Numerical Methods for
Ordinary Differential Equations

SECOND EDITION

J.C. Butcher

WILEY

Numerical Methods for Ordinary Differential Equations

Numerical Methods for Ordinary Differential Equations

Second Edition

J. C. Butcher
The University of Auckland, New Zealand

John Wiley & Sons, Ltd

Contents

Preface to the first edition

Introductory remarks

This book represents an attempt to modernize and expand my previous volume, *The Numerical Analysis of Ordinary Differential Equations: Runge–Kutta and General Linear Methods.* It is more modern in that it considers several topics that had not yet emerged as important research areas when the former book was written. It is expanded in that it contains a comprehensive treatment of linear multistep methods. This achieves a better balance than the earlier volume which made a special feature of Runge–Kutta methods.

In order to accommodate the additional topics, some sacrifices have been made. The background work which introduced the earlier book is here reduced to an introductory chapter dealing only with differential and difference equations. Several topics that seem to be still necessary as background reading are now introduced in survey form where they are actually needed. Some of the theoretical ideas are now explained in a less formal manner. It is hoped that mathematical rigour has not been seriously jeopardized by the use of this more relaxed style; if so, then there should be a corresponding gain in accessibility. It is believed that no theoretical detail has been glossed over to the extent that an interested reader would have any serious difficulty in filling in the gaps.

It is hoped that lowering the level of difficulty in the exposition will widen the range of readers who might be able to find this book interesting and useful. With the same idea in mind, exercises have been introduced at the end of each section.

Following the chapter on differential and difference equations, Chapter 2 is presented as a study of the Euler method. However, it aims for much more than this in that it also reviews many other methods and classes of methods as generalizations of the Euler method. This chapter can be used as a broad-ranging introduction to the entire subject of numerical methods for ordinary differential equations.

Chapter 3 contains a detailed analysis of Runge–Kutta methods. It includes studies of the order, stability and convergence of Runge–Kutta methods and also considers in detail the design of efficient explicit methods for non-stiff

problems. For implicit methods for stiff problems, inexpensive implementation costs must be added to accuracy and stability as a basic requirement. Recent work on each of these questions is surveyed and discussed.

Linear multistep methods, including the combination of two methods as predictor–corrector pairs, are considered in Chapter 4. The theory interrelating stability, consistency and convergence is presented together with an analysis of order conditions. This leads to a proof of the (first) 'Dahlquist barrier'. The methods in this class which are generally considered to be the most important for the practical solution of non-stiff problems are the Adams–Bashforth and Adams–Moulton formulae. These are discussed in detail, including their combined use as predictor–corrector pairs. The application of linear multistep methods to stiff problems is also of great practical importance and the treatment will include an analysis of the backward difference formulae.

In Chapter 5 the wider class of general linear methods is introduced and analysed. Questions analogous to those arising in the classical Runge–Kutta and linear multistep methods – that is, questions of consistency, stability, convergence and order – are considered and explored. Several sub-families of methods, that have a potential practical usefulness, are examined in detail. This includes the so-called DIMSIM methods and a new type of method exhibiting what is known as inherent Runge–Kutta stability.

The remarks in the following paragraphs are intended to be read following Chapter 5.

Concluding remarks

Any account of this rapidly evolving subject is bound to be incomplete. Complete books are all alike; every incomplete book is incomplete in its own way.

It has not been possible to deal adequately with implementation questions. Numerical software for evolutionary problems entered its modern phase with the DIFSUB code of Gear (1971a). 'Modern' in this sense means that most of the ingredients of subsequent codes were present. Both stiff and non-stiff problems are catered for, provision is made for Jacobian calculation either by subroutine call or by difference approximation; the choice is up to the user. Most importantly, automatic selection of stepsize and order is made dynamically as the solution develops. Compared with this early implementation of linear multistep methods, the Radau code (Hairer and Wanner, 1996) uses implicit Runge–Kutta methods for the solution of stiff problems.

In recent years, the emphasis in numerical methods for evolutionary problems has moved beyond the traditional areas of non-stiff and stiff problems. In particular, differential-algebraic equations have become the subject of intense analysis as well as the development of reliable and efficient algorithms for problems of variable difficulty, as measured for example by

the indices of the problems. Some basic references in this vibrant area are Brenan, Campbell and Petzold (1989) and Hairer, Lubich and Roche (1989) In particular, many codes are now designed for applications to stiff ordinary differential equations in which algebraic constraints also play a role. On the Runge–Kutta side, Radau is an example of this multipurpose approach. On the linear multistep side, Petzold's DASSL code is closely related to Gear's DIFSUB but has the capability of solving differential-algebraic equations, at least of low index.

Many problems derived from mechanical systems can be cast in a Hamiltonian formulation. To faithfully model the behaviour of such problems it is necessary to respect the symplectic structure. Early work on this by the late Feng Kang has led to worldwide activity in the study of this type of question. A basic reference on Hamiltonian problems is Sanz-Serna and Calvo (1994).

The emphasis on the preservation of qualitative features of a numerical solution has now grown well beyond the Hamiltonian situation and has become a mathematical discipline in its own right. We mention just two key references in this emerging subject of 'geometric integration'. They are Iserles, et al. (2000) and Hairer, Lubich and Wanner (2006).

Internet commentary

Undoubtedly there will be comments and suggestions raised by readers of this volume. A web resource has been developed to form a commentary and information exchange for issues as they arise in the future. The entry point is
$$\text{http://www.math.auckland.ac.nz/~butcher/book}$$

Acknowledgements

I acknowledge with gratitude the support and assistance of many people in the preparation of this volume. The editorial and production staff at Wiley have encouraged and guided me through the publishing process. My wife, children, grandchildren and stepchildren have treated me gently and sympathetically.

During part of the time I have been working on this book, I have received a grant from the Marsden Fund. I am very grateful for this assistance both as an expression of confidence from my scientific colleagues in New Zealand and as practical support.

The weekly workshop in numerical analysis at The University of Auckland has been an important activity in the lives of many students, colleagues and myself. We sometimes refer to this workshop as the 'Runge–Kutta Club'. Over the past five or more years especially, my participation in this workshop has greatly added to my understanding of numerical analysis through collaboration and vigorous discussions. As this book started to take shape they have provided a sounding board for many ideas, some of which

Preface to the second edition

Reintroductory remarks

The incremental changes incorporated into this edition are an acknowledgement of progress in several directions. The emphasis of structure-preserving algorithms has driven much of this recent progress, but not all of it. The classical linear multistep and Runge–Kutta methods have always been special cases of the large family of general linear methods, but this observation is of no consequence unless some good comes of it. In my opinion, there are only two good things that might be worth achieving. The first is that exceptionally good methods might come to light which would not have been found in any other way. The second is that a clearer insight and perhaps new overarching theoretical results might be expressed in the general linear setting. I believe that both these aims have been achieved but other people might not agree. However, I hope it can be accepted that some of the new methods which arise naturally as general linear methods have at least some potential in practical computation. I hope also that looking at properties of traditional methods from within the general linear framework will provide additional insight into their computational properties.

How to read this book

Of the five chapters of this book, the first two are the most introductory in nature. Chapter 1 is a review of differential and difference equations with a systematic study of their basic properties balanced against an emphasis on interesting and prototypical problems. Chapter 2 provides a broad introduction to numerical methods for ordinary differential equations. This is motivated by the simplicity of the Euler method and a view that other standard methods are systematic generalizations of this basic method. If Runge–Kutta and linear multistep methods are generalizations of Euler then so are general linear methods and it is natural to introduce a wide range of multivalue–multistage methods at this elementary level.

A reading of this book should start with these two introductory chapters. For a reader less experienced in this subject this is an obvious entry point but they also have a role for a reader who is ready to go straight into the later chapters. For such readers they will not take very long but they do set the scene for an entry into the most technical parts of the book.

Chapter 3 is intended as a comprehensive study of Runge–Kutta methods. A full theory of order and stability is presented and at least the early parts of this chapter are prerequisites for Chapter 5 and to a lesser extent for Chapter 4. The use of B-series, or the coefficients that appear in these series, is becoming more and more a standard tool for a full understanding of modern developments in this subject.

Chapter 4 is full study of linear multistep methods. It is based on Dahlquists' classic work on consistency, stability and order and includes analysis of linear and nonlinear stability. In both Chapters 3 and 4 the use of order stars to resolve order and stability questions is complemented by the introduction of order arrows. It is probably a good idea to read through most of Chapter 4 before embarking on Chapter 5. This is not because general linear methods are intrinsically inaccessible, but because an appreciation of their overarching nature hinges on an appreciation of the special cases they include.

General linear methods, the subject of Chapter 5, treat well-known methods in a unified way, but it is hoped they do more than this. There really seem to be new and useful methods buried amongst them which cannot be easily motivated in any other way. Thus, while this chapter needs to be put aside to be read as a culmination, it should not be put off too long. There is so much nice mathematics already associated with these methods, and the promise of more to come provides attraction enough. It is general linear methods, and the stability functions associated with them that really put order arrows in their rightful place.

Internet support pages

For additional information and supporting material see

        http://www.math.auckland.ac.nz/~butcher/ODE-book-2008

Reacknowledgements

I have many people to thank and to rethank in my efforts to produce an improved edition. My understanding of the stability and related properties of general linear methods has been sharpened by working with Adrian Hill and Laura Hewitt. Helmut Podhaisky has given me considerable help and advice especially on aspects of general linear method implementation. My special thanks to Jane HyoJin Lee for her assistance with the final form of the manuscript. A number of people have made comments and provided

Chapter 1

Differential and Difference Equations

10   Differential Equation Problems

100   Introduction to differential equations

As essential tools in scientific modelling, differential equations are familiar to every educated person. In this introductory discussion we do not attempt to restate what is already known, but rather to express commonly understood ideas in the style that will be used for the rest of this book.

The aim will always be to understand, as much as possible, what we expect to happen to a quantity which satisfies a differential equation. At the most obvious level, this means predicting the value this quantity will have at some future time. However, we are also interested in more general questions such as the adherence to possible conservation laws or perhaps stability of the long-term solution. Since we emphasize numerical methods, we often discuss problems with known solutions mainly to illustrate qualitative and numerical behaviour.

Even though we sometimes refer to 'time' as the independent variable, that is, as the variable on which the value of the 'solution' depends, there is no reason for insisting on this interpretation. However, we generally use x to denote the 'independent' or 'time' variable and y to denote the 'dependent variable'. Hence, differential equations will typically be written in the form

$$y'(x) = f(x, y(x)), \tag{100a}$$

where

$$y' = \frac{dy}{dx}.$$

Sometimes, for convenience, we omit the x in $y(x)$.

The terminology used in (100a) is misleadingly simple, because y could be a vector-valued function. Thus, if we are working in $\mathbb{R}^N$, and x is permitted to take on any real value, then the domain and range of the function f which

defines a differential equation and the solution to this equation are given by

$$f : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N,$$
$$y : \mathbb{R} \to \mathbb{R}^N.$$

Since we might be interested in time values that lie only in some interval $[a, b]$, we sometimes consider problems in which $y : [a, b] \to \mathbb{R}^N$, and $f : [a, b] \times \mathbb{R}^N \to \mathbb{R}^N$. When dealing with specific problems, it is often convenient to focus, not on the vector-valued functions f and y, but on individual components. Thus, instead of writing a differential equation system in the form of (100a), we can write coupled equations for the individual components:

$$
\begin{aligned}
y_1'(x) &= f_1(x, y_1, y_2, \ldots, y_N), \\
y_2'(x) &= f_2(x, y_1, y_2, \ldots, y_N), \\
&\vdots \qquad\qquad \vdots \\
y_N'(x) &= f_N(x, y_1, y_2, \ldots, y_N).
\end{aligned}
\tag{100b}
$$

A differential equation for which f is a function not of x, but of y only, is said to be 'autonomous'. Some equations arising in physical modelling are more naturally expressed in one form or the other, but we emphasize that it is always possible to write a non-autonomous equation in an equivalent autonomous form. All we need to do to change the formulation is to introduce an additional component y_{N+1} into the y vector, and ensure that this can always maintain the same value as x, by associating it with the differential equation $y_{N+1}' = 1$. Thus, the modified system is

$$
\begin{aligned}
y_1'(x) &= f_1(y_{N+1}, y_1, y_2, \ldots, y_N), \\
y_2'(x) &= f_2(y_{N+1}, y_1, y_2, \ldots, y_N), \\
&\vdots \qquad\qquad \vdots \\
y_N'(x) &= f_N(y_{N+1}, y_1, y_2, \ldots, y_N), \\
y_{N+1}'(x) &= 1.
\end{aligned}
\tag{100c}
$$

A system of differential equations alone does not generally define a unique solution, and it is necessary to add to the formulation of the problem a number of additional conditions. These are either 'boundary conditions', if further information is given at two or more values of x, or 'initial conditions', if all components of y are specified at a single value of x.

If the value of $y(x_0) = y_0$ is given, then the pair of equations

$$y'(x) = f(x, y(x)), \qquad y(x_0) = y_0, \tag{100d}$$

is known as an 'initial value problem'. Our main interest in this book is with exactly this problem, where the aim is to obtain approximate values of $y(x)$

for specific values of x, usually with $x > x_0$, corresponding to the prediction of the future states of a differential equation system.

Note that for an N-dimensional system, the individual components of an initial value vector need to be given specific values. Thus, we might write

$$y_0 = [\,\eta_1 \quad \eta_2 \quad \cdots \quad \eta_N\,]^{\mathsf{T}}.$$

When the problem is formally converted to autonomous form (100c), the value of η_{N+1} must be identical to x_0, otherwise the requirement that $y_{N+1}(x)$ should always equal x would not be satisfied.

For many naturally occurring phenomena, the most appropriate form in which to express a differential equation is as a high order system. For example, an equation might be of the form

$$y^{(n)} = \phi\left(x, y, y', y'', \ldots, y^{(n-1)}\right), \tag{100e}$$

with initial values given for $y(x_0), y'(x_0), y''(x_0), \ldots, y^{(n-1)}(x_0)$. Especially important in the modelling of the motion of physical systems subject to forces are equation systems of the form

$$\begin{aligned}
y_1''(x) &= f_1(y_1, y_2, \ldots, y_N), \\
y_2''(x) &= f_2(y_1, y_2, \ldots, y_N), \\
&\vdots \qquad\qquad \vdots \\
y_N''(x) &= f_N(y_1, y_2, \ldots, y_N),
\end{aligned} \tag{100f}$$

where the equations, though second order, do have the advantages of being autonomous and without $y_1', y_2', \ldots, y_N'$ occurring amongst the arguments of $f_1, f_2, \ldots, f_N$.

To write (100f) in what will become our standard first order system form, we can introduce additional components $y_{N+1}, y_{N+2}, \ldots, y_{2N}$. The differential equation system (100f) can now be written as the first order system

$$\begin{aligned}
y_1'(x) &= y_{N+1}, \\
y_2'(x) &= y_{N+2}, \\
&\vdots \qquad\quad \vdots \\
y_N'(x) &= y_{2N}, \\
y_{N+1}'(x) &= f_1(y_1, y_2, \ldots, y_N), \\
y_{N+2}'(x) &= f_2(y_1, y_2, \ldots, y_N), \\
&\vdots \qquad\qquad \vdots \\
y_{2N}'(x) &= f_N(y_1, y_2, \ldots, y_N).
\end{aligned} \tag{100g}$$

101   The Kepler problem

The problems discussed in this section are selected from the enormous range of possible scientific applications. The first example problem describes the motion of a single planet about a heavy sun. By this we mean that, although the sun exerts a gravitational attraction on the planet, we regard the corresponding attraction of the planet on the sun as negligible, and that the sun will be treated as being stationary. This approximation to the physical system can be interpreted in another way: even though both bodies are in motion about their centre of mass, the motion of the planet relative to the sun can be modelled using the simplification we have described. We also make a further assumption, that the motion of the planet is confined to a plane.

Let $y_1(x)$ and $y_2(x)$ denote rectangular coordinates centred at the sun, specifying at time x the position of the planet. Also let $y_3(x)$ and $y_4(x)$ denote the components of velocity in the y_1 and y_2 directions, respectively. If M denotes the mass of the sun, γ the gravitational constant and m the mass of the planet, then the attractive force on the planet will have magnitude

$$\frac{\gamma M m}{y_1^2 + y_2^2}.$$

Resolving this force in the coordinate directions, we find that the components of acceleration of the planet, due to this attraction, are $-\gamma M y_1 (y_1^2 + y_2^2)^{-3/2}$ and $-\gamma M y_2 (y_1^2 + y_2^2)^{-3/2}$, where the negative sign denotes the inward direction of the acceleration.

We can now write the equations of motion:

$$\frac{dy_1}{dx} = y_3,$$

$$\frac{dy_2}{dx} = y_4,$$

$$\frac{dy_3}{dx} = -\frac{\gamma M y_1}{(y_1^2 + y_2^2)^{3/2}},$$

$$\frac{dy_4}{dx} = -\frac{\gamma M y_2}{(y_1^2 + y_2^2)^{3/2}}.$$

By adjusting the scales of the variables, the factor γM can be removed from the formulation, and we arrive at the equations

$$\frac{dy_1}{dx} = y_3, \tag{101a}$$

$$\frac{dy_2}{dx} = y_4, \tag{101b}$$

$$\frac{dy_3}{dx} = -\frac{y_1}{(y_1^2 + y_2^2)^{3/2}}, \tag{101c}$$

$$\frac{dy_4}{dx} = -\frac{y_2}{(y_1^2 + y_2^2)^{3/2}}. \tag{101d}$$

The solutions of this system are known to be conic sections, that is, ellipses, parabolas or hyperbolas, if we ignore the possibility that the trajectory is a straight line directed either towards or away from the sun. We investigate this further after we have shown that two 'first integrals', or invariants, of the solution exist.

Theorem 101A *The quantities*

$$H = \frac{1}{2}\left(y_3^2 + y_4^2\right) - (y_1^2 + y_2^2)^{-1/2},$$

$$A = y_1 y_4 - y_2 y_3,$$

are constant.

Proof. We verify that the values of dH/dx and dA/dx are zero if y satisfies (101a)–(101d). We have

$$\frac{dH}{dx} = y_3 \frac{dy_3}{dx} + y_4 \frac{dy_4}{dx} + y_1 \frac{dy_1}{dx}(y_1^2 + y_2^2)^{-3/2} + y_2 \frac{dy_2}{dx}(y_1^2 + y_2^2)^{-3/2}$$

$$= -\frac{y_1 y_3}{(y_1^2 + y_2^2)^{3/2}} - \frac{y_2 y_4}{(y_1^2 + y_2^2)^{3/2}} + \frac{y_1 y_3}{(y_1^2 + y_2^2)^{3/2}} + \frac{y_2 y_4}{(y_1^2 + y_2^2)^{3/2}}$$

$$= 0$$

and

$$\frac{dA}{dx} = y_1 \frac{dy_4}{dx} + \frac{dy_1}{dx} y_4 - y_2 \frac{dy_3}{dx} - \frac{dy_2}{dx} y_3$$

$$= -\frac{y_1 y_2}{(y_1^2 + y_2^2)^{3/2}} + y_3 y_4 + \frac{y_2 y_1}{(y_1^2 + y_2^2)^{3/2}} - y_4 y_3$$

$$= 0. \qquad \square$$

The quantities H and A are the 'Hamiltonian' and 'angular momentum', respectively. Note that $H = T + V$, where $T = \frac{1}{2}\left(y_3^2 + y_4^2\right)$ is the kinetic energy and $V = -(y_1^2 + y_2^2)^{-1/2}$ is the potential energy.

A further property of this problem is its invariance under changes of scale of the variables:

$$y_1 = \alpha^{-2}\overline{y}_1,$$

$$y_2 = \alpha^{-2}\overline{y}_2,$$

$$y_3 = \alpha\overline{y}_3,$$

$$y_4 = \alpha\overline{y}_4,$$

$$x = \alpha^{-3}\overline{x}.$$

The Hamiltonian and angular momentum get scaled to

$$\overline{H} = \frac{1}{2}\left(\overline{y}_3^2 + \overline{y}_4^2\right) - (\overline{y}_1^2 + \overline{y}_2^2)^{-1/2} = \alpha^{-2}H,$$

$$\overline{A} = \overline{y}_1\overline{y}_4 - \overline{y}_2\overline{y}_3 \qquad\qquad = \alpha A.$$

A second type of transformation is based on a two-dimensional orthogonal transformation (that is, a rotation or a reflection or a composition of these) Q, where $Q^{-1} = Q^{\mathsf{T}}$. The time variable x is invariant, and the position and velocity variables get transformed to

$$
\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} \begin{bmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \overline{y}_3 \\ \overline{y}_4 \end{bmatrix}.
$$

It is easy to see that $A = 0$ implies that the trajectory lies entirely in a subspace defined by $\cos(\theta)y_1 = \sin(\theta)y_2$, $\cos(\theta)y_3 = \sin(\theta)y_4$ for some fixed angle θ. We move on from this simple case and assume that $A \neq 0$. The sign of H is of crucial importance: if $H \geq 0$ then it is possible to obtain arbitrarily high values of $y_1^2 + y_2^2$ without $y_3^2 + y_4^2$ vanishing. We exclude this case for the present discussion and assume that $H < 0$. Scale H so that it has a value $-\frac{1}{2}$ and at the same time A takes on a positive value. This value cannot exceed 1 because we can easily verify an identity involving the derivative of $r = \sqrt{y_1^2 + y_2^2}$. This identity is

$$
\left(r \frac{dr}{dx} \right)^2 = 2Hr^2 + 2r - A^2 = -r^2 + 2r - A^2. \tag{101e}
$$

Since the left-hand side cannot be negative, the quadratic function in r on the right-hand side must have real roots. This implies that $A \leq 1$. Write $A = \sqrt{1 - e^2}$, for $e \geq 0$, where we see that e is the eccentricity of an ellipse on which the orbit lies. The minimum and maximum values of r are found to be $1 - e$ and $1 + e$, respectively. Rotate axes so that when $r = 1 - e$, which we take as the starting point of time, $y_1 = 1 - e$ and $y_2 = 0$. At this point we find that $y_3 = 0$ and $y_4 = \sqrt{(1+e)/(1-e)}$.

Change to polar coordinates by writing $y_1 = r\cos(\theta)$, $y_2 = r\sin(\theta)$. It is found that

$$
y_3 = \frac{dy_1}{dx} = \frac{dr}{dx}\cos(\theta) - r\frac{d\theta}{dx}\sin(\theta),
$$

$$
y_4 = \frac{dy_2}{dx} = \frac{dr}{dx}\sin(\theta) + r\frac{d\theta}{dx}\cos(\theta),
$$

so that, because $y_1 y_4 - y_2 y_3 = \sqrt{1 - e^2}$, we find that

$$
r^2 \frac{d\theta}{dx} = \sqrt{1 - e^2}. \tag{101f}
$$

From (101e) and (101f) we find a differential equation for the path traced out by the orbit

$$
\left(\frac{dr}{d\theta} \right)^2 = \frac{1}{1 - e^2} r^2 \left(e^2 - (1 - r)^2 \right),
$$

and we can verify that this is satisfied by

$$\frac{1 - e^2}{r} = 1 + e\cos(\theta).$$

If we change back to Cartesian coordinates, we find that all points on the trajectory lie on the ellipse

$$(y_1 + e)^2 + \frac{y_2^2}{1 - e^2} = 1,$$

with centre $(-e, 0)$, eccentricity e, and major and minor axis lengths 1 and $\sqrt{1 - e^2}$ respectively.

As we have seen, a great deal is known about this problem. However, much less is known about the motion of a many-body gravitational system.

One of the aims of modern numerical analysis is to understand the behaviour of various geometrical properties. In some cases it is possible to preserve the value of quantities that are invariant in the exact solution. In other situations, such as problems where the Hamiltonian is theoretically conserved, it may be preferable to conserve other properties, such as what is known as 'symplectic behaviour'.

We consider further gravitational problems in Subsection 120.

102 A problem arising from the method of lines

The second initial value problem we consider is based on an approximation to a partial differential equation. Consider the parabolic system

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad (x, t) \in [0, 1] \times [0, \infty), \qquad (102a)$$

where we have used t to represent time, x to represent distance and $u(x, t)$ to represent some quantity, such as temperature, which diffuses with time. For this problem it is necessary to impose conditions on the boundaries $x = 0$ and $x = 1$ as well as at the initial time $t = 0$. We may interpret the solution as the distribution of the temperature at points in a conducting rod, given that the temperature is specified at the ends of the rod. In this case the boundary conditions would be of the form $u(0, t) = \alpha(t)$ and $u(1, t) = \beta(t)$. Equation (102a) is known as the heat or diffusion equation, and the conditions given at $x = 0$ and $x = 1$ are known as Dirichlet boundary values. This is in contrast to Neumann conditions, in which the values of $\partial u / \partial x$ are given at the ends of the x interval.

To convert this problem into an ordinary differential equation system, which mimics the behaviour of the parabolic equation, let $y_1(t), y_2(t), \ldots, y_N(t)$, denote the values of $u(\frac{1}{N+1}, t), u(\frac{2}{N+1}, t), \ldots, u(\frac{N}{N+1}, t)$, respectively. That is,

$$y_j(t) = u\left(\frac{j}{N+1}, t\right), \qquad j = 0, 1, 2, \ldots, N + 1,$$

where we have included $y_0(t) = u(0, t)$, $y_{N+1}(t) = u(1, t)$ for convenience.

For $j = 1, 2, \ldots, N$, $\partial^2 u/\partial x^2$, evaluated at $x = j/(N+1)$, is approximately equal to $(N + 1)^2(y_{j-1} - 2y_j + y_{j+1})$. Hence, the vector of derivatives of $y_1, y_2, \ldots, y_N$ is given by

$$\frac{dy_1(t)}{dt} = (N + 1)^2\big(\alpha(t) - 2y_1(t) + y_2(t)\big),$$

$$\frac{dy_2(t)}{dt} = (N + 1)^2\big(y_1(t) - 2y_2(t) + y_3(t)\big),$$

$$\frac{dy_3(t)}{dt} = (N + 1)^2\big(y_2(t) - 2y_3(t) + y_4(t)\big),$$

$$\vdots \qquad\qquad \vdots$$

$$\frac{dy_{N-1}(t)}{dt} = (N + 1)^2\big(y_{N-2}(t) - 2y_{N-1}(t) + y_N(t)\big),$$

$$\frac{dy_N(t)}{dt} = (N + 1)^2\big(y_{N-1}(t) - 2y_N(t) + \beta(t)\big).$$

This system can be written in vector–matrix form as

$$y'(t) = Ay(t) + v(t), \tag{102b}$$

where

$$A = (N+1)^2\begin{bmatrix} -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -2 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -2 \end{bmatrix}, \quad v = (N+1)^2\begin{bmatrix} \alpha(t) \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \beta(t) \end{bmatrix}.$$

The original problem is 'dissipative' in the sense that, if u and v are each solutions to the diffusion equation, which have identical boundary values but different initial values, then

$$W(t) = \frac{1}{2}\int_0^1 \big(u(x, t) - v(x, t)\big)^2 dx$$

is non-increasing as t increases. We can verify this by differentiating with respect to t and by showing, using integration by parts, that the result found

cannot be positive. We have

$$
\begin{aligned}
\frac{dW}{dt} &= \int_0^1 \Big(u(x,t) - v(x,t)\Big)\left(\frac{\partial u(x,t)}{\partial t} - \frac{\partial v(x,t)}{\partial t}\right) dx \\
&= \int_0^1 \Big(u(x,t) - v(x,t)\Big)\left(\frac{\partial^2 u(x,t)}{\partial x^2} - \frac{\partial^2 v(x,t)}{\partial x^2}\right) dx \\
&= \left[\Big(u(x,t) - v(x,t)\Big)\left(\frac{\partial u(x,t)}{\partial x} - \frac{\partial v(x,t)}{\partial x}\right)\right]_0^1 \\
&\quad - \int_0^1 \left(\frac{\partial u(x,t)}{\partial x} - \frac{\partial v(x,t)}{\partial x}\right)^2 dx \\
&= -\int_0^1 \left(\frac{\partial u(x,t)}{\partial x} - \frac{\partial v(x,t)}{\partial x}\right)^2 dx \\
&\leq 0.
\end{aligned}
$$

Even though the approximation of (102a) by (102b) is not exact, it is an advantage of the discretization we have used, that the qualitative property is still present. Let y and z be two solutions to the ordinary differential equation system. Consider the nature of

$$
\widehat{W}(t) = \frac{1}{2}\sum_{j=1}^{N}(y_j - z_j)^2.
$$

We have

$$
\begin{aligned}
\frac{d\widehat{W}}{dt} &= \sum_{i=1}^{N}(y_j - z_j)\left(\frac{dy_j}{dt} - \frac{dz_j}{dt}\right) \\
&= (N+1)^2\sum_{j=1}^{N}(y_j - z_j)\,(y_{j-1} - 2y_j + y_{j+1} - z_{j-1} + 2z_j - z_{j+1}) \\
&= 2(N+1)^2\sum_{j=1}^{N-1}(y_j - z_j)(y_{j+1} - z_{j+1}) - 2(N+1)^2\sum_{j=1}^{N}(y_j - z_j)^2 \\
&= -(N+1)^2\sum_{j=0}^{N}(y_{j+1} - y_j - z_{j+1} + z_j)^2 \\
&\leq 0.
\end{aligned}
$$

Another aspect of the discretization that might be explored is the spectrum of the matrix A, in comparison with the spectrum of the linear operator $u \mapsto \frac{d^2 u}{dx^2}$ on the space of C^2 functions on $[0,1]$ for which $u(0) = u(1) = 0$. The eigenfunctions for the continuous problem are of the form $\sin(k\pi x)$, for

$k = 1, 2, 3, \ldots$, and the corresponding eigenvalues are $-k^2\pi^2$. For the discrete problem, we need to find the solutions to the problem

$$(A - \lambda I) \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix} = 0, \tag{102c}$$

where v_1, v_2, $\ldots$, v_N are not all zero. Introducing also $v_0 = v_{N+1} = 0$, we find that it is possible to write (102c) in the form

$$v_{j-1} - qv_j + v_{j+1} = 0, \qquad j = 1, 2, \ldots, N, \tag{102d}$$

where $q = 2 + \lambda/(N+1)^2$. The difference equation (102d) has solution of the form $v_i = C(\mu^i - \mu^{-i})$, where $\mu + \mu^{-1} = q$, unless $q = \pm 2$ (which is easily seen to be impossible). Because $v_{N+1} = 0$, it follows that $\lambda^{2N+2} = 2$. Because $\mu \neq \pm 1$, it follows that

$$\mu = \exp\left(\frac{k\pi i}{N+1}\right), \qquad k = 1, 2, \ldots, N,$$

with $i = \sqrt{-1}$. Hence,

$$\lambda = -2(N+1)^2 \left(1 - \cos\left(\frac{k\pi}{N+1}\right)\right) = -4(N+1)^2 \sin^2\left(\frac{k\pi}{2N+2}\right).$$

For N much larger than k, we can use the approximation $\sin(\xi) \approx \xi$, for small ξ, to give eigenvalue number k as $\lambda_k \approx -k^2\pi^2$. On the other hand, for k small, the eigenvalue number $N + 1 - k$ is $\lambda_{N+1-k} \approx -4(N+1)^2 + k^2\pi^2$.

103   The simple pendulum

Formulation as a differential-algebraic equation

Consider a small mass m attached to a light inelastic string of length l, with the other end attached to the origin of coordinates, which can swing back and forth in a vertical plane. Let X, measured in a rightwards direction, and Y, measured in a downward direction, be the coordinates. Because the string is inelastic, the tension T in the string always matches other forces resolved in the direction of the string so as to guarantee that the length does not change.

The way these forces act on the mass is shown in Figure 103(i). Also shown is the angle θ defined by $X = l\sin(\theta)$, $Y = l\cos(\theta)$.

We denote by U and V, respectively, the velocity components in the X and
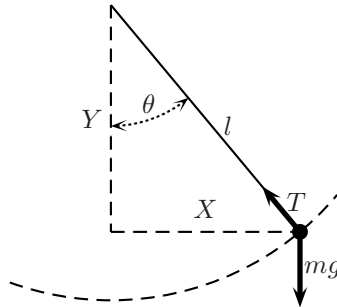
Figure 103(i)   Simple pendulum

Y directions. The motion of the pendulum is governed by the equations

$$\frac{dX}{dx} = U, \tag{103a}$$

$$\frac{dY}{dx} = V, \tag{103b}$$

$$m\frac{dU}{dx} = -\frac{TX}{l}, \tag{103c}$$

$$m\frac{dV}{dx} = -\frac{TY}{l} + mg, \tag{103d}$$

$$X^2 + Y^2 = l^2, \tag{103e}$$

where, in addition to four differential equations (103a)–(103d), the constraint (103e) expresses the constancy of the length of the string. The tension T acts as a control variable, forcing this constraint to remain satisfied. By rescaling variables in a suitable way, the 'differential-algebraic' equation system (103a)–(103e) can be rewritten with the constants m, g and l replaced by 1 in each case. In the rescaled formulation write $y_1 = X$, $y_2 = Y$, $y_3 = U$, $y_4 = V$ and $y_5 = T$, and we arrive at the system

$$\frac{dy_1}{dx} = y_3, \tag{103f}$$

$$\frac{dy_2}{dx} = y_4, \tag{103g}$$

$$\frac{dy_3}{dx} = -y_1 y_5, \tag{103h}$$

$$\frac{dy_4}{dx} = -y_2 y_5 + 1, \tag{103i}$$

$$y_1^2 + y_2^2 = 1. \tag{103j}$$

It will be convenient to choose initial values defined in terms of $\theta = \Theta$, with

the velocity equal to zero. That is,

$$y_1(0) = \sin(\Theta), \quad y_2(0) = \cos(\Theta), \quad y_3(0) = y_4(0) = 0, \quad y_5(0) = \cos(\Theta).$$

The five variables are governed by four differential equations (103f)–(103i), together with the single algebraic constraint (103j). We will say more about this below, but first we consider the classical way of simplifying the problem.

Formulation as a single second order equation

Make the substitutions $y_1 = \sin(\theta)$, $y_2 = \cos(\theta)$. Because (103j) is automatically satisfied, the value of y_5 loses its interest and we eliminate this by taking a linear combination of (103h) and (103i). This gives the equation system

$$\cos(\theta)\frac{d\theta}{dx} = y_3, \tag{103k}$$

$$-\sin(\theta)\frac{d\theta}{dx} = y_4, \tag{103l}$$

$$-\cos(\theta)\frac{dy_3}{dx} + \sin(\theta)\frac{dy_4}{dx} = \sin(\theta). \tag{103m}$$

Differentiate (103k) and (103l) and substitute into (103m) and we obtain the well-known single-equation formulation of the simple pendulum:

$$\frac{d^2\theta}{dx^2} + \sin(\theta) = 0, \tag{103n}$$

with initial values
$$\theta(0) = \Theta, \qquad \theta'(0) = 0.$$

It can be shown that the period of the pendulum is given by

$$T = 4\int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - \sin^2\phi \sin^2\frac{\Theta}{2}}}$$

and some values are given in Table 103(I).

The value for $0°$ can be interpreted as the period for small amplitudes. The fact that T increases slowly as Θ increases is the characteristic property of a simple pendulum which makes it of practical value in measuring time.

Formulation as a Hamiltonian problem

In the formulation (103n), write the H as the 'Hamiltonian'

$$H(p, q) = \tfrac{1}{2}p^2 - \cos(q),$$

Table 103(I)   Period of simple pendulum for various amplitudes

Θ	T
0°	6.2831853072
3°	6.2842620831
6°	6.2874944421
9°	6.2928884880
12°	6.3004544311
15°	6.3102066431
18°	6.3221637356
21°	6.3363486630
24°	6.3527888501
27°	6.3715163462
30°	6.3925680085

where $q = \theta$ and $p = d\theta/dx$. The second order equation (103n) is now equivalent to the first order system

$$
\begin{bmatrix} p' \\ q' \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial H}{\partial p} \\ \frac{\partial H}{\partial q} \end{bmatrix}.
$$

Differential index and index reduction

Carry out three steps, of which the first is to differentiate (103j) and substitute from (103f) and (103g) to give the result

$$y_1 y_3 + y_2 y_4 = 0. \tag{103o}$$

The second step is to differentiate (103o) and to make various substitutions from (103f)–(103i) to arrive at the equation

$$y_2 + y_3^2 + y_4^2 - y_5 = 0. \tag{103p}$$

The third and final step is to differentiate (103p) and make various substitutions to arrive at the result

$$
\frac{dy_5}{dx} = \frac{y_2}{dx} + 2y_3 \frac{dy_3}{dx} + 2y_4 \frac{dy_4}{dx} = y_4 + 2y_3(-y_1 y_5) + 2y_4(-y_2 y_5 + 1),
$$

which simplifies to

$$\frac{dy_5}{dx} = 3y_4. \tag{103q}$$

Given that consistent initial values are used, it seems that the equations (103f)–(103i) together with any of (103j), (103o), (103p) or (103q) give identical solutions.

Which of the possible formulations *should* be used? From the point of view of physical modelling, it seems to be essential to require that the length constraint (103j) should hold exactly. On the other hand, when it comes to numerical approximations to solutions, it is found that the use of this constraint in the problem description creates serious computational difficulties. It also seems desirable from a modelling point of view to insist that (103o) should hold exactly, since this simply states that the direction of motion is tangential to the arc on which it is constrained to lie.

104   A chemical kinetics problem

We next consider a model of a chemical process consisting of three species, which we denote by A, B and C. The three reactions are

$$A \to B, \tag{104a}$$

$$B + C \to A + C, \tag{104b}$$

$$B + B \to C. \tag{104c}$$

Let y_1, y_2 and y_3 denote the concentrations of A, B and C, respectively. We assume these are scaled so that the total of the three concentrations is 1, and that each of three constituent reactions will add to the concentration of any of the species exactly at the expense of corresponding amounts of the reactants. The reaction rate of (104a) will be denoted by k_1. This means that the rate at which y_1 decreases, and at which y_2 increases, because of this reaction, will be equal to $k_1 y_1$. In the second reaction (104b), C acts as a catalyst in the production of A from B and the reaction rate will be written as k_2, meaning that the increase of y_1, and the decrease of y_3, in this reaction will have a rate equal to $k_2 y_2 y_3$. Finally, the production of C from B will have a rate constant equal to k_3, meaning that the rate at which this reaction takes place will be $k_3 y_2^2$. Putting all these elements of the process together, we find the system of differential equations for the variation with time of the three concentrations to be

$$\frac{dy_1}{dx} = -k_1 y_1 + k_2 y_2 y_3, \tag{104d}$$

$$\frac{dy_2}{dx} = k_1 y_1 - k_2 y_2 y_3 - k_3 y_2^2, \tag{104e}$$

$$\frac{dy_3}{dx} = k_3 y_2^2. \tag{104f}$$

If the three reaction rates are moderately small numbers, and not greatly different in magnitude, then this is a straightforward problem. However,

vastly different magnitudes amongst k_1, k_2 and k_3 can make this problem complicated to understand as a chemical model. Also, as we shall see, the problem then becomes difficult to solve numerically. This problem was popularized by Robertson (1966), who used the reaction rates

$$k_1 = 0.04, \qquad k_2 = 10^4, \qquad k_3 = 3 \times 10^7.$$

Before looking at the problem further we note that, even though it is written as a three-dimensional system, it would be a simple matter to rewrite it in two dimensions, because $y_1 + y_2 + y_3$ is an invariant and is usually set to a value of 1, by an appropriate choice of the initial values. We always assume this value for $y_1 + y_2 + y_3$. Furthermore, if the initial value has non-negative values for each of the three components, then this situation is maintained for all positive times. To see why this is the case, write (104d), (104e) and (104f) in the forms

$$\frac{d(\exp(k_1 x)y_1)}{dx} = \exp(k_1 x)k_2 y_2 y_3,$$

$$\frac{d\big(\exp\big(\max(k_2, k_3)x\big)y_2\big)}{dx} = \exp\big(\max(k_2, k_3)x\big)F,$$

$$\frac{dy_3}{dx} = k_3 y_2^2,$$

where

$$F = k_1 y_1 + \max(k_2, k_3)y_1 y_2 + \big(\max(k_2, k_3) - k_2\big)y_2 y_3 + \big(\max(k_2, k_3) - k_3\big)y_2^2,$$

so that each of $\exp(k_1 x)y_1$, $\exp(\max(k_2, k_3)x)y_2$ and y_3 is non-decreasing.

An interesting feature of this problem is that a small perturbation that does not disturb the invariance of $y_1 + y_2 + y_3$ is damped out rapidly. To see why this is the case, eliminate y_1 so that the differential equation system in the remaining two components becomes

$$\frac{dy_2}{dx} = k_1(1 - y_2 - y_3) - k_2 y_2 y_3 - k_3 y_2^2, \qquad (104\text{g})$$

$$\frac{dy_3}{dx} = k_3 y_2^2. \qquad (104\text{h})$$

The Jacobian matrix, the matrix of partial derivatives, is given by

$$J(x) = \begin{bmatrix} -k_1 - k_2 y_3 - 2k_3 y_2 & -k_1 - k_2 y_2 \\ 2k_3 y_2 & 0 \end{bmatrix},$$

and the characteristic polynomial is

$$\lambda^2 + (k_1 + k_2 y_3 + 2k_3 y_2)\lambda + 2k_3 y_2(k_1 + k_2 y_2). \qquad (104\text{i})$$
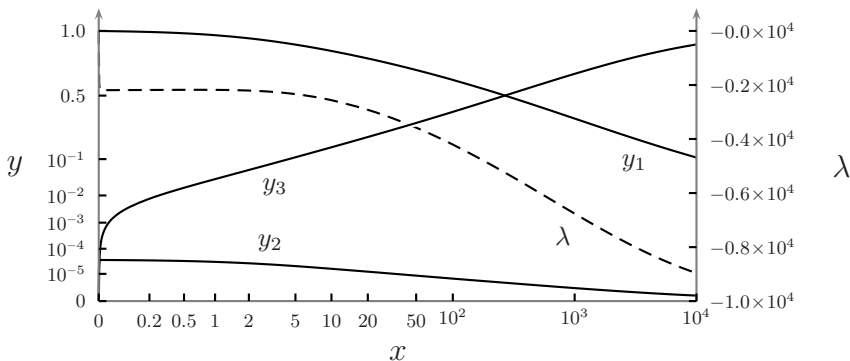
Figure 104(i)   Solution and most negative eigenvalue for the Robertson problem

An analysis of the discriminant of (104i) indicates that for $y_2, y_3 \in (0, 1]$, both zeros are real and negative. Along the actual trajectory, one of the eigenvalues of $J(x)$, denoted by λ, rapidly jumps to a very negative value, with the second eigenvalue retaining a small negative value. Consider a small perturbation z to the solution, so that the solution becomes $y + z$. Because the two components of z are small we can approximate $f(y + z)$ by $f(y) + (\partial f / \partial y)z$. Hence, the perturbation itself satisfies the equation

$$\left[\begin{array}{c} \dfrac{dz_2}{dx} \\ \dfrac{dz_3}{dx} \end{array} \right] = J(x) \left[\begin{array}{c} z_2 \\ z_3 \end{array} \right]$$

and the negative eigenvalues of $J(x)$ guarantee the decay of the components of z.

The solution to this problem, together with the value of λ, is shown in Figure 104(i).

105   The Van der Pol equation and limit cycles

The simple pendulum, which we considered in Subsection 103, is a non-linear variant of the 'harmonic oscillator' problem $y'' = -y$. We now consider another non-linear generalization of this problem, by adding a term $\mu(1 - y^2)y'$, where μ is a positive constant, to obtain the 'Van der Pol equation'

$$y''(x) = \mu(1 - y(x)^2)y'(x) - y(x).$$

This problem was originally introduced by Van der Pol (1926) in the study of electronic circuits. If μ is small and the initial values correspond to what would be oscillations of amplitude less than 1, if μ had in fact been zero, it might be expected that the values of $y(x)$ would remain small for all time.
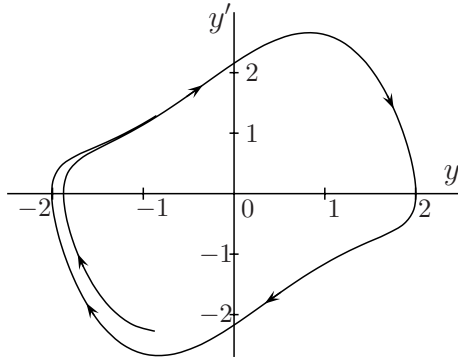
Figure 105(i)   Van der Pol problem with $\mu = 1$
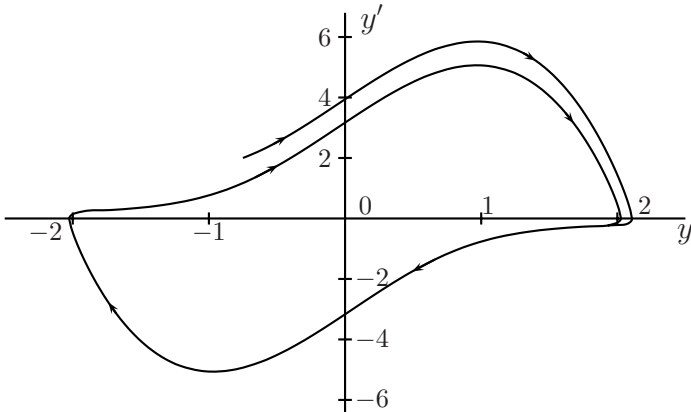
Figure 105(ii)   Van der Pol problem with $\mu = 3$

However, the non-linear term has the effect of injecting more 'energy' into the system, as we see by calculating the rate of change of $E = \frac{1}{2}y'(x)^2 + \frac{1}{2}y(x)^2$. This is found to be

$$\frac{d}{dx}\left(\frac{1}{2}y'(x)^2 + \frac{1}{2}y(x)^2\right) = \mu(1 - y(x)^2)y'(x)^2 > 0,$$

as long as $|y| < 1$.

Similarly, if $|y|$ starts with a high value, then E will decrease until $|y| = 1$. It is possible to show that the path, traced out in the (y, y') plane, loops round the origin in a clockwise direction forever, and that it converges to a 'limit cycle' – a periodic orbit. In Figure 105(i), this is illustrated for $\mu = 1$. The path traced out in the (y, y') plane moves rapidly towards the limit cycle and
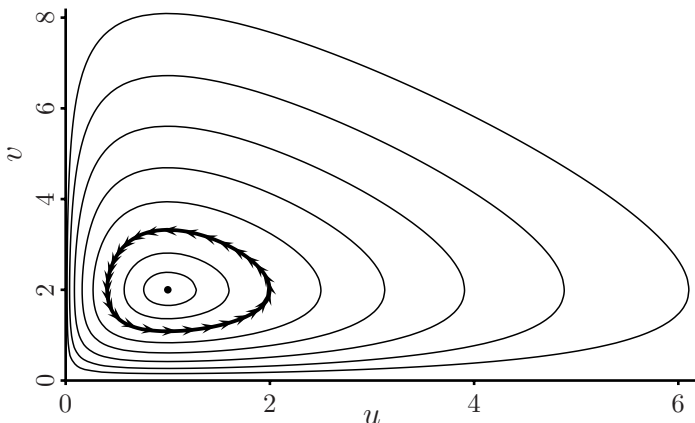
Figure 106(i)   Phase diagram for Lotka–Volterra solution with $(u_0, v_0) = (2, 2)$,
together with seven alternative orbits

is soon imperceptibly close to it. In Figure 105(ii), the case $\mu = 3$ is presented.

Of special interest in this problem, especially for large values of μ, is the fact that numerical methods attempting to solve this problem need to adjust their behaviour to take account of varying conditions, as the value of $1 - |y(x)|^2$ changes. The sharp change of direction of the path traced out near $(y, y') = (\pm 2, 0)$ for the $\mu = 3$ case, a phenomenon which becomes more pronounced as μ is further increased, is part of the numerical difficulty associated with this problem.

