ODE system.

A large proportion of the ODEs that occur in practice are stiff. Whenever equations model several processes with vastly different rates of evolution, stiffness is not far away. For example, the differential equations of chemical kinetics describe reactions that often proceed on very different time scales (think of the difference in time scales of corrosion and explosion); a stiffness ratio of 10^{17} is quite typical. Other popular sources of stiffness are control theory, reactor kinetics, weather prediction, mathematical biology and electronics: they all abound with phenomena that display variation at significantly different time scales. The world record, to the author's knowledge, is held, unsurprisingly perhaps, by the equations that describe the cosmological Big Bang: the stiffness ratio is 10^{31}.

One of the main sources of stiff equations is numerical analysis itself. As we will see in Chapter 16, parabolic partial differential equations are often approximated by large systems of stiff ODEs.

4.2 The linear stability domain and A-stability

Let us suppose that a given numerical method is applied with a constant step size $h > 0$ to the scalar linear equation

$$y' = \lambda y, \quad t \geq 0, \qquad y(0) = 1, \tag{4.7}$$

where $\lambda \in \mathbb{C}$. The exact solution of (4.7) is, of course, $y(t) = \mathrm{e}^{\lambda t}$, hence $\lim_{t\to\infty} y(t) = 0$ if and only if $\operatorname{Re}\lambda < 0$. We say that the *linear stability domain* $\mathcal{D}$ of the underlying numerical method is the set of all numbers $h\lambda \in \mathbb{C}$ such that $\lim_{n\to\infty} y_n = 0$. In other

words, $\mathcal{D}$ is the set of all $h\lambda$ for which the correct asymptotic behaviour of (4.7) is recovered, provided that the latter equation is stable.[2]

Let us commence with Euler's method (1.4). We obtain the solution sequence identically to the derivation of (4.2),

$$y_n = (1 + h\lambda)^n, \qquad n = 0, 1, \ldots \tag{4.8}$$

Therefore $\{y_n\}_{n=0,1,\ldots}$ is a geometric sequence and $\lim_{n\to\infty} y_n = 0$ if and only if $|1 + h\lambda| < 1$. We thus conclude that

$$\mathcal{D}_{\text{Euler}} = \{z \in \mathbb{C} : |1 + z| < 1\}$$

is the interior of a complex disc of unit radius, centred at $z = -1$ (see Fig. 4.2).

Before we proceed any further, let us ponder briefly the rationale behind this sudden interest in a humble scalar linear equation. After all, we do not need numerical analysis to solve (4.7)! However, for Euler's method and for all other methods that have been the theme of Chapters 1–3 we can extrapolate from scalar linear equations to linear ODE systems. Thus, suppose that we solve (4.1) *with an arbitrary* $d \times d$ *matrix* Λ. The solution sequence is given by (4.2). Suppose that Λ has a full set of eigenvectors and hence the spectral factorization $\Lambda = VDV^{-1}$, where V is a nonsingular matrix of eigenvectors and $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ contains the eigenvalues of Λ. Exactly as in (4.4), we can prove that there exist vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d \in \mathbb{C}^d$, dependent only on $\boldsymbol{y}_0$, not on n, such that

$$\boldsymbol{y}_n = \sum_{k=1}^{d} (1 + h\lambda_k)^n \boldsymbol{x}_k, \qquad n = 0, 1, \ldots \tag{4.9}$$

Let us suppose that the exact solution of the linear system is asymptotically stable. This happens if and only if $\operatorname{Re} \lambda_k < 0$ for all $k = 1, 2, \ldots, d$. To mimic this behaviour with Euler's method, we deduce from (4.9) that the step size $h > 0$ must be such that $|1 + h\lambda_k| < 1$, $k = 1, 2, \ldots, d$: all the products $h\lambda_1, h\lambda_2, \ldots, h\lambda_d$ must lie in $\mathcal{D}_{\text{Euler}}$. This means in practice that the step size is determined by the stiffest component of the system!

The restriction to systems with a full set of eigenvectors is made for ease of exposition only. In general, we may use a *Jordan factorization* (A.1.5.6) in place of a spectral factorization; see Exercise 4.2 for a simple example. Moreover, the analysis can be extended easily to inhomogeneous systems $\boldsymbol{y}' = \Lambda\boldsymbol{y} + \boldsymbol{a}$, and this is illustrated by Exercise 4.3.

The importance of $\mathcal{D}$ ranges well beyond linear systems. Given a nonlinear ODE system

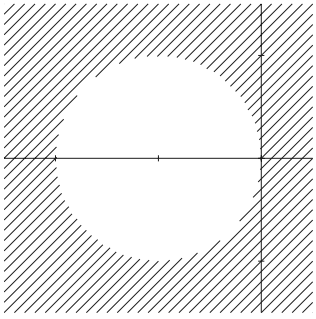
$$\boldsymbol{y}' = \boldsymbol{f}(t, \boldsymbol{y}), \quad t \geq t_0, \qquad \boldsymbol{y}(t_0) = \boldsymbol{y}_0,$$

where $\boldsymbol{f}$ is differentiable with respect to $\boldsymbol{y}$, it is usual to require that in the nth step

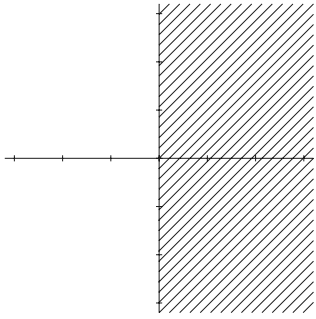
$$h\lambda_{n,1},\ h\lambda_{n,2},\ \ldots,\ h\lambda_{n,d} \ \in\ \mathcal{D},$$

[2] Our interest in (4.7) with $\operatorname{Re} \lambda > 0$ is limited, since the exact solution rapidly becomes very large. However, for nonlinear equations there is an intense interest, which we will not pursue in this volume, in those equations for which a counterpart of λ, namely the Liapunov exponent, is positive.