106   The Lotka–Volterra problem and periodic orbits

In the modelling of the two-species 'predator–prey' problem, differential equation systems of the following type arise:

$$u' = u(2 - v), \tag{106a}$$
$$v' = v(u - 1), \tag{106b}$$

where the factors $2 - v$ and $u - 1$ can be generalized in various ways. This model was proposed independently by Lotka (1925) and Volterra (1926). The two variables represent the time-dependent populations, of which v is the population of predators which feed on prey whose population is denoted by u. It is assumed that u would have been able to grow exponentially without limit, if the predator had not been present, and that the factor $2 - v$ represents the modification to its growth rate because of harvesting by the predator. The predator in turn, in the absence of prey, would die out exponentially, and requires at least a prey population of $u = 1$ to feed upon to be able to grow. Of the two stationary solutions, $(u, v) = (0, 0)$ and $(u, v) = (1, 2)$, the second

Table 106(I)   Approximations to the period T, given by (106d) for
$(u_0, v_0) = (2, 2)$

n	Approximate integral
10	4.62974838287860
20	4.61430252126987
40	4.61487057379480
80	4.61487051945097
160	4.61487051945103
320	4.61487051945103

is more interesting because small perturbations from this point will lead to periodic orbits around the stationary point. By dividing (106a) by (106b), we obtain a differential equation for the path traced out by (u, v). The solution is that $I(u, v)$ is constant, where

$$I(u, v) = \log(u) + 2\log(v) - u - v.$$

It is interesting to try to calculate values of the period T, for a given starting point (u_0, v_0). To calculate T, change to polar coordinates centred at the stationary point

$$u = 1 + r\cos(\theta), \qquad v = 2 + r\sin(\theta)$$

and calculate the integral $\int_0^{2\pi} \phi(\theta)d\theta$, where

$$\phi(\theta) = \frac{1}{v\cos^2(\theta) + u\sin^2(\theta)}. \tag{106c}$$

Starting values $(u_0, v_0) = (2, 2)$ lead to the orbit featured in Figure 106(i). Orbits with various other starting values are also shown. The period, based on the integral of (106c), has been calculated with a varying number n of equally spaced values of $\theta \in [0, 2\pi]$, using the trapezoidal rule. It is known that for certain smooth functions, the error of this type of calculation will behave, not like a power of n^{-1}, but like $\exp(-\alpha n)$, for some problem-specific parameter α. This super-convergence is evidently realized for the present problem, where the observed approximations

$$T = \int_0^{2\pi} \phi(\theta)d\theta \approx \frac{2\pi}{n}\sum_{k=0}^{n-1} \phi\left(\frac{2\pi k}{n}\right) \tag{106d}$$

are shown in Table 106(I) for $n = 10, 20, 40, \ldots, 320$. Evidently, to full machine accuracy, the approximations have converged to $T = 4.61487051945103$. An

Algorithm 106α   Computation of orbit and period for the Lotka–Volterra
problem

```
theta = linspace(0,2*pi,n+1);
co = cos(theta);
si = sin(theta);
C = u0*v0 2*exp(-u0-v0);
r = ones(size(theta));
u = 1+r.*co;
v = 2+r.*si;
carryon=1;
while carryon
   f = u.*v. 2-C*exp(u+v);
   df = -v.*r.*(v.*co. 2+u.*si. 2);
   dr = f./df;
   r = r-dr;
   u = 1+r.*co;
   v = 2+r.*si;
   carryon = norm(dr,inf) > 0.000000001;
end
phi = 1./(v.*co. 2+u.*si. 2);
period = (2*pi/n)*sum(phi(1:n));
```

explanation of the phenomenon of rapid convergence of the trapezoidal rule for
periodic functions can be found in Davis and Rabinowitz (1984), and in papers
referenced in that book.

In Algorithm 106α, MATLAB statements are presented to carry out the
computations that were used to generate Figure 106(i) and Table 106(I). To
compute the value of r for each θ, the equation $f(r) = 0$ is solved, where

$$f(r) = (\exp(I(u,v)) - C)\exp(u+v) = uv^2 - C\exp(u+v),$$

with $C = u_0 v_0^2 \exp(-u_0 - v_0)$. Note that the statement u.v. 2-C*exp(u+v)
evaluates a vector with element number i equal to $u_i v_i^2 - C \exp(u_i + v_i)$,
and that linspace(0,2*pi,n+1) generates a vector with $n+1$ components,
equally spaced in $[0, 2\pi]$.

107   The Euler equations of rigid body rotation

For a rigid body on which no moments are acting, the three components of
angular velocity, in terms of the principal directions of inertia fixed in the

body, satisfy the Euler equations:

$$I_1 \frac{dw_1}{dt} = (I_2 - I_3)w_2 w_3,$$

$$I_2 \frac{dw_2}{dt} = (I_3 - I_1)w_3 w_1, \quad\quad (107a)$$

$$I_3 \frac{dw_3}{dt} = (I_1 - I_2)w_1 w_2,$$

where the 'principal moments of inertia' I_1, I_2 and I_3 are positive. Denote the kinetic energy by $\frac{1}{2}E$ and the squared norm of the angular momentum by F. That is,

$$E = I_1 w_1^2 + I_2 w_2^2 + I_3 w_3^2, \quad\quad (107b)$$

$$F = I_1^2 w_1^2 + I_2^2 w_2^2 + I_3^2 w_3^2. \quad\quad (107c)$$

Differentiate these expressions and substitute in dw_i/dt, $i = 1, 2, 3$, to obtain a zero result in each case. Hence, E and F are invariants of the solution to (107a). This observation provides useful tests on numerical methods for this problem because there is in general no reason why these invariants should be maintained in a numerical approximation.

Exercises 10

10.1 You are given the initial value problem

$$u'''(x) - 3u''(x) + 2u(x)u'(x) = 0, \quad u(1) = 2, \ u'(1) = -1, \ u''(1) = 4.$$

Show how to reformulate this problem in the form

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

where $f : \mathbb{R}^3 \to \mathbb{R}^3$.

10.2 You are given the non-autonomous initial value problem

$$
\begin{aligned}
u' &= xu + x^2 v, & u(0) &= 3, \\
v' &= u - v + 2xw, & v(0) &= 2, \\
w' &= u + \frac{v}{1+x}, & w(0) &= 5.
\end{aligned}
$$

Show how to write this as an autonomous problem.

10.3 The matrix

$$A = (N-1)^2 \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -2 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{bmatrix}$$

arises in the numerical solution of the heat equation, but with Neumann boundary conditions. Find the eigenvalues of A.

10.4 Calculate the period of an orbit of the Lotka–Volterra problem which passes through the point $(3, 2)$.

11   Differential Equation Theory

110   Existence and uniqueness of solutions

A fundamental question that arises in scientific modelling is whether a given differential equation, together with initial conditions, can be reliably used to predict the behaviour of the trajectory at later times. We loosely use the expression 'well-posed' to describe a problem that is acceptable from this point of view. The three attributes of an initial value problem that have to be taken into account are whether there actually exists a solution, whether the solution, if it exists, is unique, and how sensitive the solution is to small perturbations to the initial information. Even though there are many alternative criteria for answering these questions in a satisfactory manner, we focus here on the existence of a Lipschitz condition. This is especially convenient because the same type of condition can be used to study the behaviour of numerical approximations.

Definition 110A *The function $f : [a, b] \times \mathbb{R}^N \rightarrow \mathbb{R}^N$ is said to satisfy a 'Lipschitz condition in its second variable' if there exists a constant L, known as a 'Lipschitz constant', such that for any $x \in [a, b]$ and $Y, Z \in \mathbb{R}^N$, $\|f(x, Y) - f(x, Z)\| \leq L\|Y - Z\|$.*

We need a basic lemma on metric spaces known as the 'contraction mapping principle'. We present this without proof.

Lemma 110B *Let M denote a complete metric space with metric ρ and let $\phi : M \rightarrow M$ denote a mapping which is a contraction, in the sense that there exists a number k, satisfying $0 \leq k < 1$, such that, for any $\eta, \zeta \in M$, $\rho(\phi(\eta), \phi(\zeta)) \leq k\rho(\eta, \zeta)$. Then there exists a unique $\xi \in M$ such that $\phi(\xi) = \xi$.*

We can now state our main result.

Theorem 110C *Consider an initial value problem*

$$y'(x) = f(x, y(x)), \tag{110a}$$
$$y(a) = y_0, \tag{110b}$$

where $f : [a, b] \times \mathbb{R}^N \to \mathbb{R}^N$ is continuous in its first variable and satisfies a Lipschitz condition in its second variable. Then there exists a unique solution to this problem.

Proof. Let M denote the complete metric space of continuous functions $y : [a, b] \to \mathbb{R}^N$, such that $y(a) = y_0$. The metric is defined by

$$\rho(y, z) = \sup_{x \in [a,b]} \exp(-K(x - a)) \|y(x) - z(x)\|,$$

where $K > L$. For given $y \in M$, define $\phi(y)$ as the solution Y on $[a, b]$ to the initial value problem

$$Y'(x) = f(x, Y(x)),$$
$$Y(a) = y_0.$$

This problem is solvable by integration as

$$\phi(y)(x) = y_0 + \int_a^x f(s, y(s)) ds.$$

This is a contraction because for any two $y, z \in M$, we have

$$\rho(\phi(y), \phi(z)) \leq \sup_{x \in [a,b]} \exp(-K(x - a)) \left\| \int_a^x \big(f(s, y(s)) - f(s, z(s))\big) ds \right\|$$

$$\leq \sup_{x \in [a,b]} \exp(-K(x - a)) \int_a^x \|f(s, y(s)) - f(s, z(s))\| \, ds$$

$$\leq L \sup_{x \in [a,b]} \exp(-K(x - a)) \int_a^x \|y(s) - z(s)\| \, ds$$

$$\leq L\rho(y, z) \sup_{x \in [a,b]} \exp(-K(x - a)) \int_a^x \exp(K(s - a)) ds$$

$$\leq \frac{L}{K} \rho(y, z).$$

The unique function y that therefore exists satisfying $\phi(y) = y$, is evidently the unique solution to the initial value problem given by (110a), (110b). $\quad\square$

The third requirement for being well-posed, that the solution is not overly sensitive to the initial condition, can be readily assessed for problems satisfying

a Lipschitz condition. If y and z each satisfy (110a) with $y(a) = y_0$ and $z(a) = z_0$, then

$$\frac{d}{dx}\|y(x) - z(x)\| \leq L\|y(x) - z(x)\|.$$

Multiply both sides by $\exp(-Lx)$ and deduce that

$$\frac{d}{dx}\big(\exp(-Lx)\|y(x) - z(x)\|\big) \leq 0,$$

implying that

$$\|y(x) - z(x)\| \leq \|y_0 - z_0\| \exp\big(L(x - a)\big). \tag{110c}$$

This bound on the growth of initial perturbations may be too pessimistic in particular circumstances. Sometimes it can be improved upon by the use of 'one-sided Lipschitz conditions'. This will be discussed in Subsection 112.

111   Linear systems of differential equations

Linear differential equations are important because of the availability of a superposition principle. That is, it is possible for a linear differential equation system to combine known solutions to construct new solutions. The standard form of a linear system is

$$\frac{dy}{dx} = A(x)y + \phi(x), \tag{111a}$$

where $A(x)$ is a possibly time-dependent linear operator. The corresponding 'homogeneous' system is

$$\frac{dy}{dx} = A(x)y. \tag{111b}$$

The superposition principle, which is trivial to verify, states that:

Theorem 111A *If $\widehat{y}$ is a solution to (111a) and $y_1, y_2, \ldots, y_k$ are solutions to (111b), then for any constants $\alpha_1, \alpha_2, \ldots, \alpha_k$, the function y given by*

$$y(x) = \widehat{y}(x) + \sum_{i=1}^{k} \alpha_i y_i(x),$$

is a solution to (111a).

The way this result is used is to attempt to find the solution which matches a given initial value, by combining known solutions.

Many linear problems are naturally formulated in the form of a single high order differential equation

$$Y^{(m)}(x) - C_1(x)Y^{(m-1)}(x) - C_2(x)Y^{(m-2)}(x) - \cdots - C_m(x)Y(x) = g(x). \tag{111c}$$

By identifying $Y(x) = y_1(x), Y'(x) = y_2(x), \ldots, Y^{(m-1)} = y_m(x)$, we can rewrite the system in the form

$$\frac{d}{dx}\begin{bmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_m(x) \end{bmatrix} = A(x)\begin{bmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_m(x) \end{bmatrix} + \phi(x),$$

where the 'companion matrix' $A(x)$ and the 'inhomogeneous term' $\phi(x)$ are given by

$$A(x) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ C_m(x) & C_{m-1}(x) & C_{m-2}(x) & \cdots & C_1(x) \end{bmatrix}, \quad \phi(x) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ g(x) \end{bmatrix}.$$

When $A(x) = A$ in (111b) is constant, then to each eigenvalue λ of A, with corresponding eigenvector v, there exists a solution given by

$$y(x) = \exp(\lambda x)v. \tag{111d}$$

When a complete set of eigenvectors does not exist, but corresponding to λ there is a chain of generalized eigenvectors

$$Av_1 = \lambda v_1 + v, \quad Av_2 = \lambda v_2 + v_1, \quad \ldots, \quad Av_{k-1} = \lambda v_{k-1} + v_{k-2},$$

then there is a chain of additional independent solutions to append to (111d):

$$y_1 = x\exp(\lambda x)v_1, \quad y_2 = x^2\exp(\lambda x)v_2, \quad \ldots, \quad y_{k-1} = x^{k-1}\exp(\lambda x)v_{k-1}.$$

In the special case in which A is a companion matrix, so that the system is equivalent to a high order equation in a single variable, as in (111c), with $C_1(x) = C_1, C_2(x) = C_2, \ldots, C_m(x) = C_m$, each a constant, the characteristic polynomial of A is

$$P(\lambda) = \lambda^m - C_1\lambda^{m-1} - C_2\lambda^{m-2} - \cdots - C_m = 0.$$

For this special case, $P(\lambda)$ is also the *minimal* polynomial, and repeated zeros *always* correspond to incomplete eigenvector spaces and the need to use generalized eigenvectors. Also, in this special case, the eigenvector corresponding to λ, together with the generalized eigenvectors if they exist, are

$$v = \begin{bmatrix} 1 \\ \lambda \\ \lambda^2 \\ \vdots \\ \lambda^{m-1} \end{bmatrix}, \quad v_1 = \begin{bmatrix} 0 \\ 1 \\ 2\lambda \\ \vdots \\ (m-1)\lambda^{m-2} \end{bmatrix}, \quad v_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ \frac{(m-1)(m-2)}{2}\lambda^{m-3} \end{bmatrix}, \quad \ldots.$$

112   Stiff differential equations

Many differential equation systems of practical importance in scientific modelling exhibit a distressing behaviour when solved by classical numerical methods. This behaviour is distressing because these systems are characterized by very high stability, which can turn into very high *instability* when approximated by standard numerical methods. We have already seen examples of stiff problems, in Subsections 102 and 104, and of course there are many more such examples. The concept of the 'one-sided Lipschitz condition' was mentioned in Subsection 110 without any explanation. Stiff problems typically have large Lipschitz constants, but many have more manageable one-sided Lipschitz constants, and this can be an aid in obtaining realistic growth estimates for the effect of perturbations.

   We confine ourselves to problems posed on an inner product space. Thus we assume that there exists an inner product on $\mathbb{R}^N$ denoted by $\langle u, v \rangle$, and that the norm is defined by $\|u\|^2 = \langle u, u \rangle$.

Definition 112A *The function f satisfies a 'one-sided Lipschitz condition', with 'one-sided Lipschitz constant' l if for all $x \in [a, b]$ and all $u, v \in \mathbb{R}^N$,*

$$\langle f(x, u) - f(x, v), u - v \rangle \le l \|u - v\|^2.$$

It is possible that the function f could have a very large Lipschitz constant but a moderately sized, or even negative, one-sided Lipschitz constant. The advantage of this is seen in the following result.

Theorem 112B *If f satisfies a one-sided Lipschitz condition with constant l, and y and z are each solutions of*

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

then for all $x \ge x_0$,

$$\|y(x) - z(x)\| \le \exp(l(x - x_0)) \|y(x_0) - z(x_0)\|.$$

Proof. We have

$$
\begin{aligned}
\frac{d}{dx}\|y(x) - z(x)\|^2 &= \frac{d}{dx}\langle y(x) - z(x), y(x) - z(x)\rangle \\
&= 2\langle f(x, y(x)) - f(x, z(x)), y(x) - z(x)\rangle \\
&\le 2l\|y(x) - z(x)\|^2.
\end{aligned}
$$

Multiply by $\exp\big(-2l(x - x_0)\big)$ and it follows that

$$\frac{d}{dx}\Big(\exp\big(-2l(x - x_0)\big)\|y(x) - z(x)\|^2\Big) \le 0,$$

so that $\exp\big(-2l(x - x_0)\big)\|y(x) - z(x)\|^2$ is non-increasing.   □

Note that the problem described in Subsection 102 possesses the one-sided Lipschitz condition with $l = 0$.

Even though stiff differential equation systems are typically non-linear, there is a natural way in which a linear system arises from a given non-linear system. Since stiffness is associated with the behaviour of perturbations to a given solution, we suppose that there is a small perturbation $\epsilon Y(x)$ to a solution $y(x)$. The parameter ϵ is small, in the sense that we are interested only in asymptotic behaviour of the perturbed solution as this quantity approaches zero. If $y(x)$ is replaced by $y(x) + \epsilon Y(x)$ in the differential equation

$$y'(x) = f(x, y(x)), \tag{112a}$$

and the solution expanded in a series in powers of ϵ, with ϵ^2 and higher powers replaced by zero, we obtain the system

$$y'(x) + \epsilon Y'(x) = f(x, y(x)) + \epsilon \frac{\partial f}{\partial y} Y(x). \tag{112b}$$

Subtract (112a) from (112b) and cancel out ϵ, and we arrive at the equation governing the behaviour of the perturbation,

$$Y'(x) = \frac{\partial f}{\partial y} Y(x) = J(x) Y(x),$$

say. The 'Jacobian matrix' $J(x)$ has a crucial role in the understanding of problems of this type; in fact its spectrum is sometimes used to characterize stiffness. In a time interval Δx, chosen so that there is a moderate change in the value of the solution to (112a), and very little change in $J(x)$, the eigenvalues of $J(x)$ determine the growth rate of components of the perturbation. The existence of one or more large and negative values of $\lambda \Delta x$, for $\lambda \in \sigma(J(x))$, the spectrum of $J(x)$, is a sign that stiffness is almost certainly present. If $J(x)$ possesses complex eigenvalues, then we interpret this test for stiffness as the existence of a $\lambda = \text{Re}\lambda + i\text{Im}\lambda \in \sigma(J(x))$ such that $\text{Re}\lambda \Delta x$ is negative with large magnitude.

Exercises 11

11.1 Show how to modify Theorem 110C so that the Lipschitz condition holds only in a neighbourhood of y_0 and where the solution is only required to exist on $[a, \widetilde{b}]$, where $\widetilde{b}$ satisfies $a < \widetilde{b} \leq b$.

11.2 By finding two vectors α and β so that the system

$$y'(x) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} y(x) + \begin{bmatrix} \sin(x) \\ 0 \\ \cos(x) \end{bmatrix},$$

has a solution of the form $\widehat{y}(x) = \sin(x)\alpha + \cos(x)\beta$, find the general solution to this problem.

12   Further Evolutionary Problems

120   *Many-body gravitational problems*

We consider a more general gravitational problem involving n mutually attracting masses $M_1, M_2, \ldots, M_n$ at position vectors $y_1(x), y_2(x), \ldots, y_n(x)$, satisfying the $3n$-dimensional second order differential equation system

$$y_i''(x) = -\sum_{j \neq i} \frac{\gamma M_j (y_i - y_j)}{\|y_i - y_j\|^3}, \qquad i = 1, 2, \ldots, n.$$

Reformulated as a first order system, the problem is $6n$-dimensional because each of the y_i has three components and the velocity vectors y_i' also have three components.

To reduce this problem to a manageable level in situations of practical interest, some simplifications can be made. For example, in models of the solar system, the most massive planets, Jupiter, Uranus, Neptune and Saturn, are typically regarded as the only bodies capable of influencing the motion of the sun and of each other. The four small planets closest to the sun, Mercury, Venus, Earth and Mars, are, in this model, regarded as part of the sun in the sense that they add to its mass in attracting the heavy outer planets towards the centre of the solar system. To study the motion of the small planets or of asteroids, they can be regarded as massless particles, moving in the gravitation fields of the sun and the four large planets, but not at the same time influencing their motion.

Another model, involving only three bodies, is useful for studying the motion of an Earth–Moon satellite or of an asteroid close enough to the Earth to be strongly influenced by it as well as by the Sun. This system, known as the restricted three–body problem, regards the two heavy bodies as revolving in fixed orbits about their common centre of mass and the small body as attracted by the two larger bodies but not affecting their motion in any way. If it is possible to approximate the large-body orbits as circles, then a further simplification can be made by working in a frame of reference that moves with them. Thus, we would regard the two large bodies as being fixed in space with their rotation in the original frame of reference translated into a modification of the equations of gravitational motion.

To simplify this discussion, we use units scaled to reduce a number of constants to unit value. We scale the masses of the two larger bodies to $1 - \mu$ and μ and their positions relative to the moving reference frame by the vectors $(\mu - 1)e_1$ and μe_1, so that their centre of mass is at the origin of coordinates. Write y_1, y_2 and y_3 as the scalar variables representing the position coordinates of the small body and y_4, y_5 and y_6 as the corresponding velocity coordinates. Under these assumptions, the equations of motion become
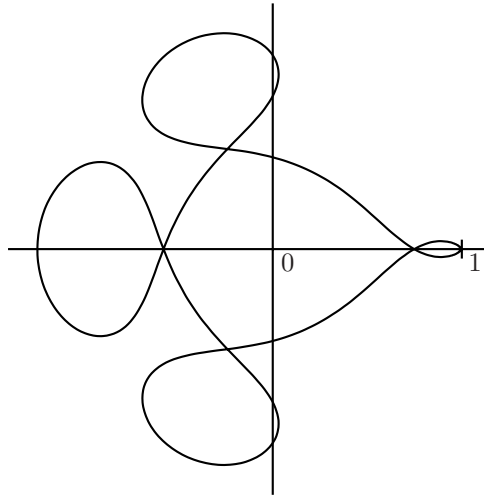
Figure 120(i)   A solution to the restricted three-body problem

$$y_1' = y_4,$$
$$y_2' = y_5,$$
$$y_3' = y_6,$$
$$y_4' = 2y_5 + y_1 - \frac{\mu(y_1 + \mu - 1)}{(y_2^2 + y_3^2 + (y_1 + \mu - 1)^2)^{3/2}} - \frac{(1 - \mu)(y_1 + \mu)}{(y_2^2 + y_3^2 + (y_1 + \mu)^2)^{3/2}},$$

$$y_5' = -2y_4 + y_2 - \frac{\mu y_2}{(y_2^2 + y_3^2 + (y_1 + \mu - 1)^2)^{3/2}} - \frac{(1 - \mu)y_2}{(y_2^2 + y_3^2 + (y_1 + \mu)^2)^{3/2}},$$

$$y_6' = -\frac{\mu y_3}{(y_2^2 + y_3^2 + (y_1 + \mu - 1)^2)^{3/2}} - \frac{(1 - \mu)y_3}{(y_2^2 + y_3^2 + (y_1 + \mu)^2)^{3/2}}.$$

Planar motion is possible; that is, solutions in which $y_3 = y_6 = 0$ at all times. One of these is shown in Figure 120(i), with the values of (y_1, y_2) plotted as the orbit evolves. The heavier mass is at the point $(\mu, 0)$ and the lighter mass is at $(1 - \mu, 0)$, where $(0, 0)$ is marked 0 and $(1, 0)$ is marked 1. For this calculation the value of $\mu = 1/81.45$ was selected, corresponding to the Earth-Moon system. The initial values for this computation were $(y_1, y_2, y_3, y_4, y_5, y_6) = (0.994, 0, 0, 0, -2.0015851063790825224, 0)$ and the period was 17.06521656015796.
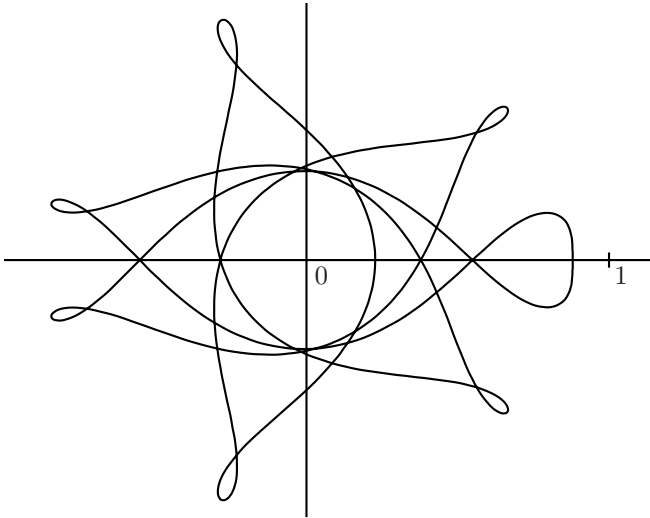
Figure 120(ii)   A second solution to the restricted three-body problem
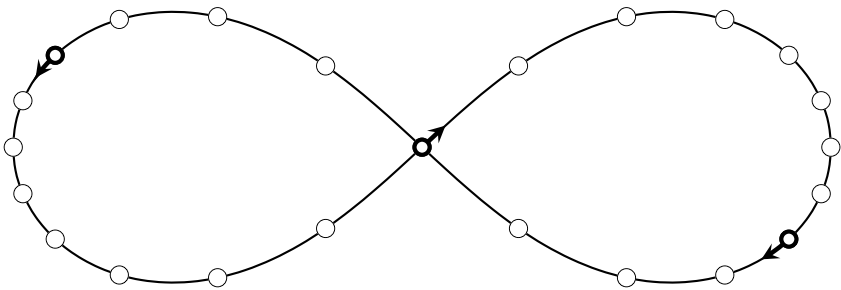
Figure 120(iii)   A figure-of-eight orbit for three equal masses

A second solution, identical except for the initial value $(y_1, y_2, y_3, y_4, y_5, y_6)$ = $(0.87978, 0, 0, 0, -0.3797, 0)$ and a period 19.14045706162071, is shown in Figure 120(ii).

If the three masses are comparable in value, then the restriction to a simpler system that we have considered is not available. However, in the case of a number of equal masses, other symmetries are possible. We consider just a single example, in which three equal, mutually attracting masses move in a figure-of-eight orbit. This is shown in Figure 120(iii).
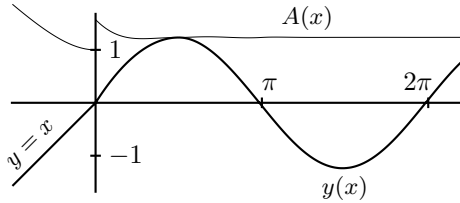
Figure 121(i) Solution to delay differential equation (121b)

121 Delay problems and discontinuous solutions

A functional differential equation is one in which the rate of change of $y(x)$ depends not just on the values of y for the same time value, but also on time values less than x. In the simplest case, this has the form

$$y'(x) = f(x, y(x), y(x - \tau)), \tag{121a}$$

where τ is a constant delay. Note that this cannot be cast as an initial value problem with the hope of actually defining a unique solution, because at an initial point x_0, the derivative depends on the value of $y(x_0 - \tau)$. What we will need to do in the case of (121a) is to specify the value of y on an initial interval $[x_0 - \tau, x_0]$.

A linear delay differential equation

We consider the problem given by

$$y'(x) = -y(x - \tfrac{\pi}{2}), \quad x > 0, \qquad y(x) = x, \quad x \in [-\tfrac{\pi}{2}, 0]. \tag{121b}$$

For x in the interval $[0, \frac{\pi}{2}]$ we find

$$y(x) = -\int_0^x (x - \tfrac{\pi}{2}) dx = \tfrac{1}{2}x(\pi - x),$$

with $y(\frac{\pi}{2}) = \frac{1}{8}\pi^2$. This process can be repeated over the sequence of intervals $[\frac{\pi}{2}, \pi]$, $[\pi, \frac{3\pi}{2}]$, ... to obtain values of $y(x)$ shown in Figure 121(i) for $x \leq 4\pi$.

It appears that the solution is attempting to approximate sinusoidal behaviour as time increases. We can verify this by estimating a local amplitude defined by

$$A(x) = (y(x)^2 + y'(x)^2)^{\frac{1}{2}}.$$

This function is also shown in Figure 121(i) and we note the discontinuity at $x = 0$, corresponding to the discontinuity in the value of $y'(x)$. Such discontinuities are to be expected because the right-derivative is given by
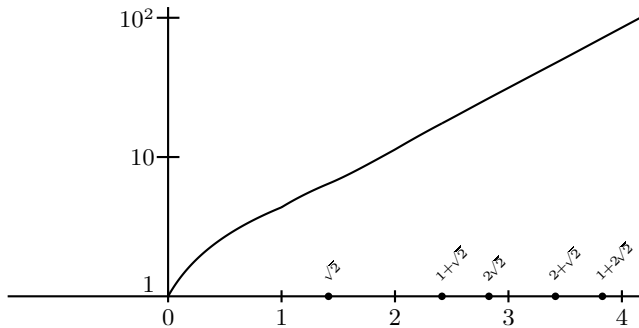
Figure 121(ii)   Solution to neutral delay differential equation (121c)

the formula for $y'(x)$ for x positive and the left-derivative is found from the derivative of the initial function. For each positive integral multiple of $\frac{1}{2}\pi$, there will always be an inherited non-smooth behaviour but this will be represented by a discontinuity in increasingly higher derivatives.

We will now consider a problem with two delays.

An example with persistent discontinuities

A delay differential equation of 'neutral type' is one in which delayed values of y' also occur in the formulation. An example of this type of problem is

$$
\begin{aligned}
y'(x) &= \tfrac{1}{2}y'(x-1) + ay(x-\sqrt{2}), & x &> 0, \\
y(x) &= 1, & x &\in [-\sqrt{2}, 0],
\end{aligned}
\tag{121c}
$$

where the constant is given by $a = \exp(\sqrt{2}) - \frac{1}{2}\exp(\sqrt{2}-1)$ and was contrived to ensure that $\exp(x)$ would have been a solution, if the initial information had been defined in terms of that function.

The solution is shown in Figure 121(ii) and we see that it seems to be approximating exponential behaviour more and more closely as x increases. However, there is a discontinuity in $y'(x)$ at every positive integer value of x. Specifically, for each n there is a jump given by

$$
\lim_{x\to n+} y'(x) - \lim_{x\to n-} y'(x) = 2^{-n}a.
$$

122   Problems evolving on a sphere

Given a function $H(y)$, we will explore situations in which solutions to $y'(x) = f(y)$ preserve the value of $H(y(x))$. In the special case in which $H(y) = \frac{1}{2}\|y\|^2$, this will correspond to motion on a sphere. We recall the standard notation

$$\nabla(H) = \begin{bmatrix} \frac{\partial H}{\partial y_1} \\ \frac{\partial H}{\partial y_2} \\ \vdots \\ \frac{\partial H}{\partial y_N} \end{bmatrix}$$

and consider problems of the 'Poisson' form

$$y' = L(x,y)\nabla(y), \tag{122a}$$

where $L(x,y)$ is always a skew-symmetric matrix. For such problems $H(y(x))$ is invariant. To verify this, calculate

$$\frac{d}{dx}H(y(x)) = \sum_{i=1}^{N}\frac{\partial H}{\partial y_i}y_i'(x) = \nabla(H)^{\mathsf{T}}L(x,y)\nabla(y) = 0,$$

because of the skew-symmetry of L.

The Euler equations, discussed in Subsection 107, provide two examples of this. To show that $E(w)$ is invariant write $H(w) = \frac{1}{2}E(w)$, and to show that $F(w)$ is invariant write $H(w) = \frac{1}{2}F(w)$. The problem reverts to the form of (122a), with y replaced by w, where $L(x,w)$ is given by

$$\begin{bmatrix} 0 & \frac{I_3 w_3}{I_1 I_2} & -\frac{I_2 w_2}{I_1 I_3} \\ -\frac{I_3 w_3}{I_1 I_2} & 0 & \frac{I_1 w_1}{I_2 I_3} \\ \frac{I_2 w_2}{I_1 I_3} & -\frac{I_1 w_1}{I_2 I_3} & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & -\frac{w_3}{I_1 I_2} & \frac{w_2}{I_1 I_3} \\ \frac{w_3}{I_1 I_2} & 0 & -\frac{w_1}{I_2 I_3} \\ -\frac{w_2}{I_1 I_3} & \frac{w_1}{I_2 I_3} & 0 \end{bmatrix},$$

respectively.

We now revert to the special case $H(x) = \frac{1}{2}y^{\mathsf{T}}y$, for which (122a) becomes

$$y' = L(x,y)y. \tag{122b}$$

An example is the contrived problem

$$\begin{bmatrix} y_1' \\ y_2' \\ y_3' \end{bmatrix} = \begin{bmatrix} 0 & -y_1 & -\sin(x) \\ y_1 & 0 & -1 \\ \sin(x) & 1 & 0 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \quad \begin{bmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \tag{122c}$$

with solution $y_1(x) = \cos(x)$, $y_2(x) = \cos(x)\sin(x)$, $y_3(x) = \sin^2(x)$. The solution values for $x \in [0, 1.4\pi]$ are shown in Figure 122(i).

Problems of the form (122b) are a special case of

$$Y' = L(x,Y)Y, \tag{122d}$$

where Y has a number of columns. In this case the inner product of two specific columns will be invariant. In particular, if $Y(x)$ is a square matrix,
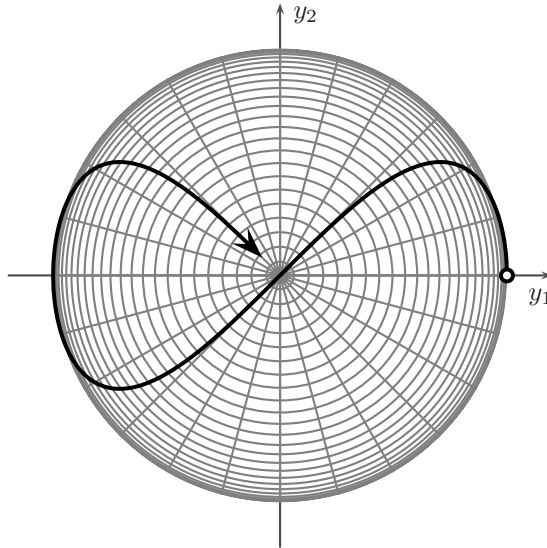
Figure 122(i)   Solution to problem (122c) with y_3 pointing out of the page

initially orthogonal, and $L(x, Y)$ is always skew-symmetric, then $Y(x)$ will remain orthogonal. Denote the elements of Y by y_{ij}. An example problem of this type is

$$Y'(x) = \begin{bmatrix} 0 & -1 & \mu y_{21} \\ 1 & 0 & -\mu y_{11} \\ -\mu y_{21} & \mu y_{11} & 0 \end{bmatrix} Y, \qquad Y(0) = I, \qquad (122e)$$

with μ a real parameter. The solution to (122e) is

$$Y(x) = \begin{bmatrix} \cos(x) & -\sin(x)\cos(\mu x) & \sin(x)\sin(\mu x) \\ \sin(x) & \cos(x)\cos(\mu x) & -\cos(x)\sin(\mu x) \\ 0 & \sin(\mu x) & \cos(\mu x) \end{bmatrix}.$$

123   Further Hamiltonian problems

In the Hamiltonian formulation of classical mechanics, generalized coordinates $q_1, q_2, \ldots, q_N$ and generalized momenta $p_1, p_2, \ldots, p_N$ are used to represent the state of a mechanical system. The equations of motion are defined in terms

of a 'Hamiltonian' function $H(p_1, p_2, \ldots, p_N, q_1, q_2, \ldots, q_N)$ by the equations

$$p_i' = -\frac{\partial H}{\partial q_i},$$

$$q_i' = \frac{\partial H}{\partial p_i}.$$

Write $y(x)$ as a vector variable, made up from N momenta followed by the N coordinates. That is,

$$y_i = \begin{cases} p_i, & 1 \le i \le N, \\ q_{i-N}, & N+1 \le i \le 2N. \end{cases}$$

With the understanding that H is regarded as a function of y, the differential equations can be written in the form $y' = f(y)$, where

$$f(y) = J\nabla(H), \qquad J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix},$$

in which I is the $N \times N$ unit matrix.

Theorem 123A $H(y(x))$ *is invariant.*

Proof. Calculate $\partial H/\partial y$ to obtain the result
$$\nabla(H)^{\mathsf{T}} J \nabla(H) = 0. \qquad \square$$

The Jacobian of this problem is equal to

$$\frac{\partial}{\partial y} f(y) = \frac{\partial}{\partial y}\left(J\nabla(H)\right) = JW(y),$$

where W is the 'Wronskian' matrix defined as the $2N \times 2N$ matrix with (i, j) element equal to $\partial^2 H/\partial y_i \partial y_j$.

If the initial value $y_0 = y(x_0)$ is perturbed by a small number ϵ multiplied by a fixed vector v_0, then, to within $O(\epsilon^2)$, the solution is modified by $\epsilon v + O(\epsilon^2)$ where
$$v'(x) = \frac{\partial f}{\partial y} v(x) = JW(y)v(x).$$

For two such perturbations u and v, it is interesting to consider the value of the scalar $u^{\mathsf{T}} J v$.

This satisfies the differential equation

$$\frac{d}{dx} u^{\mathsf{T}} J v = u^{\mathsf{T}} J J W v + (JWu)^{\mathsf{T}} J v = -u^{\mathsf{T}} W v + u^{\mathsf{T}} W v = 0.$$

Hence we have:
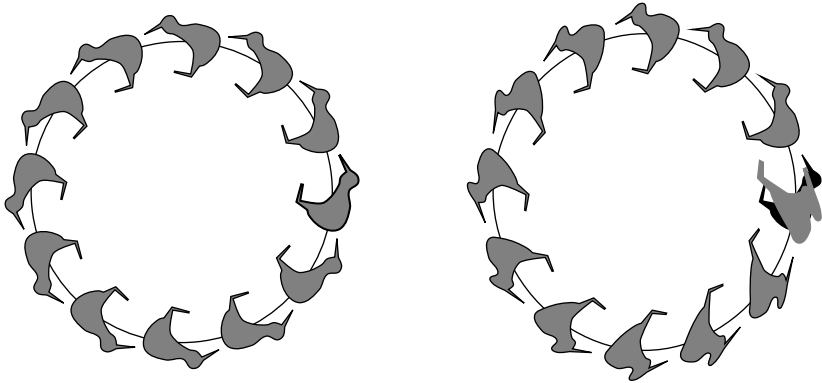
Figure 123(i)   Illustration of symplectic behaviour for $H(p,q) = p^2/2 + q^2/2$ (left) and $H(p,q) = p^2/2 - \cos(q)$ (right). The underlying image depicts the North Island brown kiwi, *Apteryx mantelli*.

Theorem 123B $u^\mathsf{T}Jv$ *is invariant with time.*

In the special case of a two-dimensional Hamiltonian problem, the value of $(\epsilon u)^\mathsf{T}J(\epsilon v)$ can be interpreted as the area of the infinitesimal parallelogram with sides in the directions u and v. As the solution evolves, u and v might change, but the area $u^\mathsf{T}Jv$ remains invariant. This is illustrated in Figure 123(i) for the two problems $H(p,q) = p^2/2 + q^2/2$ and $H(p,q) = p^2/2 - \cos(q)$ respectively.

124   Further differential-algebraic problems

Consider the initial value problem

$$y' = y + z, \tag{124a}$$

$$0 = z + z^3 - y, \tag{124b}$$

$$y(0) = 2, \quad z(0) = 1. \tag{124c}$$

This is an index 1 problem, because a single differentiation of (124b) and a substitution from (124a) converts this to a differential equation system consisting of (124b) together with $z' = (y + z)/(1 + 3z^2)$. However, this reduction does not do justice to the original formulation in the sense that a solution with slightly perturbed initial values has little to do with the original index 1 problem. This emphasizes the fact that initial conditions for the differential-algebraic equation formulation must be consistent with the algebraic constraint for it to be well-posed. A more appropriate reduction is to replace (124a) by $y' = y + \phi(y)$, where $\phi(y)$ is the real value of z which satisfies (124b).

We next introduce an initial value problem comprising two differential equations and a single algebraic constraint:

$$y_1' = -\sin(z), \tag{124d}$$
$$y_2' = 2\cos(z) - y_1, \tag{124e}$$
$$0 = y_1^2 + y_2^2 - 1, \tag{124f}$$
$$y_1(0) = 1, \quad y_2(0) = 0, \qquad z(0) = 0. \tag{124g}$$

An attempt to reduce this to an ordinary differential equation system by differentiating (124f) and substituting from (124d) and (124e), leads to a new algebraic constraint

$$-y_1\sin(z) + y_2(2\cos(z) - y_1) = 0, \tag{124h}$$

and it is clear that this will be satisfied by the solution to the original problem. However, this so-called 'hidden constraint' introduces a new complexity into this type of problem. That is, for initial values to be consistent, (124h) must be satisfied at the initial time. If, for example, the initial values $y_1(0) = 1$ and $y_2(0) = 0$ are retained, but the initial value $z(0)$ is perturbed slightly, (124h) will not be satisfied and no genuine solution exists. But the hidden constraint, as the problem has actually been posed, is satisfied, and we can take the reduction towards an ordinary differential equation system to completion. Differentiate (124h) and substitute from (124d) and (124e) and we finally arrive at

$$z'(\cos^2(z) + 2\sin^2(z)) = \sin^2(z) + y_2\sin(z) + (2\cos(z) - y_1)^2. \tag{124i}$$

Because two differentiation steps were required to reach this equation, the original system is referred to as an index 2 problem. In summary, the original index 2 problem, comprising (124d), (124e), (124f) has been reduced, first to an index 1 formulation (124d), (124e), (124h), and then to an ordinary differential equation system (124d), (124e), (124i).

Exercises 12

12.1 Show that a problem of the form

$$u' = -\alpha'(v)\gamma(u, v),$$
$$v' = \beta'(u)\gamma(u, v),$$

satisfies the assumptions of (122a) with a suitable choice of $H(u, v)$.

12.2 Write the Lotka–Volterra equations (106a), (106b) in the form given in Exercise 12.1.

13   Difference Equation Problems

130   Introduction to difference equations

While differential equations deal with functions of a continuous variable, difference equations deal with functions of a discrete variable. Instead of a formula for the derivative of a function written in terms of the function itself, we have to consider sequences for which each member is related in some specific way to its immediate predecessor or several of its most recent predecessors. Thus we may write

$$x_n = \phi_n(x_{n-1}, x_{n-2}, \ldots, x_{n-k}),$$

where k is the 'order' of this difference equation. This equation, in which x_n depends on k previous values, can be recast in a vector setting in which members of the sequence lie not in $\mathbb{R}$ but in $\mathbb{R}^k$, and depend only on *one* previous value. Thus if

$$X_n = \begin{bmatrix} x_n \\ x_{n-1} \\ \vdots \\ x_{n-k+1} \end{bmatrix},$$

then

$$X_n = \Phi_n(X_{n-1}) = \begin{bmatrix} \phi_n(x_{n-1}, x_{n-2}, \ldots, x_{n-k}) \\ x_{n-1} \\ x_{n-2} \\ \vdots \\ x_{n-k+1} \end{bmatrix}.$$

Just as for differential equations, we can use either formulation as we please.

131   A linear problem

Consider the difference equation

$$y_n = 3y_{n-1} - 2y_{n-2} + C\theta^n, \tag{131a}$$

where C and θ are constants. We do not specify an initial value, but aim instead to find the family of all solutions. As a first step, we look at the special case in which $C = 0$. In this case, the equation becomes linear in the sense that known solutions can be combined by linear combinations. The simplified equation in matrix–vector form is

$$\begin{bmatrix} y_n \\ y_{n-1} \end{bmatrix} = \begin{bmatrix} 3 & -2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_{n-1} \\ y_{n-2} \end{bmatrix},$$

which can be rewritten as

$$\begin{bmatrix} y_n - y_{n-1} \\ -y_n + 2y_{n-1} \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{n-1} - y_{n-2} \\ -y_{n-1} + 2y_{n-2} \end{bmatrix},$$

with solution defined by

$$y_n - y_{n-1} = A2^{n-1},$$
$$-y_n + 2y_{n-1} = B,$$

for constants A and B. By eliminating y_{n-1}, we find

$$y_n = A2^n + B$$

for the general solution. The fact that this combines powers of 2 and 1, the eigenvalues of the matrix

$$\begin{bmatrix} 3 & -2 \\ 1 & 0 \end{bmatrix}, \tag{131b}$$

suggests that we can look for solutions for the original formulation in the form λ^n without transforming to the matrix–vector formulation. Substitute this trial solution into (131a), with $C = 0$, and we find, apart from a factor λ^{n-2}, that the condition on λ is

$$\lambda^2 - 3\lambda + 2 = 0.$$

This is the characteristic polynomial of the matrix (131b), but it can be read off immediately from the coefficients in (131a).

To find the general solution to (131a), if $C \neq 0$, it is easy to see that we only need to find one special solution to which we can add the terms $A2^n + B$ to obtain all possible solutions. A special solution is easily found, if $\theta \neq 1$ and $\theta \neq 2$, in the form

$$y_n = \frac{C\theta^{n+2}}{(\theta - 1)(\theta - 2)}.$$

This type of special solution is not available if θ equals either 1 or 2. In these cases a special solution can be found as a multiple of n or $n2^n$, respectively. Combining these cases, we write the general solution as

$$y_n = \begin{cases} A2^n + B - Cn, & \theta = 1, \\ A2^n + B + 2Cn2^n, & \theta = 2, \\ A2^n + B + \dfrac{C\theta^2}{(\theta-1)(\theta-2)}\theta^n, & \theta \neq 1, \theta \neq 2. \end{cases}$$

132   The Fibonacci difference equation

The initial value difference equation

$$y_n = y_{n-1} + y_{n-2}, \qquad y_0 = 0, \quad y_1 = 1, \tag{132a}$$

is famous because of the mathematical, biological and even numerological significance attached to the solution values

$$1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, \ldots.$$

To find the general solution, solve the polynomial equation

$$\lambda^2 - \lambda - 1 = 0,$$

to find the terms λ_1^n and λ_2^n, where

$$\lambda_1 = \tfrac{1+\sqrt{5}}{2}, \quad \lambda_2 = \tfrac{1-\sqrt{5}}{2} = -\lambda_1^{-1}.$$

To find the coefficients A and B in the general solution

$$y_n = A \left(\tfrac{1+\sqrt{5}}{2} \right)^n + B \left(-\tfrac{1+\sqrt{5}}{2} \right)^{-n},$$

substitute $n = 0$ and $n = 1$, to find $A = -B = 5^{-1/2}$, and therefore the specific solution to the initial value problem (132a),

$$y_n = \tfrac{1}{\sqrt{5}} \left(\left(\tfrac{1+\sqrt{5}}{2} \right)^n - \left(-\tfrac{1+\sqrt{5}}{2} \right)^{-n} \right).$$

133   Three quadratic problems

We consider the solutions to the problems

$$y_n = y_{n-1}^2, \tag{133a}$$

$$y_n = y_{n-1}^2 - 2, \tag{133b}$$

$$y_n = y_{n-1} y_{n-2}. \tag{133c}$$

If $z_n = \ln(y_n)$ in (133a), then $z_n = 2z_{n-1}$ with solution $z_n = 2^n z_0$. Hence, the general solution to (133a) is

$$y_n = y_0^{2^n}.$$

To solve (133b), substitute $y_n = z_n + z_n^{-1}$, so that

$$z_n + \frac{1}{z_n} = z_{n-1}^2 + \frac{1}{z_{n-1}^2},$$

and this is satisfied by any solution to $z_n = z_{n-1}^2$. Hence, using the known solution of (133a), we find

$$y_n = z_0^{2^n} + z_0^{-2^n},$$

where z_0 is one of the solutions to the equation

$$z_0 + \frac{1}{z_0} = y_0.$$

Finally, to solve (133c), substitute $z_n = \ln(y_n)$, and we find that

$$z_n = z_{n-1} + z_{n-2}.$$

The general solution to this is found from the Fibonacci equation, so that substituting back in terms of y_n, we find

$$y_n = A^{\left(\frac{1}{2}(1+\sqrt{5})\right)^n} \cdot B^{\left(\frac{1}{2}(1-\sqrt{5})\right)^n},$$

with A and B determined from the initial values.

134 Iterative solutions of a polynomial equation

We discuss the possible solution of the polynomial equation

$$x^2 - 2 = 0.$$

Of course this is only an example, and a similar discussion would be possible with other polynomial equations. Consider the difference equations

$$y_n = y_{n-1} - \frac{1}{2}y_{n-1}^2 + 1, \qquad y_0 = 0, \tag{134a}$$

$$y_n = y_{n-1} - \frac{1}{2}y_{n-1}^2 + 1, \qquad y_0 = 4, \tag{134b}$$

$$y_n = y_{n-1} - y_{n-1}^2 + 2, \qquad y_0 = \frac{3}{2}, \tag{134c}$$

$$y_n = \frac{y_{n-1}}{2} + \frac{1}{y_{n-1}}, \qquad y_0 = 100, \tag{134d}$$

$$y_n = \frac{y_{n-1}y_{n-2} + 2}{y_{n-1} + y_{n-2}}, \qquad y_0 = 0, \quad y_1 = 1. \tag{134e}$$

Note that each of these difference equations has $\sqrt{2}$ as a stationary point. That is, each of them is satisfied by $y_n = \sqrt{2}$, for every n. Before commenting further, it is interesting to see what happens if a few values are evaluated numerically for each sequence. These are shown in Table 134(I).

Note that (134a) seems to be converging to $\sqrt{2}$, whereas (134b) seems to have no hope of ever doing so. Of course the starting value, y_0, is the distinguishing feature, and we can perhaps investigate which values converge and which ones do not. It can be shown that the fate of the iterates for various starting values can be summarized as follows:

Table 134(I)   The first few terms in the solutions of some difference equations

	Equation (134a)	Equation (134b)	Equation (134c)	Equation (134d)	Equation (134e)
y_0	0.0000000000	4.0000000000	1.5000000000	1.000000×10^2	0.0000000000
y_1	1.0000000000	-3.0000000000	1.2500000000	5.001000×10	1.0000000000
y_2	1.5000000000	-6.5000000000	1.6875000000	2.502500×10	2.0000000000
y_3	1.3750000000	-2.662500×10	0.8398437500	1.255246×10	1.3333333333
y_4	1.4296875000	-3.800703×10^2	2.1345062256	6.3558946949	1.4000000000
y_5	1.4076843262	-7.260579×10^4	-0.4216106015	3.3352816093	1.4146341463
y_6	1.4168967451	-2.635873×10^9	1.4006338992	1.9674655622	1.4142114385

$$
\begin{aligned}
y_0 \in \{-\sqrt{2}, 2+\sqrt{2}\}: &\quad \text{Convergence to } x = -\sqrt{2} \\
y_0 \in (-\sqrt{2}, 2+\sqrt{2}): &\quad \text{Convergence to } x = \sqrt{2} \\
y_0 \notin [-\sqrt{2}, 2+\sqrt{2}]: &\quad \text{Divergence}
\end{aligned}
$$

Note that the starting value $y_0 = -\sqrt{2}$, while it is a fixed point of the mapping given by (134a), is unstable; that is, any small perturbation from this initial value will send the sequence either into instability or convergence to $+\sqrt{2}$. A similar remark applies to $y_0 = 2+\sqrt{2}$, which maps immediately to $y_1 = -\sqrt{2}$.

The difference equation (134c) converges to $\pm\sqrt{2}$ in a finite number of steps for y_0 in a certain countable set; otherwise the sequence formed from this equation diverges.

Equation (134d) is the Newton method and converges quadratically to $\sqrt{2}$ for any positive y_0. By quadratic convergence, we mean that $|y_n - \sqrt{2}|$ divided by $|y_{n-1} - \sqrt{2}|^2$ is bounded. In fact, in the limit as $n \to \infty$,

$$
\frac{y_n - \sqrt{2}}{(y_{n-1} - \sqrt{2})^2} \to \frac{\sqrt{2}}{4}.
$$

The iteration scheme given by (134e) is based on the secant method for solving non-linear equations. To solve $\phi(y) = 0$, y_n is found by fitting a straight line through the two points $(y_{n-2}, \phi(y_{n-2}))$ and $(y_{n-1}, \phi(y_{n-1}))$ and defining y_n as the point where this line crosses the horizontal axis. In the case $\phi(y) = y^2 - 2$, this results in (134e).

It is interesting to ask if there exists an 'order' k for this sequence. In other words, assuming that convergence is actually achieved, does $k \geq 1$ exist such that

$$
\frac{|y_n - \sqrt{2}|}{|y_{n-1} - \sqrt{2}|^k}
$$

has a limiting value as $n \to \infty$? For the secant method k does exist, and has the value $k = \frac{1}{2}(\sqrt{5} + 1)$.

135   The arithmetic-geometric mean

Let a_0 and b_0 be real numbers chosen so that $0 < b_0 < a_0$, and define the sequence of (a_n, b_n) pairs by the formulae

$$a_n = \tfrac{1}{2}(a_{n-1} + b_{n-1}),$$
$$b_n = \sqrt{a_{n-1} b_{n-1}}, \qquad n = 1, 2, \ldots. \qquad (135a)$$

We can verify (i) that $b_{n-1} < b_n < a_n < a_{n-1}$ for all $n \geq 1$ and (ii) that the sequence $a_0 - b_0$, $a_1 - b_1$, $a_2 - b_2$, ... converges to zero. The truth of (i) follows from elementary properties of arithmetic and geometric means. Furthermore, (ii) can be proved from the identity

$$a_n - b_n = \frac{(a_{n-1} - b_{n-1})^2}{2\left(\sqrt{a_{n-1}} + \sqrt{b_{n-1}}\right)^2}.$$

The common limit of the a_n and b_n sequences is known as the 'arithmetic-geometric mean' of a_0 and b_0. We present a single application.

The quantities

$$F(a, b) = \int_0^{\pi/2} \left(a^2 \cos^2(\theta) + b^2 \sin^2(\theta)\right)^{-1/2} d\theta,$$

$$E(a, b) = \int_0^{\pi/2} \left(a^2 \cos^2(\theta) + b^2 \sin^2(\theta)\right)^{1/2} d\theta,$$

are known as 'complete elliptic integrals' of the first and second kind, respectively. The value of $4E(a, b)$ is the length of the circumference of the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

Use $a_0 = a$ and $b_0 = b$ as starting values for the computation of the sequences defined by (135a), and denote by a_∞ the arithmetic-geometric mean of a_0 and b_0. Then it can be shown that

$$F(a_0, b_0) = F(a_1, b_1),$$

and therefore that

$$F(a_0, b_0) = F(a_\infty, a_\infty) = \frac{\pi}{2 a_\infty}.$$

The value of $E(a_0, b_0)$ can also be found from the sequences that lead to the arithmetic-geometric mean. In fact

$$E(a_0, b_0) = \frac{\pi}{2 a_\infty} \left(a_0^2 - 2a_1(a_0 - a_1) - 4a_2(a_1 - a_2) - 8a_3(a_2 - a_3) - \cdots\right).$$

Exercises 13

13.1 Write the difference equation given by (134e) in the form

$$z_n = \phi(z_{n-1}),$$

with z_0 a given initial value.

13.2 Write the difference equation system

$$
\begin{aligned}
u_n &= u_{n-1} + v_{n-1}, & u_0 &= 2, \\
v_n &= 2u_{n-1} + v_{n-1}^2, & v_0 &= 1,
\end{aligned}
$$

in the form $y_n = \phi(y_{n-1}, y_{n-2})$, with y_0 and y_1 given initial values.

13.3 Use the formula for the error in linear interpolation together with the solution to (133c) to verify the order of convergence of (134e).

13.4 Calculate $\sqrt{2}$ by applying the Newton method to the equation

$$2x^{-2} - 1 = 0.$$

13.5 Calculate the value of $\sqrt{3}$ by applying the secant method to

$$x^2 - 3 = 0.$$

13.6 Calculate the circumference of the ellipse

$$\frac{x^2}{9} + \frac{y^2}{4} = 1,$$

using the arithmetic-geometric mean.

14   Difference Equation Theory

140   Linear difference equations

The standard form for linear difference equation systems is

$$X_n = A_n X_{n-1} + \phi_n, \tag{140a}$$

which becomes an initial value problem if the value of the initial vector X_0 is specified. The corresponding system in which ϕ_n is omitted is the 'homogeneous part'.

Many linear difference equations are more naturally formulated as

$$y_n = \alpha_{n1} y_{n-1} + \alpha_{n2} y_{n-2} + \cdots + \alpha_{nk} y_{n-k} + \psi_n,$$

but these are easily recast in the form (140a) by writing

$$
X_n = \begin{bmatrix} y_n \\ y_{n-1} \\ \vdots \\ y_{n-k+1} \end{bmatrix}, \quad
A_n = \begin{bmatrix} \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nk} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad
\phi_n = \begin{bmatrix} \psi_n \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.
$$

To solve (140a) as an initial value problem, we need to use products of the form

$$
\prod_{i=m}^{n} A_i = A_n A_{n-1} \cdots A_{m+1} A_m.
$$

We have:

Theorem 140A *The problem (140a), with initial value X_0 given, has the unique solution*

$$
y_n = \left(\prod_{i=1}^{n} A_i \right) X_0 + \left(\prod_{i=2}^{n} A_i \right) \phi_1 + \left(\prod_{i=3}^{n} A_i \right) \phi_2 + \cdots + A_n \phi_{n-1} + \phi_n.
$$

Proof. The result holds for $n = 0$, and the general case follows by induction. $\qquad\square$

141   Constant coefficients

We consider the solution of a linear difference equation with constant coefficients:

$$
y_n = \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + \cdots + \alpha_k y_{n-k} + \psi_n. \tag{141a}
$$

The solution is found in terms of the solution to the canonical problem in which the initial information is given in the form

$$
\begin{bmatrix} y_0 \\ y_{-1} \\ \vdots \\ y_{-k+2} \\ y_{-k+1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}.
$$

Denote the solution to this problem at step m by

$$
y_m = \theta_m, \qquad m = 0, 1, 2, \ldots, n,
$$

with $\theta_m = 0$ for $m < 0$. Given the difference equation (141a) with initial values $y_0, y_1, \ldots, y_{k-1}$, define linear combinations of this data by

$$
\begin{bmatrix}
\widetilde{y}_{k-1} \\
\widetilde{y}_{k-2} \\
\widetilde{y}_{k-3} \\
\vdots \\
\widetilde{y}_1 \\
\widetilde{y}_0
\end{bmatrix}
=
\begin{bmatrix}
1 & \theta_1 & \theta_2 & \cdots & \theta_{k-2} & \theta_{k-1} \\
0 & 1 & \theta_1 & \cdots & \theta_{k-3} & \theta_{k-2} \\
0 & 0 & 1 & \cdots & \theta_{k-4} & \theta_{k-3} \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & \theta_1 \\
0 & 0 & 0 & \cdots & 0 & 1
\end{bmatrix}^{-1}
\begin{bmatrix}
y_{k-1} \\
y_{k-2} \\
y_{k-3} \\
\vdots \\
y_1 \\
y_0
\end{bmatrix}.
\tag{141b}
$$

We are now in a position to write down the solution to (141a).

Theorem 141A *Using the notation introduced in this subsection, the solution to (141a) with given initial values $y_0, y_1, \ldots, y_{k-1}$ is given by*

$$
y_n = \sum_{i=0}^{k-1} \theta_{n-i}\widetilde{y}_i + \sum_{i=k}^{n} \theta_{n-i}\psi_i.
\tag{141c}
$$

Proof. Substitute $n = m$, for $m = 0, 1, 2, \ldots, k-1$, into (141c), and we obtain the value

$$
y_m = \widetilde{y}_m + \theta_1\widetilde{y}_{m-1} + \cdots + \theta_m\widetilde{y}_0, \quad m = 0, 1, 2, \ldots, k-1.
$$

This is equal to y_m if (141b) holds. Add the contribution to the solution from each of $m = k, k+1, \ldots, n$ and the result follows. $\qquad\square$

142   Powers of matrices

We are interested in powers of a matrix A in terms of two questions: when is the sequence of powers bounded, and when does the sequence converge to the zero matrix? There are various equivalent formulations of the criteria for these properties of A, and we state the most widely accessible of these.

Definition 142A *A square matrix A is 'stable' if there exists a constant C such that for all $n = 0, 1, 2, \ldots$, $\|A^n\| \leq C$.*

This property is often referred to as 'power-boundedness'.

Definition 142B *A square matrix A is 'convergent' if $\lim_{n\to\infty} \|A^n\| = 0$.*

Theorem 142C *Let A denote an $m \times m$ matrix. The following statements are equivalent:*

(i) A is stable.

(ii)   The minimal polynomial of A has all its zeros in the closed unit disc and all its multiple zeros in the open unit disc.

(iii) The Jordan canonical form of A has all its eigenvalues in the closed unit disc with all eigenvalues of magnitude 1 lying in 1×1 blocks.

(iv)   There exists a non-singular matrix S such that $\|S^{-1}AS\|_\infty \leq 1$.

Proof. We prove that $(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (iv) \Rightarrow (i)$. If A is stable but (ii) is not true, then either there exist λ and $v \neq 0$ such that $|\lambda| > 1$ and $Av = \lambda v$, or there exist λ, $u \neq 0$ and v such that $|\lambda| = 1$ and $Av = \lambda v + u$, with $Au = \lambda u$. In the first case, $A^n v = \lambda^n v$ and therefore $\|A^n\| \geq |\lambda|^n$ which is not bounded. In the second case, $A^n v = \lambda^n v + n\lambda^{n-1}u$ and therefore $\|A^n\| \geq n\|u\|/\|v\| - 1$, which also is not bounded. Given (ii), it is not possible that the conditions of (iii) are not satisfied, because the minimal polynomial of any of the Jordan blocks, and therefore of A itself, would have factors that contradict (ii). If (iii) is true, then S can be chosen to form J, the Jordan canonical form of A, with the off-diagonal elements chosen sufficiently small so that $\|J\|_\infty \leq 1$. Finally, if (iv) is true then $A^n = S(S^{-1}AS)^n S^{-1}$ so that $\|A^n\| \leq \|S\| \cdot \|S^{-1}AS\|^n \cdot \|S^{-1}\| \leq \|S\| \cdot \|S^{-1}\|$. $\qquad\square$

Theorem 142D *Let A denote an $m \times m$ matrix. The following statements are equivalent*

(i)   *A is convergent.*

(ii)   *The minimal polynomial of A has all its zeros in the open unit disc.*

(iii) *The Jordan canonical form of A has all its diagonal elements in the open unit disc.*

(iv)   *There exists a non-singular matrix S such that $\|S^{-1}AS\|_\infty < 1$.*

Proof. We again prove that $(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (iv) \Rightarrow (i)$. If A is convergent but (ii) is not true, then there exist λ and $u \neq 0$ such that $\lambda \geq 1$ and $Au = \lambda u$. Hence, $A^n u = \lambda^n u$ and therefore $\|A^n\| \geq |\lambda|^n$, which does not converge to zero. Given (ii), it is not possible that the conditions of (iii) are not satisfied, because the minimal polynomial of any of the Jordan blocks, and therefore of A itself, would have factors that contradict (ii). If (iii) is true, then S can be chosen to form J, the Jordan canonical form of A, with the off-diagonal elements chosen sufficiently small so that $\|J\|_\infty < 1$. Finally, if (iv) is true then $A^n = S(S^{-1}AS)^n S^{-1}$ so that $\|A^n\| \leq \|S\| \cdot \|S^{-1}\| \cdot \|S^{-1}AS\|^n \to 0$. $\qquad\square$

While the two results we have presented here are related to the convergence of difference equation solutions, the next is introduced only because of its application in later chapters.

Theorem 142E *If A is a stable $m \times m$ matrix and B an arbitrary $m \times m$ matrix, then there exists a real C such that*

$$\left\|\left(A + \frac{1}{n}B\right)^n\right\| \leq C,$$

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

Proof. Without loss of generality, assume that $\|\cdot\|$ denotes the norm $\|\cdot\|_\infty$. Because S exists so that $\|S^{-1}AS\| \leq 1$, we have

$$\left\|\left(A + \frac{1}{n}B\right)^n\right\| \leq \|S\| \cdot \|S^{-1}\| \cdot \left\|\left(S^{-1}AS + \frac{1}{n}S^{-1}BS\right)^n\right\|$$

$$\leq \|S\| \cdot \|S^{-1}\| \cdot \left(1 + \frac{1}{n}\|S^{-1}BS\|\right)^n$$

$$\leq \|S\| \cdot \|S^{-1}\| \exp(\|S^{-1}BS\|). \qquad \square$$

In applying this result to sequences of vectors, the term represented by the matrix B can be replaced by a non-linear function which satisfies suitable conditions. To widen the applicability of the result a non-homogeneous term is included.

Theorem 142F *Let A be a stable $m \times m$ matrix and $\phi : \mathbb{R}^m \to \mathbb{R}^m$ be such that $\|\phi(x)\| \leq L\|x\|$, for L a positive constant and $x \in \mathbb{R}^m$. If $w = (w_1, w_2, \ldots, w_n)$ and $v = (v_0, v_1, \ldots, v_n)$ are sequences related by*

$$v_i = Av_{i-1} + \frac{1}{n}\phi(v_{i-1}) + w_i, \quad i = 1, 2, \ldots, n, \qquad (142a)$$

then

$$\|v_n\| \leq C\left(\|v_0\| + \sum_{i=1}^{n}\|w_i\|\right),$$

where C is independent of n.

Proof. Let S be the matrix introduced in the proof of Theorem 142C. From (142a), it follows that

$$(S^{-1}v_i) = (S^{-1}AS)(S^{-1}v_{i-1}) + \frac{1}{n}(S^{-1}\phi(v_{i-1})) + (S^{-1}w_i)$$

and hence

$$\|S^{-1}v_i\| \leq \|S^{-1}AS\| \cdot \|S^{-1}v_{i-1}\| + \frac{1}{n}\|S^{-1}\phi(v_{i-1})\| + \|S^{-1}w_i\|,$$

leading to the bound

$$\|v_n\| \leq \|S\| \cdot \|S^{-1}\| \exp\left(L\|S\| \cdot \|S^{-1}\|\right)\left(\|v_0\| + \sum_{i=1}^{n}\|w_i\|\right). \qquad \square$$

Exercises 14

14.1 Find a constant C such that $\|A^n\|_\infty \leq C$, for all $n = 0, 1, \ldots$, where

$$A = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{3} & \frac{4}{3} \end{bmatrix}.$$

14.2 For what values of the complex number θ is the matrix A stable, where

$$A = \begin{bmatrix} \theta & 1 \\ 0 & 1 \end{bmatrix}.$$

Chapter 2

Numerical Differential Equation Methods

20 The Euler Method

200 Introduction to the Euler methods

The famous method of Euler was published in his three-volume work *Institutiones Calculi Integralis* in the years 1768 to 1770, republished in his collected works (Euler, 1913). This fundamental idea is based on a very simple principle. Suppose that a particle is moving in such a way that, at time x_0, its position is equal to y_0 and that, at this time, the velocity is known to be v_0. The simple principle is that, in a short period of time, so short that there has not been time for the velocity to change significantly from v_0, the change in position will be approximately equal to the change in time multiplied by v_0.

If the motion of the particle is governed by a differential equation, the value of v_0 will be known as a function of x_0 and y_0. Hence, given x_0 and y_0, the solution at x_1, assumed to be close to x_0, can be calculated as

$$y_1 = y_0 + (x_1 - x_0)v_0,$$

which can be found from known values only of x_0, x_1 and y_0. Assuming that v_1, found using the differential equation from the values x_1 and y_1, is sufficiently accurate, a second step can be taken to find y_2, an approximate solution at x_2, using the formula

$$y_2 = y_1 + (x_2 - x_1)v_1.$$

A sequence of approximations y_1, y_2, y_3, ... to the solution of the differential equation at x_1, x_2, x_3, ... is intended to lead eventually to acceptable approximations, at increasingly distant times from where the initial data was given.

Of course, the interpretation of the Euler method is much wider than in the description of the motion of a single particle, moving in time along a line. Even though the independent variable, which we denote by x, will not always have

the meaning of physical time, we will often refer to it as the 'time variable'. The dependent variable y need not have the meaning of distance and need not even be scalar. If y is vector-valued, then it can be interpreted as a collection of scalar-valued components y_1, y_2, ..., y_N. Thus, we can write

$$
y(x) = \begin{bmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_N(x) \end{bmatrix}.
$$

The differential equation, and the initial information, which together determine the values of the y components as the time variable varies, can be written in the form

$$
y'(x) = f(x, y(x)), \qquad y(x_0) = y_0. \tag{200a}
$$

In the vector-valued case, the function f is defined on $\mathbb{R} \times \mathbb{R}^N$ to $\mathbb{R}^N$. However, it is often convenient to write the individual components of f as scalar-valued functions of x and the vector $y(x)$; or, what is equivalent, of the individual components of $y(x)$. Similarly, the initial information can also be written in terms of individual components y_{10}, y_{20}, ..., y_{N0} of y_0. There is a potential for confusion in the use of subscripts to denote either individual components of y, or individual values of x, at which y is evaluated or approximated. This confusion will be avoided by using each notation only in a context which makes the meaning clear, or else, where it becomes necessary, by refining the notation.

With the freedom we have to write y as a vector or as an ensemble of scalars, we see that (200a) can be written in one of several equivalent forms. We can write out the components of $y(x)$ in $f(x, y(x))$ to emphasize the dependence of $y'(x)$ on each of these components:

$$
y'(x) = f(x, y_1(x), y_2(x), \ldots, y_N(x)), \qquad y(x_0) = y_0,
$$

or we can take this a step further by writing out the individual components of $y'(x)$:

$$
\begin{bmatrix} y_1'(x) \\ y_2'(x) \\ \vdots \\ y_N'(x) \end{bmatrix} = f(x, y_1(x), y_2(x), \ldots, y_N(x)), \qquad y(x_0) = \begin{bmatrix} y_{10} \\ y_{20} \\ \vdots \\ y_{N0} \end{bmatrix}.
$$

Finally, we obtain a very detailed formulation by writing everything in terms of individual components:
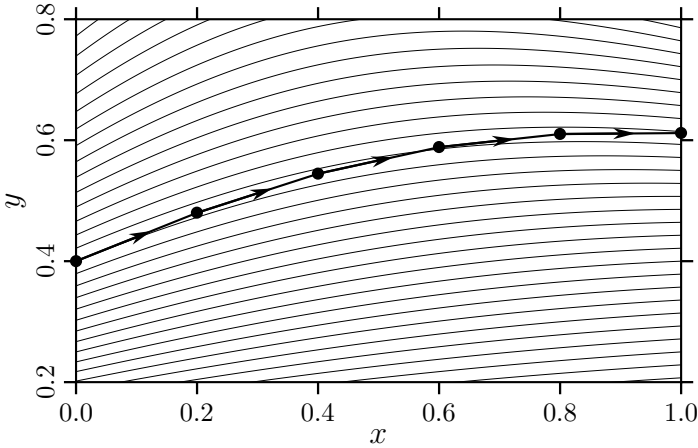
Figure 200(i)   An example of the Euler method

$$
\begin{bmatrix}
y_1'(x) \\
y_2'(x) \\
\vdots \\
y_N'(x)
\end{bmatrix}
=
\begin{bmatrix}
f_1(x, y_1(x), y_2(x), \ldots, y_N(x)) \\
f_2(x, y_1(x), y_2(x), \ldots, y_N(x)) \\
\vdots \\
f_N(x, y_1(x), y_2(x), \ldots, y_N(x))
\end{bmatrix}
,
\quad
\begin{bmatrix}
y_1(x_0) \\
y_2(x_0) \\
\vdots \\
y_N(x_0)
\end{bmatrix}
=
\begin{bmatrix}
y_{10} \\
y_{20} \\
\vdots \\
y_{N0}
\end{bmatrix}.
$$

An important special case is that f – or, for vector problems, each of the functions f_1, f_2, ..., f_N – does not depend on the time variable at all. In this case, we refer to the problem as being 'autonomous', and write it in the form

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

or in one of the expanded forms.

To conclude this subsection, we present a pictorial illustration of the use of the Euler method, for the scalar initial value problem

$$\frac{dy}{dx} = \frac{y - 2xy^2}{1 + x}, \qquad y(0) = \frac{2}{5}. \tag{200b}$$

Five steps with the method, using equally sized time steps $\frac{1}{5}$, are taken and shown against a background of solutions with varying initial values. The general solution to this problem is given by

$$y(x) = \frac{1 + x}{C + x^2},$$

for C an arbitrary constant, and the exact and approximate solutions are shown in Figure 200(i).

201   Some numerical experiments

To see how the Euler method works in practice, consider the initial value problem

$$\frac{dy}{dx} = \frac{y+x}{y-x}, \quad y(0) = 1, \tag{201a}$$

for which the exact solution is

$$y(x) = x + \sqrt{1 + 2x^2}. \tag{201b}$$

To calculate the solution at $x = 0.1$ using the Euler method, we need to use the approximation $y(0.1) \approx y(0) + 0.1y'(0)$. Since $y(0) = 1$ and $y'(0) = 1$, we find $y(0.1) \approx y(0) + 0.1y'(0) = 1 + 0.1 = 1.1$.

We can now take the calculation a second step forward, to find an approximation at $x = 0.2$ using the formula $y(0.2) \approx y(0.1) + 0.1y'(0.1)$. For the value of $y(0.1)$, we can use the result of the first Euler step and for the value of $y'(0.1)$, we can use (201a) with the approximate value of $y(0.1)$ substituted. This gives $y'(0.1) \approx (1.1 + 0.1)/(1.1 - 0.1) = 1.2$. Hence, $y(0.2) \approx y(0.1) + 0.1y'(0.1) \approx 1.1 + 0.12 = 1.22$.

In Table 201(I) these calculations are continued as far as $x = 0.5$. Steps of size 0.1 are taken throughout but, for comparison, the same results are also given for steps of sizes 0.05 and 0.025, respectively. For the three columns of approximations, the headings $h = 0.1$, $h = 0.05$ and $h = 0.025$ denote the sizes of the steps used to arrive at these approximations. The exact values of y are also given in the table.

It is interesting to compare the errors generated in the very first step, for the three values of h that we have used. For $h = 0.1$, the exact solution minus the computed solution is $1.109950 - 1.100000 = 0.009950$; for $h = 0.05$, the corresponding difference is $1.052497 - 1.050000 = 0.002497$; for $h = 0.025$, the difference is $1.025625 - 1.025000 = 0.000625$. It is seen that, approximately, when h is multiplied by a factor of $\frac{1}{2}$, the error in the first step is multiplied by a factor of $\frac{1}{4}$. This is to be expected because, according to Taylor's theorem, the exact answer at $x = h$ is $y(h) \approx y(0) + hy'(0) + (h^2/2)y''(0)$. The first two terms of this approximation are exactly what is calculated by the Euler method, so that the error should be close to $(h^2/2)y''(0)$. We can check this more closely by evaluating $y''(0) = 2$.

Of greater interest in understanding the quality of the numerical approximation is the error accumulated up to a particular x value, by a sequence of Euler steps, with varying value of h. In the case of $x = 0.5$, we see that, for the three stepsizes we have used, the errors are respectively $1.724745 - 1.687555 = 0.037190$, $1.724745 - 1.706570 = 0.018175$ and $1.724745 - 1.715760 = 0.008985$. These error values approximately drop by a factor $\frac{1}{2}$ when h is reduced by this same factor. The reason for this will be discussed more fully in Subsection 212, but it can be understood informally. Note that there is a comparable error produced in each of the steps, but there

Table 201(I) Euler method: problem (201a)

x	$h = 0.1$	$h = 0.05$	$h = 0.025$	y
0.000000	1.000000	1.000000	1.000000	1.000000
0.025000			1.025000	1.025625
0.050000		1.050000	1.051250	1.052497
0.075000			1.078747	1.080609
0.100000	1.100000	1.105000	1.107483	1.109950
0.125000			1.137446	1.140505
0.150000		1.164950	1.168619	1.172252
0.175000			1.200982	1.205170
0.200000	1.220000	1.229729	1.234510	1.239230
0.225000			1.269176	1.274405
0.250000		1.299152	1.304950	1.310660
0.275000			1.341799	1.347963
0.300000	1.359216	1.372981	1.379688	1.386278
0.325000			1.418581	1.425568
0.350000		1.450940	1.458440	1.465796
0.375000			1.499228	1.506923
0.400000	1.515862	1.532731	1.540906	1.548913
0.425000			1.583436	1.591726
0.450000		1.618044	1.626780	1.635327
0.475000			1.670900	1.679678
0.500000	1.687555	1.706570	1.715760	1.724745

are *more* of these steps, if h is small. In the case of the present calculation, the error is about h^2 in each step, but to get as far as $x = 0.5$, $n = 1/2h$ steps have to be carried out. This leads to a total error of about $nh^2 = 0.5h$. A slight refinement of this argument would replace $y''(0)$ by the mean of this quantity over the interval $[0, 0.5]$. The value of this mean is approximately 1.63299, so that the total error should be about $0.40825h$. This very crude argument leads to a prediction that is incorrect by a factor of only about 10%. In the solution of practical problems using the Euler method, or indeed a different method, it is not really feasible to estimate the total accumulated error, but it is important to know the asymptotic form of the error in terms of h. This will often make it possible to gauge the quality of approximations, by comparing the values for differing h values. It will also often make it possible to make realistic decisions as to which of various alternative numerical methods should be used for a specific problem, or even for a large class of problems.

Table 201(II)   Euler method: problem (201d) with $e = 0$

h	y_1	y_2	y_3	y_4	$\|\text{Error}\|$
$\frac{\pi}{200}$	-1.084562	0.133022	-0.159794	-0.944876	0.231124
$\frac{\pi}{400}$	-1.045566	0.067844	-0.085837	-0.973596	0.121426
$\frac{\pi}{800}$	-1.023694	0.034251	-0.044572	-0.987188	0.062333
$\frac{\pi}{1600}$	-1.012087	0.017207	-0.022723	-0.993707	0.031593
$\frac{\pi}{3200}$	-1.006106	0.008624	-0.011474	-0.996884	0.015906
$\frac{\pi}{6400}$	-1.003068	0.004317	-0.005766	-0.998450	0.007981
$\frac{\pi}{12800}$	-1.001538	0.002160	-0.002890	-0.999227	0.003998
$\frac{\pi}{25600}$	-1.000770	0.001080	-0.001447	-0.999614	0.002001

Table 201(III)   Euler method: problem (201d) with $e = \frac{1}{2}$

h	y_1	y_2	y_3	y_4	$\|\text{Error}\|$
$\frac{\pi}{200}$	-1.821037	0.351029	-0.288049	-0.454109	0.569602
$\frac{\pi}{400}$	-1.677516	0.181229	-0.163203	-0.517588	0.307510
$\frac{\pi}{800}$	-1.593867	0.091986	-0.087530	-0.548433	0.160531
$\frac{\pi}{1600}$	-1.548345	0.046319	-0.045430	-0.563227	0.082134
$\frac{\pi}{3200}$	-1.524544	0.023238	-0.023158	-0.570387	0.041559
$\frac{\pi}{6400}$	-1.512368	0.011638	-0.011693	-0.573895	0.020906
$\frac{\pi}{12800}$	-1.506208	0.005824	-0.005875	-0.575630	0.010485
$\frac{\pi}{25600}$	-1.503110	0.002913	-0.002945	-0.576491	0.005251

Table 201(IV)   Euler method: problem (201d) with $e = \frac{3}{4}$

h	y_1	y_2	y_3	y_4	$\|\text{Error}\|$
$\frac{\pi}{200}$	-2.945389	1.155781	-0.739430	0.029212	1.864761
$\frac{\pi}{400}$	-2.476741	0.622367	-0.478329	-0.168796	1.089974
$\frac{\pi}{800}$	-2.162899	0.322011	-0.284524	-0.276187	0.604557
$\frac{\pi}{1600}$	-1.972584	0.163235	-0.158055	-0.329290	0.321776
$\frac{\pi}{3200}$	-1.865987	0.082042	-0.083829	-0.354536	0.166613
$\frac{\pi}{6400}$	-1.809268	0.041102	-0.043252	-0.366542	0.084872
$\frac{\pi}{12800}$	-1.779967	0.020567	-0.021980	-0.372336	0.042847
$\frac{\pi}{25600}$	-1.765068	0.010287	-0.011081	-0.375172	0.021528

It is equally straightforward to solve problems in more than one dependent variable using the Euler method. Given the problem of inverse-square law attraction in two dimensions

$$Y''(x) = -\frac{1}{\|Y(x)\|^{3/2}} Y(x), \tag{201c}$$

where $\|Y\| = \sqrt{Y_1^2 + Y_2^2}$, it is necessary to first write the problem as a system of *first order* equations. This is done by writing y_1 and y_2 for the space coordinates Y_1 and Y_2, and writing y_3 and y_4 for the velocity coordinates, given as the first derivatives of Y_1 and Y_2. With this reformulation, the system of differential equations is written in the form

$$
\begin{aligned}
\frac{dy_1}{dx} &= y_3, \\[2mm]
\frac{dy_2}{dx} &= y_4, \\[2mm]
\frac{dy_3}{dx} &= -\frac{y_1}{(y_1^2 + y_2^2)^{3/2}}, \\[2mm]
\frac{dy_4}{dx} &= -\frac{y_2}{(y_1^2 + y_2^2)^{3/2}}.
\end{aligned}
\tag{201d}
$$

The initial value, written as a vector $y(0) = [1, 0, 0, 1]^{\mathsf{T}}$, defines the solution $y(x) = [\cos(x), \sin(x), -\sin(x), \cos(x)]^{\mathsf{T}}$. The first step of the Euler method gives a numerical result $y(h) \approx [1, h, -h, 1]^{\mathsf{T}}$; this differs from the exact result by approximately $[-\frac{1}{2}h^2, -\frac{1}{6}h^3, \frac{1}{6}h^3, -\frac{1}{2}h^2]^{\mathsf{T}}$. Rather than look at all the components of the error vector individually, it is often convenient to compute the norm of this vector and consider its behaviour as a function of h.

It will be interesting to perform many steps, sufficient to complete, for example, half of one orbit and to compare the (Euclidean) norm of the error for differing values of h. For various values of h, decreasing in sequence by a factor $\frac{1}{2}$, some calculations are presented for this experiment in Table 201(II). The approximate halving of the error, when h is halved, is easily observed in this table.

If the same problem is solved using initial values corresponding to an elliptic, rather than a circular, orbit, a similar dependence of the error on h is observed, but with errors greater in magnitude. Table 201(III) is for an orbit with eccentricity $e = \frac{1}{2}$. The starting value corresponds to the closest point on the orbit to the attracting force, and the exact value at the end of a half period is

$$y(0) = \begin{bmatrix} 1-e \\ 0 \\ 0 \\ \sqrt{\frac{1+e}{1-e}} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ 0 \\ 0 \\ \sqrt{3} \end{bmatrix}, \qquad y(\pi) = \begin{bmatrix} -1-e \\ 0 \\ 0 \\ -\sqrt{\frac{1-e}{1+e}} \end{bmatrix} = \begin{bmatrix} -\frac{3}{2} \\ 0 \\ 0 \\ -\frac{1}{\sqrt{3}} \end{bmatrix}.$$

When the eccentricity is further increased to $e = \frac{3}{4}$, the loss of accuracy in carrying out the computation is even more pronounced. Results for $e = \frac{3}{4}$ are given in Table 201(IV), where we note that, in this case, $y(\pi) = [-\frac{7}{4}, 0, 0, -1/\sqrt{7}]^{\mathsf{T}}$.

202   Calculations with stepsize control

The use of the Euler method, with constant stepsize, may not be efficient for some problems. For example, in the case of the eccentric orbits, discussed in the previous subsection, a small step should be taken for points on the orbit, close to the attracting force, and a larger step for points remote from the attracting force. In deciding how we might attempt to control the stepsize for a general problem, we need to consider how the error committed in each step can be estimated. First, however, we consider how the stepsize in a step should be chosen, to take account of this error estimate.

Because the total error is approximately the sum of the errors committed in the individual steps, at least for a limited number of steps, we look at a simple model in which the interval of integration is divided up into m subintervals, with lengths $\delta_1, \delta_2, \ldots, \delta_m$. We assume that the norms of the errors in steps carried out in these intervals are $C_1 h_1^2, C_2 h_2^2, \ldots, C_m h_m^2$, respectively, where $h_1, h_2, \ldots, h_m$ are the constant stepsizes in these subintervals. Assume that a total of N steps of integration by the Euler method are carried out and that a fraction t_i of these are performed in subinterval $i = 1, 2, \ldots, m$. This means that $t_i N$ steps are carried out in subinterval i and that $h_i = \delta_i / t_i N$. The total error committed, which we assume, in the absence of further information, to be the sum of the individual errors, is approximately

$$E = \sum_{i=1}^{m} (t_i N) C_i \left(\frac{\delta_i}{t_i N} \right)^2 = \frac{1}{N} \sum_{i=1}^{m} \delta_i^2 C_i t_i^{-1}, \qquad (202a)$$

where $\delta_i / t_i N$ is the stepsize used for every step in subinterval number i. By the Cauchy–Schwarz inequality, the minimum value of (202a) is achieved by

$$t_i = \frac{\delta_i \sqrt{C_i}}{\sum_{j=1}^{m} \delta_j \sqrt{C_j}}$$

and it follows that optimality occurs when $C_i h_i^2$ is maintained constant over every subinterval. We interpret this result to mean that the estimated values of the error should be kept as close as possible to some pre-assigned value.
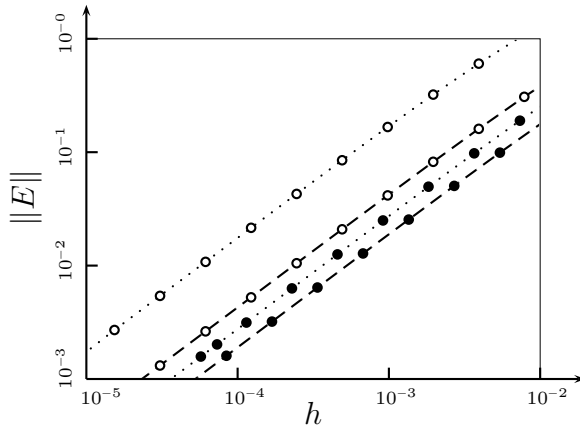
Figure 202(i)   Constant ($\circ$) and variable ($\bullet$) step for orbit with eccentricities $e = \frac{1}{2}$ ($--$) and $e = \frac{3}{4}$ ($\cdots$)

This pre-assigned value, which is under control of the user, will be regarded as the user-imposed tolerance.

To actually estimate the error committed in each step, we have a natural resource at our disposal; this is the availability of approximations to $hy'(x)$ at the beginning and end of every step. At the beginning of step n, it is, of course, the value of $hf(x_{n-1}, y_{n-1})$ used in the computation of the Euler step itself. At the end of this step we can calculate $hf(x_n, y_n)$. This might seem to be an additional calculation of the function f, but this computation needs to be done anyway, since it is needed when the *following* step is eventually carried out. From these approximations to $hy'(x_{n-1})$ and $hy'(x_n)$ we can recalculate the step from y_{n-1} using the more accurate trapezoidal rule to yield the improved approximation to $y(x_n)$, given by

$$y(x_n) \approx y(x_{n-1}) + \frac{1}{2}\big(hy'(x_{n-1}) + hy'(x_n)\big),$$

and we can use the difference between this approximation to $y(x_n)$, and the result computed by the Euler step, as our local error estimate.

Hence we have, as an estimate of the norm of the error,

$$\frac{1}{2}\big\|hf(x_{n-1}, y(x_{n-1})) - hf(x_n, y(x_n))\big\|.$$

As an illustration of how variable stepsize works in practice, the calculations of gravitational orbits with eccentricities 0.5 and 0.75 have been repeated using variable stepsize, but with the tolerances set at values that will give a total number of steps approximately the same as for the constant stepsize cases already investigated. A summary of the results is shown in Figure 202(i). To make the comparisons straightforward, only norms of errors are plotted against stepsize (or mean stepsize in the variable stepsize cases).
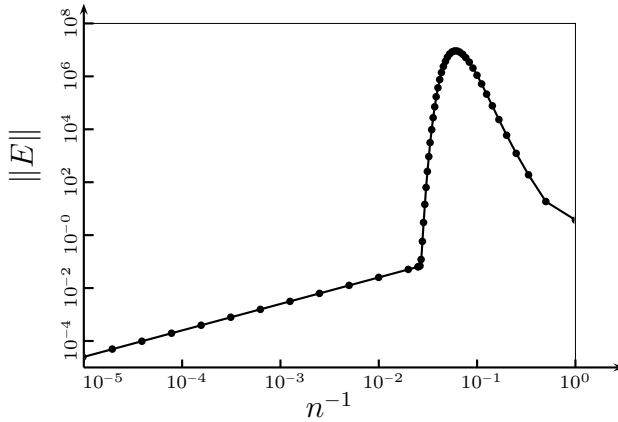
Figure 203(i)   Norm error against n^{-1} for the 'mildly stiff' problem (203a)

203   Calculations with mildly stiff problems

Consider the initial value problem

$$\begin{aligned}
\frac{dy_1}{dx} &= -16y_1 + 12y_2 + 16\cos(x) - 13\sin(x), & y_1(0) &= 1, \\
\frac{dy_2}{dx} &= 12y_1 - 9y_2 - 11\cos(x) + 9\sin(x), & y_2(0) &= 0,
\end{aligned} \tag{203a}$$

for which the exact solution is $y_1(x) = \cos(x)$, $y_2(x) = \sin(x)$. We attempt to solve this problem using the Euler method. First, we use constant stepsize. Specifically, we perform n steps with $h = \pi/n$ and with n taking on various integer values. This yields a sequence of approximations to $y(\pi)$, and results for the norm of the error are given in Figure 203(i).

The results shown here have a disturbing feature. Even though the asymptotic first order behaviour is clearly seen, this effect is recognizable only below a certain threshold, corresponding to $n = 38$. For h above the corresponding value of $\pi/38$, the errors grow sharply, until they dominate the solution itself. We consider what can be done to avoid this extreme behaviour and we turn to variable stepsize as a possible remedy. We need to be more precise than in Subsection 202, in deciding how we should apply this approach. After a step has been completed, we have to either accept or reject the step, and rejecting requires us to repeat the step, but with a scaled-down stepsize. In either case we need a policy for deciding on a stepsize to use in the new attempt at the failed step, or to use in the succeeding new step.

Because the local truncation error is asymptotically proportional to the square of h, it makes sense to scale the stepsize in the ratio $\sqrt{T/\|E\|}$, where E is the error estimate and T is the maximum permitted value of $\|E\|$. However, it is essential to insert a 'safety factor' S, less than 1, into the computation,
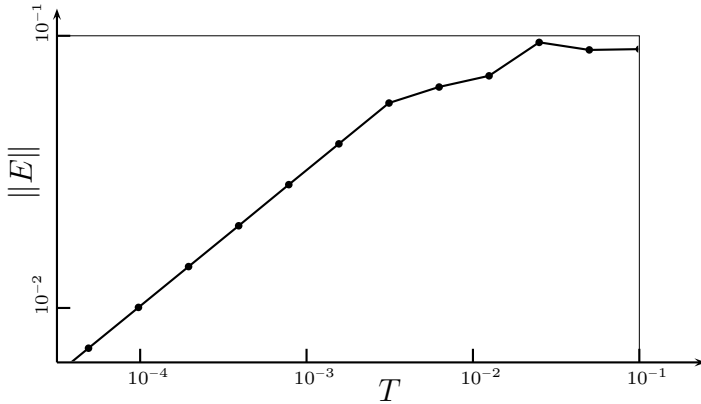
Figure 203(ii)   Norm error against tolerance T for the 'mildly stiff' problem
(203a) with variable stepsize

to guard against a rejection in a new step, because of slight variations in
the magnitude of the error estimate from step to step. It is also wise to use
two further design parameters, M and m, representing the maximum and
minimum stepsize ratios that will be permitted. Typically $M = 2$, $m = \frac{1}{2}$
and $S = 0.9$, and we adopt these values. Fortunately, this experiment of using
variable stepsize is successful, as is seen from Figure 203(ii).

There is a loss of efficiency, in that unstable behaviour typically results
in wide variations of stepsize, in sequences of adjacent steps. However, there
are relatively few steps rejected, because of excessive error estimates. For the
special choice of the tolerance $T = 0.02$, 38 successful steps were taken, in
addition to 11 failed steps. The value of the stepsize h as a function of the
value of x, at the beginning of each of the steps, is shown in Figure 203(iii).

The phenomenon experienced with this example goes under the name of
'stiffness'. To understand why this problem is stiff, and why there seems to
be a value of h such that, for values of the stepsize above this, it cannot
be solved by the Euler method, write $v_1(x)$ and $v_2(x)$ for the deviations of
$y_1(x)$ and $y_2(x)$ from the exact solution. That is, $y_1(x) = \cos(x) + v_1(x)$ and
$y_2(x) = \sin(x) + v_2(x)$. Because the system is linear, it reduces in a simple
way to

$$
\begin{bmatrix} \dfrac{dv_1}{dx} \\ \dfrac{dv_2}{dx} \end{bmatrix} = \begin{bmatrix} -16 & 12 \\ 12 & -9 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.
\tag{203b}
$$

To simplify the discussion further, find the eigenvalues, and corresponding
eigenvectors, of the matrix A occurring in (203b), where

$$
A = \begin{bmatrix} -16 & 12 \\ 12 & -9 \end{bmatrix}.
$$
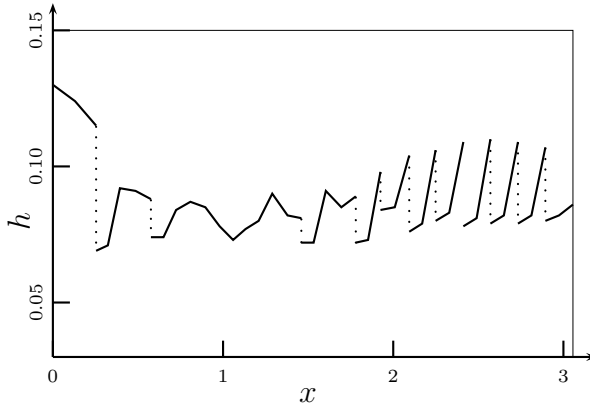
Figure 203(iii)   Stepsize h against x for the 'mildly stiff' problem (203a) with
variable stepsize for $T = 0.02$

The eigenvalues of A are $\lambda_1 = 0$ and $\lambda_2 = -25$ and the eigenvectors are the
columns of the matrix

$$T = \begin{bmatrix} 3 & 4 \\ 4 & -3 \end{bmatrix}.$$

By substituting $v = Tw$, that is,

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 4 & -3 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix},$$

we find that

$$\begin{bmatrix} \dfrac{dw_1}{dx} \\ \dfrac{dw_2}{dx} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & -25 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}.$$

The components of w each have bounded solutions, and thus the original
differential equation is stable. In particular, any perturbation in w_2 will
lead to very little change in the long term solution, because of the quickly
decaying exponential behaviour of this component. On the other hand, when
the equation for w_2 is solved *numerically*, difficulties arise. In a single step of
size h, the exact solution for w_2 should be multiplied by $\exp(-25h)$, but the
numerical approximation is multiplied by $1 - 25h$. Even though $|\exp(-25h)|$
is always less than 1 for positive h, $|1 - 25h|$ is greater than 1, so that its
powers form an unbounded sequence, unless $h \leq \frac{2}{25}$.

   This, then, is the characteristic property of stiffness: components of the
solution that should be stable become unstable when subjected to numerical
approximations in methods like the Euler method.

Table 204(I) Comparison of explicit and implicit Euler methods: problem (201a)

n	Explicit error	Implicit error	Iterations
5	0.03719000	−0.03396724	28
10	0.01817489	−0.01737078	47
20	0.00898483	−0.00878393	80
40	0.00446704	−0.00441680	149
80	0.00222721	−0.00221462	240
160	0.00111203	−0.00110889	480
320	0.00055562	−0.00055484	960
640	0.00027771	−0.00027762	1621

204   Calculations with the implicit Euler method

As we have pointed out, the Euler method approximates the integral of $y'(x)$, over each subinterval $[x_{n-1}, x_n]$, in terms of the width of the interval, multiplied by an approximation to the height of the integrand at the left-hand end. We can consider also the consequences of using the width of this interval, multiplied by the height at the *right*-hand end.

This would mean that the approximation at x_1 would be defined by $y(x_1) \approx y_1$, where $y_1 = y_0 + hf(x_1, y_1)$. This results in what is known as the 'implicit Euler method'. The complication is, of course, that the solution approximation at the end of the step is defined not by an explicit formula, but as the solution to an algebraic equation.

For some problems, we can evaluate y_1 by simple ('fixed point') iteration. That is, we calculate a sequence of approximations $Y^{[0]}$, $Y^{[1]}$, $Y^{[2]}$, … using the formula

$$Y^{[k]} = y_0 + hf(x_1, Y^{[k-1]}), \quad k = 1, 2, 3, \dots .$$

Assuming that the sequence of approximations converges, to within a required tolerance, to a limiting value Y, then we take this limit as the value of y_1. The starting value in the sequence may be taken, for simplicity and convenience, as y_0.

Some results for this method, as applied to the initial value problem (201a), are given in Table 204(I). In this table, all approximations are made for the solution at $x = 0.5$ and, for each number of steps n, the calculation is carried out using both the Euler method and the implicit form of the Euler method. The total errors for the two methods are shown. In the case of the implicit method, the total number of iterations to achieve convergence, to within a
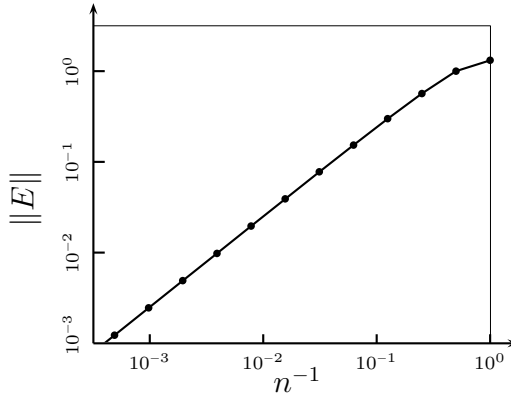
Figure 204(i)   Norm error against n^{-1} for the 'mildly stiff' problem (203a) using the method (204a)

tolerance of 10^{-6}, is also given. If a tolerance as high as 10^{-4} had been specified, there would have been only about two, rather than three, iterations per step, but the cost would still be approximately twice as great as for the explicit Euler method.

As we see from these results, there is no advantage in the implicit form of the Euler method, in the case of this problem. On the contrary, there is a serious disadvantage, because of the very much greater computing cost, as measured in terms of f evaluations, for the implicit as compared with the explicit form of the method.

For stiff problems, such as that given by (203a), the implicit Euler method shows itself to advantage. Since this problem is linear, it is possible to write the answer for the approximation computed at the end of a step explicitly. In the step going from x_0 to $x_1 = x_0 + h$, with solution approximations going from $y_0 = [(y_0)_1, (y_0)_2]^\mathsf{T}$ to $y_1 = [(y_1)_1, (y_1)_2]^\mathsf{T}$, we have the relations between these quantities given by

$$\begin{bmatrix} (y_1)_1 \\ (y_1)_2 \end{bmatrix} = h \begin{bmatrix} -16 & 12 \\ 12 & -9 \end{bmatrix} \begin{bmatrix} (y_1)_1 \\ (y_1)_2 \end{bmatrix} + \begin{bmatrix} (y_0)_1 \\ (y_0)_2 \end{bmatrix} + h \begin{bmatrix} 16\cos(x_1) - 13\sin(x_1) \\ -11\cos(x_1) + 9\sin(x_1) \end{bmatrix},$$

so that

$$\begin{bmatrix} 1 + 16h & -12h \\ -12h & 1 + 9h \end{bmatrix} \begin{bmatrix} (y_1)_1 \\ (y_1)_2 \end{bmatrix} = \begin{bmatrix} (y_0)_1 + 16h\cos(x_1) - 13h\sin(x_1) \\ (y_0)_2 - 11h\cos(x_1) + 9h\sin(x_1) \end{bmatrix}, \quad (204a)$$

and the new approximation is found using a linear equation solution.