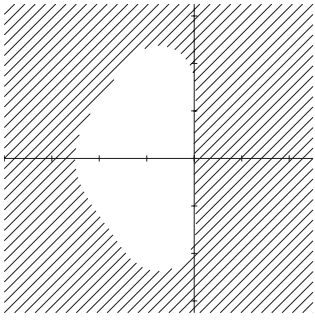
$\hat{r}_{1/0}(z) = 1 + z$



$\hat{r}_{1/1}(z) = \dfrac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}$



$\hat{r}_{3/0}(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3$



$\hat{r}_{1/2}(z) = \dfrac{1 + \frac{1}{3}z}{1 - \frac{2}{3}z + \frac{1}{6}z^2}$

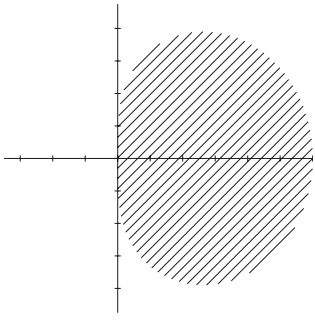


Figure 4.2 Stability domains (the unshaded areas) for various rational approximations. Note that $\hat{r}_{1/0}$ corresponds to the Euler method, while $\hat{r}_{1/1}$ corresponds both to the trapezoidal rule and the implicit midpoint rule. The $\hat{r}_{\alpha/\beta}$ notation is introduced in Section 4.3.

where the complex numbers $\lambda_{n,1}, \lambda_{n,2}, \ldots, \lambda_{n,d}$ are the eigenvalues of the *Jacobian matrix* $J_n := \partial \boldsymbol{f}(t_n, \boldsymbol{y}_n)/\partial \boldsymbol{y}$. This is based on the assumption that the local behaviour of the ODE is modelled well by the variational equation $\boldsymbol{y}' = \boldsymbol{y}_n + J_n(\boldsymbol{y} - \boldsymbol{y}_n)$. We hasten to emphasize that this practice is far from exact. Naive translation of any linear theory to a nonlinear setting can be dangerous and the correct approach is to embrace a nonlinear framework from the outset. Although in its full generality this ranges well beyond the material of this book, we provide a few pointers to modern nonlinear stability theory in Chapter 5.

Let us continue our investigation of linear stability domains. Replacing Λ by λ from (4.7) in (4.5) and bearing in mind that $y_0 = 1$, we obtain

$$y_n = \left(\frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}\right)^n, \qquad n = 0, 1, \ldots \tag{4.10}$$

Again, $\{y_n\}_{n=0,1,\ldots}$ is a geometric sequence. Therefore, we obtain for the linear stability domain in the case of the trapezoidal rule,

$$\mathcal{D}_{\rm TR} = \left\{ z \in \mathbb{C} \,:\, \left| \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z} \right| < 1 \right\}.$$

It is trivial to verify that the inequality within the braces is *identical* to $\mathrm{Re}\, z < 0$. In other words, *the trapezoidal rule mimics the asymptotic stability of linear ODE systems without any need to decrease the step size,* a property that we have already noticed in a special example in Section 4.1.

The latter feature is of sufficient importance to deserve a name of its own. We say that a method is *A-stable* if

$$\mathbb{C}^- := \{ z \in \mathbb{C} \,:\, \mathrm{Re}\, z < 0 \} \subseteq \mathcal{D}.$$

In other words, whenever a method is A-stable, we can choose the step size h (at least, for linear systems) on accuracy considerations only, without paying heed to stability constraints.

The trapezoidal rule is A-stable, whilst Euler's method is not. As is evident from Fig. 4.2, the graph labelled $\hat{r}_{1/2}(z)$ – but not the one labelled $\hat{r}_{3/0}(z)$ – corresponds to an A-stable method. It is left to the reader to ascertain in Exercise 4.4 that the theta method (1.13) is A-stable if and only if $0 \le \theta \le \frac{1}{2}$.

4.3 A-stability of Runge–Kutta methods

Applying the Runge–Kutta method (3.9) to the linear equation (4.7), we obtain

$$\xi_j = y_n + h\lambda \sum_{i=1}^{\nu} a_{j,i} \xi_i, \qquad j = 1, 2, \ldots, \nu.$$

Denote

$$\boldsymbol{\xi} := \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_\nu \end{bmatrix}, \qquad \mathbf{1} := \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^\nu;$$

then $\boldsymbol{\xi} = \mathbf{1} y_n + h\lambda A \boldsymbol{\xi}$ and the exact solution of this linear algebraic system is

$$\boldsymbol{\xi} = (I - h\lambda A)^{-1} \mathbf{1} y_n.$$

Therefore, assuming that $I - h\lambda A$ is nonsingular,

$$y_{n+1} = y_n + h\lambda \sum_{j=1}^{\nu} b_j \xi_j = \left[1 + h\lambda \boldsymbol{b}^\top (I - h\lambda A)^{-1} \mathbf{1} \right] y_n, \qquad n = 0, 1, \ldots \tag{4.11}$$

We denote by $\mathbb{P}_{\alpha/\beta}$ the set of all rational functions $\hat{p}/\hat{q}$, where $\hat{p} \in \mathbb{P}_\alpha$ and $\hat{q} \in \mathbb{P}_\beta$.

Lemma 4.1 *For every Runge–Kutta method* (3.9) *there exists* $r \in \mathbb{P}_{\nu/\nu}$ *such that*

$$y_n = [r(h\lambda)]^n, \qquad n = 0, 1, \ldots \tag{4.12}$$

Moreover, if the Runge–Kutta method is explicit then $r \in \mathbb{P}_\nu$.

Proof It follows at once from (4.11) that (4.12) is valid with

$$r(z) := 1 + z\boldsymbol{b}^\top (I - zA)^{-1}\mathbf{1}, \qquad z \in \mathbb{C}, \tag{4.13}$$

and it remains to verify that r is indeed a rational function (a polynomial for an explicit scheme) of the stipulated type.

We represent the inverse of $I - zA$ using a familiar formula from linear algebra,

$$(I - zA)^{-1} = \frac{\operatorname{adj}(I - zA)}{\det(I - zA)},$$

where $\operatorname{adj} C$ is the *adjugate* of the $\nu \times \nu$ matrix C: the (i, j)th entry of the adjugate (also known as the 'adjunct' and abbreviated in the same way) is the determinant of the (j, i)th principal minor, multiplied by $(-1)^{i+j}$. Since each entry of $I - zA$ is linear in z, we deduce that each element of $\operatorname{adj}(I - zA)$, being (up to a sign) a determinant of a $(\nu - 1) \times (\nu - 1)$ matrix, is in $\mathbb{P}_{\nu-1}$. We thus conclude that

$$\boldsymbol{b}^\top \operatorname{adj}(I - zA)\mathbf{1} \in \mathbb{P}_{\nu-1},$$

therefore $\det(I - zA) \in \mathbb{P}_\nu$ implies $r \in \mathbb{P}_{\nu/\nu}$.