The results for this calculation, presented in Figure 204(i), show that this method is completely satisfactory, for this problem. Note that the largest stepsize used is π, so that only a single step is taken.

Exercises 20

20.1 On a copy of Figure 200(i), plot the points corresponding to the solution computed by the Euler method with $y(0) = \frac{1}{4}$, $h = \frac{1}{5}$.

20.2 Write the initial value problem (200b) in the form

$$
\begin{array}{llll}
\frac{dx}{dt} &=& 1 + x, & x(0) &=& 0, \\
\frac{dy}{dt} &=& y - 2xy^2, & y(0) &=& \frac{1}{2}.
\end{array}
$$

Using this alternative formulation, recalculate the solution, using five equal steps of the Euler method, from $t = 0$ to $t = \ln 2$. Plot the solution points after each step on a graph in the (x, y) plane.

20.3 Continue the calculations in Table 201(I) to the point $x = 1$.

20.4 It is known that $E = \frac{1}{2}(y_3^2 + y_4^2) - 1/\sqrt{y_1^2 + y_2^2}$, the total energy, and $A = y_1 y_4 - y_2 y_3$, the angular momentum, are invariants of the system (201d); that is, for any value of x the values of each of these will be equal respectively to the values they had at the initial time. The quality of a numerical method for solving this problem can be measured by calculating by how much these theoretical invariants actually change in the numerical computation. Repeat the calculations in Tables 201(II), 201(III) and 201(IV) but with the deviation in the values of each of these quantities used in place of the errors.

21   Analysis of the Euler Method

210   Formulation of the Euler method

Consider a differential equation system

$$y'(x) = f(x, y(x)), \qquad y(x_0) = y_0, \tag{210a}$$

where $f : [a, b] \times \mathbb{R}^N \to \mathbb{R}^N$ is continuous and satisfies a Lipschitz condition $\|f(x, y) - f(x, z)\| \leq L\|y - z\|$, for all x in a neighbourhood of x_0 and y and z in a neighbourhood of y_0. For simplicity, we assume that the Lipschitz condition holds everywhere; this is not a serious loss of generality because the existence and uniqueness of a solution to (210a) is known to hold in a suitable interval, containing x_0, and we can extend the region where a Lipschitz condition holds to the entire N-dimensional vector space, secure in the knowledge that no practical difference will arise, because the solution will never extend beyond values in some compact set.

We assume that the solution to (210a) is required to be approximated at a point $\overline{x}$, and that a number of intermediate step points are selected. Denote these by $x_1, x_2, \ldots, x_n = \overline{x}$. Define a function, $\widetilde{y}$, on $[x_0, \overline{x}]$ by the formula

$$\widetilde{y}(x) = \widetilde{y}(x_{k-1}) + (x - x_{k-1})f(x_{k-1}, \widetilde{y}(x_{k-1})), \quad x \in (x_{k-1}, x_k], \tag{210b}$$

for $k = 1, 2, \ldots, n$. If we assume that $\widetilde{y}(x_0) = y(x_0) = y_0$, then $\widetilde{y}$ exactly agrees with the function computed using the Euler method at the points $x = x_k$, $k = 1, 2, \ldots, n$. The continuous function $\widetilde{y}$, on the interval $[x_0, \overline{x}]$, is a piecewise linear interpolant of this Euler approximation.

We are interested in the quality of $\widetilde{y}$ as an approximation to y. This will clearly depend on the values of the step points x_1, x_2, $\ldots$, and especially on the greatest of the distances between a point and the one preceding it. Denote the maximum of $x_1 - x_0$, $x_2 - x_1$, $\ldots$, $x_n - x_{n-1}$ by H.

We would like to know what happens to $\|\widetilde{y}(\overline{x}) - y(\overline{x})\|$ as $H \to 0$, given also that $\|\widetilde{y}(x_0) - y(x_0)\| \to 0$. It is also interesting to know what happens to the uniform norm of $\|\widetilde{y}(x) - y(x)\|$, for x in $[x_0, \overline{x}]$. Under very general conditions, we show that $\widetilde{y}$ converges uniformly to y, as the mesh is refined in this way.

211   Local truncation error

In a single step of the Euler method, the computed result, $y_0 + hf(x_0, y_0)$, differs from the exact answer by

$$y(x_0 + h) - y(x_0) - hf(x_0, y(x_0)) = y(x_0 + h) - y(x_0) - hy'(x_0).$$

Assuming y has continuous first and second derivatives, this can be written in the form

$$h^2 \int_0^1 (1 - s)y''(x_0 + hs)ds. \tag{211a}$$

For $i = 1, 2, \ldots, N$, component i can be written, using the mean value theorem, as $\frac{1}{2}h^2$ times component i of $y''(x_0 + hs^*)$, where s^* is in the interval $(0, 1)$. Another way of writing the error, assuming that third derivatives also exist and are bounded, is

$$\frac{1}{2}h^2 y''(x_0) + O(h^3). \tag{211b}$$

This form of the error estimate is quite convenient for interpreting numerically produced results, because if h is sufficiently small, the local error will appear to behave like a constant vector multiplied by h^2. It is also useful for determining how stepsize control should be managed.

212   Global truncation error

After many steps of the Euler method, the errors generated in these steps will accumulate and reinforce each other in a complicated manner. It is important to understand how this happens. We assume a uniform bound $h^2 m$ on the norm of the local truncation error committed in any step of length h. We aim to find a global error bound using a difference inequality. We make the standard assumption that a Lipschitz condition holds, and we write L as the Lipschitz constant.

Recall that $\widetilde{y}(x)$ denotes the computed solution on the interval $[x_0, \overline{x}]$. That is, at step values $x_0, x_1, \ldots, x_n = \overline{x}$, $\widetilde{y}$ is computed using the equation $\widetilde{y}(x_k) = y_k = y_{k-1} + (x_k - x_{k-1})f(x_{k-1}, y_{k-1})$. For 'off-step' points, $\widetilde{y}(x)$ is defined by linear interpolation; or, what is equivalent, $\widetilde{y}(x)$ is evaluated using a partial step from the most recently computed step value. That is, if $x \in (x_{k-1}, x_k)$, then

$$\widetilde{y}(x) = y_{k-1} + (x - x_{k-1})f(x_{k-1}, y_{k-1}). \tag{212a}$$

Let $\alpha(x)$ and $\beta(x)$ denote the errors in $\widetilde{y}(x)$, as an approximation to $y(x)$, and in $f(x, \widetilde{y}(x))$, as an approximation to $y'(x)$, respectively. That is,

$$\alpha(x) = y(x) - \widetilde{y}(x), \tag{212b}$$
$$\beta(x) = f(x, y(x)) - f(x, \widetilde{y}(x)), \tag{212c}$$

so that, by the Lipschitz condition,

$$\|\beta(x)\| \leq L\|\alpha(x)\|. \tag{212d}$$

Define $E(x)$ so that the exact solution satisfies

$$y(x) = y(x_{k-1}) + (x - x_{k-1})f(x_{k-1}, y(x_{k-1})) + (x - x_{k-1})^2 E(x),$$
$$x \in (x_{k-1}, x_k], \quad \text{(212e)}$$

and we assume that $\|E(x)\| \leq m$.

Subtract (212a) from (212e), and use (212b) and (212c), so that

$$\alpha(x) = \alpha(x_{k-1}) + (x - x_{k-1})\beta(x_{k-1}) + (x - x_{k-1})^2 E(x).$$

Hence,

$$\begin{aligned}
\|\alpha(x)\| &\leq \|\alpha(x_{k-1})\| + (x - x_{k-1})\|\beta(x_{k-1})\| + (x - x_{k-1})^2 m \\
&\leq \|\alpha(x_{k-1})\| + (x - x_{k-1})L\|\alpha(x_{k-1})\| + (x - x_{k-1})^2 m \\
&\leq (1 + (x - x_{k-1})L)\|\alpha(x_{k-1})\| + (x - x_{k-1})^2 m \\
&\leq (1 + (x - x_{k-1})L)\|\alpha(x_{k-1})\| + (x - x_{k-1})Hm,
\end{aligned}$$

where we have used (212d) and assumed that no step has a length greater than H. We distinguish two cases. If $L = 0$, then it follows that

$$\|\alpha(x)\| \leq \|\alpha(x_0)\| + Hm(x - x_0); \tag{212f}$$

and if $L > 0$, it follows that

$$\begin{aligned}
\left(\|\alpha(x)\| + \frac{Hm}{L}\right) &\leq (1 + (x - x_{k-1})L)\left(\|\alpha(x_{k-1})\| + \frac{Hm}{L}\right) \\
&\leq \exp((x - x_{k-1})L)\left(\|\alpha(x_{k-1})\| + \frac{Hm}{L}\right).
\end{aligned}$$

Let $\phi(x) = \exp(-(x-x_0)L)(\|\alpha(x)\| + Hm/L)$, so that $\phi(x)$ never increases. Hence,

$$\|\alpha(x)\| \leq \exp((x-x_0)L)\|\alpha(x_0)\| + \frac{\exp((x-x_0)L)-1}{L}Hm.$$

Combining the estimates found in the two cases and stating them formally, we have:

Theorem 212A *Assuming that f satisfies a Lipschitz condition, with constant L, the global error satisfies the bound*

$$\|y(x) - \widetilde{y}(x)\| \leq \begin{cases} \|y(x_0) - \widetilde{y}(x_0)\| + Hm(x-x_0), & L = 0, \\ \exp((x-x_0)L)\|y(x_0) - \widetilde{y}(x_0)\| + (\exp((x-x_0)L)-1)\frac{Hm}{L}, & \\ & L > 0. \end{cases}$$

213   Convergence of the Euler method

We consider a sequence of approximations to $y(\overline{x})$. In each of these approximations, a computation using the Euler method is performed, starting from an approximation to $y(x_0)$, and taking a sequence of positive steps. Denote approximation number n by $\widetilde{y}_n$.

The only assumption we will make about $\widetilde{y}_n$, for each specific value of n, is that the initial error $y(x_0) - \widetilde{y}_n(x_0)$ is bounded in norm by K_n and that the greatest stepsize is bounded by H_n. It is assumed that, as $n \to \infty$, $H_n \to 0$ and $K_n \to 0$. As always, we assume that f satisfies a Lipschitz condition.

Denote by D_n the value of $\|y(\overline{x}) - \widetilde{y}_n(\overline{x})\|$.

Theorem 213A *Under the conditions stated in the above discussion, $D_n \to 0$ as $n \to \infty$.*

Proof. This result follows immediately from the bound on accumulated errors given by Theorem 212A. $\qquad\square$

The property expressed in this theorem is known as 'convergence'. In searching for other numerical methods that are suitable for solving initial value problems, attention is usually limited to convergent methods. The reason for this is clear: a non-convergent method is likely to give increasingly meaningless results as greater computational effort is expended through the use of smaller stepsizes.

Because the bound used in the proof of Theorem 213A holds not only for $x = \overline{x}$, but also for all $x \in [x_0, \overline{x}]$, we can state a uniform version of this result.

Theorem 213B *Under the conditions of Theorem 213A,*

$$\sup_{x\in[x_0,\overline{x}]} \|y(x) - \widetilde{y}_n(x)\| \to 0$$

as $n \to \infty$.

Table 214(I) An example of enhanced order for problem (214a)

n	$|Error|$	Ratio
20	$1130400.0252 \times 10^{-10}$	4.4125
40	$256178.9889 \times 10^{-10}$	4.1893
80	$61150.2626 \times 10^{-10}$	4.0904
160	$14949.6176 \times 10^{-10}$	4.0442
320	$3696.5967 \times 10^{-10}$	4.0218
640	919.1362×10^{-10}	4.0108
1280	229.1629×10^{-10}	4.0054
2560	57.2134×10^{-10}	4.0026
5120	14.2941×10^{-10}	4.0003
10240	3.5733×10^{-10}	

214 Order of convergence

It is interesting to know not only that a numerical result is convergent, but also how quickly it converges. In the case of a constant stepsize h, the bound on the global error given in Theorem 212A is proportional to h. We describe this by saying that the order of the Euler method is (at least) 1.

That the order is *exactly* 1, and that it is not possible, for a general differential equation, to obtain error behaviour proportional to some higher power of h, can be seen from a simple example. Consider the initial value problem

$$y'(x) = 2x, \qquad y(0) = 0,$$

with exact solution $y(x) = x^2$. If $\overline{x} = 1$, and n steps are performed with stepsize $h = n^{-1}$, the computed solution is

$$h \sum_{k=0}^{n-1} \frac{2k}{n} = \frac{n-1}{n}.$$

This differs from the exact solution by $1/n = h$.

In spite of the fact that the order is only 1, it is possible to obtain higher order behaviour in special specific situations. Consider the initial value problem

$$y'(x) = -y(x)\tan(x) - \frac{1}{\cos(x)}, \qquad y(0) = 1, \qquad (214a)$$

with solution $y(x) = \cos(x) - \sin(x)$. Because of an exact cancellation of the most significant terms in the error contributions, at different parts of the
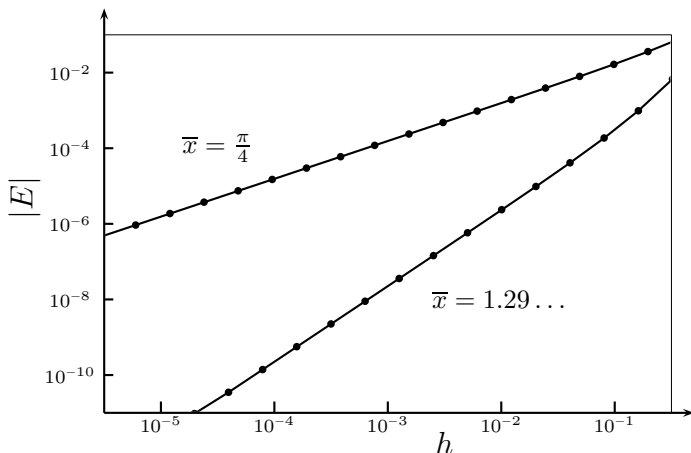
Figure 214(i)   Error versus stepsize for problem (214a) at two alternative
output points

trajectory, the computed results for this problem are consistent with the order
being 2 rather than 1, if the output value is taken as $\overline{x} \approx 1.292695719373$.
Note that $\overline{x}$ was chosen to be a zero of $\exp(x)\cos(x) = 1$. As can be seen from
Table 214(I), as the number of steps doubles, the error reduces by a factor
approximately equal to 2^{-2}. This is consistent with second order, rather than
first order, behaviour. The errors are also plotted in Figure 214(i).

An analysis of the apparent cancellation of the most significant component
of the global truncation error is easy to carry out if we are willing to do
the estimation with terms, which decrease rapidly as $h \to 0$, omitted from
the calculation. A more refined analysis would take these additional terms
into account, but would obtain bounds on their effect on the final result. In
step k, from a total of n steps, the local truncation error is approximately
$-\frac{1}{2}h^2(\cos(x_k) - \sin(x_k))$. To find the contribution this error makes to the
accumulated error at $x_n = \overline{x}$, multiply by the product

$$\big(1 - h\tan(x_{n-1})\big)\big(1 - h\tan(x_{n-2})\big)\cdots\big(1 - h\tan(x_k)\big). \qquad (214b)$$

We have the approximation

$$\frac{\cos(x + h)}{\cos(x)} = \cos(h) - \sin(h)\tan(x) \approx 1 - h\tan(x),$$

so that (214b) can be written approximately as

$$\frac{\cos(x_n)}{\cos(x_{n-1})}\frac{\cos(x_{n-1})}{\cos(x_{n-2})}\cdots\frac{\cos(x_{k+1})}{\cos(x_k)} = \frac{\cos(x_n)}{\cos(x_k)}.$$

Table 214(II)   An example of reduced order for problem (214c)

n	|Error|	Ratio
8	0.3012018700	1.4532
16	0.2072697687	1.4376
32	0.1441738248	1.4279
64	0.1009724646	1.4220
128	0.0710078789	1.4186
256	0.0500556444	1.4166
512	0.0353341890	1.4155
1024	0.0249615684	1.4149
2048	0.0176414532	1.4146
4096	0.0124709320	1.4144
8192	0.0088169646	1.4143
16384	0.0062340372	1.4143
32768	0.0044079422	

Multiply this by the error in step k and add over all steps. The result is

$$-\frac{1}{2}h^2\cos(\overline{x})\sum_{k=1}^{n}\frac{\cos(x_k)-\sin(x_k)}{\cos(x_k)},$$

which is approximately equal to the integral

$$-\frac{1}{2}h\cos(\overline{x})\int_0^{\overline{x}}\frac{\cos(x)-\sin(x)}{\cos(x)}dx = -\frac{1}{2}h\cos(\overline{x})(\overline{x}+\ln\cos(\overline{x})).$$

This vanishes when $\exp(\overline{x})\cos(\overline{x})=1$.

For comparison, results are also given in Figure 214(i) for a similar sequence of h values, but at the output point $\overline{x}=\pi/4$. This case is unsurprising, in that it shows typical order 1 behaviour.

Finally, we present a problem for which an order, even as high as 1, is not observed. The initial value problem is

$$y'(x) = -\frac{xy}{1-x^2}, \qquad y(0) = 1, \qquad (214c)$$

with exact solution $y=\sqrt{1-x^2}$. The solution is sought at $\overline{x}=1$ and the numerical results are shown in Table 214(II). It is seen that, as the number of steps doubles, the error reduces by a factor of approximately $2^{-1/2}$. Thus,
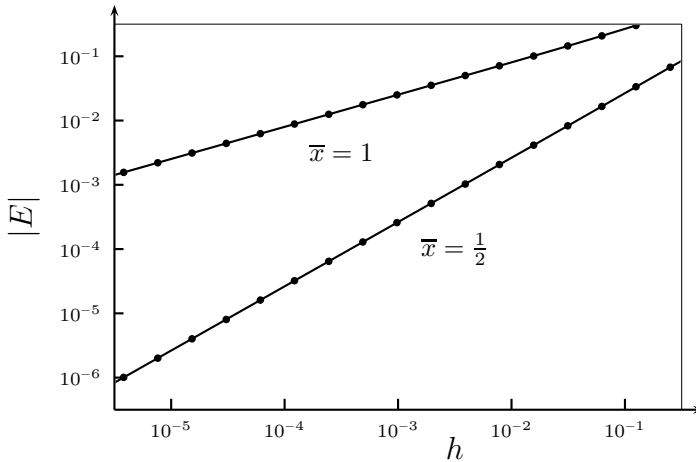
Figure 214(ii)   Error versus stepsize for problem (214c) at two alternative
output points

the order seems to have been reduced from 1 to $\frac{1}{2}$. The reason for the loss of
order for this problem is that the Lipschitz condition does not hold at the end
of the trajectory (at $x = 1$, $y = 0$). As for any initial value problem, the error
in the approximate solution at this point develops from errors generated at
every time step. However, in this case, the local truncation error in the very
last step is enough to overwhelm the contributions to the error inherited from
all previous steps. In fact the local truncation error for the final step is

$$y(1) - y(1 - h) - hf(1 - h, y(1 - h))$$

$$= -\sqrt{1 - (1 - h)^2} + h(1 - h)\frac{\sqrt{1 - (1 - h)^2}}{1 - (1 - h)^2},$$

which simplifies to

$$-\frac{1}{\sqrt{2 - h}}h^{\frac{1}{2}} \approx -2^{-\frac{1}{2}}h^{\frac{1}{2}}.$$

Thus, the order $\frac{1}{2}$ behaviour can be explained just by the error contributed
by the last step.

A second computation, for the solution at $\overline{x} = \frac{1}{2}$, causes no difficulty and
both results are shown in Figure 214(ii).

215   Asymptotic error formula

In a numerical approximation to the solution to a differential equation, using
the Euler method, contributions to the total error are typically produced in
every step. In addition to this, there may be errors introduced at the very

start of the integration process, due to an inaccuracy in the numerical initial value. We attempt to model the development of this error using an asymptotic approach. That is, we assume that the magnitude of all contributions to the error are bounded in terms of some small parameter. We consider only the limiting case, as all stepsizes tend to zero. Consider a step which advances the approximate solution from x to $x + h$. Because the local truncation error in this step is approximately $\frac{1}{2}y''(x)h^2$, the rate at which errors are being generated, as x increases, will be approximately $y''(x)h$.

We suppose that for a step starting at x, the stepsize is equal to $Hs(x)$, where $0 < s(x) \leq 1$ throughout the integration. We use H as the small parameter, referred to above, and assume that the initial error is equal to a constant, which we denote by v_0, times H. Using the integrated form of the differential equation,

$$y(x) = y(x_0) + \int_{x_0}^{x} f(x, y(x))dx, \qquad (215a)$$

we write the perturbation to y, defining the numerical approximation, as $y(x) + Hv(x)$. Thus $y(x) + Hv(x)$ is approximately equal to

$$y(x) + Hv(x) = y(x_0) + Hv_0 + \int_{x_0}^{x} \left(f(x, y(x) + Hv(x)) + \tfrac{1}{2}Hs(x)y''(x) \right)dx.$$

Because H is small, we approximate $f\big(x, y(x) + Hv(x)\big)$ by $f(x, y(x)) + H(\partial f/\partial y)v(x)$:

$$y(x) + Hv(x) = y(x_0) + Hv_0$$
$$+ \int_{x_0}^{x} \left(f(x, y(x)) + H\frac{\partial f}{\partial y}v(x) + \tfrac{1}{2}Hs(x)y''(x) \right)dx. \quad (215b)$$

Subtract (215a) from (215b), divide the difference by H, and we find

$$v(x) = v_0 + \int_{x_0}^{x} \left(\frac{\partial f}{\partial y}v(x) + \frac{1}{2}s(x)y''(x) \right)dx,$$

so that v satisfies the initial value problem

$$v'(x) = \frac{\partial f}{\partial y}v(x) + \frac{1}{2}s(x)y''(x), \quad v(x_0) = v_0. \qquad (215c)$$

We use this result in an attempt to understand the contribution to the total error of local errors introduced at various points on the trajectory. This is done by writing $\Phi(\xi, \overline{x})$ for the solution at $\overline{x}$ to the differential equation

$$w'(x) = \frac{\partial f}{\partial y}w(x), \quad w(\xi) = I,$$

where w takes values in the space of $N \times N$ matrices. In the special case where $\partial f/\partial y$ is a constant matrix M, the solution is

$$\Phi(\xi, \overline{x}) = \exp((\overline{x} - \xi)M).$$

We can now write the solution at $x = \overline{x}$ of (215c) in the form

$$v(\overline{x}) = \Phi(x_0, \overline{x})v_0 + \frac{1}{2} \int_{x_0}^{\overline{x}} \Phi(x, \overline{x})s(x)y''(x)dx.$$

This suggests that s should be chosen, as closely as possible, to maintain a constant value of $\|\Phi(x, \overline{x})s(x)y''(x)\|$, if the norm of the total error is to be kept low for a given number of steps performed.

216   Stability characteristics

In addition to knowing that a numerical method converges to the true solution over a bounded interval, it is interesting to know how errors behave over an unbounded interval. Obtaining quantitative results is difficult, because we are no longer able to take limits, as stepsizes tend to zero. Hence, our attention will move towards qualitative questions, such as whether or not a computed result remains bounded. By comparing the answer to questions like this with the known behaviour of the exact solution, we obtain further insight into the appropriateness of the numerical approximation to model the differential equation.

A further reason for carrying out this type of qualitative analysis is that so-called 'stiff problems' frequently arise in practice. For such problems, qualitative or 'stability' analysis is vital in assessing the fitness of the method to be used in the numerical solution.

Because of the great complexity of this type of analysis, we need to restrict ourselves to purely linear problems with constant coefficients. Thus, we could consider a system of differential equations of the form

$$y'(x) = My(x), \tag{216a}$$

with the matrix M constant. Using fixed stepsize h, the Euler method gives as the approximate solution at $x_n = x_0 + nh$,

$$y_n = (I + hM)y_{n-1},$$

leading to the numerical solution

$$y_n = (I + hM)^n y_0. \tag{216b}$$

For this problem, the exact solution is

$$y(x_n) = \exp(nhM)y(x_0). \tag{216c}$$

We wish to examine some features of the approximate solution (216b) by comparing these features with corresponding features of the exact solution (216c).

By making a change of basis, so that $y(x) = S\widehat{y}(x)$, and $y_n = S\widehat{y}_n$, where S is a constant non-singular matrix, we can rewrite the differential equation in the form

$$\widehat{y}'(x) = \widehat{M}\widehat{y}(x), \tag{216d}$$

where $\widehat{M} = S^{-1}MS$. The solution is

$$\widehat{y}(x_n) = \exp(nh\widehat{M})\widehat{y}(x_0).$$

The solution computed by the Euler method transforms to

$$\widehat{y}_n = (I + h\widehat{M})^n \widehat{y}_0.$$

If the transformed matrix $\widehat{M}$ is chosen as the Jordan canonical form of M, then the differential equation system (216d) and the numerical approximation become, to some extent, decoupled. This means that, for each distinct eigenvalue q, one of the equations in the system (216d) has the simple form

$$y'(x) = qy(x), \tag{216e}$$

and other components that correspond to the same Jordan block will depend on this solution, but will not contribute to its behaviour.

Hence, to obtain acceptable behaviour, for the type of linear problem given by (216a), it is essential that we obtain acceptable behaviour for (216e). All this will mean is that $(1 + hq)^n$ will be an acceptable approximation to $\exp(nhq)$. At very least, we want bounded behaviour for $(1+hq)^n$, as $n \to \infty$, whenever $\exp(nhq)$ is bounded. This, in turn, implies that $|1+hq|$ is bounded by 1, if $\operatorname{Re} q \leq 0$ and q is an eigenvalue of M. Because any analysis of this type will involve the product of h and q, it is convenient to write this product as $z = hq$. We allow the possibility that z is complex, because there is no reason for M to have only real eigenvalues.

The set of points in the complex plane, in which z may lie for this stable behaviour, is known as the 'stability region'. Because it is the set for which $|1 + z| \leq 1$, this stability region is the disc with centre at -1 and radius 1. This is shown as the unshaded region in Figure 216(i). By contrast, we can find the stability region of the implicit Euler method by replacing $hf(x_n, y_n)$ by zy_n in the formula defining this method. That is, $y_n = y_{n-1} + hf(x_n, y_n)$ becomes

$$y_n = y_{n-1} + zy_n.$$

Hence, $y_n = (1-z)^{-1}y_{n-1}$, and the sequence formed by this relation is bounded if and only if $|1 - z| \geq 1$. This is the complement in the complex plane of the interior of the disc with centre 1 and radius 1, shown as the unshaded region of Figure 216(ii).
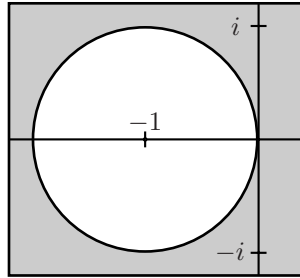
Figure 216(i)   Stability region: Euler method
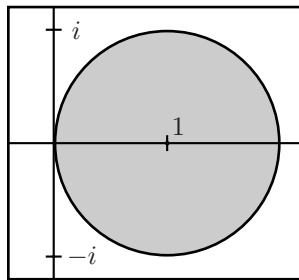
Figure 216(ii)   Stability region: implicit Euler method

Even if we cannot obtain accurate approximations to the solution to equations like (216e), we frequently wish to guarantee that the numerical approximation is bounded in cases when the *exact* solution is bounded. This means that we are especially interested in numerical methods, for which the stability region includes all of the left half-plane. This is the case for the implicit Euler method (Figure 216(ii)) but, as we clearly see from Figure 216(i), not for the Euler method itself. Methods with this desirable property are said to be 'A-stable'. It is widely accepted that this property is close to being essential for stiff problems.

For these two one-step methods, the ratio y_n/y_{n-1} is known as the 'stability function'. Denote this by $R(z)$ so that

$$R(z) = \begin{cases} 1 + z, & \text{(Euler method)} \\ \dfrac{1}{1-z}. & \text{(implicit Euler method)} \end{cases}$$

From a consideration of elementary complex analysis, the property of A-stability can be expressed slightly differently. Obviously, for a method to be A-stable, the stability function must have no poles in the left half-plane. Also the magnitude $|R(z)|$ must be bounded by 1, for z on the imaginary axis.
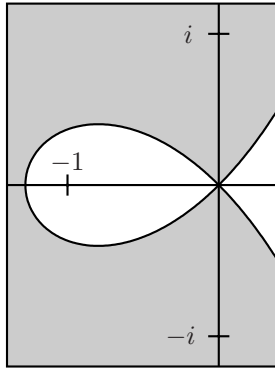
Figure 216(iii)   Order star: Euler method
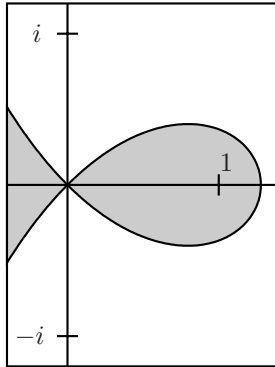
Figure 216(iv)   Order star: implicit Euler method

The interesting thing is that these two conditions are also sufficient for A-stability. If a method with these properties were *not* A-stable, then this would be contrary to the maximum modulus principle.

Multiplying $R(z)$ by $\exp(-z)$ should make no difference to these conclusions. That is, if the set in the complex plane for which $|R(z)\exp(-z)| \leq 1$ is plotted instead, A-stability can still be categorized by this set, including the imaginary axis, together with there being no poles in the left half-plane. The reason for this assertion is that the factor $\exp(-z)$ does not add to, or take away from, the set of poles. Furthermore, its magnitude is precisely 1 when the real part of z is zero.

The modified plots for the two methods are shown in Figures 216(iii) and 216(iv). These were named 'order stars' by their inventors, Wanner, Hairer and Nørsett (1978). The important new feature, introduced by the insertion of
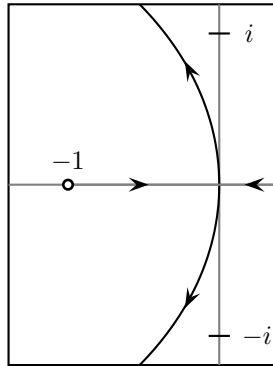
Figure 216(v)   Order arrows: Euler method
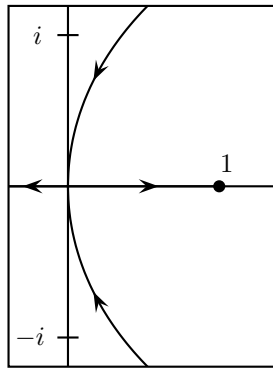
Figure 216(vi)   Order arrows: implicit Euler method

the exponential factor, is the pattern that has appeared near zero. Because, for each of these methods, $R(z)\exp(-z) = 1 + Cz^2 + O(z^3)$, the magnitude of this will be greater than, equal to, or less than 1 for $|z|$ small, approximately when the real part of Cz^2 is positive, zero or negative, respectively. The regions adjoining zero for which $\mathrm{Re}(R(z)\exp(-z)) > 0$ are known as 'fingers', and those for which $\mathrm{Re}(R(z)\exp(-z)) < 0$, are 'dual fingers'. The bounded fingers necessarily contain poles and the bounded dual fingers necessarily contain zeros. For both the Euler method and the implicit Euler method, there is an exact pairing between zeros and bounded dual fingers, and between poles and bounded fingers. Since this pairing also generalizes to other large classes of methods, specifically those methods for which the order is maximal, given the degrees of the numerator and denominator in the stability function, it is possible to relate the angles, at which fingers come out from zero, to the positions of the poles. It will be shown in Subsection 354 how this can be

used to determine the possible A-stability of specific methods, and classes of methods.

Although less well known, order arrows have a role similar to that of order stars, in the analysis of stability questions. For a given stability function $R(z)$, we plot the paths in the complex plane where $w(z) = \exp(-z)R(z)$ is real and positive. Arrows are attached to the paths to show the direction of increasing w. For the Euler and implicit Euler methods, order arrow diagrams are shown in Figures 216(v) and 216(vi) respectively.

217 *Local truncation error estimation*

We recall from Subsection 202 that stepsize control based on a local error estimate was useful in forcing the Euler method to devote computational effort to those parts of the trajectory where it is most needed. We discuss here the principles behind this idea.

Let y_1, y_2, ..., y_{n-1}, y_n, ... denote a sequence of approximations to the solution to an initial value problem, computed using the Euler method. For our present purposes, we can assume that the stepsize takes a constant value h, since we are discussing the estimation of the local truncation error only over a single interval. Because we are considering the *local* error, we treat the incoming approximation for step n as though it were exact. That is, we introduce a solution $\widehat{y}$ to the initial value problem

$$\widehat{y}'(x) = f(x, \widehat{y}(x)), \qquad \widehat{y}(x_{n-1}) = y_{n-1}.$$

We can then interpret $\widehat{y}(x_n) - y_n$ as the error introduced in step n alone.

Although it is not feasible to obtain convenient and useful bounds on this quantity, it is possible to obtain asymptotically correct approximations without additional cost. These will often be useful for the purpose of controlling the stepsize, to produce efficient numerical algorithms, although they cannot be used to obtain rigorous error bounds.

An approximation for $\widehat{y}(x_n)$, to within $O(h^3)$, is found using a truncated Taylor series

$$\widehat{y}(x_{n-1} + h) \approx \widehat{y}(x_{n-1}) + h\widehat{y}'(x_{n-1}) + \frac{h^2}{2!}\widehat{y}''(x_{n-1}),$$

and the first two terms are

$$\widehat{y}(x_{n-1}) + h\widehat{y}'(x_{n-1}) = y_{n-1} + hf(x_{n-1}, y_{n-1}) = y_n.$$

Hence, we see that the truncation error is approximately

$$\frac{h^2}{2!}\widehat{y}''(x_{n-1}).$$

An alternative interpretation of this quantity, at least asymptotically, with terms involving third and higher powers of h ignored, is as the difference

between the result computed by the Euler method and a result computed, at least for the current step, using a method which has a higher order.

As we will see in Section 22, there are many ways in which such a higher order method can be found. One method is to evaluate $hf(x_n, y_n)$, and to recompute the step as

$$y_{n-1} + \frac{1}{2}\Big(hf(x_n, y_n) + hf(x_{n-1}, y_{n-1})\Big). \tag{217a}$$

If we were intending to actually use this more accurate approximation, then the second computation of the function f in each step would approximately double the work that needs to be done to complete each step. However, all we intend to do is to estimate the error and, for this reason, the cost is unchanged, because we need the value of $hf(x_n, y_n)$ to proceed to the next step in any case.

Thus, we see that a convenient, and essentially cost-free, method for estimating local truncation errors is as the difference of the result found by the Euler method itself, and the result found from (217a). This leads to the error estimate

$$\frac{1}{2}\Big(hf(x_n, y_n) - hf(x_{n-1}, y_{n-1})\Big).$$

We already know this estimate can be used, quite satisfactorily, to control stepsize, because of its evident success in Subsection 202.

218   Rounding error

The mathematical analysis of the behaviour of a numerical method, such as the Euler method, is usually idealized to exact arithmetic. However, in practical computation, the nature of computer arithmetic can play a significant, and possibly overwhelming, part. Thus the discussion of error growth, given in Subsection 212, is deficient in this respect. Let α_n denote the total error in the result, computed at step n, and β_n the corresponding error in the derivative, computed at this step. Thus,

$$\alpha_n = y(x_n) - y_n,$$
$$\beta_n = f(x_n, y(x_n)) - f(x_n, y_n).$$

The sequences of exact and approximate values are interrelated by

$$y_n = y_{n-1} + hf(x_{n-1}, y_{n-1}) - r_n,$$
$$y(x_n) = y(x_{n-1}) + hf(x_{n-1}, y(x_{n-1})) + l_n,$$

where r_n is the rounding error, otherwise known as the round-off error, committed in this step, and l_n is the truncation error that we have already discussed.
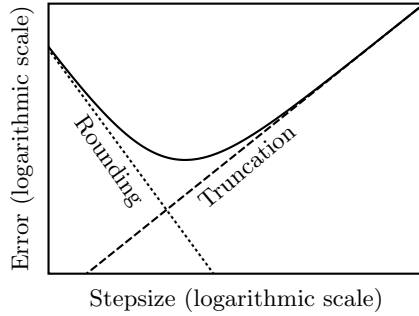
Figure 218(i)   Schema showing effects of rounding error

These lead to the difference equation

$$\alpha_n = \alpha_{n-1} + h\beta_{n-1} + l_n + r_n.$$

Even though we know something about l_n, in particular that it behaves asymptotically like a constant times h^2, very little is known about r_n.

A somewhat pessimistic model of rounding error would bound its magnitude in terms of the magnitude of y_n. It would also assume that its sign (or direction, in the high-dimensional case) is always such as to reinforce errors already accumulated. Bounding the magnitude of the rounding error, in terms of the magnitude of y_n, is quite reasonable, because the greatest contribution to the total rounding error will usually arise from the final addition of $hf(x_{n-1}, y_{n-1})$ to y_{n-1}. Of these two terms, y_{n-1} is usually far the greater in magnitude. Thus, the rounding error will have a magnitude approximately equal to $\|y_{n-1}\|\epsilon \approx \|y_n\|\epsilon$, where ϵ is the machine round-off constant defined as the smallest positive number which satisfies the inequality $1 + \epsilon > 1$, in computer arithmetic.

The other aspect of this model, that rounding errors always conspire to produce the worst possible outcome, is, of course, too severe an assumption. An alternative is to treat the rounding errors arising in different steps as being independently and randomly distributed.

The pessimistic assumption adds an additional term to the accumulated error of Ch^{-1}, for C a constant, because the local error will be more or less the same in each step and the number of steps is inversely proportional to h. The randomness assumption will lead to the rounding error contribution being replaced by a term of the form $Ch^{-1/2}$. A detailed analysis of the probabilistic model of rounding error in initial value problem calculations is presented in Henrici (1962).

Under either the deterministic or the probabilistic model, it is clear that the conclusion of the convergence of computed solutions to the exact solution, as the stepsize tends to zero, will have to be reconsidered. If truncation error alone was significant, the error behaviour would be very much as shown by the dashed line in Figure 218(i). On the other hand, if there were no appreciable

Algorithm 218α   Simple version of Euler

```
for i = 1: n
    term = h*f(y);
    y = y + term;
end
```

Algorithm 218β   Sophisticated version of Euler using compensated summation

```
z = zeros(size(y));
for i = 1: n
    term = h*f(y) + z;
    newy = y + term;
    z = term - (newy - y);
    y = newy;
end
```

truncation error, the accumulated rounding error would be very much like the dotted line. The solid line shows the combined effect of these two sources of error. Since a logarithmic scale is used for both stepsize and error, the two individual error components will be approximately straight lines whose slope will depend on the order of the method, one in the case of Euler, and whether the pessimistic or the probabilistic model of rounding error growth is assumed.

Rather than attempting to carry out this analysis, we remark that it is possible, to a large extent, to overcome the worst effects of the accumulation of rounding errors, as steps are computed in sequence. This is done by estimating the value of r_n in any particular step, and then adding this to the value of $hf(x_n, y_n)$, before this is added in the following step. This improved technique, which can be used for many situations involving the summation of a large number of small numbers, is sometimes known as the Gill–Møller algorithm (Gill, 1951; Møller, 1965, 1965a), but is now more often referred to as 'compensated summation'. An analysis, in the context of floating point arithmetic, was carried out by Kahan (1965) and particular applications to initial value problems were considered in Vitasek (1969). A modern survey of compensated summation, with further references, is available in Higham (1993).

We show how this is done by presenting two fragments of MATLAB code, of which the first, referred to as Algorithm 218α, computes the solution naively, and the second, Algorithm 218β, makes the improvement that we have referred to. In each case, the problem is assumed to be written in autonomous form; this is convenient because, if it were not the case, the updating of the x variable would need to be done in a similar way to the y variable. It is assumed that the statement f(y) yields the value of the derivative vector for given y.
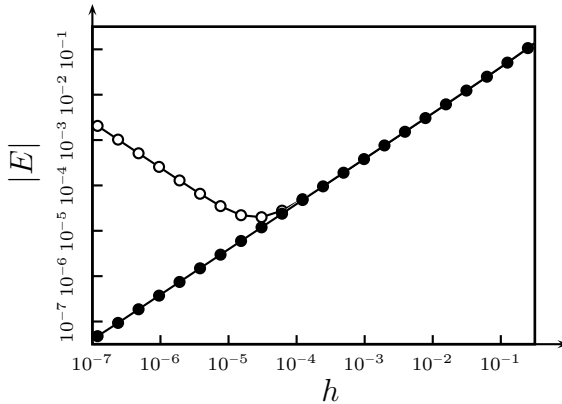
Figure 218(ii)   Errors for naive (○) and sophisticated (●) forms of the Euler method

Although each of these algorithms is coded to work in a vector setting, it will be adequate, for illustrative purposes, to confine ourselves to numerical experiments with a scalar problem. Specifically, we use the problem given by (201a), using a sequence of stepsizes, $h = 2^{-2}$, $h = 2^{-3}$, ..., $h = 2^{-24}$. Each of the two algorithms was used, and the errors were plotted on the same graph, which is presented in Figure 218(ii). To avoid the necessity of using abnormally small stepsizes, before rounding error becomes significant, the calculations were performed in an arithmetic system in which it was possible to force an accuracy of only nine significant decimal digits. It is seen that the naive form of the method produces results that are increasingly infected by rounding for stepsizes less than 2^{-15}. For the Gill–Møller (compensated summation) algorithm, on the other hand, there is no sign of accumulated rounding error at all. It can also be seen that the naive version of the method gives results much as was anticipated in Figure 218(i).

To give additional insight into how compensated summation works, a further calculation on the initial value problem (201a) was performed, using modified arithmetic in which the computations were consistently rounded to three significant decimal digits. Using the notation in Algorithm 218β, these results are shown in Table 218(I) for the first ten steps, using stepsize 0.01. The crucial step in the calculation, the evaluation of z, can be expected to be performed with little or no error. The reason for this is that each of the two subtractions, newy - y and term - (newy - y), has operands which are close to being equal, and these subtractions are usually performed without rounding error. Exceptions may occur when two operands are almost equal, but where the exponent parts of the floating point representations differ by one; but this situation will be relatively rare. If we also concede that the errors generated in the addition of two small quantities, in the statement term = h*f(y) + z, are not of great significance, then we see that, although y might

Table 218(I)   Ten steps of sophisticated Euler to three significant decimals

x	y	h*f(y)	term	newy	z
0.00	1.00	0.0100	0.0100	1.01	0.0000
0.01	1.01	0.0102	0.0102	1.02	0.0002
0.02	1.02	0.0104	0.0106	1.03	0.0006
0.03	1.03	0.0106	0.0112	1.04	0.0012
0.04	1.04	0.0108	0.0120	1.05	0.0020
0.05	1.05	0.0110	0.0130	1.06	0.0030
0.06	1.06	0.0112	0.0142	1.07	0.0042
0.07	1.07	0.0114	0.0156	1.09	−0.0044
0.08	1.09	0.0116	0.0072	1.10	−0.0028
0.09	1.10	0.0118	0.0090	1.11	−0.0010
0.10	1.11				

not be accurate as an approximation to y at the end of a step, the value of y
+ z, if it could be evaluated accurately, would be a very good approximation,
because the statement term - (newy - y) effectively increases the old value
of y + z by h*f(y), to form the new value of y + z.

As further evidence in support of the use of compensated summation, we
present the results of an extended calculation, with the same three decimal
arithmetic system used to produce Table 218(I). In this calculation, 100 steps
were taken, so that the numerical approximations are now extended to the
interval $[0, 1]$. Shown in Figure 218(iii) are the computed values of y, found
using each of Algorithms 218α and 218β. In each case a rounding-free version
of the same results was subtracted to isolate the error due to rounding alone.
The sum of y and z, for the sophisticated algorithm, is also given. Because
the values of these quantities vary widely, a scale is used for which a value
ϵ corresponds to a rounding error of $\epsilon \exp(10^4|\epsilon|)$. It is clear that, in this
example, the sophisticated version of Euler performs overwhelmingly better
than the crude version.

Exercises 21

21.1 For the differential equation $y' = y$, $y(0) = 1$, find the function $\widetilde{y}$,
given by (212a), where $n = 4$ and $[x_0, x_1, x_2, x_3, x_4] = [0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1]$ and
$\widetilde{y}(0) = y(0)$.

21.2 For the same problem as in Exercise 21.1, but with n an arbitrary
positive integer and $x_k = k/n$, for $k = 0, 1, 2, \ldots, n$, find the value
of $\widetilde{y}(1) - y(1)$ and show that this converges to 0 as $n \to \infty$.
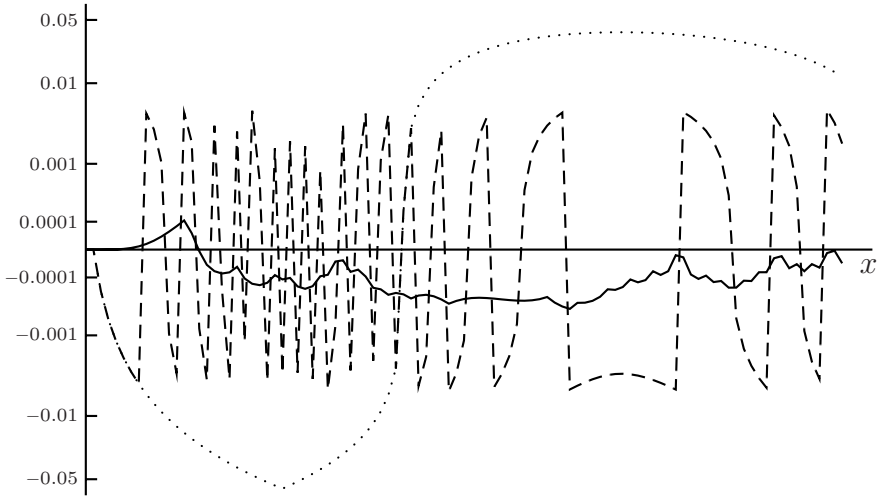
Figure 218(iii)   Accumulation of rounding errors in low accuracy calculations with sophisticated Euler, showing y (dashed line) and y+z (solid line); also, for comparison, crude Euler (dotted line)

21.3 Prove (211a), using integration by parts.

21.4 Assuming that $L = 0$, prove (212f), using induction on k.

21.5 Repeat the calculation in Subsection 218, but making the correction in the Gill–Møller algorithm only every second step.

22   Generalizations of the Euler Method

220   Introduction

As we have seen, in our discussion of the Euler method in Sections 20 and 21, this simplest of all numerical methods enjoys many desirable properties but, at the same time, suffers from some limitations. In the present section, we consider generalizations, which will yield improved numerical behaviour but will retain, as much as possible, its characteristic property of simplicity.

An important aim will be to obtain methods for which the asymptotic errors behave like high powers of the stepsize h. For such methods, the gain in accuracy, resulting from a given reduction in stepsize, would be greater than for the Euler method, because for this method, the error behaves only like the first power of h. We also examine the stability characteristics of these various more general methods. As we saw in Subsection 216, the Euler method does

Table 221(I)   Errors in the numerical solution of the orbital problem (201d)
with zero eccentricity through a half period using (221a) and (221b)

n	y_1 error	Ratio	y_2 error	Ratio
32	0.01479021		−0.04016858	
		3.9676		3.9688
64	0.00372781		−0.01012098	
		4.0417		4.0001
128	0.00092233		−0.00253020	
		4.0361		4.0041
256	0.00022852		−0.00063190	
		4.0219		4.0031
512	0.00005682		−0.00015785	
		4.0119		4.0018
1024	0.00001416		−0.00003945	

n	y_3 error	Ratio	y_4 error	Ratio
32	0.04038636		−0.01548159	
		3.9497		4.1552
64	0.01022525		−0.00372585	
		4.0132		4.0659
128	0.00254793		−0.00091636	
		4.0163		4.0294
256	0.00063440		−0.00022742	
		4.0105		4.0138
512	0.00015818		−0.00005666	
		4.0059		4.0067
1024	0.00003949		−0.00001414	

not work well for stiff problems, because of stability considerations. We would
like to find methods that have better stability.

The two major aims, greater accuracy and better stability, have to be
balanced against the need to avoid additional computational costs, associated
for example, with starting and stepsize-changing mechanisms. In the next few
subsections, we explore some of the approaches used to achieve these aims.

221   *More computations in a step*

Instead of computing f only once in each time step, as in the Euler method,
we might look for methods which evaluate f (with different arguments, of
course) two or more times. We consider a single example of this idea in which
f is evaluated twice.

Since the Euler method is based on a left-hand quadrature rule, we might
ask how it is possible to base a method on the trapezoidal rule. The difficulty
with this is that the derivative at the beginning of the step is known, but at the
end it is not known. To overcome this difficulty, one of the two f evaluations
can be used to approximate the solution value at the end of the step, using the
same approximation that is used in the Euler method. From this first order
approximation, an approximation to the derivative at the end of the step is

computed. The quota of two f evaluations has now been exhausted, but there is now data available to apply the trapezoidal rule formula.

Putting all these stages of the computation together, we write the algorithm for computing y_n in the form

$$y_n^* = y_{n-1} + hf(x_{n-1}, y_{n-1}), \tag{221a}$$

$$y_n = y_{n-1} + \frac{h}{2}\big(f(x_n, y_n^*) + f(x_{n-1}, y_{n-1})\big). \tag{221b}$$

This is an example of a Runge–Kutta method.

As an example of the use of this method, refer to Table 221(I), where the Kepler problem (201d), with zero eccentricity, is integrated through a half period. The number of steps, n, takes on successive values 2^i, $i = 5, 6, \ldots, 10$, so that h takes on values $\pi 2^{-i}$, $i = 5, 6, \ldots, 10$, respectively. The second order nature of the approximations is suggested by the rate at which errors decrease in each of the four components, as n is repeatedly doubled.