Finally, if the method is explicit then A is strictly lower triangular and $I - zA$ is, regardless of $z \in \mathbb{C}$, a lower triangular matrix with ones along the diagonal. Therefore $\det(I - zA) \equiv 1$ and r is a polynomial. ∎

Lemma 4.2 *Suppose that an application of a numerical method to the linear equation* (4.7) *produces a geometric solution sequence,* $y_n = [r(h\lambda)]^n$, $n = 0, 1, \ldots$, *where* r *is an arbitrary function. Then*

$$\mathcal{D} = \{z \in \mathbb{C} : |r(z)| < 1\}. \tag{4.14}$$

Proof This follows at once from the definition of the set $\mathcal{D}$. ∎

Corollary *No explicit Runge–Kutta (ERK) method* (3.5) *can be A-stable.*

Proof Given an ERK method, Lemma 4.1 states that the function r is a polynomial and (4.13) implies that $r(0) = 1$. No polynomial, except for the constant function $r(z) \equiv c \in (-1, 1)$, may be uniformly bounded by the value unity in $\mathbb{C}^-$, and this excludes A-stability. ∎

For both Euler's method and the trapezoidal rule we have observed already that the solution sequence obeys the conditions of Lemma 4.2. This is hardly surprising, since both methods can be written in a Runge–Kutta formalism.

◇ **The function r for specific IRK schemes** Let us consider the methods

$$\begin{array}{c|cc} 0 & \frac{1}{4} & -\frac{1}{4} \\ \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array} \qquad \text{and} \qquad \begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}.$$

We have already encountered both in Chapter 3: the first is (3.10), whereas the second corresponds to collocation at $c_1 = \frac{1}{3}$, $c_2 = 1$.

Substitution into (4.13) confirms that the function r is identical for the two methods:

$$r(z) = \frac{1 + \frac{1}{3}z}{1 - \frac{2}{3}z + \frac{1}{6}z^2}. \tag{4.15}$$

To check A-stability we employ (4.14). Representing $z \in \mathbb{C}$ in polar coordinates, $z = \rho \mathrm{e}^{\mathrm{i}\theta}$, where $\rho > 0$ and $|\theta + \pi| < \frac{1}{2}\pi$, we query whether $|r(\rho \mathrm{e}^{\mathrm{i}\theta})| < 1$. This would be equivalent to

$$\left|1 + \tfrac{1}{3}\rho \mathrm{e}^{\mathrm{i}\theta}\right|^2 < \left|1 - \tfrac{2}{3}\rho \mathrm{e}^{\mathrm{i}\theta} + \tfrac{1}{6}\rho^2 \mathrm{e}^{2\mathrm{i}\theta}\right|^2$$

and hence to

$$1 + \tfrac{2}{3}\rho\cos\theta + \tfrac{1}{9}\rho^2 < 1 - \tfrac{4}{3}\rho\cos\theta + \rho^2\left(\tfrac{1}{3}\cos 2\theta + \tfrac{4}{9}\right) - \tfrac{2}{9}\rho^3\cos\theta + \tfrac{1}{36}\rho^4.$$

Rearranging terms, the condition for $\rho \mathrm{e}^{\mathrm{i}\theta} \in \mathcal{D}$ becomes

$$2\rho\left(1 + \tfrac{1}{9}\rho^2\right)\cos\theta < \tfrac{1}{3}\rho^2(1 + \cos 2\theta) + \tfrac{1}{36}\rho^4 = \tfrac{2}{3}\rho^2\cos^2\theta + \tfrac{1}{36}\rho^4,$$

and this is obeyed for all $z \in \mathbb{C}^-$ since $\cos\theta < 0$ for all such z. Both methods are therefore A-stable.

A similar analysis can be applied to the Gauss–Legendre methods of Section 3.4, but the calculations become increasingly labour intensive for large values of ν. Fortunately, we are just about to identify a few shortcuts that render this job significantly easier. ◇

Our first observation is that there is no need to check every $z \in \mathbb{C}^-$ to verify that a given rational function r originates in an A-stable method (such an r is called *A-acceptable*).

Lemma 4.3 *Let r be an arbitrary rational function that is not a constant. Then $|r(z)| < 1$ for all $z \in \mathbb{C}^-$ if and only if all the poles of r have positive real parts and $|r(\mathrm{i}t)| \le 1$ for all $t \in \mathbb{R}$.*

Proof If $|r(z)| < 1$ for all $z \in \mathbb{C}^-$ then, by continuity, $|r(z)| \le 1$ for all $z \in \operatorname{cl}\mathbb{C}^-$. In particular, r is not allowed to have poles in the closed left half-plane and $|r(\mathrm{i}t)| \le 1$, $t \in \mathbb{R}$.

To prove the converse we note that, provided its poles reside to the right of $\mathrm{i}\mathbb{R}$, the rational function r is analytic in the closed set $\operatorname{cl}\mathbb{C}^-$. Therefore, and since r is

not constant, it attains its maximum along the boundary. In other words $|r(\mathrm{i}t)| \leq 1$, $t \in \mathbb{R}$, implies $|r(z)| < 1$, $z \in \mathbb{C}^-$, and the proof is complete. ∎

The benefits of the lemma are apparent in the case of the function (4.15): the poles reside at $2 \pm \mathrm{i}\sqrt{2}$, hence at the open right half-plane. Moreover $|r(\mathrm{i}t)| \leq 1$, $t \in \mathbb{R}$, is equivalent to

$$\left|1 + \tfrac{1}{3}\mathrm{i}t\right|^2 \leq \left|1 - \tfrac{2}{3}\mathrm{i}t - \tfrac{1}{6}t^2\right|^2, \qquad t \in \mathbb{R},$$

and hence to

$$1 + \tfrac{1}{9}t^2 \leq 1 + \tfrac{1}{9}t^2 + \tfrac{1}{36}t^4, \qquad t \in \mathbb{R}.$$

The gain is even more spectacular for the two-stage Gauss–Legendre method, since in this case

$$r(z) = \frac{1 + \frac{1}{2}z + \frac{1}{12}z^2}{1 - \frac{1}{2}z + \frac{1}{12}z^2}$$

(although it is possible to evaluate this from the RK tableau in Section 3.4, a considerably easier derivation follows from the proof of the corollary to Theorem 4.6). Since the poles $3 \pm \mathrm{i}\sqrt{3}$ are in the open right half-plane and $|r(\mathrm{i}t)| \equiv 1$, $t \in \mathbb{R}$, the method is A-stable.

Our next result focuses on the kind of rational functions r likely to feature in (4.12).