222 *Greater dependence on previous values*

After the first step of a numerical method has been completed, approximations are available, to be used in the computation of y_n, not only for $y(x_{n-1})$ and $y'(x_{n-1})$ but also for $y(x_{n-2})$ and $y'(x_{n-2})$. After further steps, even more previous information is available. Instead of computing y_n in a complicated manner from just the value of y_{n-1}, we could consider making more use of the values computed in past steps, as they become available.

In the generalization of the Euler method, introduced in Subsection 221, we were, in effect, using an approximation to the derivative not at x_{n-1}, but at $x_{n-\frac{1}{2}} = x_{n-1} + \frac{1}{2}h$. One way of doing a similar adjustment, but using past information, is to note that existing data indicates that the value of $y'(x)$ is changing by about $f(x_{n-1}, y_{n-1}) - f(x_{n-2}, y_{n-2})$ per step. It therefore seems reasonable to assume that, as x advances from x_{n-1} to $x_{n-\frac{1}{2}}$, the approximation to the derivative at x_{n-1}, given as $f(x_{n-1}, y_{n-1})$, should be increased by $\frac{1}{2}\big(f(x_{n-1}, y_{n-1}) - f(x_{n-2}, y_{n-2})\big)$ to obtain a usable approximation to $y'(x_{n-\frac{1}{2}})$. This means that we could approximate the derivative at $x_{n-\frac{1}{2}}$, the mid-point of the interval, by $\frac{3}{2}f(x_{n-1}, y_{n-1}) - \frac{1}{2}f(x_{n-2}, y_{n-2})$, to yield the numerical method

$$y_n = y_{n-1} + h\big(\tfrac{3}{2}f(x_{n-1}, y_{n-1}) - \tfrac{1}{2}f(x_{n-2}, y_{n-2})\big). \tag{222a}$$

This method is an example of a 'linear multistep method'.

Before we can carry out numerical tests with this method, we first need some procedure for carrying out the first step of the computation. Once y_1 is calculated, the information that is needed for the computation of y_2, and subsequently the solution at later steps, will be available as needed. In the

Table 222(I)   Errors in the numerical solution of the orbital problem (201d)
with zero eccentricity through a half period using (222a)

n	y_1 error	Ratio	y_2 error	Ratio
32	0.00295976		0.00537347	
		7.8987		2.3976
64	0.00037472		0.00224114	
		8.0168		3.3219
128	0.00004674		0.00067465	
		8.0217		3.6879
256	0.00000583		0.00018294	
		8.0136		3.8503
512	0.00000073		0.00004751	
		8.0074		3.9267
1024	0.00000009		0.00001210	

n	y_3 error	Ratio	y_4 error	Ratio
32	−0.00471581		−0.00154957	
		2.1899		7.9797
64	−0.00215339		−0.00019419	
		3.2451		8.1221
128	−0.00066358		−0.00002391	
		3.6551		8.1017
256	−0.00018155		−0.00000295	
		3.8351		8.0620
512	−0.00004734		−0.00000037	
		3.9194		8.0339
1024	−0.00001208		−0.00000005	

experiments we report here, the first step is taken using the Runge–Kutta
method introduced in the previous subsection.

The errors are shown in Table 222(I) and we see that, for this problem at
least, the results are just as good as for the Runge–Kutta method (221a) and
(221b), even though only one derivative is computed in each step. In fact, for
components 1 and 4, better than second order convergence is observed.

223   Use of higher derivatives

For many practical problems, it is possible to derive formulae for the second
and higher derivatives of y, making use of the formula for y' given by
a differential equation. This opens up many computational options, which
can be used to enhance the performance of multistage (Runge–Kutta) and
multivalue (multistep) methods. If these higher derivatives are available, then
the most popular option is to use them to evaluate a number of terms in
Taylor's theorem. Even though we consider this idea further in Section 25, we
present a simple illustrative example here.

Consider the initial value problem

$$y' = yx + y^2, \qquad y(0) = \frac{1}{2}, \tag{223a}$$
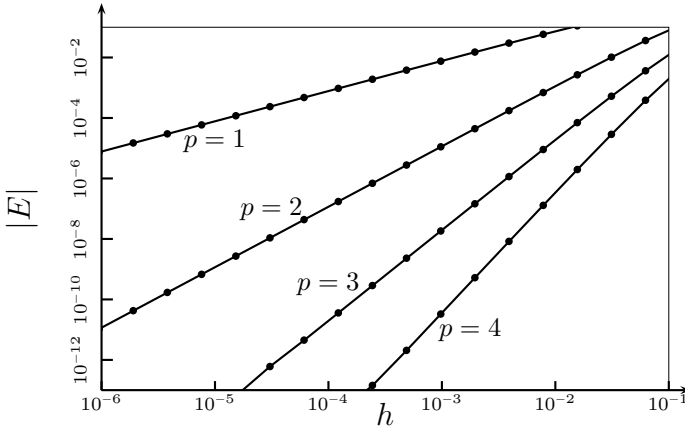
Figure 223(i) Errors in problem (223a) using Taylor series with orders $p = 1, 2, 3, 4$

with solution

$$y(x) = \frac{\exp(\frac{1}{2}x^2)}{2 - \int_0^x \exp(\frac{1}{2}x^2)dx}.$$

By differentiating (223a) once, twice and a third time, it is found that

$$y'' = (x + 2y)y' + y, \tag{223b}$$
$$y''' = (x + 2y)y'' + (2 + 2y')y', \tag{223c}$$
$$y^{(4)} = (x + 2y)y''' + (3 + 6y')y''. \tag{223d}$$

We illustrate the Taylor series method by solving (223a) with output point $\overline{x} = 1$. Using n steps and stepsize $h = 1/n$, for $n = 8, 16, 32, \ldots, 2^{20}$, the method was used with orders $p = 1, 2, 3$ and 4. For example, if $p = 4$, then

$$y_n = y_{n-1} + hy' + \frac{h^2}{2}y'' + \frac{h^3}{6}y''' + \frac{h^2}{24}y^{(4)},$$

where y', y'', y''' and $y^{(4)}$ are given by (223a), (223b), (223c) and (223d) with x_{n-1} and y_{n-1} substituted for x and y, respectively.

The results for these experiments are shown in Figure 223(i). In each case the error is plotted, where we note that the exact result is

$$\exp(\tfrac{1}{2})/\big(2 - \int_0^1 \exp(\tfrac{1}{2}x^2)dx\big),$$

with numerical value 2.04799324543883.
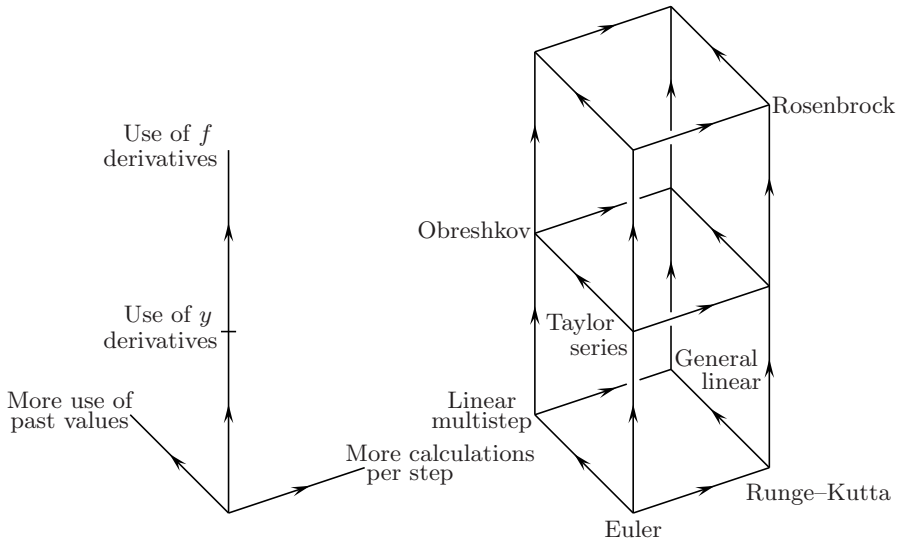
Figure 224(i)   Classification of general method types

224   *Multistep–multistage–multiderivative methods*

While multistep methods, multistage methods and multiderivative methods
all exist in their own right, many attempts have been made to combine their
attributes so as to obtain new methods of greater power. By introducing higher
y derivatives into multistep methods, a new class of methods is found. These
are known as Obreshkov methods, after their discoverer Obreshkov (1940).

The best-known combination of the use of higher derivatives with Runge–
Kutta methods is in Rosenbrock methods (Rosenbrock, 1963). This is actually
a greater generalization, in the sense that derivatives of f are used. These
must be regarded as more general, because y'' can be found in the case of
an autonomous problem as $y''(x) = f'(y(x))(f(y(x)))$. On the other hand, it
is not possible to compute $f'(y(x))$ from values of the various y derivatives.
Rosenbrock methods have a role in the solution of stiff problems.

Other potentially useful combinations certainly exist but, in this book, we
mainly confine ourselves to combinations of multistage and multiderivative
methods. These we refer to as 'general linear methods'. The various methods
that come under the classifications we have discussed here can be seen in
a diagrammatic representation in Figure 224(i). The Euler method can be
thought of as the infimum of all the method classes, and is shown at the lowest
point of this diagram. On the other hand, the class of general linear methods
is the supremum of all multistage and multivalue methods. The supremum of
all methods, including also those with a multiderivative nature, is represented
by the highest point in Figure 224(i).

225 *Implicit methods*

We have already seen, in Subsection 204, that the implicit Euler method has a role in the solution of stiff problems. Implicitness also exists in the case of linear multistep and Runge–Kutta methods. For example, the second order backward difference formula (also known as BDF2),

$$y_n = \frac{2}{3}hf(x_n, y_n) + \frac{4}{3}y_{n-1} - \frac{1}{3}y_{n-2}, \tag{225a}$$

is also used for stiff problems. There are also implicit Runge–Kutta methods, suitable for the solution of stiff problems.

Another example of an implicit method is the 'implicit trapezoidal rule', given by

$$y_n = y_{n-1} + \frac{h}{2}\big(f(x_n, y_n) + f(x_{n-1}, y_{n-1})\big). \tag{225b}$$

Like the Euler method itself, and its implicit variant, (225b) is, at the same time, a linear multistep method and a Runge–Kutta method. As a linear multistep method, it can be regarded as a member of the Adams–Moulton family of methods. As a Runge–Kutta method, it can be regarded as a member of the Lobatto IIIA family.

Implicit methods carry with them the need to solve the nonlinear equation on which the solution, at a new step value, depends. For non-stiff problems, this can be conveniently carried out by fixed-point iteration. For example, the solution of the implicit equation (225b) is usually found by evaluating a starting approximation $\eta^{[0]}$, given as y_n in (222a). A sequence of approximations $\eta^{[k]}$, $k = 1, 2, \ldots$, is then formed by inserting $\eta^{[k]}$ in place of y_n on the left-hand side of (225b), and $\eta^{[k-1]}$ in place of y_n on the right-hand side. That is,

$$\eta^{[k]} = y_{n-1} + \frac{h}{2}\left(f\big(x_n, \eta^{[k-1]}\big) + f(x_{n-1}, y_{n-1})\right), \quad k = 1, 2, \ldots. \tag{225c}$$

The value of y_n actually used for the solution is the numerically computed limit to this sequence.

For stiff problems, unless h is chosen abnormally small, this sequence will not converge, and more elaborate schemes are needed to evaluate the solution to the implicit equations. These schemes are generally variants of the Newton–Raphson method, and will be discussed further in reference to the particular methods as they arise.

226 *Local error estimates*

It is usually regarded as necessary to have, as an accompaniment to any numerical method, a means of assessing its accuracy, in completing each step it takes. The main reason for this is that the work devoted to each step,

and the accuracy that is achieved in the step, should be balanced for overall efficiency. If the cost of each step is approximately constant, this means that the error committed in the steps should be approximately equal.

A second reason for assessing the accuracy of a method, along with the computation of the solution itself, is that it may be more efficient to change to a higher, or lower, member of the family of methods being used. The only way that this can really be decided is for the accuracy of a current method to be assessed and, at the same time, for some sort of assessment to be made of the alternative method under consideration. We discuss here only the local error of the current method.

It is not known how much a computed answer differs from what would correspond to the exact answer, defined locally. What is often available, instead, is a second approximation to the solution at the end of each step. The difference of these two approximations can sometimes be used to give quantitative information on the error in one of the two solution approximations.

We illustrate this idea in a single case. Suppose the method given by (222a) is used to give a starting value for the iterative solution of (225b). It is possible to estimate local errors by using the difference of these two approximations. We discuss this in more detail in the context of predictor–corrector Adams methods.

Exercises 22

22.1 Assuming the function f satisfies a Lipschitz condition and that y, y', y'' and y''' are continuous, explain why the method given by (221a) and (221b) has order 2.

22.2 Explain why the method given by (222a) has order 2.

22.3 Find a method similar to (221a) and (221b), except that it is based on the mid-point rule, rather than the trapezoidal rule.

22.4 For a 'quadrature problem', $f(x, y) = \phi(x)$, compare the likely accuracies of the methods given in Subsections 221 and 222.

22.5 Verify your conclusion in Exercise 22.4 using the problem $y'(x) = \cos(x)$ on the interval $[0, \pi/2]$.

22.6 Show that the backward difference method (225a) has order 2.

22.7 Calculate the solution to (203a) using the backward difference method (225a). Use n steps with constant stepsize $h = \pi/n$ for $n = 2^0, 2^1, 2^2, \ldots, 2^{10}$. Verify that second order behaviour is observed.

23 Runge–Kutta Methods

230 Historical introduction

The idea of generalizing the Euler method, by allowing for a number of evaluations of the derivative to take place in a step, is generally attributed to Runge (1895). Further contributions were made by Heun (1900) and Kutta (1901). The latter completely characterized the set of Runge–Kutta methods of order 4, and proposed the first methods of order 5. Special methods for second order differential equations were proposed by Nyström (1925), who also contributed to the development of methods for first order equations. It was not until the work of Huťa (1956, 1957) that sixth order methods were introduced.

Since the advent of digital computers, fresh interest has been focused on Runge–Kutta methods, and a large number of research workers have contributed to recent extensions to the theory, and to the development of particular methods. Although early studies were devoted entirely to explicit Runge–Kutta methods, interest has now moved to include implicit methods, which have become recognized as appropriate for the solution of stiff differential equations.

A number of different approaches have been used in the analysis of Runge–Kutta methods, but the one used in this section, and in the more detailed analysis of Chapter 3, is that developed by the present author (Butcher, 1963), following on from the work of Gill (1951) and Merson (1957).

231 Second order methods

In Subsection 221, a method was introduced based on the trapezoidal rule quadrature formula. It turns out that for any non-zero choice of a parameter θ, it is possible to construct a method with two stages and this same order. All that is required is a first partial step to form an approximation a distance θh into the step. Using the derivative at this point, together with the derivative at the beginning of the step, the solution at the end of the step is then found using the second order quadrature formula

$$\int_0^1 \phi(x)dx \approx \left(1 - \frac{1}{2\theta}\right)\phi(0) + \frac{1}{2\theta}\phi(\theta).$$

Thus, to advance the solution from x_{n-1} to $x_n = x_{n-1} + h$, the result is found from

$$Y = y_{n-1} + \theta h f(x_{n-1}, y_{n-1}), \tag{231a}$$

$$y_n = y_{n-1} + \left(1 - \frac{1}{2\theta}\right)h f(x_{n-1}, y_{n-1}) + \frac{1}{2\theta}h f(x_{n-1} + \theta h, Y). \tag{231b}$$

Note that the intermediate stage value Y is an approximation to the solution at the 'off-step' point $x_{n-1} + \theta h$, and is equal to y_n^*, in the special case we have already considered, given by (221a) and (221b), in which $\theta = 1$. The other most commonly used value is $\theta = \frac{1}{2}$, as in the 'mid-point rule'.

232   The coefficient tableau

It is convenient to represent a Runge–Kutta method by a partitioned tableau, of the form

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

in which the vector c indicates the positions, within the step, of the stage values, the matrix A indicates the dependence of the stages on the derivatives found at other stages, and b^{T} is a vector of quadrature weights, showing how the final result depends on the derivatives, computed at the various stages.

In the case of explicit methods, such as those we have considered so far in this section, the upper triangular components of A are left blank, because they have zero value.

The first two of the following examples of Runge–Kutta tableaux are, respectively, for the Euler method and the general second order method, parameterized by an arbitrary non-zero θ. The special cases, which are also given, are for the trapezoidal rule method, designated here as RK21 and the mid-point rule method, RK22, correspond to $\theta = 1$ and $\theta = \frac{1}{2}$, respectively:

$$
\begin{array}{c|c}
0 & \\
\hline
 & 1
\end{array}
$$

$$
\begin{array}{c|cc}
0 & & \\
\theta & \theta & \\
\hline
 & 1 - \frac{1}{2\theta} & \frac{1}{2\theta}
\end{array}
$$

$$
\text{RK21}: \quad
\begin{array}{c|cc}
0 & & \\
1 & 1 & \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}
\tag{232a}
$$

$$
\text{RK22}: \quad
\begin{array}{c|cc}
0 & & \\
\frac{1}{2} & \frac{1}{2} & \\
\hline
 & 0 & 1
\end{array}
\tag{232b}
$$

233   Third order methods

It is possible to construct methods with three stages, which have order 3 numerical behaviour. These have the form

$$
\begin{array}{c|ccc}
0 & & & \\
c_2 & a_{21} & & \\
c_3 & a_{31} & a_{32} & \\
\hline
& b_1 & b_2 & b_3
\end{array} \quad ,
$$

where $a_{21} = c_2$ and $a_{31} + a_{32} = c_3$. The conditions for order 3, taken from results that will be summarized in Subsection 234, are

$$b_1 + b_2 + b_3 = 1, \tag{233a}$$

$$b_2 c_2 + b_3 c_3 = \frac{1}{2}, \tag{233b}$$

$$b_2 c_2^2 + b_3 c_3^2 = \frac{1}{3}, \tag{233c}$$

$$b_3 a_{32} c_2 = \frac{1}{6}. \tag{233d}$$

The following tableaux

$$
\text{RK31}: \quad
\begin{array}{c|ccc}
0 & & & \\
\frac{2}{3} & \frac{2}{3} & & \\
\frac{2}{3} & \frac{1}{3} & \frac{1}{3} & \\
\hline
& \frac{1}{4} & 0 & \frac{3}{4}
\end{array}
\tag{233e}
$$

and

$$
\text{RK32}: \quad
\begin{array}{c|ccc}
0 & & & \\
\frac{1}{2} & \frac{1}{2} & & \\
1 & -1 & 2 & \\
\hline
& \frac{1}{6} & \frac{2}{3} & \frac{1}{6}
\end{array}
\tag{233f}
$$

give two possible solutions to (233a)–(233d).

234   Introduction to order conditions

As the order being sought increases, the algebraic conditions on the coefficients of the method become increasingly complicated. The pattern behind these conditions is known and, in this brief introduction to the order conditions, we state the results without any justification and show, by examples, how they are used.
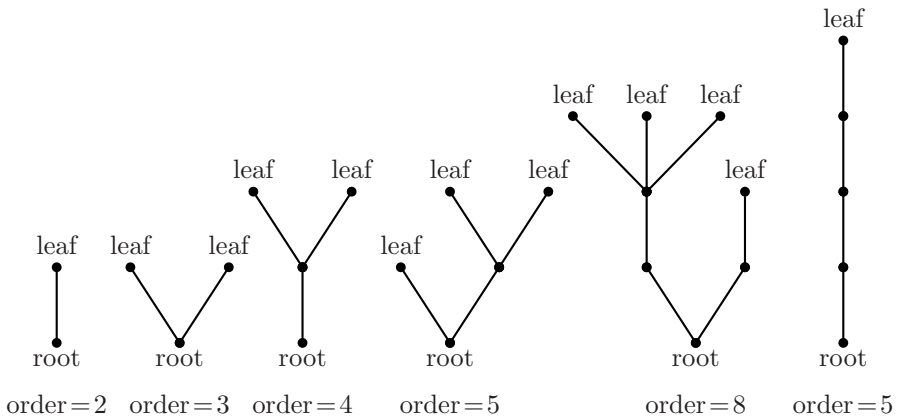
Figure 234(i)   Some illustrative rooted trees

Let T denote the set of all 'rooted trees'. These are simple combinatorial graphs, which have the properties of being connected, having no cycles, and having a specific vertex designated as the root. The 'order' of a tree is the number of vertices in this tree. If the order is greater than 1, then the 'leaves' of a tree are the vertices from which there are no outward-growing arcs; in other words, a leaf is a vertex, other than the root, which has exactly one other vertex joined to it.

An assortment of trees of various orders, with leaves and the root indicated in each case, is shown in Figure 234(i). In pictorial representations of particular rooted trees, as in this figure, we use the convention of placing the root at the lowest point in the picture.

For each tree t, a corresponding polynomial in the coefficients of the method can be written down. Denote this by $\Phi(t)$. Also associated with each tree t is an integer $\gamma(t)$. We now explain how $\Phi(t)$ and $\gamma(t)$ are constructed.

In the case of $\Phi(t)$, associate with each vertex of the tree, except the leaves, a label $i, j, \ldots$, and assume that i is the label attached to the root. Write down a sequence of factors of which the first is b_i. For each arc of the tree, other than an arc that terminates in a leaf, write down a factor, say a_{jk}, where j and k are the beginning and end of the arc (assuming that all directions are in the sense of movement away from the root). Finally, for each arc terminating at a leaf, write down a factor, say c_j, where j is the label attached to the beginning of this arc. Having written down this sequence of factors, sum their product for all possible choices of each of the labels, in the set $\{1, 2, \ldots, s\}$.

To find the value of $\gamma(t)$, associate a factor with each vertex of the tree. For

Table 234(I) The rooted trees up to order 4

Tree	•	!	V	!
Order	1	2	3	3
Φ	$\sum_i b_i$	$\sum_i b_i c_i$	$\sum_i b_i c_i^2$	$\sum_{ij} b_i a_{ij} c_j$
γ	1	2	3	6

Tree	ᵥᵥ	ᵥ	Y	!
Order	4	4	4	4
Φ	$\sum_i b_i c_i^3$	$\sum_{ij} b_i c_i a_{ij} c_j$	$\sum_{ij} b_i a_{ij} c_j^2$	$\sum_{ijk} b_i a_{ij} a_{jk} c_k$
γ	4	8	12	24

the leaves this factor is 1, and for all other vertices it is equal to the sum of the factors attached to all outward-growing neighbours, plus 1. The product of the factors, for all vertices of the tree, is the value of $\gamma(t)$.

The values of these quantities are shown in Table 234(I), for each of the eight trees with orders up to 4. A further illustrative example is given by the tree

$$t = \text{ᵥᵥᵛ}$$
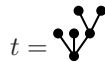
for which $\Phi(t) = \sum_{ij} b_i c_i^2 a_{ij} c_j^2$ and $\gamma(t) = 18$. Details of the calculation of these quantities are presented in Figure 234(ii). On the left-hand diagram labels i and j are attached to the non-terminal vertices, as used in the formula for $\Phi(t)$, using the factors shown in the middle diagram. On the right-hand diagram, the factors are shown whose product gives the value of $\gamma(t)$.

$$\Phi(t) = \sum_{ij} b_i c_i^2 a_{ij} c_j^2 \qquad \gamma(t) = 1 \cdot 1 \cdot 3 \cdot 1 \cdot 1 \cdot 6 = 18$$
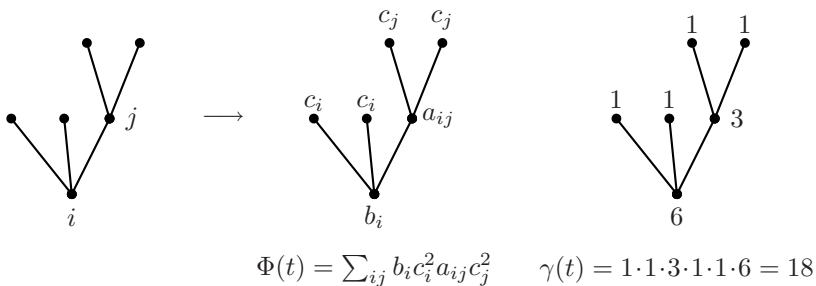
Figure 234(ii) Calculation details for $\Phi(t)$ and $\gamma(t)$, where $t = \text{ᵥᵥᵛ}$

235   Fourth order methods

Write the order conditions presented in the previous subsection, in the special case $s = 4$, assuming, because the method will be explicit, that $a_{ij} = 0$ unless $i > j$. This yields the conditions

$$b_1 + b_2 + b_3 + b_4 = 1, \tag{235a}$$

$$b_2 c_2 + b_3 c_3 + b_4 c_4 = \frac{1}{2}, \tag{235b}$$

$$b_2 c_2^2 + b_3 c_3^2 + b_4 c_4^2 = \frac{1}{3}, \tag{235c}$$

$$b_3 a_{32} c_2 + b_4 a_{42} c_2 + b_4 a_{43} c_3 = \frac{1}{6}, \tag{235d}$$

$$b_2 c_2^3 + b_3 c_3^3 + b_4 c_4^3 = \frac{1}{4}, \tag{235e}$$

$$b_3 c_3 a_{32} c_2 + b_4 c_4 a_{42} c_2 + b_4 c_4 a_{43} c_3 = \frac{1}{8}, \tag{235f}$$

$$b_3 a_{32} c_2^2 + b_4 a_{42} c_2^2 + b_4 a_{43} c_3^2 = \frac{1}{12}, \tag{235g}$$

$$b_4 a_{43} a_{32} c_2 = \frac{1}{24}. \tag{235h}$$

That $c_4 = 1$ can be shown, by solving for b_2, b_3 and b_4, from equations (235b), (235c) and (235e); by then solving for a_{32}, a_{42} and a_{43} from (235d), (235f) and (235g); and then by substituting into (235h). Many solutions and families of solutions are known to these conditions; the following are two examples:

$$
\text{RK41}: \quad
\begin{array}{c|cccc}
0 \\
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & 0 & \frac{1}{2} \\
1 & 0 & 0 & 1 \\
\hline
& \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6}
\end{array}
\tag{235i}
$$

$$
\text{RK42}: \quad
\begin{array}{c|cccc}
0 \\
\frac{1}{4} & \frac{1}{4} \\
\frac{1}{2} & 0 & \frac{1}{2} \\
1 & 1 & -2 & 2 \\
\hline
& \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6}
\end{array}
\tag{235j}
$$

236 Higher orders

Because the number of rooted trees of various orders increases rapidly for orders greater than 4, the complexity of the order conditions also increases. Above order 4, it is no longer possible to obtain order s with just s stages. For order 5, six stages are required, and for order 6, seven stages are required. Above this order, there are even sharper increases in the required numbers of stages. We give a single example of a fifth order method:

$$
\text{RK5}:
\begin{array}{c|cccccc}
0 \\
\frac{1}{4} & \frac{1}{4} \\
\frac{1}{4} & \frac{1}{8} & \frac{1}{8} \\
\frac{1}{2} & 0 & 0 & \frac{1}{2} \\
\frac{3}{4} & \frac{3}{16} & -\frac{3}{8} & \frac{3}{8} & \frac{9}{16} \\
1 & -\frac{3}{7} & \frac{8}{7} & \frac{6}{7} & -\frac{12}{7} & \frac{8}{7} \\
\hline
 & \frac{7}{90} & 0 & \frac{32}{90} & \frac{12}{90} & \frac{32}{90} & \frac{7}{90}
\end{array}
\tag{236a}
$$

237 Implicit Runge–Kutta methods

Implicit methods have the potential advantage, compared with explicit methods, that there will be fewer stages for the same order. The disadvantage is in the implicit nature of at least some of the stages. This makes it impossible to avoid iterative methods of evaluation. For the purpose of experimental comparison with explicit methods, we present here just three methods:

$$
\begin{array}{c|cc}
\frac{1}{3} & \frac{1}{3} & 0 \\
1 & 1 & 0 \\
\hline
 & \frac{3}{4} & \frac{1}{4}
\end{array}
\tag{237a}
$$

$$
\begin{array}{c|cc}
3-2\sqrt{2} & \frac{5-3\sqrt{2}}{4} & \frac{7-5\sqrt{2}}{4} \\
1 & \frac{1+\sqrt{2}}{4} & \frac{3-\sqrt{2}}{4} \\
\hline
 & \frac{1+\sqrt{2}}{4} & \frac{3-\sqrt{2}}{4}
\end{array}
\tag{237b}
$$

$$
\begin{array}{c|cc}
\frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\
\frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}
\tag{237c}
$$

It can be verified that (237a) has order 3, (237b) has order only 2 and (237c) has order 4. In the implicit case, the cost of using a specific method depends not so much on the number of stages, as on the difficulty in evaluating the

stages. From this point of view, (237a) is the easiest to use because only one of the stages is implicit; (237b) and (237c) each have two interconnected implicit stages but, as we will see in Subsection 363, the order 2 method (237b) can be implemented more cheaply than (237c).

238   Stability characteristics

In Subsection 216, we discussed the stability of the Euler method when solving a linear problem of the form

$$y'(x) = qy(x).$$

If $z = hq$, then in a single step of length h, the exact solution will be multiplied by the factor $\exp(z)$. In the same time interval the approximate solution computed using a Runge–Kutta method will be multiplied by a function of z, specific to the particular Runge–Kutta method. As in Subsection 216, we denote this 'stability function' by $R(z)$. The 'stability region', defined as $\{z \in \mathbb{C} : |R(z)| \leq 1\}$, is the set of points in the complex plane such that the computed solution remains bounded after many steps of computation. There is special interest in values of z in the left half-plane, because in this case the exact solution is bounded and good modelling of the problem would require the computed solution to behave in a similar manner.

For an s-stage Runge–Kutta method, defined by the tableau

$$\begin{array}{c|c} c & A \\ \hline & b^{\mathsf{T}} \end{array} \tag{238a}$$

the vector Y, made up from the s stage values, satisfies

$$Y = \mathbf{1}y_0 + hAqY = \mathbf{1}y_0 + zAY,$$

where y_0 is the incoming approximation. It follows that

$$Y = (I - zA)^{-1}\mathbf{1}y_0.$$

Substitute this into the solution approximation found at the end of the step, and we find

$$y_1 = y_0 + hb^{\mathsf{T}}qY = y_0 + zb^{\mathsf{T}}(I - zA)^{-1}\mathbf{1}y_0 = R(z)y_0,$$

where

$$R(z) = 1 + zb^{\mathsf{T}}(I - zA)^{-1}\mathbf{1}. \tag{238b}$$

If (238a) represents an explicit Runge–Kutta method with order $p = s = 1, 2, 3$ or 4, then we can evaluate $R(z)$ very simply as the exponential series truncated
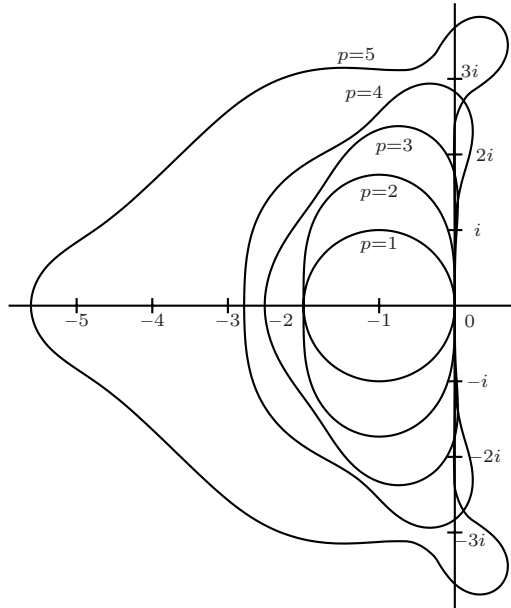
Figure 238(i)   Stability regions for some explicit Runge–Kutta methods

at the z^s term. To see why this should be the case, expand $(I - zA)^{-1}$ by the geometric series and evaluate the terms using the order condition

$$b^\mathsf{T}A^{k-1}\mathbf{1} = b^\mathsf{T}A^{k-2}c = \frac{1}{k!}, \qquad k = 1, 2, \dots, p.$$

Hence, we have for the four cases for which $s = p$ is possible,

$$R(z) = \begin{cases} 1 + z, & p = 1, \\ ]1 + z + \frac{1}{2}z^2, & p = 2, \\ 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3, & p = 3, \\ 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4, & p = 4. \end{cases}$$

The boundaries of the stability regions defined by these functions are shown in Figure 238(i). In each case the stability region is the bounded set enclosed by these curves.

For explicit methods with $s = 6$ and $p = 5$, the stability function takes the form

$$R(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4 + \frac{1}{120}z^5 + Cz^6,$$

where C depends on the particular method. In the case of the method given by the tableau (236a), $C = \frac{1}{1280}$, and the stability region for this is also shown in Figure 238(i). In each case, the value of p is attached to the curve.
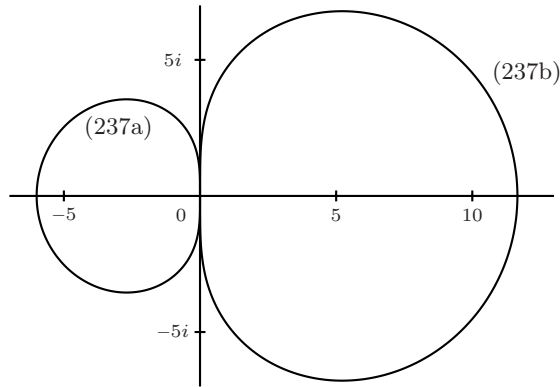
Figure 238(ii)   Stability regions for some implicit Runge–Kutta methods

It is never possible, even by the addition of extra stages, to devise explicit methods with order at least 1, which have unbounded stability regions, because $R(z)$ is always a polynomial equal to $1 + z + O(z^2)$. However, as we saw in the case of the implicit Euler method, there is no such barrier for implicit Runge–Kutta methods.

For the three methods quoted in Subsection 237, the stability functions are found to be

$$
R(z) = \begin{cases}
\dfrac{1 + \dfrac{2z}{3} + \dfrac{z^2}{6}}{1 - \dfrac{z}{3}}, & \text{method (237a)}, \\[3ex]
\dfrac{1 + (\sqrt{2} - 1)z}{\left(1 - \left(1 - \frac{1}{2}\sqrt{2}\right)z\right)^2}, & \text{method (237b)}, \\[3ex]
\dfrac{1 + \dfrac{z}{2} + \dfrac{z^2}{12}}{1 - \dfrac{z}{2} + \dfrac{z^2}{12}}, & \text{method (237c)},
\end{cases}
$$

and the three stability regions are shown in Figure 238(ii). Note that for the fourth order method (237c), the stability region is exactly the closed left half-plane. The method (237a) shares the property of explicit Runge–Kutta methods of having a bounded stability region, whereas (237b) has an unbounded stability region which includes the left half-plane.
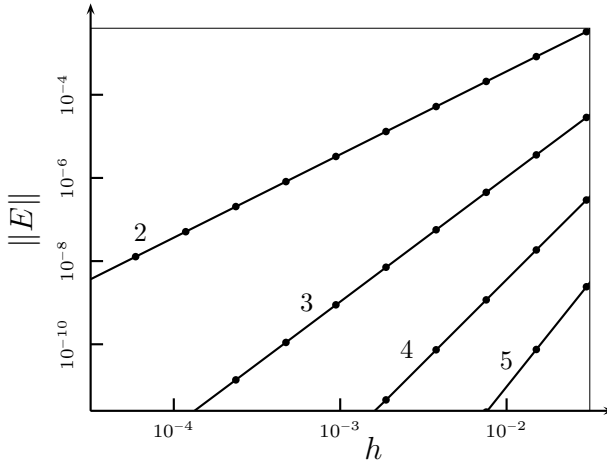
Figure 239(i)   Orbital calculations for various Runge–Kutta methods

239   Numerical examples

High order methods generally perform better than low order methods if
sufficiently small stepsizes are used. We illustrate this by attempting, with the
methods introduced in this section, a solution to the gravitational problem
(201d) with initial values corresponding to an eccentricity $e = \frac{1}{2}$. Although
calculations were performed with each of the seven methods RK21, RK22,
RK31, RK32, RK41, RK42, RK5, only results for the four methods RK22,
RK31, RK42 and RK5 are actually presented in Figure 239(i). It was observed
that for the two methods with each of the orders 2, 3 and 4, there was very
little difference between the accuracy achieved and a representative of each
order – in fact the slightly more accurate method was chosen in each case –
is sufficient to illustrate the phenomenon of h^p dependence. In Figure 239(i),
methods RK22, RK31, RK42 and RK5 are denoted by 2, 3, 4 and 5.

For this problem, high order methods are always more accurate than low
order methods. However, the relative advantage is exaggerated in that no
account is taken of the greater work in completing a step as the order increases.
Assuming that the total computational work is proportional to the number
of stages in the method, it is a simple matter to compensate for this; all that
needs to be done is to multiply the number of steps by the number of stages in
each method. The comparisons with this correction made are shown in Figure
239(ii). The general conclusion, that high order is more efficient than low
order, still follows from these comparisons, but not to such a marked extent.

Numerical tests, not reported here, indicate similar behaviour for implicit
methods. For the initial value problem (201a), with output computed at $x = 1$,
(237a) and (237b) gave slightly worse results than for corresponding explicit
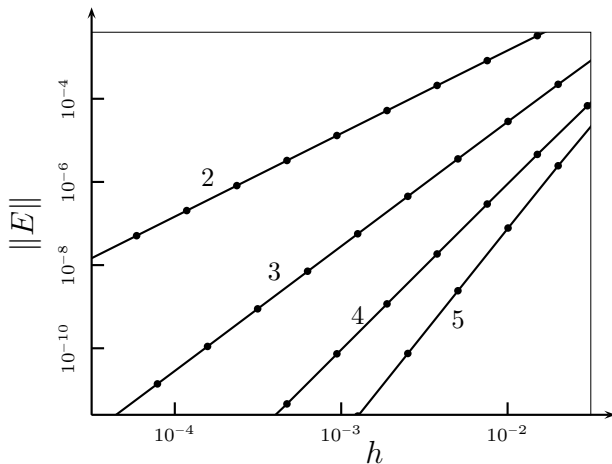
Figure 239(ii)   Runge–Kutta methods with cost corrections

methods. However, for the fourth order method (237c), the results were approximately six decimal places better. This suggests that, even though the cost of evaluating the result in each step of an implicit method is significantly higher, the extra cost is sometimes worthwhile for this method.

Exercises 23

23.1 Repeat the calculation that led to Table 221(I) but using the method given by (231a) and (231b) with $\theta = \frac{1}{2}$.

23.2 Find a solution to the third order conditions (233a), (233b), (233c) and (233d) such that $b_1 = 0$, $c_3 = 1$.

23.3 Continue Table 234(I) to include trees of order 5.

23.4 Write down the formula for $\Phi(t)$ and the value of $\gamma(t)$ for t the order 7 tree

$$t = \skew{}$$

23.5 By noting that $b_4 a_{43} a_{32} c_2 \cdot b_3 (c_4 - c_3)(c_3 - c_2)c_3 = b_4 a_{43}(c_3 - c_2)c_3 \cdot b_3(c_4 - c_3)a_{32}c_2$, prove that $c_4 = 1$ for any solution to the fourth order conditions (235a)–(235h).

23.6 Find the order of the implicit method given by the tableau (237a).

23.7 Solve the orbital problem with eccentricity $e = 0$ using the implicit method (237a).

24 Linear Multistep Methods

240 Historical introduction

The idea of extending the Euler method by allowing the approximate solution at a point to depend on the solution values and the derivative values at several previous step values was originally proposed by Bashforth and Adams (1883). Not only was this special type of method, now known as the Adams–Bashforth method, introduced, but a further idea was suggested. This further idea was developed in detail by Moulton (1926). Other special types of linear multistep methods were proposed by Nyström (1925) and Milne (1926, 1953). The idea of predictor–corrector methods is associated with the name of Milne, especially because of a simple type of error estimate available with such methods. The 'backward difference' methods were introduced by Curtiss and Hirschfelder (1952), and these have a special role in the solution of stiff problems.

The modern theory of linear multistep methods was developed in large measure by Dahlquist (1956), and has become widely known through the exposition by Henrici (1962, 1963).

241 Adams methods

The most important linear multistep methods for non-stiff problems are of Adams type. That is, the solution approximation at x_n is defined either as

$$y_n = y_{n-1} + h(\beta_1 f(x_{n-1}, y_{n-1}) + \beta_2 f(x_{n-2}, y_{n-2})$$
$$+ \cdots + \beta_k f(x_{n-k}, y_{n-k})), \quad \text{(241a)}$$

or as

$$y_n = y_{n-1} + h(\beta_0 f(x_n, y_n) + \beta_1 f(x_{n-1}, y_{n-1}) + \beta_2 f(x_{n-2}, y_{n-2})$$
$$+ \cdots + \beta_k f(x_{n-k}, y_{n-k})), \quad \text{(241b)}$$

where, in each case, the constants (β_0), β_1, β_2, ..., β_k are chosen to give the highest possible order.

The meaning of order, and how it is achieved in particular cases, is straightforward in the case of methods of the form (241a), which are known as 'Adams–Bashforth' methods. Assuming that no errors have yet been introduced when the approximation at x_n is about to be calculated, we can replace the terms on the right-hand side by the quantities they are supposed to approximate, that is, by $y(x_{n-1})$, $y'(x_{n-1})$, $y'(x_{n-2})$, ..., $y'(x_{n-k})$, respectively. The amount by which the approximation, written in this form, differs from $y(x_n)$ is the error generated in this particular step. If this error can be estimated for a smooth problem as $O(h^{p+1})$, then the method is regarded as having order p.

For the methods given by (241b), which are known as 'Adams–Moulton' methods, the term involving $f(x_n, y_n)$ is a complication in this understanding

of order. However, the conclusion turns out to be exactly the same as for Adams–Bashforth methods: if every term in (241b) is replaced by the quantity it is supposed to be approximating and the two sides of this equation differ by an amount that can be estimated as $O(h^{p+1})$, then the method has order p.

To obtain a simple criterion for a given order, we can write all terms in

$$y(x_n) - y(x_{n-1}) - h\big(\beta_0 y'(x_n) + \beta_1 y'(x_{n-1}) + \beta_2 y'(x_{n-2})$$
$$+ \cdots + \beta_k y'(x_{n-k})\big) \quad (241c)$$

as Taylor series about, for example, x_n. This gives an expression of the form

$$C_1 h y'(x_n) + C_2 h^2 y''(x_n) + \cdots + C_p h^p y^{(p)}(x_n) + O(h^{p+1}),$$

and the conditions for order p will be that $C_1 = C_2 = \cdots = C_p = 0$.

It can be shown that an equivalent criterion is that (241c) vanishes whenever y is a polynomial of degree not exceeding p.

We will use these criteria to derive Adams–Bashforth methods with $p = k$ for $k = 2, 3, 4$, and Adams–Moulton methods with $p = k + 1$ for $k = 1, 2, 3$.

For $k = 4$, the Taylor expansion of (241c) takes the form

$$hy'(x_n)(1 - \beta_0 - \beta_1 - \beta_2 - \beta_3 - \beta_4)$$
$$+ h^2 y''(x_n)\big(-\tfrac{1}{2} + \beta_1 + 2\beta_2 + 3\beta_3 + 4\beta_4\big)$$
$$+ h^3 y^{(3)}(x_n)\big(\tfrac{1}{6} - \tfrac{1}{2}(\beta_1 + 4\beta_2 + 9\beta_3 + 16\beta_4)\big)$$
$$+ h^4 y^{(4)}(x_n)\big(-\tfrac{1}{24} + \tfrac{1}{6}(\beta_1 + 8\beta_2 + 27\beta_3 + 64\beta_4)\big) + O(h^5),$$

so that

$$C_1 = 1 - \beta_0 - \beta_1 - \beta_2 - \beta_3 - \beta_4,$$
$$C_2 = -\tfrac{1}{2} + \beta_1 + 2\beta_2 + 3\beta_3 + 4\beta_4,$$
$$C_3 = \tfrac{1}{6} - \tfrac{1}{2}(\beta_1 + 4\beta_2 + 9\beta_3 + 16\beta_4),$$
$$C_4 = -\tfrac{1}{24} + \tfrac{1}{6}(\beta_1 + 8\beta_2 + 27\beta_3 + 64\beta_4).$$

For the Adams–Bashforth methods the value of β_0 is zero; for $k = 2$ we also have $\beta_3 = \beta_4 = 0$ and we must solve the equations $C_1 = C_2 = 0$. This gives $\beta_1 = \tfrac{3}{2}$ and $\beta_2 = -\tfrac{1}{2}$. For $k = 3$ we allow β_3 to be non-zero and we require that $C_1 = C_2 = C_3 = 0$. The solutions of these equations is $\beta_1 = \tfrac{23}{12}$, $\beta_2 = -\tfrac{4}{3}$, $\beta_3 = \tfrac{5}{12}$. For $k = 4$, we solve $C_1 = C_2 = C_3 = C_4 = 0$ to find $\beta_1 = \tfrac{55}{24}$, $\beta_2 = -\tfrac{59}{24}$, $\beta_3 = \tfrac{37}{24}$, $\beta_4 = -\tfrac{3}{8}$.

For the Adams–Moulton methods we allow β_0 to be non-zero. For $k = 1$ ($p = 2$) we have $\beta_2 = \beta_3 = \beta_4 = 0$ and $C_1 = C_2 = 0$; this gives $\beta_0 = \beta_1 = \tfrac{1}{2}$. In a similar manner we find for $k = 2$ ($p = 3$) that $\beta_0 = \tfrac{5}{12}$, $\beta_1 = \tfrac{2}{3}$, $\beta_2 = -\tfrac{1}{12}$; and for $k = 3$ ($p = 4$) that $\beta_0 = \tfrac{3}{8}$, $\beta_1 = \tfrac{19}{24}$, $\beta_2 = -\tfrac{5}{24}$, $\beta_3 = \tfrac{1}{24}$.

242 *General form of linear multistep methods*

Even though Adams methods are amongst the most commonly used classes of linear multistep methods, they are very specialized in that the dependence of y_n on previously computed values ignores y_{n-1}, y_{n-2}, $\ldots$, y_{n-k}. The general form of the method includes additional terms to take these into account. It thus has the form

$$
\begin{aligned}
y_n = {}& \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + \cdots + \alpha_k y_{n-k} \\
& + h\big(\beta_0 f(x_n, y_n) + \beta_1 f(x_{n-1}, y_{n-1}) + \beta_2 f(x_{n-2}, y_{n-2}) \\
& \hspace{4cm} + \cdots + \beta_k f(x_{n-k}, y_{n-k})\big). \quad \text{(242a)}
\end{aligned}
$$

It is customary to characterize this method by polynomials whose coefficients are the numbers α_1, α_2, $\ldots$, α_k, β_0, β_1, β_2, $\ldots$, β_k. The standard terminology is to use polynomials $\rho(z)$ and $\sigma(z)$ defined by

$$
\begin{aligned}
\rho(z) &= z^k - \alpha_1 z^{k-1} - \alpha_2 z^{k-2} - \cdots - \alpha_k, \\
\sigma(z) &= \beta_0 z^k + \beta_1 z^{k-1} + \beta_2 z^{k-2} + \cdots + \beta_k.
\end{aligned}
$$

The style we are adopting in this book makes it more convenient to use a slightly different pair of polynomials,

$$
\begin{aligned}
\alpha(z) &= 1 - \alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_k z^k, \\
\beta(z) &= \beta_0 + \beta_1 z + \beta_2 z^2 + \cdots + \beta_k z^k.
\end{aligned}
$$

Of course, it really makes little difference whether we use (ρ, σ) or $[\alpha, \beta]$ to characterize a method because, once the value of k is known, we can move between them by the relations

$$
\alpha(z) = z^k \rho\Big(\frac{1}{z}\Big), \quad \beta(z) = z^k \sigma\Big(\frac{1}{z}\Big), \quad \rho(z) = z^k \alpha\Big(\frac{1}{z}\Big), \quad \sigma(z) = z^k \beta\Big(\frac{1}{z}\Big).
$$

For all eligible α polynomials, $\alpha(0) = 1$, and for Adams methods, $\alpha(z) = 1 - z$. Using the $[\alpha, \beta]$ representation, we can distinguish Adams–Bashforth from Adams–Moulton by the fact that $\beta(0) = 0$ for the Bashforth variety.

243 *Consistency, stability and convergence*

Suppose we attempt the numerical solution of the simple differential equation $y'(x) = 0$, with exact solution $y(x) = 1$, using the linear multistep method characterized by the pair of polynomials $[\alpha, \beta]$. If the exact answer has already been found for k steps in a row, it seems to be a desirable property of the method that the exact value is also found in one further step. This computed value is equal to $\alpha_1 + \alpha_2 + \cdots + \alpha_k$. For this expression to have the value 1 is equivalent to the assumption that $\alpha(1) = 0$ or, what is equivalent, that

$\rho(1) = 0$. Because of its fundamental importance, this property will be given the name 'preconsistency'.

Another interpretation of preconsistency can be found in terms of the covariance of the numerical method with respect to a translation. By a translation we mean the replacing of an autonomous initial value problem $y'(x) = f(y(x))$, $y(x_0) = y_0$, by a related problem $z'(x) = f(z(x) + v)$, $z(x_0) = y_0 - v$. For the exact solutions to these problems, the value of z will always equal the value of y with the vector v subtracted. In considering a numerical solution to each of these problems, we can do the calculation in terms of y and then carry out the translation afterwards; or we can do the transformation first and carry out the numerical approximation using the z values. By 'covariance' we mean that the two numerical results are exactly the same.