Lemma 4.4 *Suppose that the solution sequence $\{y_n\}_{n=0}^{\infty}$, which is produced by applying a method of order p to the linear equation* (4.7) *with a constant step size, obeys* (4.12)*. Then necessarily*

$$r(z) = \mathrm{e}^z + \mathcal{O}\left(z^{p+1}\right), \qquad z \to 0. \tag{4.16}$$

Proof Since $y_{n+1} = r(h\lambda)y_n$ and the exact solution, subject to the initial condition $y(t_n) = y_n$, is $\mathrm{e}^{h\lambda}y_n$, the relation (4.16) follows from the definition of order. ∎

We say that a function r that obeys (4.16) is of *order* p. This should not be confused with the order of a numerical method: it is easy to construct pth-order methods with a function r whose order exceeds p, in other words, methods that exhibit superior order when applied to linear equations.

The lemma narrows down considerably the field of rational functions r that might occur in A-stability analysis. The most important functions exploit all available degrees of freedom to increase the order.

Theorem 4.5 *Given any integers α, $\beta \geq 0$, there exists a unique function $\hat{r}_{\alpha/\beta} \in \mathbb{P}_{\alpha/\beta}$ such that*

$$\hat{r}_{\alpha/\beta} = \frac{\hat{p}_{\alpha/\beta}}{\hat{q}_{\alpha/\beta}}, \qquad \hat{q}_{\alpha/\beta}(0) = 1$$

and $\hat{r}_{\alpha/\beta}$ is of order $\alpha+\beta$. The explicit forms of the numerator and the denominator are respectively

$$\begin{aligned}\hat{p}_{\alpha/\beta}(z) &= \sum_{k=0}^{\alpha}\binom{\alpha}{k}\frac{(\alpha+\beta-k)!}{(\alpha+\beta)!}z^k,\\ \hat{q}_{\alpha/\beta}(z) &= \sum_{k=0}^{\beta}\binom{\beta}{k}\frac{(\alpha+\beta-k)!}{(\alpha+\beta)!}(-z)^k = \hat{p}_{\beta/\alpha}(-z).\end{aligned} \tag{4.17}$$

Moreover $\hat{r}_{\alpha/\beta}$ is (up to a rescaling of the numerator and the denominator by a non-zero multiplicative constant) the only member of $\mathbb{P}_{\alpha/\beta}$ of order $\alpha+\beta$, and no function in $\mathbb{P}_{\alpha/\beta}$ may exceed this order. ■

The functions $\hat{r}_{\alpha/\beta}$ are called *Padé approximations* to the exponential. Most of the functions r that have been encountered so far are of this kind; thus (compare with (4.8), (4.10) and (4.15))

$$\hat{r}_{1/0}(z) = 1+z, \qquad \hat{r}_{1/1} = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}, \qquad \hat{r}_{1/2}(z) = \frac{1+\frac{1}{3}z}{1-\frac{2}{3}z+\frac{1}{6}z^2}.$$

Padé approximations can be classified according to whether they are A-acceptable. Obviously, we need $\alpha \le \beta$ otherwise $\hat{r}_{\alpha/\beta}$ cannot be bounded in $\mathbb{C}^-$. Surprisingly, the latter condition is not sufficient. It is not difficult to prove, for example, that $\hat{r}_{0/3}$ is not A-acceptable!

Theorem 4.6 (The Wanner–Hairer–Nørsett theorem) *The Padé approximation $\hat{r}_{\alpha/\beta}$ is A-acceptable if and only if $\alpha \le \beta \le \alpha+2$.* ■

Corollary *The Gauss–Legendre IRK methods are A-stable for every $\nu \ge 1$.*

Proof We know from Section 3.4 that a ν-stage Gauss–Legendre method is of order 2ν. By Lemma 4.1 the underlying function r belongs to $\mathbb{P}_{\nu/\nu}$ and, by Lemma 4.4, it approximates the exponential function to order 2ν. Therefore, according to Theorem 4.5, $r = \hat{r}_{\nu/\nu}$, a function that is A-acceptable by Theorem 4.6. It follows that the Gauss–Legendre method is A-stable. ■

4.4 A-stability of multistep methods

Attempting to extend the definition of A-stability to the multistep method (2.8), we are faced with a problem: the implementation of an s-step method requires the provision of s values and only one of these is supplied by the initial condition. We will see in Chapter 7 how such values are derived in realistic computation. Here we adopt the attitude that a stable solution of the linear equation (4.7) is required *for all possible values of* $y_1, y_2, \ldots, y_{s-1}$. The justification of this pessimistic approach is that otherwise, even were we somehow to choose 'good' starting values, a small perturbation (e.g., a roundoff error) might well divert the solution trajectory toward instability. The reasons are similar to those already discussed in Section 4.1 in the context of the Euler method.

Let us suppose that the method (2.8) is applied to the solution of (4.7). The outcome is

$$\sum_{m=0}^{s} a_m y_{n+m} = h\lambda \sum_{m=0}^{s} b_m y_{n+m}, \qquad n = 0, 1, \ldots,$$

which we write in the form

$$\sum_{m=0}^{s} (a_m - h\lambda b_m) y_{n+m} = 0, \qquad n = 0, 1, \ldots \tag{4.18}$$

The equation (4.18) is an example of a *linear difference equation,*

$$\sum_{m=0}^{s} g_m x_{n+m} = 0, \qquad n = 0, 1, \ldots, \tag{4.19}$$

and it can be solved similarly to the more familiar linear differential equation

$$\sum_{m=0}^{s} g_m x^{(m)} = 0, \qquad t \geq t_0,$$

where the superscript indicates differentiation m times. Specifically, we form the characteristic polynomial

$$\eta(w) := \sum_{m=0}^{s} g_m w^m.$$

Let the zeros of η be $w_1, w_2, \ldots, w_q$, say, with multiplicities $k_1, k_2, \ldots, k_q$ respectively, where $\sum_{i=1}^{q} k_i = s$. The general solution of (4.19) is

$$x_n = \sum_{i=1}^{q} \left(\sum_{j=0}^{k_i - 1} c_{i,j} n^j \right) w_i^n, \qquad n = 0, 1, \ldots \tag{4.20}$$

The s constants $c_{i,j}$ are uniquely determined by the s starting values $x_0, x_1, \ldots, x_{s-1}$.

Lemma 4.7 *Let us suppose that the zeros (as a function of w) of*

$$\eta(z, w) := \sum_{m=0}^{s} (a_m - b_m z) w^m, \qquad z \in \mathbb{C},$$

are $w_1(z), w_2(z), \ldots, w_{q(z)}(z)$, while their multiplicities are $k_1(z), k_2(z), \ldots, k_{q(z)}(z)$ respectively. The multistep method (2.8) *is A-stable if and only if*

$$|w_i(z)| < 1, \quad i = 1, 2, \ldots, q(z) \qquad \textit{for every} \qquad z \in \mathbb{C}^-. \tag{4.21}$$

Proof As for (4.20), the behaviour of y_n is determined by the magnitude of the numbers $w_i(h\lambda)$, $i = 1, 2, \ldots, q(h\lambda)$. If all reside inside the complex unit disc then their powers decay faster than any polynomial in n, therefore $y_n \to 0$. Hence, (4.21) is sufficient for A-stability.

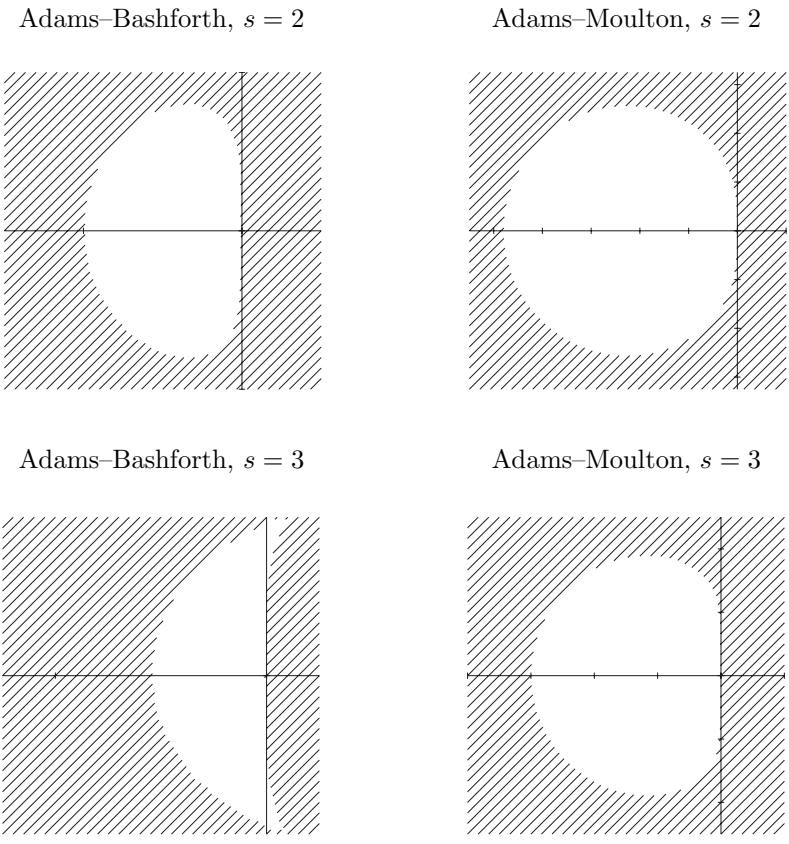


Figure 4.3 Linear stability domains $\mathcal{D}$ of Adams methods, explicit on the left and implicit on the right.

However, if $|w_1(h\lambda)| \geq 1$, say, then there exist starting values such that $c_{1,0} \neq 0$; therefore it is impossible for y_n to tend to zero as $n \to \infty$. We deduce that (4.21) is necessary for A-stability and so conclude the proof. ■

Instead of a single geometric component in (4.11), we have now a linear combination of several (in general, s) components to reckon with. This is the *quid pro quo* for using $s - 1$ starting values in addition to the initial condition, a practice whose perils have been highlighted already in the introduction to Chapter 2. According to Exercise 2.2, if a method is convergent then one of these components approximates the exponential function to the same order as the order of the method: this is similar to Lemma 4.4. However, the remaining zeros are purely parasitic: we can attribute no meaning to them so far as approximation is concerned.

Fig. 4.3 displays the linear stability domains of Adams methods, all at the same scale. Notice first how small they are and that they are reduced in size for the larger

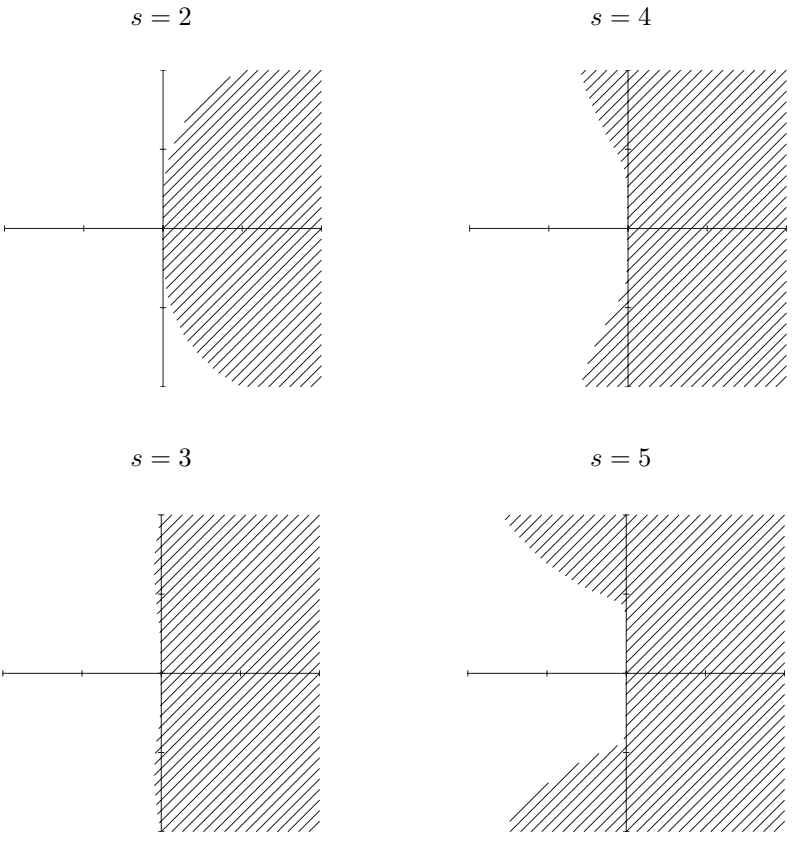


Figure 4.4 Linear stability domains $\mathcal{D}$ of BDF methods of orders $s = 2, 3, 4, 5$, shown at the same scale. Note that only $s = 2$ is A-stable.

s value. Next, pay attention to the difference between the explicit Adams–Bashforth and the implicit Adams–Moulton. In the latter case the stability domain, although not very impressive compared with those for other methods of Section 4.3, is substantially larger than for the explicit counterpart. This goes some way toward explaining the interest in implicit Adams methods, but more important reasons will be presented in Chapter 6.