It is easy to verify that the only way this can be guaranteed to happen, if the calculations are carried out using a linear multistep method, is for the method to be preconsistent.

For a preconsistent method it is desirable that the exact solution can also be found for another simple differential initial value problem: the problem given by $y'(x) = 1$, $y(0) = 0$. For every step, the value of $f(y_n)$ is precisely 1. Substitute these into (242a), and it is found that

$$nh = \sum_{i=1}^{k} \alpha_i h(n - i) + h \sum_{i=1}^{k} \beta_i,$$

implying that

$$n \left(1 - \sum_{i=1}^{k} \alpha_i \right) = \sum_{i=1}^{k} \beta_i - \sum_{i=1}^{k} i\alpha_i.$$

The left-hand side vanishes for a preconsistent method, whereas the right-hand side can be written in the form $\beta(1) + \alpha'(1)$. A 'consistent method' is a method that satisfies the condition that $\beta(1) + \alpha'(1) = 0$, in addition to satisfying the preconsistency condition $\alpha(1) = 0$.

No matter how precise numerical approximations to the solution to a differential equation might be, this precision has no ultimate benefit unless the effect on later step values of small errors is bounded. Later steps are effected by the introduction of a perturbation in step m both through their dependence on y_m itself and through their dependence on $hf(x_m, y_m)$. To simplify the discussion we exclude the second cause of error dependence by restricting ourselves to a simple 'quadrature' type of problem in which $y'(x) = f(x)$. This will mean that the difference between the unperturbed and perturbed problem will satisfy the even simpler equation $y'(x) = 0$. Consider the difference equation satisfied by the numerical solution just for the perturbation itself. This difference equation is

$$y_n = \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + \cdots + \alpha_k y_{n-k}. \tag{243a}$$

A linear multistep method is said to be 'stable' if all solutions to the difference equation (243a) are bounded as $n \to \infty$.

From the theory of linear difference equations, we know exactly when this will be the case. It is necessary and sufficient that all zeros of the polynomial ρ lie in the closed unit disc $\{z : |z| \leq 1\}$ and that all repeated zeros lie in the open unit disc $\{z : |z| < 1\}$. Because the zeros of α are the reciprocals of those of ρ we can equally state these conditions as (i) all zeros of α lie outside the open unit disc, and (ii) all repeated zeros of α lie outside the closed unit disc.

'Convergence' refers to the ability of a method to approximate the solution to a differential equation to any required accuracy, if sufficiently many small steps are taken. Of course, any numerical result computed by a linear multistep method will depend not only on the particular coefficients of the method and the differential equation, but also on the procedure used to obtain starting values. In the formal definition of this concept, we will not impose any conditions on how the starting values are approximated except to require that, as $h \to 0$, the errors in the starting values tend to zero. Because the exact solution is continuous, this is equivalent to requiring that the starting values all converge to the initial value specified for the problem.

Divide the interval $[x_0, \overline{x}]$ into n steps each of size $h = (\overline{x} - x_0)/n$, for every positive integer n. Solve a standard initial value problem using starting values $y_0, y_1, \ldots, y_{k-1}$ which depend on h and converge to $y(x_0)$ as $h \to 0$. Let the error in the approximation computed at $\overline{x}$ be denoted by ϵ_n. The method is convergent if necessarily $\epsilon_n \to 0$ as $n \to \infty$.

We discuss this property, and its relationship to other concepts introduced in this subsection, in Chapter 4. In the meantime, we state without proof the important result expressed in the following.

Theorem 243A *A linear multistep method is convergent if and only if it is stable and consistent.*

244   Predictor–corrector Adams methods

Continuing the discussion of Adams–Bashforth and Adams–Moulton methods from Subsection 241, we present in tabular form the coefficients of these methods for orders as high as 8. In the Adams–Bashforth case this means presenting the methods as far as $k = 8$ and in the Moulton case as far as $k = 7$.

Along with the coefficients of the methods, the value is given of the error constants. For example, in the case of the Adams–Bashforth method with order 2 we can write

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

Table 244(I) Coefficients and error constants for Adams–Bashforth methods

k	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	C
1	1								$-\frac{1}{2}$
2	$\frac{3}{2}$	$-\frac{1}{2}$							$\frac{5}{12}$
3	$\frac{23}{12}$	$-\frac{4}{3}$	$\frac{5}{12}$						$-\frac{3}{8}$
4	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{3}{8}$					$\frac{251}{720}$
5	$\frac{1901}{720}$	$-\frac{1387}{360}$	$\frac{109}{30}$	$-\frac{637}{360}$	$\frac{251}{720}$				$-\frac{95}{288}$
6	$\frac{4277}{1440}$	$-\frac{2641}{480}$	$\frac{4991}{720}$	$-\frac{3649}{720}$	$\frac{959}{480}$	$-\frac{95}{288}$			$\frac{19087}{60480}$
7	$\frac{198721}{60480}$	$-\frac{18637}{2520}$	$\frac{235183}{20160}$	$-\frac{10754}{945}$	$\frac{135713}{20160}$	$-\frac{5603}{2520}$	$\frac{19087}{60480}$		$-\frac{5257}{17280}$
8	$\frac{16083}{4480}$	$-\frac{1152169}{120960}$	$\frac{242653}{13440}$	$-\frac{296053}{13440}$	$\frac{2102243}{120960}$	$-\frac{115747}{13440}$	$\frac{32863}{13440}$	$-\frac{5257}{17280}$	$\frac{1070017}{3628800}$

Table 244(II) Coefficients and error constants for Adams–Moulton methods

k	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	C
0	1								$\frac{1}{2}$
1	$\frac{1}{2}$	$\frac{1}{2}$							$-\frac{1}{12}$
2	$\frac{5}{12}$	$\frac{2}{3}$	$-\frac{1}{12}$						$\frac{1}{24}$
3	$\frac{3}{8}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$					$-\frac{19}{720}$
4	$\frac{251}{720}$	$\frac{323}{360}$	$-\frac{11}{30}$	$\frac{53}{360}$	$-\frac{19}{720}$				$\frac{3}{160}$
5	$\frac{95}{288}$	$\frac{1427}{1440}$	$-\frac{133}{240}$	$\frac{241}{720}$	$-\frac{173}{1440}$	$\frac{3}{160}$			$-\frac{863}{60480}$
6	$\frac{19087}{60480}$	$\frac{2713}{2520}$	$-\frac{15487}{20160}$	$\frac{586}{945}$	$-\frac{6737}{20160}$	$\frac{263}{2520}$	$-\frac{863}{60480}$		$\frac{275}{24192}$
7	$\frac{5257}{17280}$	$\frac{139849}{120960}$	$-\frac{4511}{4480}$	$\frac{123133}{120960}$	$-\frac{88547}{120960}$	$\frac{1537}{4480}$	$-\frac{11351}{120960}$	$\frac{275}{24192}$	$-\frac{33953}{3628800}$

where the error constant is equal to $C = \frac{5}{12}$. The values for the Adams–Bashforth methods are given in Table 244(I) and for the Adams–Moulton methods in Table 244(II).

The Adams methods are usually implemented in 'predictor–corrector' form. That is, a preliminary calculation is carried out using the Bashforth form of the method. The approximate solution at a new step value is then used to evaluate an approximation to the derivative value at the new point. This derivative approximation is then used in the Moulton formula in place of the derivative at the new point. There are many alternatives as to what is done next, and we will describe just one of them. Let y_n^* denote the approximation to $y(x_n)$ found during the Bashforth part of the step calculation and y_n the improved approximation found in the Moulton part of the step. Temporarily denote by β_i^* the value of β_i in the Bashforth formula so that β_i will denote only the Moulton coefficient. The value of k corresponding to the Bashforth formula will be denoted here by k^*. Usually k and k^* are related by $k^* = k+1$ so that both formulae have the same order $p = k + 1$.

In the Bashforth stage of the calculation we compute

$$y_n^* = y_{n-1} + h \sum_{i=1}^{k^*} \beta_i^* f(x_{n-i}, y_{n-i}), \qquad (244a)$$

and in the Moulton stage

$$y_n = y_{n-1} + h\beta_0 f(x_n, y_n^*) + h \sum_{i=1}^{k} \beta_i f(x_{n-i}, y_{n-i}). \qquad (244b)$$

Methods of this type are referred to as 'predictor–corrector' methods because the overall computation in a step consists of a preliminary prediction of the answer followed by a correction of this first predicted value. The use of (244a) and (244b) requires two calculations of the function f in each step of the computation. Such a scheme is referred to as being in 'predict–evaluate–correct–evaluate' or 'PECE' mode. An alternative scheme in which the second evaluation is never performed is said to be in 'predict–evaluate–correct' or 'PEC' mode. In this mode, every occurrence of $f(x_{n-i}, y_{n-i})$ would need to be replaced by $f(x_{n-i}, y_{n-i}^*)$, and would represent the value of a derivative evaluated in a previous step but based on the *predicted* approximation to $y(x_{n-i})$ in that step. Thus, (244a) and (244b) would be replaced by

$$y_n^* = y_{n-1} + h \sum_{i=1}^{k^*} \beta_i^* f(x_{n-i}, y_{n-i}^*)$$

and

$$y_n = y_{n-1} + h\beta_0 f(x_n, y_n^*) + h \sum_{i=1}^{k} \beta_i f(x_{n-i}, y_{n-i}^*).$$

In addition to PEC and PECE modes it is also possible to have PECEC and PECECE and, more generally $P(EC)^k$ and $P(EC)^k E$, modes, in which corrections and evaluations are done repeatedly. Using this same type of terminology, $P(EC)^\infty$ indicates iteration to convergence.

245 *The Milne device*

A feature of predictor–corrector methods is that two approximations to $y(x_n)$ are found in each step and each of these possesses different error constants, even though they might have the same order p. Denote the error constant for the Adams–Bashforth p-step method, as given in Table 244(I), by C_p^*, and the corresponding error constant for the $(p-1)$-step Adams–Moulton method, as given in Table 244(II), by C_{p-1}. This means that the error in y_n^*, assuming that previous step values are exact, is equal to

$$y_n^* = y(x_n) - h^{p+1}C_p^* y^{(p+1)}(x_n) + O(h^{p+2}). \tag{245a}$$

Of course, the previous values will not be exact, but we can interpret (245a) in the general case as the new error introduced into step n. Similarly, we can interpret the corresponding formula for the error in the $(p-1)$-step Adams–Moulton method as representing the error introduced into the corrected value of step n. The formula for the Adams–Moulton method is

$$y_n = y(x_n) - h^{p+1}C_{p-1} y^{(p+1)}(x_n) + O(h^{p+2}). \tag{245b}$$

By calculating the difference of the predicted and corrected approximations and multiplying by an appropriate factor, we can estimate the error in the corrected value. That is,

$$y(x_n) - y_n \approx \frac{C_{p-1}}{C_{p-1} - C_p^*}(y_n^* - y_n). \tag{245c}$$

This device, credited to Milne (1926), is used in practical algorithms to estimate local truncation errors for stepsize control. In some modern implementations, the order of the predictor is one lower than that of the corrector, and the Milne device loses the natural significance that we have described. However, it is still found to be a useful tool for adapting a numerical computation to the behaviour of the solution.

246   Starting methods

For a k-step method, where $k > 1$, something special has to be done in the first $k-1$ steps. The method itself gives an algorithm for computing y_k in terms of y_0, y_1, ..., y_{k-1}, and then y_{k+1} in terms of y_1, y_2, ..., y_k, with all subsequent approximations found in a similar manner. However, it must be considered how y_1, y_2, ..., y_{k-1} are to be found before the later steps can be evaluated.

It would be possible to evaluate the first $k-1$ approximations using a sequence of low order methods. However, this would introduce serious errors which would nullify all the advantages of the later use of a method of high order. It would also be possible to use a Runge–Kutta method for the first $k-1$ steps. As long as the Runge–Kutta method has the same order as the linear k-step method to be used for the later steps, then there will be no overall order loss.

In the numerical experiments to be reported in the following subsection, a simple technique is used to retain the use of a single predictor–corrector method, and at the same time to maintain the long term order during the starting process. It is intended that the results should be straightforward and easy to understand, without the influence of alternative methods used in
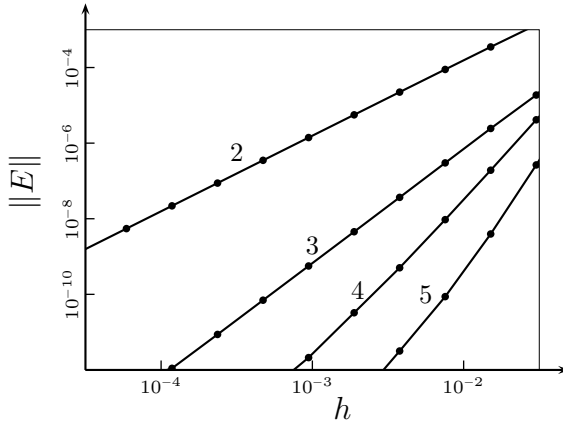
Figure 247(i)   Orbital calculations for various PEC methods

the early steps. What we do is to introduce, as unknowns to be computed, approximations to the values of $f(x_i, y_i)$, for $i = -(k-1), -(k-2), \ldots, -1$. Initial values for these quantities are chosen as $f(x_{i-1}, y_{i-1}) = f(x_0, y_0)$. With these values available, it is possible to carry out the computations in turn of y_i and of $f(x_i, y_i)$ for $i = 1, 2, \ldots, k-1$. This then makes it possible to reverse the direction of integration, by changing the sign of h used in the computations, and to recompute y_i and $f(x_i, y_i)$ for $i = -1, -2, \ldots, -(k-1)$. This process of alternately integrating forwards and backwards can be repeated until convergence is achieved. Once this has happened, acceptable starting values will have been found to permit the step values numbered $i = k$, $i = k+1$, $\ldots$ to be evaluated in turn.

247   Numerical examples

Using the starting process described in Subsection 246, and a range of orders, the same test problem as was used in Subsection 239, that is, (201d) with $e = \frac{1}{2}$, was solved for PEC and PECE Adams methods. The errors generated for these methods are shown in Figures 247(i) (PEC methods) and 247(ii) (PECE methods). The orders are attached to the curves. Note that, at least for this problem, the two modes have almost identical errors. This means, perhaps, that the extra cost of PECE methods is not justified. However, for large stepsizes, there is an advantage in PECE methods because many types of unstable behaviour exhibit themselves more severely for PEC methods. For example, the iterative starting procedure that we have used, failed to converge for large stepsizes (not shown in the diagrams). This effect persisted for a larger range of stepsizes for PEC methods than was the case for PECE methods.
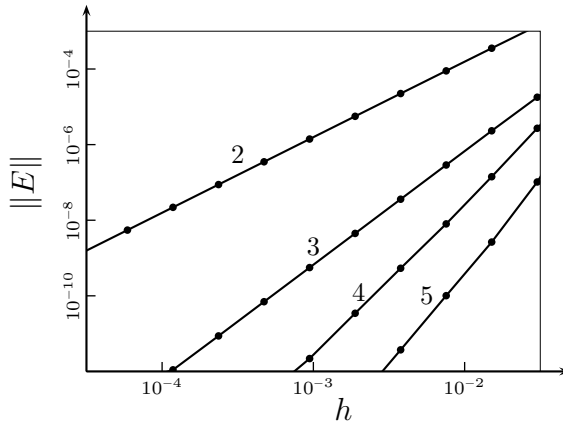
Figure 247(ii)   Orbital calculations for various PECE methods

Exercises 24

24.1 Find a linear multistep method of order 3 of the form

$$y_n = y_{n-2} + \beta_1 h f(x_{n-1}, y_{n-1}) + \beta_2 h f(x_{n-2}, y_{n-2}) + \beta_3 h f(x_{n-3}, y_{n-2}).$$

24.2 Find a linear multistep method of order 3 of the form

$$y_n = y_{n-2} + \beta_0 h f(x_n, y_n) + \beta_1 h f(x_{n-1}, y_{n-1}) + \beta_2 h f(x_{n-2}, y_{n-2}).$$

24.3 If the differential equation $y' = y$ is solved using the implicit method $y_n = y_{n-2} + 2h f(x_{n-1}, y_{n-1})$, show that the resulting difference equation has a solution which grows in powers of $1 + h + \frac{1}{2}h^2 + O(h^3)$ and a second solution that grows in powers of a quantity with *smaller* magnitude.

24.4 If the differential equation $y' = -y$ is solved using the same method, show that the resulting difference equation has a solution which grows in powers of $1 - h + \frac{1}{2}h^2 + O(h^3)$ but has a second solution that grows in powers of a quantity with *greater* magnitude.

25   Taylor Series Methods

250   Introduction to Taylor series methods

A differential equation $y'(x) = f(x, y(x))$, characterized by the function f, is presented to a computer in the form of a procedure, function or subroutine for computing values of $f(u, v)$ for given arguments u and v. The program carries out the evaluation of this procedure in a manner that exactly corresponds to the occurrence of the function f, in the mathematical formulation of the

numerical method. In this brief introduction, we consider the use of procedures that evaluate, for given values of x and $y(x)$, not only the value of $y'(x)$, but also the value of $y''(x)$ and possibly also $y'''(x)$ and other higher derivatives.

With such facilities available, there is a wide range of possible methods, but the natural and straightforward choice of Taylor series is almost always followed. By repeated differentiation, we can find functions $f_2(x, y(x))$, $f_3(x, y(x))$, ..., $f_m(x, y(x))$, which give values, respectively, of $y''(x)$, $y'''(x)$, ..., $y^{(m)}(x)$.

The order m formula for computing $y(x_n) = y(x_{n-1} + h)$ using these functions, evaluated at $x = x_{n-1}$ and $y = y_{n-1}$, is

$$y_n = y_{n-1} + hf(x_{n-1}, y_{n-1}) + \frac{h^2}{2!} f_2(x_{n-1}, y_{n-1})$$
$$+ \cdots + \frac{h^m}{m!} f_m(x_{n-1}, y_{n-1}). \quad \text{(250a)}$$

Most serious investigations of this method have been concerned, above all, with the automatic generation of procedures for generating the second, third, ... derivative functions f_2, f_3, ... from a given first derivative function f. While this aspect of the Taylor series method is more within the scope of algebraic manipulation than of numerical analysis, there are other important aspects which arise, just as for other methods. These include error estimation, order selection and stepsize control.

Although many individuals and teams have made important contributions to the use of Taylor series methods, we mention three in particular. The program of Gibbons (1960), using a computer with the limited memory available at that time, used a recursive technique to generate the Taylor coefficients automatically. A similar approach using greater sophistication and more powerful computational tools was used by Barton, Willers and Zahar (1971). The work of Moore (1964) is especially interesting, in that it uses interval arithmetic and supplies rigorous error bounds for the computed solution.

251   *Manipulation of power series*

We consider problems for which the components of the function f are rational in x and in the components of y. This means that the terms occurring in (250a) can all be computed by the use of addition (and subtraction), multiplication and division.

We use power series with the $1/i!$ factor absorbed into the coefficient of $f_i(x_{n-1}, y_{n-1})$. Hence each component takes the form $a_0 + a_1 h + a_2 h^2 + \cdots + a_m h^m$. If a second such expansion, $b_0 + b_1 h + b_2 h^2 + \cdots + b_m h^m$, is added or subtracted, then we simply add or subtract corresponding coefficients. The product of two terms is found by expanding the formal product but truncating

after the h^m term. This means that the product of $a_0 + a_1 h + a_2 h^2 + \cdots + a_m h^m$ and $b_0 + b_1 h + b_2 h^2 + \cdots + b_m h^m$ would be $c_0 + c_1 h + c_2 h^2 + \cdots + c_m h^m$, where

$$c_i = \sum_{j=0}^{i} a_{i-j} b_j, \qquad i = 0, 1, \ldots, m. \tag{251a}$$

The formula for the quotient

$$a_0 + a_1 h + a_2 h^2 + \cdots + a_m h^m \approx \frac{c_0 + c_1 h + c_2 h^2 + \cdots + c_m h^m}{b_0 + b_1 h + b_2 h^2 + \cdots + b_m h^m}$$

is found by reinterpreting the relationship between the a_i, b_i and c_i coefficients in (251a) to give

$$a_i = \begin{cases} \dfrac{c_0}{b_0}, & i = 0, \\[2ex] \dfrac{c_i - \sum_{j=1}^{i} a_{i-j} b_j}{b_0}, & i = 1, 2, \ldots, m. \end{cases} \tag{251b}$$

Given a system of differential equations with dependent variables y^1, y^2, $\ldots$, y^N, write the truncated power series for $y^k(x_{n-1} + h)$ in the form $y_0^k + h y_1^k + \cdots + h^m y_m^k$, $k = 1, 2, \ldots, N$. Also denote the power series for component k of $f(x_{n-1}, Y)$ by $f_0^k + h f_1^k + \cdots + h^m f_m^k$, where the vector Y has its components substituted by the series $y_0^l + h y_1^l + \cdots + h^m y_m^l$, $l = 1, 2, \ldots, N$.

We consider how to evaluate in turn the y_i^k coefficients for each $k = 1, 2, \ldots, N$, with i taking on values from 0 to m. For $i = 0$, all the y_i^k are known from initial information at the start of the current step. For each value of $i > 0$ we already know the coefficients y_j^k for all k and for all $j < i$. It is thus possible to evaluate the h^{i-1} terms in the components in the power series for $f(x_{n-1} + h, y_{n-1})$. Writing the differential equation in the form

$$\frac{d}{dh}(y_0^k + h y_1^k + \cdots + h^m y_m^k) = y_1^k + 2h y_2^k + \cdots + m h^{m-1} y_m^k$$
$$= f_0^k + h f_1^k + \cdots + h^{m-1} f_{m-1}^k,$$

where the last term on the right-hand side has been deleted, we see that $y_i^k = f_{i-1}^k / i$.

When we have reached $i = m$, all the required coefficients are known at $x = x_{n-1}$, and it is possible to take the step to $x = x_n$.

This method of solution will be illustrated in the next subsection.

252   An example of a Taylor series solution

We consider the example problem, already introduced in Subsection 201,

$$y'(x) = \frac{y + x}{y - x}, \qquad y(0) = 1. \tag{252a}$$

Algorithm 252α A Taylor step for problem (252a)

```
a(1) = y;
b(1) = y + x;
c(1) = y - x;
for i = 0: m - 1,
   temp = b(i+1);
   for j = 1: i,
      temp = temp - d(1+i-j)*c(1+j);
   end;
   d(i+1) = temp/c(1);
   a(i+2) = d(i+1)/(i+1);
   if i == 0,
      b(i+2) = a(i+2) + 1;
      c(i+2) = a(i+2) - 1;
   else
      b(i+2) = a(i+2);
      c(i+2) = a(i+2);
   end;
end;
x = x + h;
y = a(m+1);
for i = m-1:-1:0,
   y = a(i+1) + h*y;
end;
```

Let $a_0, a_1, \ldots, a_m$ denote Taylor coefficients for $y(x_{n-1} + h)$, $b_0, b_1, \ldots, b_m$ be the corresponding coefficients for $y+x$, and $c_0, c_1, \ldots, c_m$ be the coefficients for $y - x$. If $d_0, d_1, \ldots, d_m$ are the coefficients for $(y+x)/(y-x)$, then Algorithm 252α, written in MATLAB, can be used to update the value of $x = x_{n-1}$ and $y = y_{n-1}$ to the values at the end of a step, $x = x_n$, $y = y_n$. Note that $a_0, a_1, \ldots, a_m$ are represented in this program by $\mathtt{a(1)}$, $\mathtt{a(2)}$, ... , $\mathtt{a(m+1)}$, because MATLAB subscripts start from 1 (and similarly for the b_i, etc.).

Numerical experiments based on this program have been made for a sequence of m values from 1 to 10 and using stepsizes $h = 0.10 \times 2^{-k}$, with $k = 1, 2, \ldots$. The errors in the approximations to $y(0.5)$ are presented in Figure 252(i). It can be seen that the rate of increase in accuracy, as smaller and smaller steps are taken, becomes more and more impressive as m increases. The results found for $m = 9$ and $m = 10$ are not included because, even for 10 steps with $h = 0.05$, the numerical results in these cases are accurate to approximately 15 decimal places.
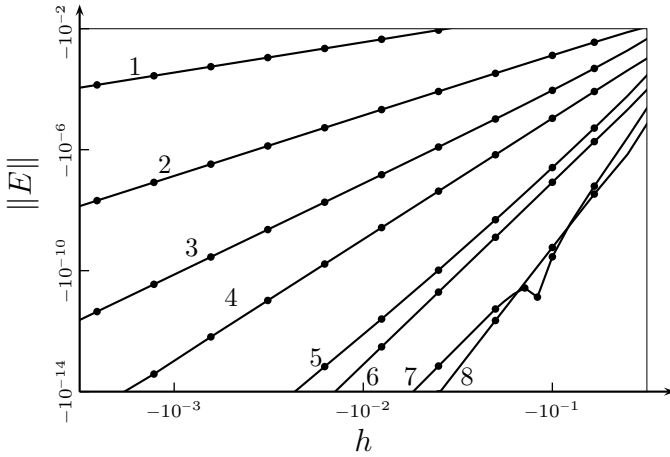
Figure 252(i)   Taylor series calculations
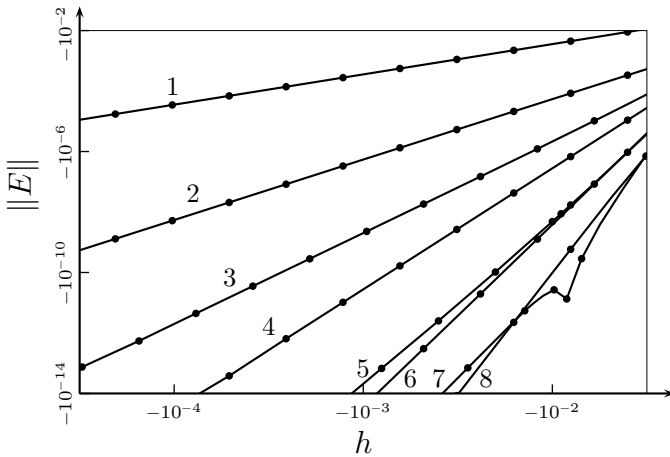
Figure 252(ii)   Taylor series calculations with cost correction

Since high values of m are more time-consuming, the favourable impression of their advantages shown in this figure is an exaggeration. Since the cost is approximately proportional to m, a fairer comparison would be to plot the errors against h/m. This weighted comparison is shown in Figure 252(ii).

The advantage of high order methods over low order methods is still evident from this more balanced comparison.

Table 253(I)   Coefficients defined by (253c)

m	α_1	α_2	α_3	α_4	β_1	β_2	β_3	β_4
1	$\frac{3}{2}$				$-\frac{1}{2}$			
2	$-\frac{1}{2}$	$\frac{17}{12}$			$\frac{3}{2}$	$\frac{7}{12}$		
3	$\frac{15}{2}$	$-\frac{31}{10}$	$\frac{37}{40}$		$-\frac{13}{2}$	$-\frac{29}{10}$	$-\frac{49}{120}$	
4	$-\frac{65}{2}$	$\frac{515}{28}$	$-\frac{107}{28}$	$\frac{769}{1680}$	$\frac{67}{2}$	$\frac{437}{28}$	$\frac{239}{84}$	$\frac{117}{560}$

Table 253(II)   Coefficients defined by (253d)

m	γ_1	γ_2	γ_3	γ_4	δ_1	δ_2	δ_3	δ_4
1	$\frac{1}{2}$				$\frac{1}{2}$			
2	$\frac{1}{2}$	$-\frac{1}{12}$			$\frac{1}{2}$	$\frac{1}{12}$		
3	$\frac{1}{2}$	$-\frac{1}{10}$	$\frac{1}{120}$		$\frac{1}{2}$	$\frac{1}{10}$	$\frac{1}{120}$	
4	$\frac{1}{2}$	$-\frac{3}{28}$	$\frac{1}{84}$	$-\frac{1}{1680}$	$\frac{1}{2}$	$\frac{3}{28}$	$\frac{1}{84}$	$\frac{1}{1680}$

253   Other methods using higher derivatives

We consider the possibility of using higher derivative information at more than one step value. In particular, we consider two special schemes of the form

$$y_n = y_{n-1} + h\alpha_1 f(x_{n-1}, y_{n-1}) + h^2\alpha_2 f_2(x_{n-1}, y_{n-1}) + \cdots$$
$$+ h^m \alpha_m f_m(x_{n-1}, y_{n-1}) + h\beta_1 f(x_{n-2}, y_{n-2})$$
$$+ h^2\beta_2 f_2(x_{n-2}, y_{n-2}) + \cdots + h^m\beta_m f_m(x_{n-2}, y_{n-2}) \quad (253a)$$

and

$$y_n = y_{n-1} + h\gamma_1 f(x_n, y_n) + h^2\gamma_2 f_2(x_n, y_n) + \cdots$$
$$+ h^m \gamma_m f_m(x_n, y_n) + h\delta_1 f(x_{n-1}, y_{n-1})$$
$$+ h^2\delta_2 f_2(x_{n-1}, y_{n-1}) + \cdots + h^m\delta_m f_m(x_{n-1}, y_{n-1}). \quad (253b)$$

The scheme (253a) uses information already available before step n is attempted. Thus it can be regarded as a generalization of an Adams–Bashforth method. In contrast, the scheme (253b) is fully implicit, and thus corresponds to an Adams–Moulton method. Using Taylor series analyses, conditions for order $2m$ can readily be found. These are equivalent to the conditions

$$\exp(z) - (1 + \alpha_1 z + \alpha_2 z^2 + \cdots + \alpha_m z^m)$$
$$- (\beta_1 z + \beta_2 z^2 + \cdots + \beta_m z^m) \exp(-z) = O(z^{2m+1}) \quad \text{(253c)}$$

and

$$\exp(z)(1 - \gamma_1 z - \gamma_2 z^2 - \cdots - \gamma_m z^m)$$
$$- (1 + \delta_1 z + \delta_2 z^2 + \cdots + \delta_m z^m) = O(z^{2m+1}). \quad \text{(253d)}$$

Note that the rational function

$$\frac{N(z)}{D(z)} = \frac{1 + \delta_1 z + \delta_2 z^2 + \cdots + \delta_m z^m}{1 - \gamma_1 z - \gamma_2 z^2 - \cdots - \gamma_m z^m},$$

is known as a Padé approximation to the exponential function. It is the unique rational function with degree m in both numerator and denominator, which maximizes the order of approximation of $N(z)/D(Z)$ to $\exp(z)$.

For easy reference, the coefficients α_i, β_i, γ_i and δ_i are shown in Tables 253(I) and 253(II) up to $m = 4$.

An example of the use of the methods discussed here, in a predictor–corrector mode, will be presented in Subsection 255.

254   The use of f derivatives

In this subsection, we consider an autonomous differential equation system $y'(x) = f(y(x))$. As we remarked in Subsection 224, the use of derivatives of f is more general than the use of higher derivatives of y. Methods that use f' directly have mainly been proposed for the solution of stiff problems by one-step methods. If an implicit Runge–Kutta method is used, the implementation requires the solution of non-linear equations, typically by a Newton-type method. It was proposed by Rosenbrock (1963) that the Newton iterations could be replaced by a single iteration involving the inverse of a matrix such as $I - h\gamma f'(y(x_{n-1}))$. Methods formed in this way use this linear operation as an intrinsic part of the order requirement for the method. We give a single example in which modified derivatives F_1 and F_2, and the final result at the end of a step, are computed by the formulae

$$\left(I - h\left(1 - \tfrac{\sqrt{2}}{2}\right)f'(y_{n-1})\right)F_1 = f(y_{n-1}), \quad \text{(254a)}$$

$$\left(I - h\left(1 - \tfrac{\sqrt{2}}{2}\right)f'(y_{n-1})\right)F_2 = f\left(y_{n-1} + h\left(\tfrac{\sqrt{2}}{2} - \tfrac{1}{2}\right)F_1\right), \quad \text{(254b)}$$

$$y_n = y_{n-1} + hF_2. \quad \text{(254c)}$$

Methods of various orders have been derived by Rosenbrock and others. These are known collectively as Rosenbrock methods, although the ambiguous name 'implicit Runge–Kutta methods' is sometimes applied to them.
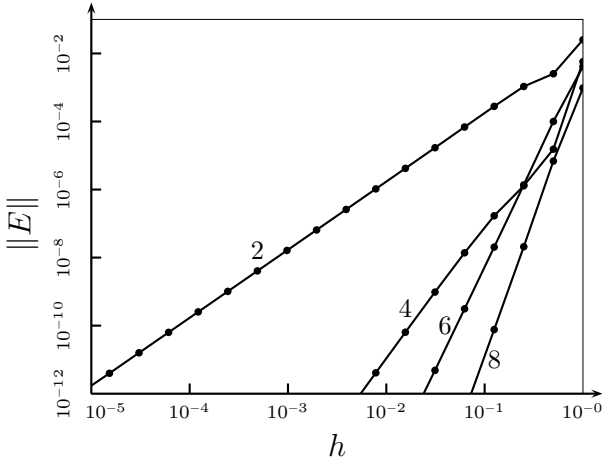
Figure 255(i)   Predictor–corrector multiderivative methods
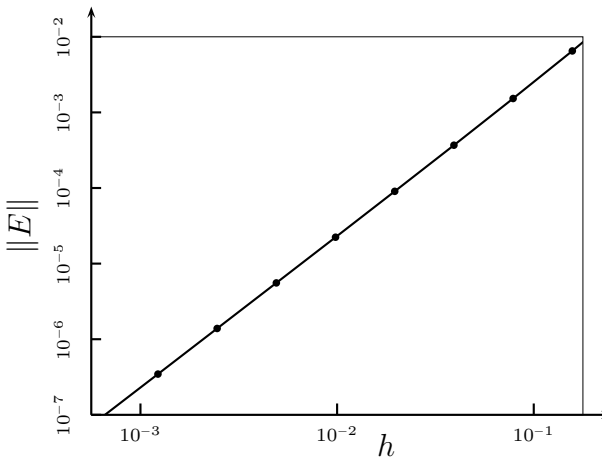
Figure 255(ii)   Rosenbrock method given by (254a)–(254c)

255   Further numerical examples

We consider the solution of the same problem discussed in Subsection 252, but using the methods of Subsection 253. The two methods discussed there, for various values of m, implying orders $2m$, attached to the curves, are used together in predictor–corrector mode in Figure 255(i). A comparison with Figure 252(i) shows the new methods to be slightly more accurate for the same stepsizes.

The final numerical result in this subsection is based on the mildly stiff problem (203a), written in the form

$$\frac{dy_1}{dx} = -16y_1 + 12y_2 + 16\cos(y_3) - 13\sin(y_3), \qquad y_1(0) = 1,$$

$$\frac{dy_2}{dx} = 12y_1 - 9y_2 - 11\cos(y_3) + 9\sin(y_3), \qquad y_2(0) = 0,$$

$$\frac{dy_3}{dx} = 1, \qquad y_3(0) = 0.$$

The norm errors for the approximate solution at $x = \pi$ are given for various h in Figure 255(ii).

Exercises 25

25.1 Consider the function $f(x, y) = x^2 + y^2$ and the differential equation $y'(x) = f(x, y(x))$. Derive formulae for the second, third and fourth derivatives.

25.2 Solve the initial value problem $y'(x) = x^2 + y(x)^2$, $y(0) = 1$ by the fourth order Taylor series method using n steps with constant stepsize $h = 1/n$ to yield approximations to the solution at $x = 1$. Use $n = 1, 2, 4, \ldots, 2^{10}$. Are the results consistent with the order 4 nature of the method?

25.3 Use the eighth order predictor–corrector method discussed in Subsection 253 to solve this problem.

25.4 Show that the Rosenbrock method given by (254a), (254b) and (254c) has order 2.

26   Hybrid Methods

260   Historical introduction

The idea of combining the ideas behind Runge–Kutta methods with those behind linear multistep methods dates from the period 1960–1970. One approach is to make use of stage derivatives computed in one or more previous steps in the computation of the approximation at the end of a current step. Methods based on this idea are referred to as pseudo Runge–Kutta methods. The earliest work on these methods is that of Byrne and Lambert (1966).

Another type of generalization of existing methods was proposed in three independent publications (Gragg and Stetter, 1964; Butcher, 1965; Gear, 1965). The most commonly used name for these is that introduced by Gear, 'hybrid methods', although we use here the name 'modified multistep methods' introduced by Butcher. A consideration of these various generalizations has led to the construction of comprehensive theories. We consider one of the earliest of these formulations in this section, and refer to the wide class of multivalue–multistage methods as 'general linear methods'.

261   Pseudo Runge–Kutta methods

The paper by Byrne and Lambert suggests a generalization of Runge–Kutta methods in which stage derivatives computed in *earlier* steps are used alongside stage derivatives found in the current step, to compute the output value in the step. The stages themselves are evaluated in exactly the same way as for a Runge–Kutta method. We consider the case where the derivatives found only in the immediately previous step are used. Denote these by $F_i^{[n-1]}$, $i = 1, 2, \ldots, s$, so that the derivatives evaluated in the current step, n, are $F_i^{[n]}$, $i = 1, 2, \ldots, s$.

The defining equations for a single step of the method will now be

$$Y_i = y_{n-1} + h \sum_{j=1}^{s} a_{ij} F_j^{[n]},$$

$$F_i^{[n]} = f(x_{n-1} + hc_i, Y_i),$$

$$y_n = y_{n-1} + h \left(\sum_{i=1}^{s} b_i F_i^{[n]} + \sum_{i=1}^{s} \overline{b}_i F_i^{[n-1]} \right).$$

We consider a single example of a pseudo Runge–Kutta method in which there are $s = 3$ stages and the order is $p = 4$. The coefficients are given by the tableau

$$
\begin{array}{c|ccc}
0 \\
\frac{1}{2} & \frac{1}{2} \\
1 & -\frac{1}{3} & \frac{4}{3} \\
\hline
& \frac{11}{12} & \frac{1}{3} & \frac{1}{4} \\
\hline
& \frac{1}{12} & -\frac{1}{3} & -\frac{1}{4}
\end{array}
\tag{261a}
$$

where the additional vector contains the $\overline{b}^{\mathsf{T}}$ components.

Characteristic handicaps with this sort of method are starting and changing stepsize. Starting in this case can be accomplished by taking the first step with the classical Runge–Kutta method but inserting an additional stage Y_5, with the role of $Y_3^{(1)}$, to provide, along with $Y_2^{(2)} = Y_2$, the derivatives in step 1 required to complete step 2. Thus the starting step is based on the Runge–Kutta method

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

262   Generalized linear multistep methods

These methods, known also as hybrid methods or modified linear multistep methods, generalize linear multistep methods, interpreted as predictor–corrector pairs, by inserting one or more additional predictors, typically at off-step points. Although many examples of these methods are known, we give just a single example for which the off-step point is $\frac{8}{15}$ of the way through the step. That is, the first predictor computes an approximation to $y(x_{n-1} + \frac{8}{15}h) = y(x_n - \frac{7}{15}h)$. We denote this first predicted value by the symbol $\widetilde{y}_{n-7/15}$ and the corresponding derivative by $\widetilde{f}_{n-7/15} = f(x_n - \frac{7}{15}h, \widetilde{y}_{n-7/15})$. Similarly, the second predictor, which gives an initial approximation to $y(x_n)$, will be denoted by $\widetilde{y}_n$ and the corresponding derivative by $\widetilde{f}_n = f(x_n, \widetilde{y}_n)$. This notation is in contrast to y_n and f_n, which denote the corrected step approximation to $y(x_n)$ and the corresponding derivative $f(x_n, y_n)$, respectively. The relationships between these quantities are

$$\widetilde{y}_{n-7/15} = -\frac{529}{3375}y_{n-1} + \frac{3904}{3375}y_{n-2} + h\left(\frac{4232}{3375}f_{n-1} + \frac{1472}{3375}f_{n-2}\right),$$

$$\widetilde{y}_n = \frac{152}{25}y_{n-1} - \frac{127}{25}y_{n-2} + h\left(\frac{189}{92}\widetilde{f}_{n-7/15} - \frac{419}{100}f_{n-1} - \frac{1118}{575}f_{n-2}\right),$$

$$y_n = y_{n-1} + h\left(\frac{25}{168}\widetilde{f}_n + \frac{3375}{5152}\widetilde{f}_{n-7/15} + \frac{19}{96}f_{n-1} - \frac{1}{552}f_{n-2}\right).$$

263   General linear methods

To obtain a general formulation of methods that possess the multivalue attributes of linear multistep methods, as well as the multistage attributes of Runge–Kutta methods, general linear methods were introduced by the present author (Butcher, 1966). However, the formulation we present, while formally different, is equivalent in terms of the range of methods it can represent, and was introduced in Burrage and Butcher (1980).

Suppose that r quantities are passed from step to step. At the start of step n, these will be denoted by $y_1^{[n-1]}$, $y_2^{[n-1]}$, ..., $y_r^{[n-1]}$, and after the step is completed, the corresponding quantities available for use in the subsequent step will be $y_1^{[n]}$, $y_2^{[n]}$, ..., $y_r^{[n]}$. During the computation of the step, s stage values Y_1, Y_2, ..., Y_s are computed, along with the corresponding stage derivatives F_1, F_2, ..., F_s. For convenience of notation, we can create supervectors containing either r or s subvectors as follows:

$$y^{[n-1]} = \begin{bmatrix} y_1^{[n-1]} \\ y_2^{[n-1]} \\ \vdots \\ y_r^{[n-1]} \end{bmatrix}, \quad y^{[n]} = \begin{bmatrix} y_1^{[n]} \\ y_2^{[n]} \\ \vdots \\ y_r^{[n]} \end{bmatrix}, \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix}, \quad F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_s \end{bmatrix}.$$

Just as for Runge–Kutta methods, the stages are computed making use of linear combinations of the stage derivatives but, since there are now a collection of input approximations, further linear combinations are needed to express the dependence on this input information. Similarly, the output quantities depend linearly on both the stage derivatives and the input quantities. All in all, four matrices are required to express all the details of these computations, and we denote these by $A = [a_{ij}]_{s,s}$, $U = [u_{ij}]_{s,r}$, $B = [b_{ij}]_{r,s}$ and $V = [v_{ij}]_{r,r}$.

The formulae for the stage values and the output values are

$$Y_i = \sum_{j=1}^{s} h a_{ij} F_j + \sum_{j=1}^{r} u_{ij} y_j^{[n-1]}, \quad i = 1, 2, \ldots, s,$$

$$y_i^{[n]} = \sum_{j=1}^{s} h b_{ij} F_j + \sum_{j=1}^{r} v_{ij} y_j^{[n-1]}, \quad i = 1, 2, \ldots, r,$$

or, using Kronecker product notation for an N-dimensional problem,

$$Y = h(A \otimes I_N)F + (U \otimes I_N)y^{[n-1]},$$

$$y^{[n]} = h(B \otimes I_N)F + (V \otimes I_N)y^{[n-1]}.$$

We devote Chapter 5 to a detailed study of general linear methods but, for the present, we illustrate the all-encompassing nature of the methods included in this family by presenting a number of sample methods written in this terminology.

In each case, the coefficients of the general linear formulation are presented in the $(s+r) \times (s+r)$ partitioned matrix

$$\left[\begin{array}{cc} A & U \\ B & V \end{array} \right].$$

The Euler method and implicit Euler methods are, respectively,

$$\left[\begin{array}{c|c} 0 & 1 \\ \hline 1 & 1 \end{array} \right] \quad \text{and} \quad \left[\begin{array}{c|c} 1 & 1 \\ \hline 1 & 1 \end{array} \right].$$

The Runge–Kutta methods (232a) and (233f) and (235i) are, respectively,

$$\left[\begin{array}{cc|c} 0 & 0 & 1 \\ 1 & 0 & 1 \\ \hline \frac{1}{2} & \frac{1}{2} & 1 \end{array} \right] \quad \text{and} \quad \left[\begin{array}{ccc|c} 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & 1 \\ -1 & 2 & 0 & 1 \\ \hline \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 1 \end{array} \right] \quad \text{and} \quad \left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ \hline \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & 1 \end{array} \right].$$

The second order Adams–Bashforth and Adams–Moulton and PECE methods based on these are, respectively,

$$
\left[\begin{array}{c|ccc}
0 & 1 & \frac{3}{2} & -\frac{1}{2} \\
\hline
0 & 1 & \frac{3}{2} & -\frac{1}{2} \\
1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0
\end{array}\right]
\quad \text{and} \quad
\left[\begin{array}{c|c}
\frac{1}{2} & 1 \\
\hline
\frac{1}{2} & 1
\end{array}\right]
\quad \text{and} \quad
\left[\begin{array}{cc|ccc}
0 & 0 & 1 & \frac{3}{2} & -\frac{1}{2} \\
\frac{1}{2} & 0 & 1 & \frac{1}{2} & 0 \\
\hline
\frac{1}{2} & 0 & 1 & \frac{1}{2} & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0
\end{array}\right],
$$

where for each of the Adams–Bashforth and PECE methods, the output quantities are approximations to $y(x_n)$, $hy'(x_n)$ and $hy'(x_{n-1})$, respectively.

Finally, we re-present two methods derived in this section. The first is the pseudo Runge–Kutta method (261a), for which the general linear representation is

$$
\left[\begin{array}{ccc|cccc}
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\frac{1}{2} & 0 & 0 & 1 & 0 & 0 & 0 \\
-\frac{1}{3} & \frac{4}{3} & 0 & 1 & 0 & 0 & 0 \\
\hline
\frac{11}{12} & \frac{1}{3} & \frac{1}{4} & 1 & \frac{1}{12} & -\frac{1}{3} & -\frac{1}{4} \\
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0
\end{array}\right].
$$

The four output quantities for this method are the approximate solution found at the end of the step, together with h multiplied by each of the three stage derivatives. The second of the two general linear methods, that do not fit into any of the classical families, is the method introduced in Subsection 262. Its general linear method coefficient matrix is

$$
\left[\begin{array}{ccc|cccc}
0 & 0 & 0 & -\frac{529}{3375} & \frac{3904}{3375} & \frac{4232}{3375} & \frac{1472}{3375} \\
\frac{189}{92} & 0 & 0 & \frac{152}{25} & -\frac{127}{25} & -\frac{419}{100} & -\frac{1118}{575} \\
\frac{3375}{5152} & \frac{25}{168} & 0 & 1 & 0 & \frac{19}{96} & -\frac{1}{552} \\
\hline
\frac{3375}{5152} & \frac{25}{168} & 0 & 1 & 0 & \frac{19}{96} & -\frac{1}{552} \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0
\end{array}\right].
$$

For this method, the output quantities are given by $y_1^{[n]} \approx y(x_n)$, $y_2^{[n]} \approx y(x_{n-1})$, $y_3^{[n]} \approx hy'(x_n)$ and $y_4^{[n]} \approx hy'(x_{n-1})$.
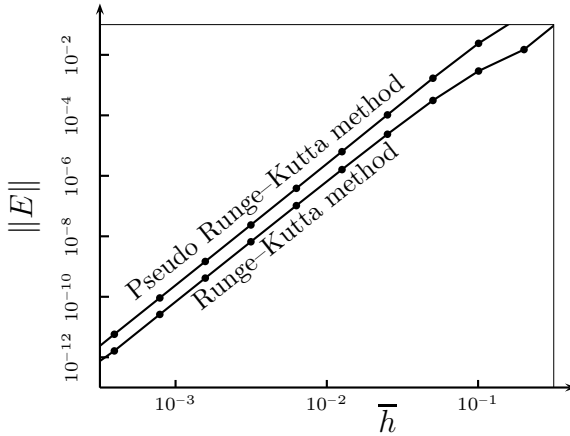
Figure 264(i)   Comparison of Runge–Kutta with pseudo Runge–Kutta method

264   Numerical examples

The limited numerical testing performed here does not give a great deal of support to the use of pseudo Runge–Kutta or hybrid methods. Using the Kepler problem with eccentricity $e = \frac{1}{2}$ over a half period, the pseudo Runge–Kutta method (261a) was compared with the classical Runge–Kutta method and the results are summarized in Figure 264(i). To make the comparison as fair as possible, the axis denoted by $\overline{h}$ shows the stepsize per function evaluation. That is, for the Runge–Kutta method, $h = 4\overline{h}$, and for the pseudo Runge–Kutta method, $h = 3\overline{h}$. The classical Runge–Kutta is significantly more accurate for this problem.

   A similar comparison has been made between the hybrid method discussed in Subsection 262 and a fifth order Runge–Kutta method, but the results, which are not presented here, show almost identical performance for the two methods.

Exercises 26

26.1 Find the error computed in a single step using the method (261a) for the problem

$$y'(x) = x^4$$

and show that this is 16 times the error for the classical Runge–Kutta method.

26.2 Find a fifth order method similar to the one discussed in Subsection 262, but with first predictor giving an approximation to $y(x_n - \frac{1}{2}h)$.

26.3 Show how to represent the PEC method based on the second order Adams–Bashforth predictor and the third order Adams–Moulton corrector as a general linear method.

26.4 Show how to represent the PECEC method based on second order Adams–Bashforth and Adams–Moulton methods as a general linear method.

27   Introduction to Implementation

270   Choice of method

Many differential equation solvers have been constructed, based on a variety of computational schemes, from Runge–Kutta and linear multistep methods, to Taylor series and extrapolation methods. In this introduction to implementation of initial value solvers, we will use an 'Almost Runge–Kutta' (ARK) method. We will equip this method with local error estimation, variable stepsize and interpolation. It is intended for non-stiff problems but can be used also for delay problems, because of its reliable and accurate built-in interpolation.