However, as already mentioned in Chapter 2, Adams methods were never intended to cope with stiff equations. After all, this was the motivation for the introduction of backward differentiation formulae in Section 2.3. We turn therefore to Fig. 4.4, which displays linear stability domains for BDF methods – and are disappointed ... True, the set $\mathcal{D}$ is larger than was the case for, say, the Adams–Moulton method. However, only the two-step method displays any prospects of A-stability.

Let us commence with the good news: the BDF is indeed A-stable in the case $s = 2$. To demonstrate this we require two technical lemmas, which will be presented

with a comment in lieu of a complete proof.

Lemma 4.8 *The multistep method* (2.8) *is A-stable if and only if* $b_s > 0$ *and*

$$|w_1(\mathrm{i}t)|,\ |w_2(\mathrm{i}t)|,\ \ldots,\ |w_{q(\mathrm{i}t)}(\mathrm{i}t)| \le 1, \qquad t \in \mathbb{R},$$

where $w_1, w_2, \ldots, w_{q(z)}$ *are the zeros of* $\eta(z, \cdot)$ *from Lemma 4.7.*

Proof On the face of it, this is an exact counterpart of Lemma 4.3: $b_s > 0$ implies analyticity in $\mathrm{cl}\,\mathbb{C}^-$ and the condition on the moduli of zeros extends the inequality on $|r(z)|$. This is deceptive, since the zeros of $\eta(z, \cdot)$ do not reside in the complex plane but in an s-sheeted *Riemann surface* over $\mathbb{C}$. This does not preclude the application of the maximum principle, except that somewhat more sophisticated mathematical machinery is required. ∎

Lemma 4.9 (The Cohn–Schur criterion) *Both zeros of the quadratic* $\alpha w^2 + \beta w + \gamma$, *where* $\alpha, \beta, \gamma \in \mathbb{C}$, $\alpha \neq 0$, *reside in the closed complex unit disc if and only if*

$$|\alpha| \ge |\gamma|, \quad \big||\alpha|^2 - |\gamma|^2\big| \ge \big|\alpha\bar{\beta} - \beta\bar{\gamma}\big| \quad \textit{and} \quad \alpha = \gamma \neq 0 \ \Rightarrow\ |\beta| \le 2|\alpha|. \tag{4.22}$$

Proof This is a special case of a more general result, the *Cohn–Lehmer–Schur* criterion. The latter provides a finite algorithm to check whether a given complex polynomial (of any degree) has all its zeros in any closed disc in $\mathbb{C}$. ∎

Theorem 4.10 *The two-step BDF* (2.15) *is A-stable.*

Proof We have

$$\eta(z, w) = (1 - \tfrac{2}{3}z)w^2 - \tfrac{4}{3}w + \tfrac{1}{3}.$$

Therefore $b_2 = \frac{2}{3}$ and the first A-stability condition of Lemma 4.8 is satisfied. To verify the second condition we choose $t \in \mathbb{R}$ and use Lemma 4.9 to ascertain that neither of the moduli of the zeros of $\eta(\mathrm{i}t, \cdot)$ exceeds unity. Consequently $\alpha = 1 - \frac{2}{3}\mathrm{i}t$, $\beta = -\frac{4}{3}$, $\gamma = \frac{1}{3}$ and we obtain

$$|\alpha|^2 - |\gamma|^2 = \tfrac{4}{9}(2 + t^2) > 0$$

and

$$(|\alpha|^2 - |\gamma|^2)^2 - |\alpha\bar{\beta} - \beta\bar{\gamma}|^2 = \tfrac{16}{81}t^4 \ge 0.$$

Consequently, (4.22) is satisfied and we deduce A-stability. ∎

Unfortunately, not only the 'positive' deduction from Fig. 4.4 is true. The absence of A-stability in the BDF for $s \ge 2$ (of course, $s \le 6$, otherwise the method would not be convergent and we would never use it!) is a consequence of a more general and fundamental result.

Theorem 4.11 (The Dahlquist second barrier) *The highest order of an A-stable multistep method* (2.8) *is 2.* ∎

Comparing the Dahlquist second barrier with the corollary to Theorem 4.6, it is difficult to escape the impression that multistep methods are inferior to Runge–Kutta

methods when it comes to A-stability. This, however, does not mean that they should not be used with stiff equations! Let us look again at Fig. 4.4. Although the cases $s = 3, 4, 5$ fail A-stability, it is apparent that for each stability domain $\mathcal{D}$ there exists $\alpha \in (0, \pi]$ such that the infinite wedge

$$\mathcal{V}_\alpha := \left\{ \rho \mathrm{e}^{\mathrm{i}\theta} \,:\, \rho > 0,\ |\theta + \pi| < \alpha \right\} \subseteq \mathbb{C}^-$$

belongs to $\mathcal{D}$. In other words, provided that all the eigenvalues of a linear ODE system reside in $\mathcal{V}_\alpha$, no matter how far away they are from the origin, there is no need to depress the step size in response to stability restrictions. Methods with $\mathcal{V}_\alpha \subseteq \mathcal{D}$ are called A(α)-stable.[3] All BDF methods for $s \leq 6$ are A(α)-stable: in particular $s = 3$ corresponds to $\alpha = 86°2'$; as Fig. 4.4 implies, almost all the region $\mathbb{C}^-$ resides in the linear stability domain.

Comments and bibliography

Different aspects of stiff equations and A-stability form the theme of several monographs of varying degrees of sophistication and detail. Gear (1971) and Lambert (1991) are the most elementary, whereas Hairer & Wanner (1991) is a compendium of just about everything known in the subject area *circa* 1991. (No text, however, for obvious reasons, abbreviates the phrase 'linear stability domain' ...)

Before we comment on a few themes connected with stability analysis, let us mention briefly two topics which, while tangential to the subject matter of this chapter, deserve proper reference. Firstly, the functions $\hat{r}_{\alpha/\beta}$, which have played a substantial role in Section 4.3, are a special case of general Padé approximation. Let f be an arbitrary function that is analytic in the neighbourhood of the origin. The function $\hat{r} \in \mathbb{P}_{\alpha/\beta}$ is said to be an $[\alpha/\beta]$ *Padé approximant* of f if

$$\hat{r}(z) = f(z) + \mathcal{O}\big(z^{\alpha+\beta+1}\big), \qquad z \to 0.$$

Padé approximations possess a beautiful theory and have numerous applications, not just in the more obvious fields – the approximation of functions, numerical analysis etc. – but also in analytic number theory: they are a powerful tool in many transcendentality proofs. Baker & Graves-Morris (1981) presented a useful account of the Padé theory. Secondly, the Cohn–Schur criterion (Lemma 4.9) is a special case of a substantially more general body of knowledge that allows us to locate the zeros of polynomials in specific portions of the complex plane by a finite number of operations on the coefficients (Marden, 1966). A familiar example is the *Routh–Hurwitz* criterion, which tests whether all the zeros reside in $\mathbb{C}^-$ and is an important tool in control theory.