Many methods are designed for variable order, but this is a level of complexity which we will avoid in this introduction. The method to be presented has order 3 and, because it is a multivalue method, it might be expected to require an elaborate starting sequence. However, it is a characteristic property of ARK methods that starting will present a negligible overhead on the overall costs and will involve negligible complication in the design of the solver.

Recall from Subsection 263 the notation used for formulating a general linear method. In the case of the new experimental method, the coefficient matrix is

$$
\begin{bmatrix} A & U \\ B & V \end{bmatrix} =
\left[
\begin{array}{ccc|ccc}
0 & 0 & 0 & 1 & \frac{1}{3} & \frac{1}{18} \\
\frac{1}{2} & 0 & 0 & 1 & \frac{1}{6} & \frac{1}{18} \\
0 & \frac{3}{4} & 0 & 1 & \frac{1}{4} & 0 \\
\hline
0 & \frac{3}{4} & 0 & 1 & \frac{1}{4} & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
3 & -3 & 2 & 0 & -2 & 0
\end{array}
\right].
$$

Because general linear methods have no specific interpretation, we need to state the meaning of the various quantities which play a role in the formulation of the method. Approximate values of these are as follows:

Algorithm 270α   A single step using an ARK method

```
function [xout, yout] = ARKstep(x,y,f,h)
Uy = y*[1,1,1;1/3,1/6,1/4;1/18,1/18,0];
hF = h*f(x+(1/3)*h,Uy(:,1));
hF = [hF,h*f(x+(2/3)*h,Uy(:,2)+(1/2)*hF)];
xout = x+h;
y1out = Uy(:,3)+hF*[0;3/4];
hF = [hF,h*f(xout,y1out)];
y3out = hF*[3;-3;2]-2*y(:,2);
yout = [y1out,hF(:,3),y3out];
```

$$y_1^{[n-1]} = y(x_{n-1}),$$
$$y_2^{[n-1]} = hy'(x_{n-1}),$$
$$y_3^{[n-1]} = h^2y''(x_{n-1}),$$
$$Y_1 = y(x_{n-1} + \tfrac{1}{3}h),$$
$$Y_2 = y(x_{n-1} + \tfrac{2}{3}h),$$
$$Y_3 = y(x_{n-1} + h),$$
$$y_1^{[n]} = y(x_n),$$
$$y_2^{[n]} = hy'(x_n),$$
$$y_3^{[n]} = h^2y''(x_n).$$

The method is third order and we would expect that, with precise input values, the output after a single step would be correct to within $O(h^4)$. With the interpretation we have introduced, this is not quite correct because the third output value is in error by $O(h^3)$ from its target value. We can correct this by writing down a more precise formula for $y_3^{[n-1]}$, and correspondingly for $y_3^{[n]}$. However, we can avoid having to do this, by remarking that the method satisfies what are called 'annihilation conditions' which cause errors $O(h^3)$ in the input $y_3^{[n-1]}$ to be cancelled out in the values computed for $y_1^{[n]}$ and $y_2^{[n]}$. For this method, the stages are all computed correctly to within $O(h^3)$, rather than only to first order accuracy as in an explicit Runge–Kutta method. The computations constituting a single step of the method in the solution of a differential equation $y' = f(x,y)$ are shown in Algorithm 270α. The array y as a parameter for the function ARKstep consists of three columns with the values of $y_1^{[n-1]}$, $y_2^{[n-1]}$, $y_3^{[n-1]}$, respectively. The updated values of these quantities, at the end of step n, are embedded in a similar way in the output result yout.

271   Variable stepsize

Variation in the stepsize as the integration proceeds, is needed to deal with changes in behaviour in the apparent accuracy of individual steps. If, in addition to computing the output results, an approximation is computed to the error committed in each step, a suitable strategy is to adjust h to maintain the error estimates close to a fixed value, specified by a user-imposed tolerance.

In the case of the ARK method introduced in Subsection 270, we propose to compute an alternative approximation to y at the end of the step and to regard their difference as an error estimate. This alternative approximation will be defined as

$$\widehat{y}_n = y_1^{[n-1]} + \tfrac{1}{8}y_2^{[n-1]} + \tfrac{3}{8}(hF_1 + hF_2) + \tfrac{1}{8}hF_3, \qquad (271a)$$

based on the three-eighths rule quadrature formula. It is known that the difference between $\widehat{y}_n$ and $y_1^{[n]}$ is $O(h^4)$, and this fact will be used in stepsize adjustments.

Because of the asymptotic behaviour of the error estimate, we can increase or decrease the error predicted in the following step, by multiplying h by

$$r = \left(\frac{T}{\|\widehat{y} - y_1^{[n]}\|} \right)^{1/4}. \qquad (271b)$$

This assumes that the error, or at least the quantity we are estimating, is changing slowly from step to step. If $\|\widehat{y} - y_1^{[n]}\| \le T$ is used as a criterion for accepting the current step, then the use of (271b) to predict the *next* stepsize allows the possibility of obtaining an unwanted rejection in the new step. Hence it is customary to insert a safety factor, equal to 0.9 for example, in (271a). Furthermore, to avoid violent swings of h in exceptional circumstances, the stepsize ratio is usually forced to lie between two bounds, such as 0.5 and 2.0. Thus we should refine (271b) by multiplying h not by r, but by $\min(\max(0.5, 0.9r), 2.0)$. For robust program design, the division in (271b) must be avoided when the denominator becomes accidentally small.

In modern solvers, a more sophisticated stepsize adjustment is used, based on PI control (Gustafsson, Lundh and Söderlind, 1988; Gustafsson, 1991). In the terminology of control theory, P control refers to 'proportional control', whereas PI or 'proportional integral control' uses an accumulation of values of the controller, in this case a controller based on error estimates, over recent time steps.

To illustrate the ideas of error estimation and stepsize control, a modified version of Algorithm 270α is presented as Algorithm 271α. The additional parameter T denotes the tolerance; the additional outputs hout and reject are, respectively, the proposed stepsize in the succeeding step and an indicator as to whether the current step apparently achieved sufficient accuracy. In the case reject $= 1$, signifying failure, the variables xout and yout retain the corresponding input values x and y.

Algorithm 271α An ARK method step with stepsize control

```
function [xout,yout,hout,reject] = ARKstep(x,y,f,h,T)
Uy = y*[1,1,1;1/3,1/6,1/4;1/18,1/18,0];
hF = h*f(x+(1/3)*h,Uy(:,1));
hF = [hF,h*f(x+(2/3)*h,Uy(:,2)+(1/2)*hF)];
xout = x+h;
y1out = Uy(:,3)+hF*[0;3/4];
hF = [hF,h*f(xout,y1out)];
y3out = hF*[3;-3;2]-2*y(:,2);
yout = [y1out,hF(:,3),y3out];
err = norm(hF*[3/8;-3/8;1/8]-y(:,2)/8);
reject = err > T;
if err < 0.04*T
   r = 2;
else
   r = (T/err)^0.25;
   r = min(max(0.5, 0.9*r),2.0);
end
if reject
   xout = x;
   yout = y;
end
hout = r*h;
yout=yout*diag([1,r,r^2]);
```

272   Interpolation

To obtain an approximation solution for a specific value of x, it is possible
to shorten the final step, if necessary, to complete the step exactly at the
right place. However, it is usually more convenient to rely on a stepsize
control mechanism that is independent of output requirements, and to produce
required output results by interpolation, as the opportunity arises. The use
of interpolation makes it also possible to produce output at multiple and
arbitrary points. For the third order method introduced in Subsection 270, a
suitable interpolation scheme is based on the third order Hermite interpolation
formula using both solution and derivative data at the beginning and end of
each step. It is usually considered to be an advantage for the interpolated
solution to have a reasonably high order of continuity at the step points and
the use of third order Hermite will give first order continuity. We will write
the interpolation formula in the form

$$y(x_{n-1} + ht) \approx (1 + 2t)(1 - t)^2 y(x_{n-1}) + (3 - 2t)t^2 y(x_n)$$
$$+ t(1 - t)^2 hy'(x_{n-1}) - t^2(1 - t)hy'(x_n).$$
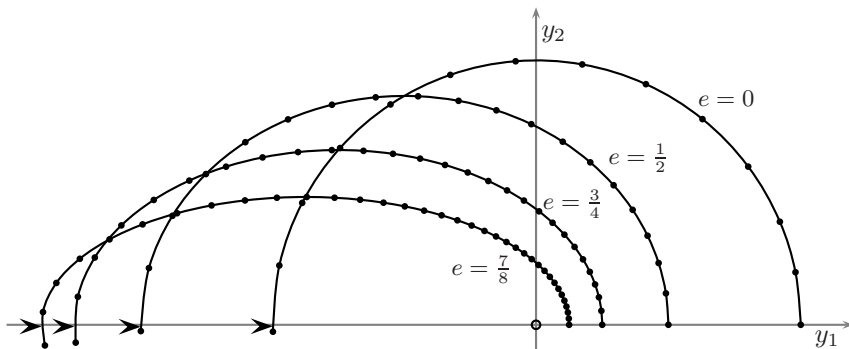
Figure 273(i)   Third order ARK method computations for the Kepler problem

273   Experiments with the Kepler problem

To see how well the numerical method discussed in this section works in practice, it has been applied to the Kepler problem introduced in Subsection 101. For each of the eccentricity values chosen, denoted by e, the problem has been scaled to an initial value

$$y(0) = \begin{bmatrix} 1 - e & 0 & 0 & \sqrt{(1+e)/(1-e)} \end{bmatrix}^{\mathsf{T}},$$

so that the period will be 2π. The aim is to approximate the solution at $x = \pi$ for which the exact result is

$$y(\pi) = \begin{bmatrix} -1 - e & 0 & 0 & -\sqrt{(1-e)/(1+e)} \end{bmatrix}^{\mathsf{T}}.$$

In the first experiment, the problem was solved for a range of eccentricities $e = 0, \frac{1}{2}, \frac{3}{4}, \frac{7}{8}$ with a tolerance of $T = 10^{-4}$. The results are shown in Figure 273(i) with all step points marked. The computed result for $x = \pi$ cannot be found from the variable stepsize schemes unless interpolation is carried out or the final step is forced to arrive exactly at the right value of x. There was no discernible difference between these two half-period approximations, and their common values are indicated on the results.

   The second experiment performed with this problem is to investigate the dependence on the accuracy actually achieved, as the tolerance is varied. The results achieved are almost identical for each of the eccentricities considered and the results will be reported only for $e = \frac{7}{8}$. Before reporting the outcome of this experiment, we might ask what might be expected. If we really were controlling locally committed errors, the stepsize would, approximately, be proportional to $T^{1/(p+1)}$; however, the contribution to global error, of errors

Table 273(I) Global error and numbers of steps for varying tolerance with the Kepler problem

T	$\|\text{Error}\|$	Ratio	Steps
8^0	4.84285		7
		3.94773	
8^{-1}	1.22674		8
		3.71289	
8^{-2}	3.30401×10^{-1}		8
		3.98876	
8^{-3}	8.28328×10^{-2}		10
		3.54007	
8^{-4}	2.33986×10^{-2}		13
		4.72504	
8^{-5}	4.95205×10^{-3}		19
		4.73180	
8^{-6}	1.04655×10^{-3}		30
		4.65786	
8^{-7}	2.24684×10^{-4}		50
		4.58854	
8^{-8}	4.89663×10^{-5}		82
		4.78350	
8^{-9}	1.02365×10^{-5}		137
		4.75845	
8^{-10}	2.15123×10^{-6}		228
		4.74429	
8^{-11}	4.53436×10^{-7}		382
		4.73529	
8^{-12}	9.57567×10^{-8}		642
		4.76011	
8^{-13}	2.01165×10^{-8}		1078
		4.75737	
8^{-14}	4.22848×10^{-9}		1810

committed within each small time interval, is proportional to h^p. Hence we should expect that, for very small tolerances, the total error will be proportional to $T^{p/(p+1)}$. But the controller we are using for the ARK method is not based on an asymptotically correct error estimate, and this will alter the outcome.

In fact the results given in Table 273(I), for this third order method, do show an approximately two-thirds power behaviour. We see this by looking at the ratios of successive norm errors as T is reduced by a factor of 8. Also included in the table is the number of steps. As T becomes small, the number of steps should approximately double each time T is decreased by a factor $\frac{1}{8}$.

274   *Experiments with a discontinuous problem*

The stepsize control mechanism, coded into Algorithm 271α, contains upper and lower bounds on the stepsize ratios. The choice of these bounds acquires crucial importance when low order discontinuities arise in the solution. When a step straddles a point at which there is a sudden change in one of the low order derivatives, this will be recognized by the solver as a massive error estimate, unless the stepsize is abnormally short. Consider, for example, the two-dimensional problem
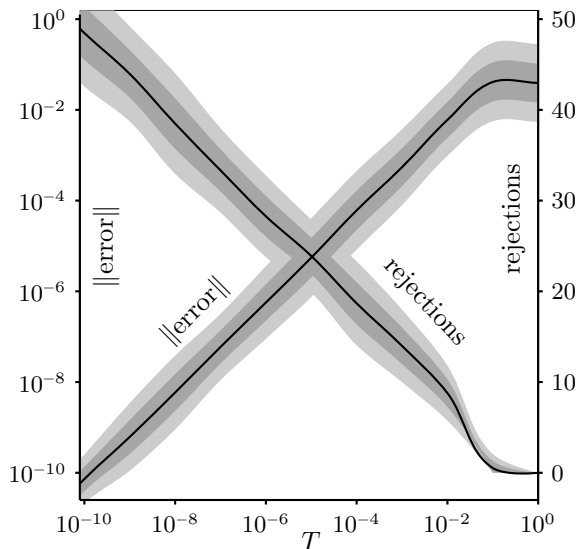
Figure 274(i)   Errors and number of rejections for (274a)

$$y'(x) = \begin{cases} \begin{bmatrix} 0 \\ 1 \end{bmatrix}, & y_1 > y_2, \\ \begin{bmatrix} -1 \\ 0 \end{bmatrix}, & y_2 \geq y_1, \end{cases} \qquad y(0) = \begin{bmatrix} 1 \\ \pi/6 \end{bmatrix}. \qquad (274a)$$

The solution to this problem is very simple: $y(x) = [1, x + \pi/6]^\mathsf{T}$ for $x < 1 - \pi/6$ and $y(x) = [2 - \pi/6 - x, 1]^\mathsf{T}$ for $x > 1 - \pi/6$. Because we are interested in how well the method deals with discontinuous behaviour, we will not take into account our knowledge of where this point is located. What should we expect to happen? We would expect the first step, which jumps over $x = 1 - \pi/6$, to fail and for the stepsize to be reduced as much as the stepsize controller permits. There will then be a sequence of successes (followed by step increases), or failures (followed by step decreases). This sequence will terminate only when the stepsize is small enough for the quantity used as the error estimate to be less than T. Numerical results for this problem using Algorithm 271α are presented in Figure 274(i).

These show the dependence on the accuracy achieved, measured in terms of the error in the component of y_2 after the trajectory has turned the corner at $y = [1, 1]^\mathsf{T}$, together with the number of steps rejected in the whole process of locating the discontinuity in y' and getting past it.

The results will be sensitive to the initial stepsize and, to guarantee we have represented typical and representative behaviour, a large number of initial stepsizes were used with each tolerance. For both the error calculations and the rejected step totals, the results indicate mean values over this range of initial h with shading showing the mean values plus or minus the standard deviation and plus or minus twice the standard deviations. The results suggest that, for this and similar problems, we should expect the error to have a similar magnitude to the tolerance and the number of rejections to be proportional to the logarithm of the tolerance.

Exercises 27

27.1 By computing the scaled derivative of the output from the classical fourth order Runge–Kutta method RK41 (235i), within the current step, rather than from the first stage of the following step, show that the method becomes the general linear method

$$
\left[
\begin{array}{cccc|cc}
0 & 0 & 0 & 0 & 1 & \frac{1}{2} \\
\frac{1}{2} & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{6} & 0 & 1 & \frac{1}{6} \\
\hline
\frac{1}{3} & \frac{1}{3} & \frac{1}{6} & 0 & 1 & \frac{1}{6} \\
0 & 0 & 0 & 1 & 0 & 0
\end{array}
\right].
$$

27.2 Write a fourth order method, with stepsize control, based on the method in Exercise 27.1 which is equivalent to two steps of RK41, each with stepsize h, combined with a single step from the same input, with stepsize $2h$. Use the difference between the two-step result and the double-step result as an error estimator.

27.3 Denote the starting point in Exercise 27.2 as x_{-1} so that the results are computed at $x_0 = x_{-1} + h$ and $x_1 = x_0 + h$. Find a suitable interpolator for this method based on approximations to $y(x_{-1})$, $hy'(x_{-1})$, $y(x_0)$, $y(x_1)$, $hy'(x_1)$ to yield an approximation to $y(x_0 + ht)$, for $t \in [-1, 1]$. Add this interpolator to the variable step method discussed in Exercise 27.2.

Chapter 3

Runge–Kutta Methods

30   Preliminaries

300   Rooted trees

We saw in Section 23 that the graphs known as 'rooted trees' play a central role in the analysis of the accuracy of Runge–Kutta methods. We regard a rooted tree as a pair (V, E), where V is a finite set of 'vertices' and E a set of 'edges'. The edges consist of ordered pairs of members of V, subject to certain conditions. The first condition is that every member of V, except one element known as the 'root', occurs exactly once amongst the second member in each pair in E. The special root vertex does not occur as the second member of any pair. For the final condition, for (V, E) to be a rooted tree, there are two alternatives, which are known to be equivalent: the first is that the graph defined by (V, E) is connected; and the second is that (V, E) defines a partial ordering.

It will be convenient, throughout this discussion, to refer to members of V which do not occur as the first member of any pair in V. For a given edge $[x, y] \in E$, x will be referred to as the 'parent' of y and y will be referred to as a 'child' of x. Thus, a vertex may have one or more children but, if it has none, it is a leaf. Similarly every vertex, except the root, has exactly one parent, whereas the root has no parent.

We do not pursue the formal properties of graphs, and of rooted trees in particular, because they are formulated in specialist books on this subject and are easily appreciated through examples and diagrams. In diagrammatic depictions of a directed graph, the vertices are represented as points and the edges by arrowed line segments joining pairs of points, with the arrow pointing from the first to second member of the pair. We illustrate these ideas in Figure 300(i), where a number of rooted trees are shown. In contrast, Figure 300(ii) shows some graphs which are *not* rooted trees. In these figures, the members of V are chosen to be positive integers. Wherever possible, the diagrams are arranged so that the root, if it exists, is at the bottom of the picture and so that all arrows are pointing in a direction with an upwards component.

Even though we are representing rooted trees using points, labelled by

V	$\{1,2,3,4\}$	$\{1,2,3,4\}$	$\{1,2,3,4,5\}$
E	$\{[1,2],[1,3],[1,4]\}$	$\{[1,2],[1,3],[3,4]\}$	$\{[1,2],[1,3],[3,4],[3,5]\}$
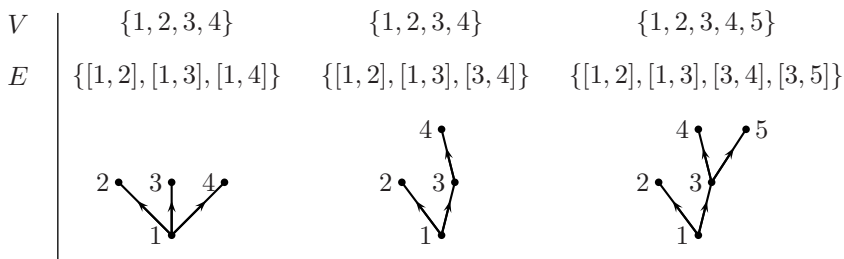
Figure 300(i)   Some directed graphs which are rooted trees

V	$\{1,2,3,4\}$	$\{1,2,3,4\}$	$\{1,2,3,4,5\}$
E	$\{[1,2],[1,3],[4,2],[4,3]\}$	$\{[1,3],[3,4]\}$	$\{[1,2],[1,3],[1,4],[1,5],[3,4]\}$
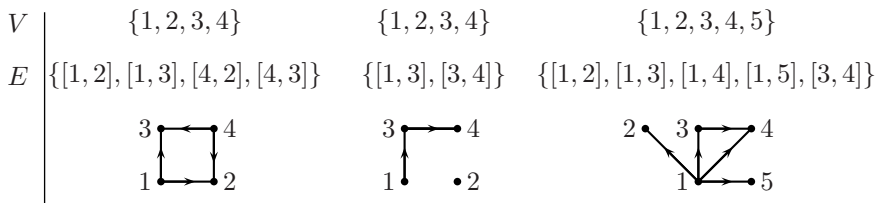
Figure 300(ii)   Some directed graphs which are not rooted trees

members of a vertex set, we are interested in the abstract structure behind this definition. That is, if (V,E) and (V',E') are rooted trees and there exists a bijection $\varphi : V \to V'$ such that $[x,y] \in E$ if and only if $[\varphi(x),\varphi(y)] \in E'$, then the two rooted trees are identical, when represented as diagrams, except for the labels attached to the points. We can thus regard an 'abstract rooted tree' as an equivalence class under this type of isomorphism. We use each interpretation from time to time, according to our convenience; where it is not clear from the context which is intended, we add some words of clarification. For a labelled tree $\mathbf{t}$, the corresponding abstract tree will be denoted by $|\mathbf{t}|$.

To conclude this introduction to rooted trees, we present two alternative notations for trees. In each notation, we denote the single tree, with only one vertex, by the symbol τ. In the first notation, we consider a tree t such that, when the root is removed, there remain a number of disconnected trees, say t_1, t_2, ..., t_m, where m is the number of 'children' of the root of t. We then write $t = [t_1 t_2 \cdots t_m]$. This gives a recursion for constructing a symbolic denotation for any particular tree. When some of t_1, t_2, ..., t_m are equal to each other, it will be convenient to represent these repetitions using a power notation. For example, $[t_1 t_1 t_2 t_2 t_2 t_3]$ will also be written as $[t_1^2 t_2^3 t_3]$.

The second notation builds up a symbolic representation of all trees by using a non-associative product of rooted trees, such that $t_1 t_2$ is formed by joining them at the roots, with an additional edge from the root v_1 of t_1 to

Table 300(I)   Trees, notations for trees and various functions on trees

$r(t)$	t			$\sigma(t)$	$\gamma(t)$
1	•	τ	τ	1	1
2		$[\tau]$	$\tau\tau$	1	2
3		$[\tau^2]$	$\tau\tau.\tau$	2	3
3		$[[\tau]]$	$\tau.\tau\tau$	1	6
4		$[\tau^3]$	$(\tau\tau.\tau)\tau$	6	4
4		$[\tau[\tau]]$	$\tau\tau.\tau\tau = (\tau.\tau\tau)\tau$	1	8
4		$[[\tau^2]]$	$\tau(\tau\tau.\tau)$	2	12
4		$[[[\tau]]]$	$\tau(\tau.\tau\tau)$	1	24
5		$[\tau^4]$	$(\tau\tau.\tau)\tau.\tau$	24	5
5		$[\tau^2[\tau]]$	$(\tau.\tau\tau)\tau.\tau = (\tau\tau.\tau\tau)\tau = (\tau\tau.\tau).\tau\tau$	2	10
5		$[\tau[\tau^2]]$	$\tau\tau.(\tau\tau.\tau) = \tau(\tau\tau.\tau).\tau$	2	15
5		$[\tau[[\tau]]]$	$\tau(\tau.\tau\tau).\tau = \tau\tau.(\tau.\tau\tau)$	1	30
5		$[[\tau]^2]$	$(\tau.\tau\tau).\tau\tau$	2	20
5		$[[\tau^3]]$	$\tau.(\tau\tau.\tau)\tau$	6	20
5		$[[\tau[\tau]]]$	$\tau(\tau\tau.\tau\tau) = \tau.(\tau.\tau\tau)\tau$	1	40
5		$[[[\tau^2]]]$	$\tau.\tau(\tau\tau.\tau)$	2	60
5		$[[[[\tau]]]]$	$\tau.\tau(\tau.\tau\tau)$	1	120

the root v_2 of t_2. Thus if $t_1 = |(V_1, E_1)|$ and $t_2 = |(V_2, E_2)|$, and V_1 and V_2 are disjoint sets, then $t_1 t_2$ is the tree $|(V_1 \cup V_2, E_1 \cup E_2 \cup [v_1, v_2])|$. Because the product is not associative, we need to distinguish between $(t_1 t_2)t_3$ and $t_1(t_2 t_3)$ without introducing more parentheses than necessary. Hence, we sometimes write $(t_1 t_2)t_3 = t_1 t_2.t_3$ and $t_1(t_2 t_3) = t_1.t_2 t_3$.

We illustrate these notations in Table 300(I), where all trees with up to five vertices are shown. Also shown are the functions $r(t)$, $\sigma(t)$ and $\gamma(t)$ to be introduced in the next subsection.

301   Functions on trees

For a rooted tree t, define $r(t)$, the 'order' of t, as the number of vertices in t. That is, if t is labelled as (V, E), then $r(t) = \#V$, the cardinality of the set V. Let $A(t)$ denote the group of automorphisms on a particular labelling

of t. That is, $A(t)$ is the set of mappings $\varphi : V \to V$ such that $[x, y] \in E$ if and only if $[\varphi(x), \varphi(y)] \in E$. The group $A(t)$ will be known as the 'symmetry group' of t; its order will be known as the 'symmetry', and denoted by $\sigma(t)$. The 'density' of t, $\gamma(t)$, is defined as the product over all vertices of the order of the subtree rooted at that vertex. We illustrate these definitions using a specific tree (V, E) with nine vertices given by

$$V = \{a, b, c, d, e, f, g, h, i\},$$
$$E = \{[a, b], [a, c], [b, d], [b, e], [b, f], [c, g], [c, h], [c, i]\}.$$

The diagram representing this tree, with the vertex labels attached, is
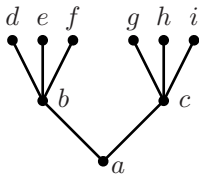
The value of $r(t)$ is, of course, 9. The symmetry group is the set of permutations generated by all members of the symmetric group on $\{d, e, f\}$, by all members of the symmetric group on $\{g, h, i\}$, and the group S_2 generated by the single permutation, in which b and c are interchanged, d and g are interchanged, e and h are interchanged, and f and i are interchanged. Thus the order of the symmetry group is $\sigma(t) = 3!3!2! = 72$. To calculate $\gamma(t)$, attach integers to the vertices as follows:
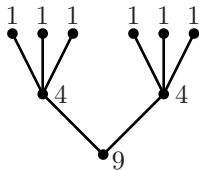
leading to $\gamma(t) = 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 4 \cdot 4 \cdot 9 = 144$.

We complete this subsection with a formal statement of recursions for the computation of $r(t)$, $\sigma(t)$ and $\gamma(t)$.

Theorem 301A *Let $t = [t_1^{m_1} t_2^{m_2} \cdots t_k^{m_k}]$, where $t_1, t_2, \ldots, t_k$ are distinct trees. Then*

$$r(t) = 1 + \sum_{i=1}^{k} m_i r(t_i), \tag{301a}$$

$$\sigma(t) = \prod_{i=1}^{k} m_i! \sigma(t_i)^{m_i}, \tag{301b}$$

$$\gamma(t) = r(t) \prod_{i=1}^{k} \gamma(t_i)^{m_i}. \tag{301c}$$

Furthermore,

$$r(\tau) = \sigma(\tau) = \gamma(\tau) = 1. \tag{301d}$$

Proof. To verify (301d), calculate r, σ and γ for the single tree with one vertex. To prove (301a), add the numbers of vertices in the $m_1 + m_2 + \cdots + m_k$ trees attached to the new root, and add one extra for the new root. In the calculation of $\gamma(t)$, the integers attached to the vertices in the $m_1 + m_2 + \cdots + m_k$ trees joined to the new root are the same as in the constituent trees themselves. The product of these integers, and the integer $r(t)$, gives the result (301c). Finally, (301b) follows by noting that the permutations which leave the vertex pairs, making up the list of edges, are just as in the individual attached trees, together with the additional permutations of the label sets amongst identical subtrees. $\square$

302   *Some combinatorial questions*

We consider the question of labelling a tree t with $r(t)$ vertices, using the symbols $\{1, 2, \ldots, r(t)\}$, under the following conditions:

   (i)   Each vertex receives one and only one label.
   (ii)   Labellings that are equivalent under the symmetry group are counted only once.
   (iii) If (i, j) is a labelled edge then $i < j$.

The number of distinct ways of labelling the given tree t will be denoted by $\alpha(t)$. A similar question, in which conditions (i) and (ii) apply, but (iii) does not, leads to a function $\beta(t)$. We have:

Theorem 302A

$$\alpha(t) = \frac{r(t)!}{\sigma(t)\gamma(t)}, \tag{302a}$$

$$\beta(t) = \frac{r(t)!}{\sigma(t)}. \tag{302b}$$

Proof. The value of $\beta(t)$ is found by labelling the vertices of t with all permutations and then dividing by $\sigma(t)$ so as to count, only once, sets of labellings which are equivalent under symmetry. In the case of $\alpha(t)$, we are restricted by the requirement that, of the labels assigned to any vertex v and to its descendants, only the lowest may be assigned to v. The product of the factors that must be divided out to satisfy this constraint is $\gamma(t)$. $\square$

We now look at the enumeration question of the number of rooted trees of various orders.

Theorem 302B *Let θ_k, $k = 1, 2, 3, \ldots$ denote the number of rooted trees with exactly k vertices. Then,*

$$\theta_1 + \theta_2 x + \theta_3 x^2 + \cdots = (1 - x)^{-\theta_1} (1 - x^2)^{-\theta_2} (1 - x^3)^{-\theta_3} \cdots . \qquad (302c)$$

Before proving this result, we consider how (302c) is to be interpreted. The right-hand side can be formally expanded as a power series, and it can be seen that the coefficient of x^k depends only on $\theta_1, \theta_2, \ldots, \theta_k$ (and is independent of any of $\theta_1, \theta_2, \ldots$ if $k = 0$). Equate this to the coefficient of x^k on the left-hand side and the result is a formula for θ_{k+1} in terms of previous members of the θ sequence. In particular, $k = 0$ gives $\theta_1 = 1$. We now turn to the justification of the result.

Proof. Let $\Theta_k(U)$ denote the number of trees of order k that can be formed using the operation $(t_1, t_2, \ldots, t_n) \mapsto [t_1, t_2, \ldots, t_n]$, where t_1, t_2, $\ldots$, t_n are all members of U which is assumed to be a subset of T. In particular, $\Theta_k(T)$ is identical to θ_k. Let V denote the set $U \cup \{\hat{t}\}$, where $\hat{t} \notin U$. Every tree of the form $[\hat{t}^m, \ldots]$, with order k, is included in a set with $\Theta_k(V) - \Theta_k(U)$ members. However, there are the same number of members of this set as there are trees of order $k - r(\hat{t})$ of the form $[\hat{t}^{m-1}, \ldots]$. Thus, $\Theta_k(V) - \Theta_k(U) = \Theta_{k-r(\hat{t})}(V)$, which is equivalent to

$$\Theta_1(U) + \Theta_2(U)x + \cdots = (1 - x^{r(\hat{t})})(\Theta_1(V) + \Theta_2(V)x + \cdots). \qquad (302d)$$

Since

$$\Theta_1(U) + \Theta_2(U)x + \cdots = 1,$$

when U is the empty set, we can successively compute the value of this expression when $U = \{t_1, t_2, \ldots, t_n\}$ using (302d) as

$$\Theta_1(U) + \Theta_2(U)x + \cdots = \prod_{k=1}^{n} (1 - x^{r(t_k)})^{-1}. \qquad (302e)$$

Now assume that t_1, t_2, $\ldots$ consist of all trees of orders up to some integer p, and we can write (302e) as

$$\Theta_1(U) + \Theta_2(U)x + \cdots = \prod_{k=1}^{p} (1 - x^k)^{-\theta_k}.$$

Since $\Theta_i(U) = \theta_i$ if $i \le p + 1$, we obtain the result by replacing $\prod_{k=1}^{p}$ by $\prod_{k=1}^{\infty}$. $\qquad \square$

The values of θ_k, computed using Theorem 302B, are shown in Table 302(I) up to order 10. Also shown are the total numbers of trees up to a given order, and two further functions equal to the totals of the $\alpha(t)$ and $\beta(t)$ values for each order.

Table 302(I)   Various enumerations of rooted trees up to order 10

n	θ_n	$\sum_{i=1}^{n} \theta_i$	$\sum_{r(t)=n} \alpha(t)$	$\sum_{r(t)=n} \beta(t)$
1	1	1	1	1
2	1	2	1	2
3	2	4	2	9
4	4	8	6	64
5	9	17	24	625
6	20	37	120	7776
7	48	85	720	117649
8	115	200	5040	2097152
9	286	486	40320	43046721
10	719	1205	362880	1000000000

The entries in last two columns of Table 302(I) are important in classical combinatorics, although their roles in our work is only incidental. The sum of the $\beta(t)$ for $r(t) = n$ is the number of fully labelled rooted trees with n vertices, whereas the corresponding sum for $\alpha(t)$ is the number of monotonically labelled rooted trees. It is easy to guess a formula for each of these totals, and we now verify these.

Theorem 302C *Let* $A_n = \sum_{r(t)=n} \alpha(t)$, $B_n = \sum_{r(t)=n} \beta(t)$. *Then*

$$A_n = (n-1)!, \quad B_n = n^{n-1}.$$

Proof. Let X_n denote the set of vectors of the form $[x_1, x_2, \ldots, x_{n-1}]$ and Y_n the set of vectors of the form $[y_1, y_2, \ldots, y_{n-1}]$, where $x_i \in \{1, 2, \ldots, i\}$ and $y_i \in \{1, 2, \ldots, n\}$, for $i = 1, 2, \ldots, n$. It is easy to see that the cardinalities of these sets are $\#X_n = (n-1)!$, $\#Y_n = n^{n-1}$. We conclude the proof by showing how to define bijections between the monotonically labelled rooted trees of order n and X_n and between the fully labelled rooted trees of order n and Y_n. In each case, given a labelled rooted tree, let v denote the leaf with greatest label and assign, as the value of x_{n-1} or y_{n-1}, respectively, the label attached to the parent of v. Delete the leaf v and continue the process until only the root remains. That is, in step $i = 1, 2, \ldots, n-1$, we work with a tree with $n+1-i$ vertices. We assign to x_{n-i} (or to y_{n-i}, respectively) the label attached to the parent of the leaf with the highest remaining label, and then delete this leaf to yield a tree with $n - i$ vertices. □

Although we have not included details of the bijections involved in this summarized proof, we illustrate these in the cases $n = 4$, for monotonically labelled trees in Table 302(II), and $n = 3$, for fully labelled trees in Table 302(III).

Table 302(II)   The bijection relating a monotonically labelled fourth order tree t and $x \in X_4$

x	t	x	t	x	t
$[1,1,1]$	(tree: 3, 4 over 1, with 2)	$[1,1,2]$	(tree: 4, 3, 2 over 1)	$[1,1,3]$	(tree: 4, 3 over 1, with 2)
$[1,2,1]$	(tree: 3 over 2 over 1, with 4)	$[1,2,2]$	(tree: 3, 4 over 2 over 1)	$[1,2,3]$	(tree: 4-3-2-1 chain)

Table 302(III)   The bijection relating a fully labelled third order tree t and $y \in Y_3$

y	t	y	t	y	t
$[1,1]$	(tree: 2, 3 over 1)	$[1,2]$	(tree: 3 over 2 over 1)	$[1,3]$	(tree: 2 over 3 over 1)
$[2,1]$	(tree: 3 over 1 over 2)	$[2,2]$	(tree: 1, 3 over 2)	$[2,3]$	(tree: 1 over 3 over 2)
$[3,1]$	(tree: 2 over 1 over 3)	$[3,2]$	(tree: 1 over 2 over 3)	$[3,3]$	(tree: 1, 2 over 3)

303   The use of labelled trees

We have seen that $\alpha(t)$, introduced in Subsection 302, is the number of distinct ways of labelling the vertices of t with the integers $\{1, 2, \ldots, r(t)\}$, on condition that for each edge (i,j), $i < j$. It is convenient to generalize this by writing S for an finite ordered set such that the cardinality is $\#S = r(t)$, and counting trees labelled with members of S such that $i < j$ for each edge (i,j). Let T_S^* denote the set of trees labelled in this way and let $|\mathbf{t}|$ denote the member of T corresponding to $\mathbf{t} \in T_S^*$, but with the vertex labels removed. This means that $\alpha(t)$ can be interpreted as the number of members of T_S^* such that $|\cdot|$ maps them to $t \in T$. Similarly, we write T_S for the set of trees labelled by a set with cardinality $r(t)$, where no assumption is made about order. In this case $\beta(t)$ is the number of $\mathbf{t} \in T_S$, such that $|\mathbf{t}| = t$.

304   Enumerating non-rooted trees

Recall the generating function for the numbers of rooted trees of various orders

$$\theta(x) = \theta_1 + \theta_2 x + \theta_3 x^2 + \cdots,$$
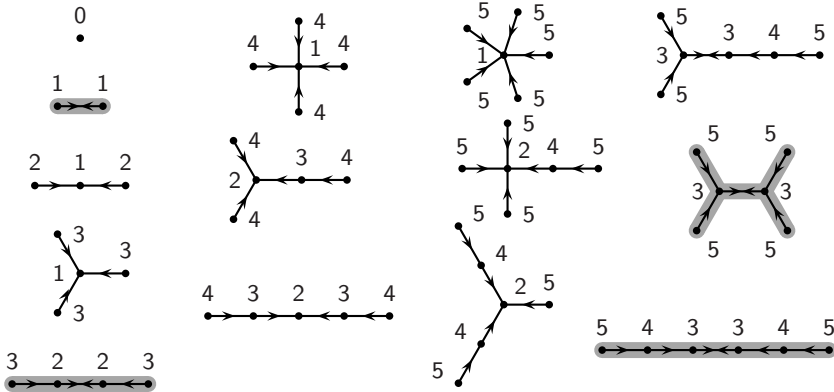
Figure 304(i)   Trees with up to six vertices

where θ_1, θ_2, ... are given in (302c). Also write

$$\phi(x) = \phi_1 + \phi_2 x + \phi_3 x^2 + \cdots,$$
$$\psi(x) = \psi_1 + \psi_2 x + \psi_3 x^2 + \cdots,$$

as the generating functions for the numbers of trees ϕ_i of orders $i = 1, 2, \ldots$ and the numbers of non-superfluous trees ψ_i. The meaning of 'superfluous tree' will become clear from the discussion which follows.

Given a tree, we can form a family of rooted trees by designating one of its vertices to be the root. We will refer to two such rooted trees as adjacent if the two roots are at the ends of the same edge in the underlying tree. For any particular vertex v, let $t = [t_1, t_2, \ldots, t_m]$ and write $\phi(v) = \max_{i=1}^{m} r(t_i)$. There will be at most one vertex adjacent to v for which the value of ϕ is lower. However, for some trees with even order there will be two adjacent vertices for which the values of ϕ are each equal to $r(t)/2$. The 14 trees with up to six vertices are shown in Figure 304(i). The value of ϕ is attached to each vertex, with arrows showing the direction of decreasing ϕ. In the cases of two adjacent vertices v and w with $\phi(v) = \phi(w)$, two arrows are shown meeting midway through the edge.

For a rooted tree formed from a tree by selecting a vertex as the root, we can move along an arrow to obtain a vertex with a lower value of ϕ. Thus we should subtract from the total number of rooted trees of a given order n, the number of pairs or trees with unequal orders. This means subtracting the number of rooted trees of the form tu, where $r(t) < r(u)$. In the case of trees where $n = 2m$ is even, and for two adjacent vertices, the rooted trees tu and ut occur, where $r(t) = r(u)$, we need to subtract half the number of such trees unless $t = u$, in which case no subtraction is performed.

For a tree of order $n = 2m + 1$, the number of trees will thus be $\theta_n - \sum_{i=1}^{m} \theta_i \theta_{n-i}$, which is the coefficient of x^{n-1} in

$$\theta(x) - \frac{x}{2} \Big(\theta(x)^2 \mp \theta(x^2) \Big), \tag{304a}$$

where the term involving $\theta(x^2)$ does not actually contribute to this case of odd n. In the case of even $n = 2m$, the number of trees will be

$$\theta_n - \sum_{i=1}^{m-1} \theta_i \theta_{n-i} - \frac{1}{2} \theta_m (\theta_m \mp 1),$$

where $\mp$ is interpreted as $-$, and this is again equal to the coefficient of x^{n-1} in (304a).

Counting non-superfluous trees is the same except that we need to subtract from the totals the number of trees of the form tt, and this gives the same result as (304a) but with $\mp$ replaced by $+$. Putting these results together we formally state:

Theorem 304A *The generating functions for trees and non-superfluous trees are*

$$\phi(x) = \theta(x) - \frac{x}{2} \Big(\theta(x)^2 - \theta(x^2) \Big), \tag{304b}$$

$$\psi(x) = \theta(x) - \frac{x}{2} \Big(\theta(x)^2 + \theta(x^2) \Big). \tag{304c}$$

305   Differentiation

We need to develop fairly intricate formulae involving derivatives of vector-valued functions of vector arguments. Hence, in this subsection and the next, we review basic calculus ideas in a vector setting. We start with the elementary notions of the derivative of a real-valued function of a single real variable, and the partial derivatives of a real-valued function of several real variables. A real-valued function f, whose domain contains an open interval around the real number a, is differentiable at a if there exists a number $f'(a)$, referred to as the derivative of f at a, such that $|f(a + \delta) - f(a) - f'(a)\delta|/|\delta| \to 0$ as $|\delta| \to 0$. This definition is extended in two ways. First, f can take values in $\mathbb{R}^N$, in which case f is differentiable if each of its components is differentiable. Furthermore, $f'(a) \in \mathbb{R}^N$ is equal to the vector made up from the derivatives of the components of f. Another way of writing this is

$$\frac{\|f(a + \delta) - f(a) - f'(a)\delta\|}{|\delta|} \to 0 \quad \text{as } |\delta| \to 0.$$

When the domain of f is generalized to $X \subset \mathbb{R}^M$, such that $a \in O \subset X$, where O is an open set, such as a product of open intervals, then $f'(a)$, if it

exists, is a linear operator, $f'(a) : \mathbb{R}^M \to \mathbb{R}^N$, such that

$$\frac{\|f(a + \delta) - f(a) - f'(a)\delta\|}{\|\delta\|} \to 0 \quad \text{as } \|\delta\| \to 0.$$

If the components of a and f are written as

$$a = \begin{bmatrix} a^1 \\ a^2 \\ \vdots \\ a^M \end{bmatrix}, \qquad f = \begin{bmatrix} f^1 \\ f^2 \\ \vdots \\ f^N \end{bmatrix},$$

then the linear operator $f'(a)$ is represented by the matrix of partial derivatives

$$f'(a) = \begin{bmatrix} f^1_1(a) & f^1_2(a) & \cdots & f^1_M(a) \\ f^2_1(a) & f^2_2(a) & \cdots & f^2_M(a) \\ \vdots & \vdots & & \vdots \\ f^N_1(a) & f^N_2(a) & \cdots & f^N_M(a) \end{bmatrix} = \begin{bmatrix} \frac{\partial f^1}{\partial a^1} & \frac{\partial f^1}{\partial a^2} & \cdots & \frac{\partial f^1}{\partial a^M} \\ \frac{\partial f^2}{\partial a^1} & \frac{\partial f^2}{\partial a^2} & \cdots & \frac{\partial f^2}{\partial a^M} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f^N}{\partial a^1} & \frac{\partial f^N}{\partial a^2} & \cdots & \frac{\partial f^N}{\partial a^M} \end{bmatrix}.$$

Second and higher derivatives are bilinear and multilinear operators. In the tensor representation

$$f^i_{j_1 j_2 \cdots j_k}(a) = \frac{\partial^k f^i}{\partial a_{j_1} \partial a_{j_2} \cdots \partial a_{j_k}},$$

the argument (a) is omitted, for convenience, if its value is understood.

Finally, in this subsection, we remark that evaluation of the result of operating with the order k derivative $f^i_{j_1 j_2 \cdots j_k}$, as a k-linear operator, on the collection of k arguments $v_1, v_2, \ldots, v_k \in \mathbb{R}^M$, gives the expression

$$\sum_{j_1=1}^{M} \sum_{j_2=1}^{M} \cdots \sum_{j_k=1}^{M} f^i_{j_1 j_2 \cdots j_k} v_1^{j_1} v_2^{j_2} \cdots v_k^{j_k}. \tag{305a}$$

The complicated appearance of (305a) can be alleviated by omitting all the summation symbols and regarding them as implied. This is the well-known 'summation convention', and we use this notational simplification freely throughout this book. Thus we write, instead of (305a),

$$f^i_{j_1 j_2 \cdots j_k} v_1^{j_1} v_2^{j_2} \cdots v_k^{j_k}.$$

Table 306(I)   Members of $\mathcal{I}_2$ and their symmetries

I	( )	(1)	(2)	(1,1)	(1,2)	(2,2)	(1,1,1)	(1,1,2)	(1,2,2)	(2,2,2)
$\sigma(I)$	1	1	1	2	1	2	6	2	2	6

306   Taylor's theorem

We start from the identity,

$$f(a + \delta) = f(a) + f'(a)(\delta) + \tfrac{1}{2!}f''(a)(\delta, \delta)$$

$$+ \cdots + \frac{1}{n!}f^{(n)}(a)(\delta, \delta, \dots, \delta) + R_n, \quad (306\text{a})$$

where the 'remainder' R_n is given by

$$R_n = \frac{1}{n!}\int_0^1 f^{(n+1)}(a + \xi\delta)(\delta, \delta, \dots, \delta)(1 - \xi)^n d\xi.$$

This is proved by induction, with the key step being

$$R_{n-1} = \frac{1}{n!}f^{(n)}(a)(\delta, \delta, \dots, \delta) + R_n,$$

which is verified by integration by parts. With Taylor's theorem written in the form (306a), the result is quite versatile and applies if $f : X \subset \mathbb{R}^M \to \mathbb{R}^N$, where $a + \xi\delta \in O \subset X$, for all $\xi \in [0,1]$. Assuming that $\|f^{(n+1)}(x)\|$ exists and is bounded for $x \in O$, then

$$\|R_n\| = O(\|\delta\|^{n+1}).$$

We consider a slight variation of the theorem, in which δ is replaced by the sum of a finite number of vectors, δ_i, $i = 1, 2, \dots, m$. We assume that f is analytic in a neighbourhood of a and that each of the δ_i is *small*. The formal result we present can be interpreted as a finite series, together with remainder, with the details dependent on the relative magnitudes of the δ_i. Let I denote a sequence of integers from the set $\{1, 2, \dots, m\}$ and $\mathcal{I}_m$ the set of all such sequences. Two sequences I and I' will be regarded as identical if the members of I' are a permutation of the members of I. The 'symmetry' of I is the order of the group of permutations of the elements of $\{1, 2, \dots, \#I\}$, which maps the ordered members of I to themselves. That is, if I contains k_i occurrences of i, for each $i = 1, 2, \dots, m$, then

$$\sigma(I) = k_1!k_2!\cdots k_m!. \quad (306\text{b})$$

For $m = 2$, the first few $I \in \mathcal{I}_m$, together with the corresponding symmetries, are given in Table 306(I).

For $I = (i_1, i_2, \ldots, i_k) \in \mathcal{I}_m$, we denote by δ_I the quantity

$$(\delta_{i_1}, \delta_{i_2}, \ldots, \delta_{i_m}) \in (\mathbb{R}^N)^m.$$

These will be used as operands for multilinear operators, such as $f^{(m)}(a)$, and, in the case $I = ()$, we interpret $f(a)()$ as being simply $f(a)$. We are now in a position to state the form of the Taylor expansion (306a), when δ is replaced by $\sum_{i=1}^m \delta_i$.

Theorem 306A

$$f\left(a + \sum_{i=1}^m \delta_i\right) = \sum_{I \in \mathcal{I}_m} \frac{1}{\sigma(I)} f^{(\#I)}(a)\delta_I.$$

Proof. Continue to write k_i for the number of occurrences of i in I, so that $\sigma(I)$ is given by (306b). The coefficient of $f^{(\#I)}(a)\delta_I$ is equal to the coefficient of $\prod_{i=1}^m x^{k_i}$ in $\exp\left(\sum_{i=1}^m x_i\right)$. This equals the coefficient of $\prod_{i=1}^m x^{k_i}$ in

$$\left(1 + x_1 + \tfrac{1}{2!}x_1^2 + \cdots\right)\left(1 + x_2 + \tfrac{1}{2!}x_2^2 + \cdots\right)\cdots\left(1 + x_m + \tfrac{1}{2!}x_m^2 + \cdots\right)$$

and is equal to $1/\sigma(I)$. $\qquad\qquad\square$

We illustrate this result by applying (306A) to the case $m = 2$, using Table 306(I):

$$\begin{aligned}
f(a + \delta_1 + \delta_2) = {}& f(a) + f'(a)\delta_1 + f'(a)\delta_2 + \tfrac{1}{2}f''(a)(\delta_1, \delta_1) \\
& + f''(a)(\delta_1, \delta_2) + \tfrac{1}{2}f''(a)(\delta_2, \delta_2) + \tfrac{1}{6}f'''(a)(\delta_1, \delta_1, \delta_1) \\
& + \tfrac{1}{2}f'''(a)(\delta_1, \delta_1, \delta_2) + \tfrac{1}{2}f'''(a)(\delta_1, \delta_2, \delta_2) + \tfrac{1}{6}f'''(a)(\delta_2, \delta_2, \delta_2) + \cdots.
\end{aligned}$$

Exercises 30

30.1 Find $r(t)$, $\sigma(t)$, $\gamma(t)$, $\alpha(t)$ and $\beta(t)$ for the tree $t = |\mathbf{t}|$, where $|\mathbf{t}| = (V, E)$, with

$$V = \{a, b, c, d, e, f, g\} \text{ and } E = \{(a, b), (b, c), (b, d), (a, e), (e, f), (e, g)\}.$$

30.2 Find $r(t)$, $\sigma(t)$, $\gamma(t)$, $\alpha(t)$ and $\beta(t)$ for the tree $t = [[\tau]^2 \tau^3]$.

30.3 Find $r(t)$, $\sigma(t)$, $\gamma(t)$, $\alpha(t)$ and $\beta(t)$ for the tree $t = \tau\tau \cdot (\tau\tau \cdot \tau)\tau$.