The characterization of all A-acceptable Padé approximations to the exponential function was the subject of a long-standing conjecture. Its resolution in 1978 by Gerhard Wanner, Ernst Hairer and Syvert Nørsett introduced the novel technique of *order stars* and was one of the great heroic tales of modern numerical mathematics. This technique can be also used to prove a far-reaching generalization of Theorem 4.11, as well as many other interesting results in the numerical analysis of differential equations. A comprehensive account of order stars features in Iserles & Nørsett (1991).

As far as A-stability for multistep equations is concerned, Theorem 4.11 implies that not much can be done. One obvious alternative, which has been mentioned in Section 4.4,

[3] Numerical analysts, being (mostly) human, tend to express α in degrees rather than radians.

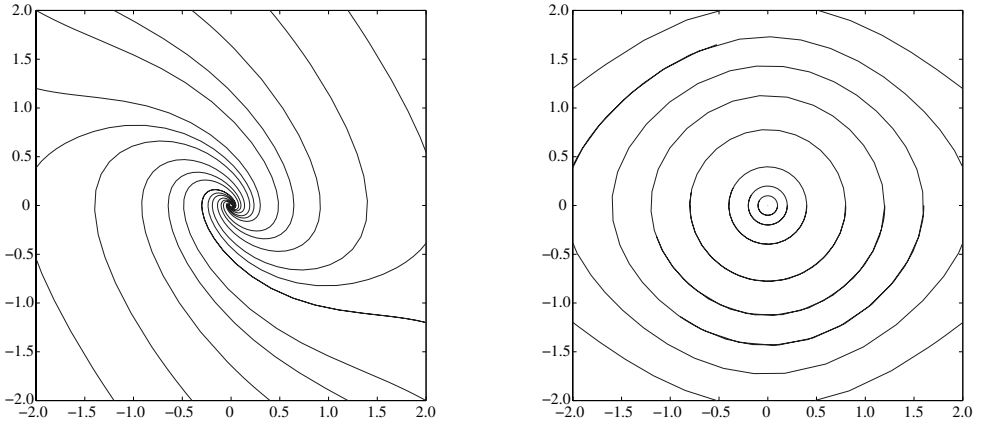


Figure 4.5 Phase planes for the damped oscillator $y'' + y' + \sin y = 0$ (on the left) and the undamped oscillator $y'' + \sin y = 0$ (on the right).

is to relax the stability requirement, in which case the order barrier disappears altogether. Another possibility is to combine the multistep rationale with the Runge–Kutta approach and possibly to incorporate higher derivatives as well. The outcome, a *general linear method* (Butcher, 2006), circumvents the barrier of Theorem 4.11.

We have mentioned in Section 4.2 that the justification of the linear model, which has led us into the concept of A-stability, is open to question when it comes to nonlinear equations. It is, however, a convenient starting point. The stability analysis of discretized nonlinear ODEs is these days a thriving industry! One model of nonlinear stability analysis is addressed in the next chapter but we make no pretence that it represents anything but a taster for a considerably more extensive theory.

And this is a convenient moment for a confession. Stiff ODEs might seem 'difficult' and indeed have been considered as such for a long time. Yet, once you get the hang of them, use the right methods and take care of stability issues, you are highly unlikely ever to go wrong. To understand why is this so and to get yourself in the right frame of mind for the next chapter, examine the phase plane of the damped nonlinear oscillator $y'' + y' + \sin y = 0$ on the left of Fig. 4.5.[4] (Of course, we convert this second-order ODE into a system of two coupled first-order ODEs $y_1' = y_2$, $y_2' = -\sin y_1 - y_2$.) No matter where we start within the displayed range, the destination is the same, the origin. Now, applying a numerical method means that our next step is typically misdirected to a neighbouring trajectory in the phase plane, but it is obvious from the figure that the flow itself is 'self correcting'. Unless we are committing errors which are both large and biased, a hallmark of an unstable method, ultimately our global picture will be at the very least of the right qualitative character: the numerical trajectory will tend to the origin. Small errors will correct themselves, provided that the method is stable enough.

Compare this with the undamped nonlinear oscillator $y'' + \sin y = 0$ on the right of Fig. 4.5. Except when it starts at the origin, in which case not much happens, the flow

[4]This system is not stiff but even this gentle damping is sufficient to convey our point, while a *real* stiff system, e.g. $y'' + 1000y' + \sin y = 0$, would have led to a plot that was considerably less intelligible.

(again, within the range of displayed initial values) progresses in periodic orbits. Now, no matter how accurate our method and no matter how stable it is, small errors can 'kick' us to the wrong trajectory; and repeated 'kicks', no matter how minute, are likely to produce ultimately a numerical trajectory that exhibits completely the wrong qualitative behaviour. Instead of a periodic orbit, the numerical solution might tend to a fixed point, diverge to infinity or, if our step size is too large, even exhibit spurious chaotic behaviour.

Stiff differential equations allow the possibility of redemption. As long as you recognise your sinful ways, correct your behaviour and adopt the right method and the right step size, your misdemeanours wil be forgiven and your solution will prosper. Not so the nonlinear oscillator $y'' + \sin y = 0$. Your numerical sins stay forever with you and accumulate forever. Or at least until you learn in the next chapter how to deal with this situation.

Baker, G.A. and Graves-Morris, P. (1981), *Padé Approximants,* Addison–Wesley, Reading, MA.

Butcher, J.C. (2006), General linear methods, *Acta Numerica* **15**, 157–256.

Gear, C.W. (1971), *Numerical Initial Value Problems in Ordinary Differential Equations,* Prentice–Hall, Englewood Cliffs, NJ.

Hairer, E. and Wanner, G. (1991), *Solving Ordinary Differential Equations II: Stiff Problems and Differential-Algebraic Equations,* Springer-Verlag, Berlin.

Iserles, A. and Nørsett, S.P. (1991), *Order Stars,* Chapman & Hall, London.

Lambert, J.D. (1991), *Numerical Methods for Ordinary Differential Systems,* Wiley, London.

Marden, M. (1966), *Geometry of Polynomials,* American Mathematical Society, Providence, RI.

Exercises

4.1 Let $\boldsymbol{y}' = \Lambda\boldsymbol{y}$, $\boldsymbol{y}(t_0) = \boldsymbol{y}_0$, be solved (with a constant step size $h > 0$) by a one-step method with a function r that obeys the relation (4.12). Suppose that a nonsingular matrix V and a diagonal matrix D exist such that $\Lambda = VDV^{-1}$. Prove that there exist vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d \in \mathbb{R}^d$ such that

$$\boldsymbol{y}(t_n) = \sum_{j=1}^{d} \mathrm{e}^{t_n\lambda_j}\boldsymbol{x}_j, \qquad n = 0, 1, \ldots,$$

and

$$\boldsymbol{y}_n = \sum_{j=1}^{d} [r(h\lambda)]^n \boldsymbol{x}_j, \qquad n = 0, 1, \ldots,$$

where $\lambda_1, \lambda_2, \ldots, \lambda_d$ are the eigenvalues of Λ. Deduce that the values of $\boldsymbol{x}_1$ and of $\boldsymbol{x}_2$, given in (4.3) and (4.4) are identical.

4.2* Consider the solution of $\boldsymbol{y}' = \Lambda\boldsymbol{y}$ where

$$\Lambda = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}, \qquad \lambda \in \mathbb{C}^-.$$

a Prove that

$$\Lambda^n = \begin{bmatrix} \lambda^n & n\lambda^{n-1} \\ 0 & \lambda^n \end{bmatrix}, \qquad n = 0, 1, \ldots$$

b Let g be an arbitrary function that is analytic about the origin. The 2×2 matrix $g(\Lambda)$ can be defined by substituting powers of Λ into the Taylor expansion of g. Prove that

$$g(t\Lambda) = \begin{bmatrix} g(t\lambda) & tg'(t\lambda) \\ 0 & g(t\lambda) \end{bmatrix}.$$

c By letting $g(z) = \mathrm{e}^z$ prove that $\lim_{t\to\infty} \boldsymbol{y}(t) = \boldsymbol{0}$.

d Suppose that $\boldsymbol{y}' = \Lambda\boldsymbol{y}$ is solved with a Runge–Kutta method, using a constant step size $h > 0$. Let r be the function from Lemma 4.1. Letting $g = r$, obtain the explicit form of $[r(h\Lambda)]^n$, $n = 0, 1, \ldots$

e Prove that if $h\lambda \in \mathcal{D}$, where $\mathcal{D}$ is the linear stability domain of the Runge–Kutta method, then $\lim_{n\to\infty} \boldsymbol{y}_n = \boldsymbol{0}$.

4.3* This question is concerned with the relevance of the linear stability domain to the numerical solution of *inhomogeneous* linear systems.

a Let Λ be a nonsingular matrix. Prove that the solution of $\boldsymbol{y}' = \Lambda\boldsymbol{y} + \boldsymbol{a}$, $\boldsymbol{y}(t_0) = \boldsymbol{y}_0$, is

$$\boldsymbol{y}(t) = \mathrm{e}^{(t-t_0)\Lambda}\boldsymbol{y}_0 + \Lambda^{-1}[\mathrm{e}^{(t-t_0)\Lambda} - I]\boldsymbol{a}, \qquad t \geq t_0.$$

Thus, deduce that if Λ has a full set of eigenvectors and all its eigenvalues reside in $\mathbb{C}^-$ then $\lim_{t\to\infty} \boldsymbol{y}(t) = -\Lambda^{-1}\boldsymbol{a}$.

b Assuming for simplicity's sake that the underlying equation is scalar, i.e. $y' = \lambda y + a$, $y(t_0) = y_0$, prove that a single step of the Runge–Kutta method (3.9) results in

$$y_{n+1} = r(h\lambda)y_n + q(h\lambda), \qquad n = 0, 1, \ldots,$$

where r is given by (4.13) and

$$q(z) := ha\boldsymbol{b}^\top(I - zA)^{-1}\boldsymbol{1} \in \mathbb{P}_{(\nu-1)/\nu}, \qquad z \in \mathbb{C}.$$

c Deduce, by induction or otherwise, that

$$y_n = [r(h\lambda)]^n y_0 + \left\{ \frac{[r(h\lambda)]^n - 1}{r(h\lambda) - 1} \right\} q(h\lambda), \qquad n = 0, 1, \ldots$$

d Assuming that $h\lambda \in \mathcal{D}$, prove that $\lim_{n\to\infty} y_n$ exists and is bounded.

4.4 Determine all values of θ such that the theta method (1.13) is A-stable.

4.5 Prove that for every ν-stage explicit Runge–Kutta method (3.5) of order ν it is true that

$$r(z) = \sum_{k=0}^{\nu} \frac{1}{k!} z^k, \qquad z \in \mathbb{C}.$$

4.6 Evaluate explicitly the function r for the following Runge–Kutta methods:

a
$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array},$$

b
$$\begin{array}{c|cc} \frac{1}{6} & \frac{1}{6} & 0 \\ \frac{5}{6} & \frac{2}{3} & \frac{1}{6} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array},$$

c
$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ 1 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}.$$

Are these methods A-stable?

4.7 Prove that the Padé approximation $\hat{r}_{0/3}$ is not A-acceptable.

4.8 Determine the order of the two-step method

$$\boldsymbol{y}_{n+2} - \boldsymbol{y}_n = \tfrac{2}{3} h \left[\boldsymbol{f}(t_{n+2}, \boldsymbol{y}_{n+2}) + \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) + \boldsymbol{f}(t_n, \boldsymbol{y}_n) \right], \quad n = 0, 1, \ldots$$

Is it A-stable?

4.9 The two-step method

$$\boldsymbol{y}_{n+2} - \boldsymbol{y}_n = 2h \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}), \qquad n = 0, 1, \ldots \tag{4.23}$$

is called the *explicit midpoint rule.*

a Denoting by $w_1(z)$ and $w_2(z)$ the zeros of the underlying function $\eta(z, \cdot\,)$, prove that $w_1(z) w_2(z) \equiv -1$ for all $z \in \mathbb{C}$.

b Show that $\mathcal{D} = \emptyset$.

c We say that $\tilde{\mathcal{D}}$ is a *weak linear stability domain* of a numerical method if, when applied to the scalar linear test equation, it produces a uniformly bounded solution sequence. (It is easy to see that $\tilde{\mathcal{D}} = \operatorname{cl} \mathcal{D}$ for most methods of interest.) Determine explicitly $\tilde{\mathcal{D}}$ for the method (4.23).

The method (4.23) will feature again in Chapters 16 and 17, in the guise of the *leapfrog* scheme.

4.10 Prove that if the multistep method (2.8) is convergent then $0 \in \partial \tilde{\mathcal{D}}$.