30.4 Define $f : \mathbb{R}^3 \to \mathbb{R}^3$ by

$$f(y^1, y^2, y^3) = \begin{bmatrix} y^1 + y^2 y^3 \\ (y^1)^2 + 2y^1 y^2 \\ 1 + (y^2 + y^3)^2 \end{bmatrix}.$$

Find formulae for f_j^i, f_{jk}^i and f_{jkl}^i, for $i, j, k, l = 1, 2, 3$.

30.5 Expand $f(a + \delta_1\xi + \delta_2\xi^2 + \delta_3\xi^3)$ up to terms in ξ^3 using Theorem 306A.

31   Order Conditions

310   Elementary differentials

To investigate the error in carrying out a single step of a Runge–Kutta method, we need to compare successive terms in the Taylor expansions of the exact and the computed solutions. These involve expressions whose structures are related to rooted trees. In the case of the exact solution, it is possible to evaluate the Taylor coefficients by repeated differentiation. We start with a differential equation, assumed for convenience to be autonomous,

$$y'(x) = f(y(x)). \tag{310a}$$

We also write (310a) in component-by-component form, with arguments omitted for brevity, as

$$(y^i)' = f^i. \tag{310b}$$

To obtain the second derivative, use the chain rule

$$y''(x) = \frac{d}{dx} f(y(x)) = f'(y(x))y'(x) = f'(y(x))f(y(x)) \tag{310c}$$

or, using (310b) as the starting point,

$$\frac{d}{dx}(y^i)' = \frac{d}{dx} f^i = f^i_j f^j. \tag{310d}$$

Note that in (310d) we have used the summation convention. We continue to use this convention without further comment. The third derivative can be found in a similar manner, but is complicated by the fact that $y(x)$ is present in both factors in $f'(y(x))f(y(x))$. Even though we are omitting arguments, $y(x)$ is also implicitly present in the tensor form $f^i_j f^j$. The two forms of the third derivative are

$$\frac{d^3}{dx^3} y(x) = f''(y(x))\big(f(y(x)), f(y(x))\big) + f'(y(x))(f'(y(x))f(y(x))), \tag{310e}$$

$$\frac{d^3}{dx^3} y^i = f^i_{jk} f^j f^k + f^i_j f^j_k f^k. \tag{310f}$$

We can find a pattern in the terms occurring in the first, second and third derivatives, using rooted trees. In the total derivative form, (310a), (310c), (310e), we relate $f(y(x))$ to a leaf in a tree, we relate $f'(y(x))$ to a vertex with a single outwardly directed edge, and we relate $f''(y(x))$ to a vertex with two outward edges. In the case of f' and f'', the outward edges are joined to subtrees, as representatives of the operands of these linear and bilinear operators, respectively.

For the tensor representations of the terms in the first three derivatives of y^i, we treat the superscripts in f^i, f^j, f^k as members of the vertex set V, and

Table 310(I) Relation between terms in y derivatives and rooted trees

Tree	Operator diagram	Term	Labelled tree	Tensor term
	$\bullet\mathbf{f}$	$f(y(x))$	$\bullet\, i$	f^i
	$\mathbf{f}$ $\mathbf{f}'$	$f'(y(x))f(y(x))$	j i	$f^i_j f^j$
	$\mathbf{f}$ $\mathbf{f}$ $\mathbf{f}''$ $\mathbf{f}$	$f''(y(x))(f(y(x)), f(y(x)))$	j k i k	$f^i_{jk} f^j f^k$
	$\mathbf{f}'$ $\mathbf{f}'$	$f'(y(x))(f'(y(x))f(y(x)))$	j i	$f^i_j f^j_k f^k$

we define the edge set E in terms of the pairs, such as (i, j) that occur in f^i_j, f^i_{jk}.

Thus, we can identify four trees as representatives of the terms that occur in the first, second and third derivatives of y. In Table 310(I) we illustrate this correspondence using both formulations. Note that we write $\mathbf{f}$, $\mathbf{f}'$ and $\mathbf{f}''$ as abbreviations for $f(y(x))$, $f'(y(x))$ and $f''(y(x))$, respectively.

We can expect this pattern to continue, because the operation of differentiating adds an additional vertex to an existing tree, in a number of different ways, and each of these corresponds to a further tree.

Definition 310A *Given a tree t and a function $f : \mathbb{R}^N \to \mathbb{R}^N$, analytic in a neighbourhood of y, the 'elementary differential' $F(t)(y)$ is defined by*

$$F(\tau)(y) = f(y), \tag{310g}$$

$$F([t_1, t_2, \ldots, t_m]) = f^{(m)}(y)(F(t_1)(y), F(t_2)(y), \ldots, F(t_m)(y)). \tag{310h}$$

Note that the tensor interpretation of (310h) is written as

$$F^i([t_1, t_2, \ldots, t_m]) = f^i_{j_1, j_2, \ldots, j_m} F^{j_1}(t_1) F^{j_2}(t_2) \cdots F^{j_m}(t_m).$$

The elementary differentials up to order 5 are shown in Table 310(II). Note that we use the same abbreviation as in Table 310(I), in which $\mathbf{f}$, $\mathbf{f}'$, $\ldots$ denote $f(y(x))$, $f(y(x))'$, $\ldots$. The values of $\alpha(t)$ are also shown; their significance will be explained in the next subsection.

As part of the equipment we need to manipulate expressions involving elementary differentials we consider the value of

$$hf\left(y_0 + \sum_{t \in T} \theta(t) \frac{h^{r(t)}}{\sigma(t)} F(t)(y_0)\right). \tag{310i}$$

Table 310(II) Elementary differentials for orders 1 to 5

$r(t)$	t	$\alpha(t)$	$F(t)(y)$	$F(t)(y)^i$
1		1	$\mathbf{f}$	f^i
2		1	$\mathbf{f}'\mathbf{f}$	$f^i_j f^j$
3		1	$\mathbf{f}''(\mathbf{f},\mathbf{f})$	$f^i_{jk} f^j f^k$
3		1	$\mathbf{f}'\mathbf{f}'\mathbf{f}$	$f^i_j f^j_k f^k$
4		1	$\mathbf{f}'''(\mathbf{f},\mathbf{f},\mathbf{f})$	$f^i_{jkl} f^j f^k f^l$
4		3	$\mathbf{f}''(\mathbf{f},\mathbf{f}'\mathbf{f})$	$f^i_{jk} f^j f^k_l f^l$
4		1	$\mathbf{f}'\mathbf{f}''(\mathbf{f},\mathbf{f})$	$f^i_j f^j_{kl} f^k f^l$
4		1	$\mathbf{f}'\mathbf{f}'\mathbf{f}'\mathbf{f}$	$f^i_j f^j_k f^k_l f^l$
5		1	$\mathbf{f}^{(4)}(\mathbf{f},\mathbf{f},\mathbf{f},\mathbf{f})$	$f^i_{jklm} f^j f^k f^l f^m$
5		6	$\mathbf{f}^{(3)}(\mathbf{f},\mathbf{f},\mathbf{f}'\mathbf{f})$	$f^i_{jkl} f^j f^k f^l_m f^m$
5		4	$\mathbf{f}''(\mathbf{f},\mathbf{f}''(\mathbf{f},\mathbf{f}))$	$f^i_{jk} f^j f^k_{lm} f^l f^m$
5		4	$\mathbf{f}''(\mathbf{f},\mathbf{f}'\mathbf{f}'\mathbf{f})$	$f^i_{jk} f^j f^k_l f^l_m f^m$
5		3	$\mathbf{f}''(\mathbf{f}'\mathbf{f},\mathbf{f}'\mathbf{f})$	$f^i_{jk} f^j_l f^l f^k_m f^m$
5		1	$\mathbf{f}'\mathbf{f}'''(\mathbf{f},\mathbf{f},\mathbf{f})$	$f^i_j f^j_{klm} f^k f^l f^m$
5		3	$\mathbf{f}'\mathbf{f}''(\mathbf{f},\mathbf{f}'\mathbf{f})$	$f^i_j f^j_{kl} f^k f^l_m f^m$
5		1	$\mathbf{f}'\mathbf{f}'\mathbf{f}''(\mathbf{f},\mathbf{f})$	$f^i_j f^j_k f^k_{lm} f^l f^m$
5		1	$\mathbf{f}'\mathbf{f}'\mathbf{f}'\mathbf{f}'\mathbf{f}$	$f^i_j f^j_k f^k_l f^l_m f^m$

As a formal series, this can be evaluated using the following result:

Lemma 310B *The value of* (310i) *is*

$$\sum_{t\in T} \widetilde{\theta}(t)\frac{h^{r(t)}}{\sigma(t)}F(t)(y_0),$$

where $\widetilde{\theta}$ *is defined by*

$$\widetilde{\theta}(t) = \begin{cases} 1, & t = \tau, \\ \displaystyle\prod_{i=1}^{k} \theta(t_i), & t = [t_1 t_2 \cdots t_k]. \end{cases}$$

Proof. Use Theorem 306A. The case $t = \tau$ is obvious. For $t = [t_1^{m_1} t_2^{m_2} \cdots t_j^{m_j}]$, where $t_1, t_2, \ldots, t_j$ are distinct, the factor

$$\left(\sigma(I) \prod_{i=1}^{j} \sigma(t_j)^{m_j} \right)^{-1},$$

where I is the index set consisting of m_1 copies of 1, m_2 copies of 2, $\ldots$ and m_j copies of j, is equal to $\sigma(t)^{-1}$. $\qquad\qquad\square$

311   The Taylor expansion of the exact solution

We approach the question of the Taylor series of the exact solution from two points of view. In the first, we evaluate the Taylor coefficients by repeated differentiation, as we have illustrated in Subsection 310. In the second, we successively find Taylor series for the Picard iterates.

   The central result in the first approach is an expression for the derivatives written in terms of labelled trees. Throughout the discussion it will be assumed, without further comment, that y is a solution to $y'(x) = f(y(x))$ and that y is differentiable arbitrarily often. First, we need a formula for the derivative of a single elementary differential.

Lemma 311A *Let $S = S_0 \cup \{s\}$ be an ordered set, where every member of S_0 is less than s. Let $\mathbf{t}$ be a member of $T_{S_0}^*$. Then*

$$\frac{d}{dx} F(|\mathbf{t}|)(y(x))$$

is the sum of $F(|\mathbf{u}|)(y(x))$ over all $\mathbf{u} \in T_S^*$ such that the subtree formed by removing s from the set of vertices is $\mathbf{t}$.

Proof. If $S = \{s_0, s\}$, then the result is equivalent to

$$\frac{d}{dx} f(y(x)) = f'(y(x)) f(y(x)).$$

We now complete the proof by induction in the case $S = \{s_0\} \cup S_1 \cup S_2 \cup \cdots \cup S_k \cup \{s\}$, where $\{s_0\}$, S_1, S_2, $\ldots$, S_k, $\{s\}$ are disjoint subsets of the ordered set S. By the induction hypothesis, assume that the result of the lemma is true, when S is replaced by S_i, $i = 1, 2, \ldots, k$. If $\mathbf{t} \in T_{S_0}^*$, then

$$|\mathbf{t}| = [|\mathbf{t}_1| \, |\mathbf{t}_2| \, \cdots \, |\mathbf{t}_k|],$$

where $\mathbf{t}_i \in T_{S_i}^*$, $i = 1, 2, \ldots, k$. Differentiate

$$F(|\mathbf{t}|)(y(x))$$
$$= f^{(k)}(y(x)) \big(F(|\mathbf{t}_1|)(y(x)), F(|\mathbf{t}_2|)(y(x)), \ldots, F(|\mathbf{t}_k|)(y(x)) \big), \quad (311a)$$

to obtain

$$Q_0 + Q_1 + Q_2 + \cdots + Q_k,$$

where

$$Q_0 = f^{(k+1)}(y(x))\big(F(|\mathbf{t}_1|)(y(x)), F(|\mathbf{t}_2|)(y(x)), \ldots, F(|\mathbf{t}_k|)(y(x)), f(y(x))\big)$$

and, for $i = 1, 2, \ldots, k$,

$$Q_i = f^{(k)}(y(x))\big(F(|\mathbf{t}_1|)(y(x)), \ldots, \frac{d}{dx}F(|\mathbf{t}_i|)(y(x)), \ldots, F(|\mathbf{t}_k|)(y(x))\big).$$

The value of Q_0 is

$$F([|\mathbf{t}_1|\,|\mathbf{t}_2|\,\cdots\,|\mathbf{t}_k|\,|\mathbf{t}_0|])(y(x)),$$

where $|\mathbf{t}_0|$ is τ labelled with the single label s. For $i = 1, 2, \ldots, k$, the value of Q_i is the sum of all terms of the form (311a), with $F(|\mathbf{t}_i|)(y(x))$ replaced by terms of the form $F(|\mathbf{u}_i|)(y(x))$, where $\mathbf{u}_i$ is formed from $\mathbf{t}_i$ by adding an additional leaf labelled by s. The result of the lemma follows by combining all terms contributing to the derivative of (311a). □

Theorem 311B *Let S denote a finite ordered set. Then*

$$y^{(\#S)}(y_0) = \sum_{\mathbf{t} \in T_S} F(|\mathbf{t}|)(y_0).$$

Proof. In the case $|\mathbf{t}| = \tau$, the result is obvious. For the case $\#S > 1$, apply Lemma 311A repeatedly by adding additional (and increasing) members to S. □

We rewrite this result in terms of unlabelled trees, by noting that the number of times that a tree t with order $\#S$ occurs as the unlabelled counterpart of a member of T_S^*, is exactly $\alpha(t)$.

Theorem 311C

$$y^{(n)}(y(x)) = \sum_{t \in T_n} \alpha(t)F(t)(y(x)).$$

The alternative approach to finding the Taylor coefficients is based on the Picard integral equation

$$y(x_0 + h\xi) = y(x_0) + h \int_0^\xi f(y(x_0 + h\xi))d\xi,$$

which, written in terms of Picard iterations, becomes

$$y_n(x_0 + h\xi) = y(x_0) + h \int_0^\xi f(y_{n-1}(x_0 + h\xi))d\xi, \qquad (311b)$$

where the initial iterate is given by

$$y_0(x + h\xi) = y(x_0). \tag{311c}$$

For $n = 1, 2, \ldots$, we expand $y_n(x_0 + h\xi)$ for $\xi \in [0, 1]$, omitting terms that are $O(h^{n+1})$.

Theorem 311D *The Taylor expansion of y_n given by (311b) and (311c) is equal to*

$$y_n = y(x_0) + \sum_{i=1}^{n} h^i \xi^i \sum_{t \in T_i} \frac{1}{\sigma(t)\gamma(t)} F(t)(y(x_0)) + O(h^{n+1}). \tag{311d}$$

Proof. The case $n = 0$ is obvious. We now use induction and suppose that (311d) is true with n replaced by $n - 1$. By Lemma 310B, with

$$\theta(t) = \frac{1}{\gamma(t)},$$

we have as the coefficient of $F(t)(y(x_0))h^{r(t)}$, the expression

$$\int_0^{\xi} \frac{1}{\prod_{i=1}^{k} \gamma(t_i)} \xi^{r(t)-1} d\xi = \frac{1}{r(t) \prod_{i=1}^{k} \gamma(t_i)} \xi^{r(t)} = \frac{1}{\gamma(t)} \xi^{r(t)},$$

where $t = [t_1 t_2 \cdots t_k]$.   □

312   Elementary weights

Having found the Taylor expansion of the exact solution to an initial value problem, we now find the corresponding expansion for the approximation computed by a Runge–Kutta method. A term-by-term comparison of these will provide criteria for the error generated in a single step to be zero, except for terms that can be estimated in terms of high powers of the stepsize h.

As a prelude, we consider a three-stage explicit Runge–Kutta method. We find the Taylor expansion in this simple case up to terms in h^3. As the standard problem that we use for studying Runge–Kutta methods, we consider the autonomous initial value system

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

where $f : \mathbb{R}^N \to \mathbb{R}^N$. The method has the tableau

$$\begin{array}{c|ccc}
0 & & & \\
c_2 & c_2 & & \\
c_3 & c_3 - a_{32} & a_{32} & \\
\hline
 & b_1 & b_2 & b_3
\end{array}.$$

Denote by Y_1, Y_2 and Y_3 the results computed at the stages and y_1 for the result computed at the end of the step.

We can in turn find truncated Taylor expansions for the stages and the output result. We also use Theorem 306A to evaluate the Taylor expansions of $hF_i = hf(Y_i)$, for $i = 1, 2, 3$. Note that the stage values need only terms up to h^2, because the extra h in hF_i takes away the need to find h^3 terms except for the hF_i terms and the final result:

$$
\begin{aligned}
Y_1 &= y_0, \\
hF_1 &= hf(y_0), \\
Y_2 &= y_0 + c_2 hf(y_0), \\
hF_2 &= hf(y_0) + c_2 h^2 f'(y_0)f(y_0) + \tfrac{1}{2}c_2^2 h^3 f''(y_0)(f(y_0), f(y_0)) + O(h^3), \\
Y_3 &= y_0 + (c_3 - a_{32})hf(y_0) + a_{32}\left(hf(y_0) + c_2 h^2 f'(y_0)f(y_0)\right) + O(h^3) \\
&= y_0 + c_3 hf(y_0) + a_{32}c_2 h^2 f'(y_0)f(y_0) + O(h^3), \\
hF_3 &= hf(y_0) + c_3 h^2 f'(y_0)f(y_0) + a_{32}c_2 h^3 f'(y_0)f'(y_0)f(y_0) \\
&\qquad + \tfrac{1}{2}c_3^2 h^3 f''(y_0)(f(y_0), f(y_0)) + O(h^4), \\
y_1 &= y_0 + (b_1 + b_2 + b_3)hf(y_0) + (b_2 c_2 + b_3 c_3)h^2 f'(y_0)f(y_0) \\
&\qquad + \tfrac{1}{2}(b_2 c_2^2 + b_3 c_3^2)h^3 f''(y_0)(f(y_0), f(y_0)) \\
&\qquad\qquad + b_3 a_{32}c_2 h^3 f'(y_0)f'(y_0)f(y_0) + O(h^4).
\end{aligned}
$$

We recognize elementary differentials, evaluated at y_0, appearing in these expansions and we rewrite y_1 as

$$
\begin{aligned}
y_1 = y_0 &+ h\Phi(\bullet)F(\bullet)(y_0) + h^2 \Phi(\mathbf{!})F(\mathbf{!})(y_0) \\
&+ \tfrac{1}{2}h^3 \Phi(\mathbf{V})F(\mathbf{V})(y_0) + h^3 \Phi\left(\mathbf{!}\atop\right) F\left(\mathbf{!}\atop\right)(y_0) + O(h^4),
\end{aligned}
$$

where the coefficients associated with the four trees of orders up to 3 are given by

$$
\begin{aligned}
\Phi(\bullet) &= b_1 + b_2 + b_3, \\
\Phi(\mathbf{!}) &= b_2 c_2 + b_3 c_3, \\
\Phi(\mathbf{V}) &= b_2 c_2^2 + b_3 c_3^2, \\
\Phi\left(\mathbf{!}\atop\right) &= b_3 a_{32} c_2.
\end{aligned}
$$

It is obvious that these expressions, which we have already introduced in Section 234, are of vital importance in understanding the accuracy of Runge–Kutta methods. We name them 'elementary weights' and define them formally, along with similar expressions associated with the individual stages, in the next definition. At the same time we define 'derivative weights' associated with the stages.

Table 312(I) Relation between elementary weights and rooted trees

labelled tree t	Elementary weight $\Phi(t)$
$\bullet\, i$	$\displaystyle\sum_{i=1}^{s} b_i$
(tree)	$\displaystyle\sum_{i,j=1}^{s} b_i a_{ij} = \sum_{i=1}^{s} b_i c_i$
(tree)	$\displaystyle\sum_{i,j,k=1}^{s} b_i a_{ij} a_{ik} = \sum_{i=1}^{s} b_i c_i^2$
(tree)	$\displaystyle\sum_{i,j,k=1}^{s} b_i a_{ij} a_{jk} = \sum_{i,j=1}^{s} b_i a_{ij} c_j$

Definition 312A *Let*

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

denote the tableau for an s-stage Runge–Kutta method. Then the 'elementary weights' $\Phi(t)$, the 'internal weights' $\Phi_i(t)$ and the 'derivative weights' $(\Phi_i D)(t)$ for $t \in T$ and $i = 1, 2, \ldots, s$ are defined by

$$(\Phi_i D)(\tau) = 1, \tag{312a}$$

$$\Phi_i(t) = \sum_{j=1}^{s} a_{ij} (\Phi_j D)(t), \tag{312b}$$

$$(\Phi_i D)([t_1 t_2 \cdots t_k]) = \prod_{j=1}^{k} \Phi_i(t_j), \tag{312c}$$

$$\Phi(t) = \sum_{i=1}^{s} b_i (\Phi_i D)(t). \tag{312d}$$

This definition is used recursively. First $\Phi_i D$ is found for $t = \tau$, using (312a), then Φ_i is evaluated for this single vertex tree, using (312b). This enables $(\Phi_i D)([\tau])$, using (312c), and then $\Phi_i([\tau])$ to be found for each stage. The order is built up in this way until $(\Phi_i D)(t)$ is known for any required tree. Finally, (312d) is used to evaluate $\Phi(t)$.

The notation $\Phi_i D$ is part of a more general scheme, which we introduce in Subsection 387. In the meantime, D should be thought of as an operator to be applied to Φ_i, which replaces the sequence of Taylor coefficient weights in a stage value by the set of coefficient weights for the stage derivatives.

Table 312(II)   Elementary weights for orders 1 to 5

$r(t)$	t	$\Phi(t)$
1		$\sum_{i=1}^{s} b_i$
2		$\sum_{i=1}^{s} b_i c_i$
3		$\sum_{i=1}^{s} b_i c_i^2$
3		$\sum_{i,j=1}^{s} b_i a_{ij} c_j$
4		$\sum_{i=1}^{s} b_i c_i^3$
4		$\sum_{i,j=1}^{s} b_i c_i a_{ij} c_j$
4		$\sum_{i,j=1}^{s} b_i a_{ij} c_j^2$
4		$\sum_{i,j,k=1}^{s} b_i a_{ij} a_{jk} c_k$
5		$\sum_{i=1}^{s} b_i c_i^4$
5		$\sum_{i,j=1}^{s} b_i c_i^2 a_{ij} c_j$
5		$\sum_{i,j=1}^{s} b_i c_i a_{ij} c_j^2$
5		$\sum_{i,j,k=1}^{s} b_i c_i a_{ij} a_{jk} c_k$
5		$\sum_{i=1}^{s} b_i \left(\sum_{j=1}^{s} a_{ij} c_j \right)^2$
5		$\sum_{i,j=1}^{s} b_i a_{ij} c_j^3$
5		$\sum_{i,j,k=1}^{s} b_i a_{ij} c_j a_{jk} c_k$
5		$\sum_{i,j,k=1}^{s} b_i a_{ij} a_{jk} c_k^2$
5		$\sum_{i,j,k,l=1}^{s} b_i a_{ij} a_{jk} a_{kl} c_l$

An alternative formula for $\Phi(t)$, which uses the vertex and edge characterization of each tree t, is given in the following lemma, which we state without proof.

Lemma 312B *Denote the vertex set V of the tree t by the set of index symbols $V = \{j, k, l, \ldots\}$, where j is the root of t. Let the corresponding edge set be E. Form the expression*

$$b_j \prod_{(k,l) \in E} a_{kl} \tag{312e}$$

and sum this over each member of V ranging over the index set $\{1, 2, \ldots, s\}$.

*The resulting sum is the value of $\Phi(t)$. A similar formula for $\Phi_i(t)$, where i
is not a member of V, is found by replacing (312e) by*

$$a_{ij} \prod_{(k,l)\in E} a_{kl} \qquad\qquad (312\text{f})$$

and summing this as for $\Phi(t)$.

Note that, although c does explicitly appear in Definition 312A or Lemma
312B, it is usually convenient to carry out the summations $\sum_{l=1}^{s} a_{kl}$ to yield
a result c_k if l denotes a leaf (terminal vertex) of V. This is possible because
l occurs only once in (312e) and (312f).

We illustrate the relationship between the trees and the corresponding
elementary weights in Table 312(I). For each of the four trees, we write $\Phi(t)$
in the form given directly by Lemma 312B, and also with the summation
over leaves explicitly carried out. Finally, we present in Table 312(II) the
elementary weights up to order 5.

313 The Taylor expansion of the approximate solution

We show that the result output by a Runge–Kutta methods is exactly the
same as (311d), except that the factor $\gamma(t)^{-1}$ is replaced by $\Phi(t)$. We first
establish a preliminary result.

Lemma 313A *Let $k = 1, 2, \ldots,$. If*

$$Y_i = y_0 + \sum_{r(t)\leq k-1} \frac{1}{\sigma(t)}\Phi_i(t)h^{r(t)}F(t)(y_0) + O(h^k), \qquad (313\text{a})$$

then

$$hf(Y_i) = \sum_{r(t)\leq k} \frac{1}{\sigma(t)}(\Phi_i D)(t)h^{r(t)}F(t)(y_0) + O(h^{k+1}). \qquad (313\text{b})$$

Proof. Use Lemma 310B. The coefficient of $\sigma(t)^{-1}F(t)(y_0)h^{r(t)}$ in $hf(Y_i)$ is
$\prod_{j=1}^{n}\Phi_i(t_j)$, where $t = [t_1 t_2 \cdots t_k]$. □

We are now in a position to derive the formal Taylor expansion for the
computed solution. The proof we give for this result is for a general Runge–
Kutta method that may be implicit. In the case of an explicit method, the
iterations used in the proof can be replaced by a sequence of expansions for
Y_1, for $hf(Y_1)$, for Y_2, for $hf(Y_2)$, and so on until we reach Y_s, $hf(Y_s)$ and
finally y_1.

Theorem 313B *The Taylor expansions for the stages, stage derivatives and output result for a Runge–Kutta method are*

$$Y_i = y_0 + \sum_{r(t)\leq n} \frac{1}{\sigma(t)} \Phi_i(t) h^{r(t)} F(t)(y_0) + O(h^{n+1}), \; i = 1, 2, \ldots, s, \; (313c)$$

$$hf(Y_i) = \sum_{r(t)\leq n} \frac{1}{\sigma(t)} (\Phi_i D)(t) h^{r(t)} F(t)(y_0) + O(h^{n+1}), \; i = 1, 2, \ldots, s, \; (313d)$$

$$y_1 = y_0 + \sum_{r(t)\leq n} \frac{1}{\sigma(t)} \Phi(t) h^{r(t)} F(t)(y_0) + O(h^{n+1}). \quad\quad (313e)$$

Proof. In a preliminary part of the proof, we consider the sequence of approximations to Y_i given by

$$Y_i^{[0]} = y_0, \quad\quad\quad\quad\quad\quad i = 1, 2, \ldots, s, \quad\quad (313f)$$

$$Y_i^{[k]} = y_0 + h \sum_{j=1}^{s} a_{ij} f\left(Y_j^{[k-1]}\right), \quad\quad i = 1, 2, \ldots, s. \quad\quad (313g)$$

We prove by induction that $Y_i^{[n]}$ agrees with the expression given for Y_i to within $O(h^{n+1})$. For $n = 0$ this is clear. For $n > 0$, suppose it has been proved for n replaced by $n - 1$. From Lemma 313A with $k = n - 1$ and Y_i replaced by $Y_i^{[n-1]}$, we see that

$$hf(Y_i^{[n-1]}) = \sum_{r(t)\leq n} \frac{1}{\sigma(t)} (\Phi_i D)(t) h^{r(t)} F(t)(y_0) + O(h^{n+1}), \quad i = 1, 2, \ldots, s.$$

Calculate $Y_i^{[n]}$ using (313c) and the preliminary result follows. Assume that h is sufficiently small to guarantee convergence of the sequence $(Y_i^{[0]}, Y_i^{[1]}, Y_i^{[2]}, \ldots)$ to Y_i and (313c) follows. Finally, (313d) follows from Lemma 313A and (313e) from (312d). $\quad\quad\quad\square$

314   Independence of the elementary differentials

Our aim of comparing the Taylor expansions of the exact and computed solutions to an initial value problem will give an inconclusive answer unless the terms involving the various elementary differentials can be regarded as independent. We introduce a special type of differential equation for which any finite number of elementary differentials evaluate to independent vectors.

Let U denote any finite subset of T, such that if

$$t_i = [t_1^{m_1}, t_2^{m_2}, \ldots, t_k^{m_k}] \in U, \quad\quad\quad\quad (314a)$$

Table 314(I)   Trees to order 4 with corresponding differential equations

i	t_i	y_i'	$= f_i$
1	•	$[\,]$	$y_1' = 1,$
2		$[t_1]$	$y_2' = y_1,$
3		$[t_1^2]$	$y_3' = \frac{1}{2}y_1^2,$
4		$[t_2]$	$y_4' = y_2,$
5		$[t_1^3]$	$y_5' = \frac{1}{6}y_1^3,$
6		$[t_1 t_2]$	$y_6' = y_1 y_2,$
7		$[t_3]$	$y_7' = y_3,$
8		$[t_4]$	$y_8' = y_4.$

then each of $t_1, t_2, \ldots, t_k$ is also a member of U. For example, U might consist of all trees with orders up to some specified integer. Assume that when we write a tree in this way, the t_i, $i = 1, 2, \ldots, k$, are all distinct. Suppose that N is the number of members of U, and consider the m-dimensional differential equation system in which

$$y_i' = \prod_{j=1}^{k} \frac{y_j^{m_j}}{m_j!}, \tag{314b}$$

corresponding to tree number i defined in (314a). The initial values are supposed to be $y_i(0) = y_i(x_0) = 0$, for $i = 1, 2, \ldots, N$. The interesting property of this initial value problem is encapsulated in the following result:

Theorem 314A *The values of the elementary differentials for the differential equation* (314b), *evaluated at the initial value, are given by*

$$F(t_i)(y(x_0)) = e_i, \qquad i = 1, 2, \ldots, N.$$

Because the natural basis vectors $e_1, e_2, \ldots, e_N$ are independent, there cannot be any linear relation between the elementary differentials for an arbitrary differential equation system.

We illustrate this theorem in the case where U consists of the eight trees with up to four vertices. Table 314(I) shows the trees numbered from $i = 1$ to $i = 8$, together with their recursive definitions in the form (314a) and the corresponding differential equations. Note that the construction given here is given as an exercise in Hairer, Nørsett and Wanner (1993) .

315   Conditions for order

Now that we have expressions for the Taylor expansions of the exact solution, and also of the computed solution, we have all we need to find conditions for order. If the exact solution has Taylor series given by (311d) and the approximate solution has Taylor series given by (313e), then we need only compare these term by term to arrive at the principal result on the order of Runge–Kutta methods.

Theorem 315A *A Runge–Kutta method with elementary weights*

$$\Phi : T \to \mathbb{R},$$

has order p if and only if

$$\Phi(t) = \frac{1}{\gamma(t)}, \quad \text{for all } t \in T \text{ such that } r(t) \le p. \tag{315a}$$

Proof. The coefficient of $F(t)(y_0)h^{r(t)}$ in (313e) is $\frac{1}{\sigma(t)}\Phi(t)$, compared with the coefficient in (311d), which is $\frac{1}{\sigma(t)\gamma(t)}$. Equate these coefficients and we obtain (315a). $\qquad\square$

316   Order conditions for scalar problems

Early studies of Runge–Kutta methods were built around the single scalar equation

$$y'(x) = f(x, y(x)). \tag{316a}$$

Even though it was always intended that methods derived for (316a) should be interpreted, where appropriate, in a vector setting, a subtle difficulty arises for orders greater than 4.

We adopt the notation f_x, f_y for partial derivatives of f with respect to the first and second arguments, with similar notations for higher derivatives. Also, for simplicity, we omit the arguments in expressions like $f_x(x, y)$. By straightforward differentiation of (316a), we have

$$y'' = f_x + f_y y' = f_x + f_y f,$$

where the two terms together correspond to the elementary differential associated with $t = \mathord{\uparrow}$. Similarly, for the third derivative we have

$$y''' = \bigl(f_{xx} + 2f_{xy}f + f_{yy}f^2\bigr) + \bigl(f_y(f_x + f_y f)\bigr),$$

where the grouped terms correspond to $t = \mathord{\vee}$ and $t = \mathord{\uparrow}$, respectively.

The expressions that arise here, and for the fourth derivative, are more complicated, because of the presence of derivatives with respect to x. However, the terms can be grouped together according to the elementary differentials to which they correspond. Furthermore, the order conditions are identical to those found in the general vector case. When similar expressions are worked out for the 17 elementary differentials of order 5, we find a confusion between the results for two particular trees. In fact for each of $t_1 = $ and $t_2 = $, $F(t)$ reduces to

$$F(t) = f_y(f_{yy}f + f_{xy})(f_yf + f_x),$$

and instead of two order conditions

$$\Phi(t_1) = \sum b_i c_i a_{ij} a_{jk} c_k = \frac{1}{\gamma(t_1)} = \frac{1}{30}$$

and

$$\Phi(t_2) = \sum b_i a_{ij} c_j a_{jk} c_k = \frac{1}{\gamma(t_2)} = \frac{1}{40}, \tag{316b}$$

we have the single condition

$$\sum b_i c_i a_{ij} a_{jk} c_k + \sum b_i a_{ij} c_j a_{jk} c_k = \frac{7}{120}.$$

We discuss in Subsection 325 the construction of fifth order methods. These usually satisfy the so-called $D(1)$ condition, which we introduce in Subsection 321. This simplifying assumption has, as one of its consequences, the dependence of (316b) on other conditions, for which there is no confusion. Hence, for methods satisfying $D(1)$, scalar and vector order 5 conditions are equivalent.

For orders 6 and higher, the confusion between the order conditions for the scalar case becomes more pronounced. The first published methods of this order(Huťa, 1956, 1957) were derived for scalar problems but, nevertheless, have order 6 for the general vector case (Butcher, 1963a).

317   Independence of elementary weights

We show in Subsection 324 that, given a positive integer p, there exists an integer s such that there is a Runge–Kutta method with s stages with order p. We now present a more general result on the independence of the elementary weights but without a specific value of s given.

Theorem 317A *Given a finite subset T_0, of T and a mapping $\phi : T_0 \to \mathbb{R}$, there exists a Runge–Kutta method such that the elementary weights satisfy*

$$\Phi(t) = \phi(t), \qquad \text{for all } t \in T_0.$$

Proof. Let $\#T_0 = n$. The set of possible values that can be taken by the vector of $\Phi(t)$ values, for all $t \in T_0$, is a vector space. To see why this is the case, consider Runge–Kutta methods given by the tableaux

$$
\begin{array}{c|c}
c & A \\
\hline
 & b^{\mathsf{T}}
\end{array}
\qquad \text{and} \qquad
\begin{array}{c|c}
\overline{c} & \overline{A} \\
\hline
 & \overline{b}^{\mathsf{T}}
\end{array}
\tag{317a}
$$

with s and $\overline{s}$ stages, respectively. If the elementary weight functions for these two Runge–Kutta methods are Φ and $\overline{\Phi}$, then the method given by the tableau

$$
\begin{array}{c|cc}
c & A & 0 \\
\overline{c} & 0 & \overline{A} \\
\hline
 & \theta b^{\mathsf{T}} & \overline{\theta} \overline{b}^{\mathsf{T}}
\end{array}
$$

has elementary weight function $\theta\Phi + \overline{\theta}\overline{\Phi}$. Let $V \subset \mathbb{R}^n$ denote this vector space. We complete the proof by showing that $V = \mathbb{R}^n$. If this were not the case, there would exist a non-zero function $\psi : T_0 \to \mathbb{R}$ such that $\sum_{t \in T_0} \psi(t)\Phi(t) = 0$, for all Runge–Kutta methods. Because every coefficient in a Runge–Kutta tableau can be multiplied by an arbitrary scalar θ to give a new method for which $\Phi(t)$ is replaced by $\theta^{r(t)}\Phi(t)$, we may assume that every non-zero value of ψ corresponds to trees with the same order k. This is impossible for $k = 1$, because in this case there is only a single tree τ. Suppose the impossibility of this has been proved for all orders less than k, but that there exist trees t_1, t_2, ..., t_m, each of order k, such that $\sum_{i=1}^{m} \psi(t_i)\Phi(t_i) = 0$, for all Runge–Kutta methods with $\psi(t_i) \neq 0$, for $i = 1, 2, \ldots, m$. Write $t_i = [t_{i1}^{l_{i1}} t_{i2}^{l_{i2}} \cdots]$, for $i = 1, 2, \ldots, m$. Let $\hat{t}$ denote a tree appearing amongst the t_{ij} which does not occur with the same exponent in each of the t_i. Construct an s-stage Runge–Kutta method

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

for which each of $\Phi(t_{ij}) = 1$, except for $\Phi(\hat{t}) = \theta$. Define second Runge–Kutta tableau with $s + 1$ stages of the form

$$
\begin{array}{c|cc}
c & A & 0 \\
1 & b^{\mathsf{T}} & 0 \\
\hline
 & 0 & 1
\end{array} \; .
$$

If q_i is the exponent of $\hat{t}$ in t_i, then it follows that

$$
\sum_{i=1}^{m} \psi(t_i)\theta^{q_i} = 0.
$$

Since θ can take any value and since q_i is not constant, it is not possible that ψ is never zero. $\qquad \square$

318   *Local truncation error*

The conditions for order give guarantees that the Taylor expansions of the exact and computed solutions agree up to terms in h^p. Obtaining an understanding of the respective terms in h^{p+1} is regarded as a key to deriving methods that not only have a specific order, but also have a small truncation error. Because the number of terms of this order rises rapidly as p increases, it is extremely difficult to know how this sort of optimality should be arrived at. Picking out just the terms of order $p+1$, we can write the local truncation error in a single step as

$$h^{p+1} \sum_{r(t)=p+1} \frac{1}{\sigma(t)} \left(\frac{1}{\gamma(t)} - \Phi(t) \right) F(t)(y_0) + O(h^{p+2}). \qquad (318\text{a})$$

Since we are interested in asymptotic behaviour, that is, limiting behaviour for h small, we do not devote much attention to the term $O(h^{p+2})$. The coefficient of h^{p+1} in (318a) is bounded in magnitude by

$$\sum_{r(t)=p+1} \frac{1}{\sigma(t)} \left| \Phi(t) - \frac{1}{\gamma(t)} \right| \cdot \| F(t)(y_0) \|, \qquad (318\text{b})$$

and this should somehow be made small. There is simply no general rule interrelating the magnitudes of the various elementary differentials, and some assumptions need to be made.

The first approach that can be considered is to compare, term by term, the expression for $\frac{1}{(p+1)!} y^{(p+1)}(x_0)$, which is proportional to the local truncation error coefficient for linear multistep methods or for implicit Runge–Kutta methods of collocation type. The coefficient in this expression, corresponding to t, is

$$\frac{1}{\sigma(t)\gamma(t)},$$

so that the corresponding multiplier to yield the corresponding term in (318b) is

$$|\gamma(t)\Phi(t) - 1|.$$

Hence, we can bound (318b) by

$$\max_{r(t)=p+1} |\gamma(t)\Phi(t) - 1| \sum_{r(t)=p+1} \frac{1}{\sigma(t)\gamma(t)} \cdot \| F(t)(y_0) \|$$

and hence, it might be desirable to minimize

$$\max_{r(t)=p+1} |\gamma(t)\Phi(t) - 1|$$

in seeking an efficient method.

Another approach would be to assume a bound M on $\|f\|$, a bound L on the linear operator $\|f'\|$, and further bounds to make up the sequence

$$\|f\| \leq M,$$
$$\|f'\| \leq L,$$
$$\|f''\| \leq \frac{L^2}{M},$$
$$\|f'''\| \leq \frac{L^3}{M^2},$$
$$\vdots \qquad \vdots$$
$$\|f^{(p)}\| \leq \frac{L^p}{M^{p-1}}.$$

This will mean that for any tree of order $p+1$, $\|F(t)(y_0)\| \leq L^p M$ and that

$$\sum_{r(t)=p+1} \frac{1}{\sigma(t)} \left| \Phi(t) - \frac{1}{\gamma(t)} \right| \cdot \|F(t)(y_0)\| \leq \sum_{r(t)=p+1} \frac{1}{\sigma(t)} \left| \Phi(t) - \frac{1}{\gamma(t)} \right| \cdot L^p M.$$

In studying the behaviour of a particular method of order p when used to solve a particular initial value problem, we wish to assume that the local truncation error is bounded asymptotically by some constant multiplied by h^{p+1}. This assumption will hinge on smoothness of the solution and the differentiability, sufficiently many times, of f.

319   Global truncation error

We consider the cumulative effect of errors in many steps leading to an error in a final output point. Suppose that n steps are performed to carry the solution from an initial point x_0 to a final point $\overline{x}$. If a constant stepsize is used, this would need to be equal to $(\overline{x} - x_0)/n$ to exactly reach the final point. Denote the approximations computed by a Runge–Kutta method by y_1, y_2, ..., y_n, with $y_0 = y(x_0)$. If the error committed in each of the n steps is bounded by Ch^{p+1} then the total contribution to the error would seem to be

$$nCh^{p+1} = C(\overline{x} - x_0)h^p.$$

We attempt to make this argument more precise by noting that an error in the initial value input to a step will lead to an error in the output value consisting of two terms. The first of these is the perturbation to the output due to the error in the input, and the second is the truncation error due to the method itself.

In the statement of a preliminary lemma that we need, $|A|$ and $|b^{\mathsf{T}}|$ will denote the matrix A and the vector b^{T}, respectively, with every term replaced by its magnitude.
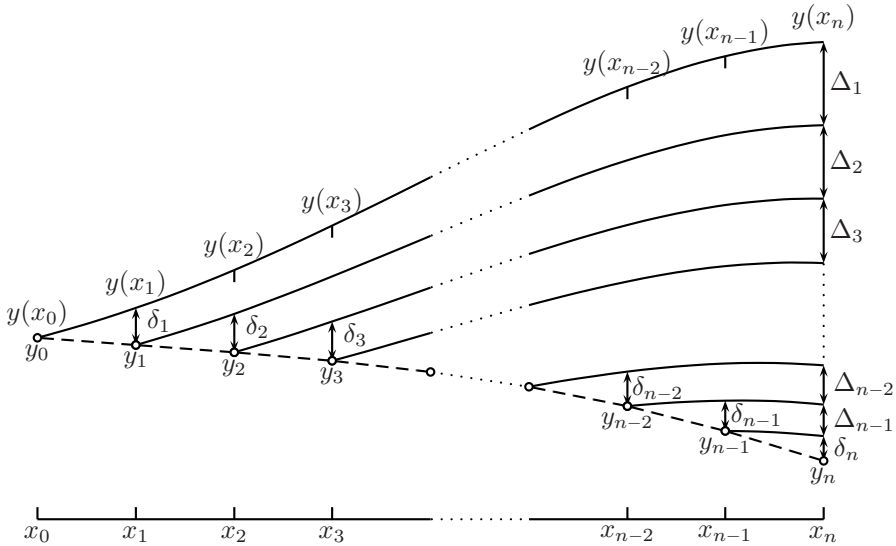
Figure 319(i)   Growth of global errors from local errors referred to the
computed solution

Lemma 319A *Let f denote a function $\mathbb{R}^m \to \mathbb{R}^m$, assumed to satisfy a Lipschitz condition with constant L. Let $y_0 \in \mathbb{R}^m$ and $z_0 \in \mathbb{R}^m$ be two input values to a step with the Runge–Kutta method (A, b^{T}, c), using stepsize $h \leq h_0$, where $h_0 L \rho(|A|) < 1$, and let y_1 and z_1 be the corresponding output values. Then*

$$\|y_1 - z_1\| \leq (1 + hL^{\star})\|y_0 - z_0\|,$$

where

$$L^{\star} = L|b^{\mathsf{T}}|(I - h_0 L|A|)^{-1}\mathbf{1}.$$

Proof. Denote the stage values by Y_i and Z_i, $i = 1, 2, \ldots, s$, respectively. From the equation $Y_i - Z_i = (y_0 - z_0) + h\sum_{j=1}^{s} a_{ij}(f(Y_j) - f(Z_j))$, we deduce that

$$\|Y_i - Z_i\| \leq \|y_0 - z_0\| + h_0 L \sum_{j=1}^{s} |a_{ij}|\|Y_j - Z_j\|,$$

so that, substituting into

$$\|y_1 - z_1\| \leq \|y_0 - z_0\| + hL \sum_{j=1}^{s} |b_j|\|Y_j - Z_j\|,$$

we obtain the result. $\square$
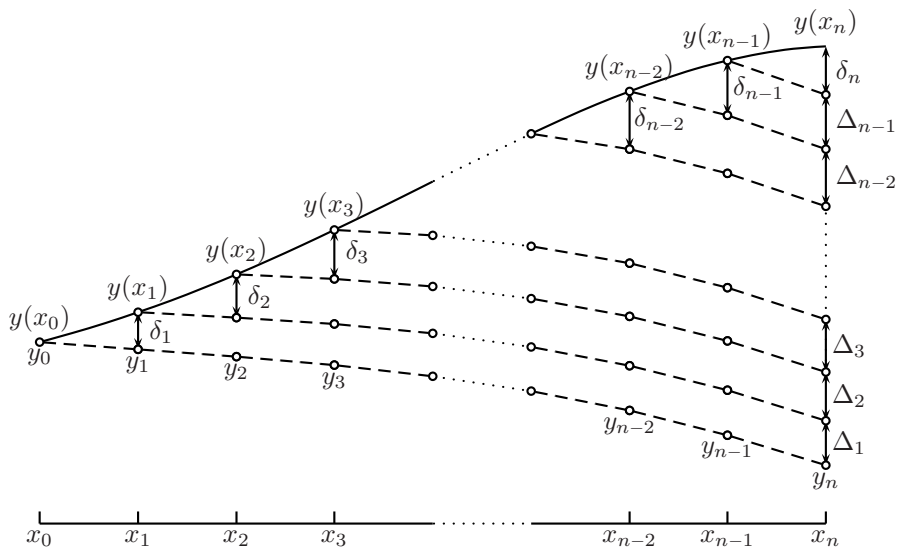
Figure 319(ii)   Growth of global errors from local errors referred to the exact solution

To see how to use this result, consider Figures 319(i) and 319(ii). Each of these shows the development of global errors generated by local truncation errors in individual steps. In Figure 319(i), the local truncation errors are referred to the computed solution. That is, in this figure, δ_k is the difference between the exact solution defined by an initial value at the start of step k and the numerical solution computed in this step. Furthermore, Δ_k is the contribution to the global error resulting from the error δ_k in step k. An alternative view of the growth of errors is seen from Figure 319(ii), where δ_k is now the difference between the exact solution at x_k and the computed solution found by using an input value y_{k-1} at the start of this step exactly equal to $y(x_{k-1})$. As in the previous figure, Δ_k is the contribution to the global error resulting from the local error δ_k. To obtain a bound on the global truncation error we first need an estimate on $\delta_1, \delta_2, \ldots, \delta_n$ using these bounds. We then estimate by how much δ_k can grow to Δ_k, $k = 1, 2, \ldots, n$. The global error is then bounded in norm by $\sum_{k=1}^{n} \Delta_k$. We have a bound already from (110c) on how much a perturbation in the exact solution can grow. If we were basing our global error bound on Figure 319(i) then this would be exactly what we need. However, we use Figure 319(ii), and in this case we obtain the same growth factor but with L replaced by $L^\star$. The advantage of using an argument based on this figure, rather than on Figure 319(i), is that we can then use local truncation error defined in the standard way, by comparing the exact solution at step value x_n with the numerically computed result over a single step with initial value $y(x_{n-1})$.

Theorem 319B *Let h_0 and $L^\star$ be such that the local truncation error at step $k = 1, 2, \ldots, n$ is bounded by*

$$\delta_k \leq Ch^{p+1}, \qquad h \leq h_0.$$

Then the global truncation error is bounded by

$$\|y(x_n) - y_n\| \leq \begin{cases} \frac{\exp(L^\star(\overline{x}-x_0))-1}{L^\star}Ch^p, & L^\star > 0, \\ (\overline{x} - x_0)Ch^p, & L^\star = 0. \end{cases}$$

Proof. Use Figure 319(ii) and obtain the estimate

$$\|y(x_n) - y_n\| \leq Ch^{p+1} \sum_{k=1}^{n}(1 + hL^\star)^k.$$

The case $L^\star = 0$ is obvious. For the case $L^\star > 0$, calculate the sum and use the bound

$$(1 + hL^\star)^n \leq \exp(L^\star hn) = \exp(L^\star(\overline{x} - x_0)). \qquad \square$$

Exercises 31

31.1 Define $f : \mathbb{R}^3 \to \mathbb{R}^3$ by

$$f(y^1, y^2, y^3) = \begin{bmatrix} y^1 + y^2 y^3 \\ (y^1)^2 + 2y^1 y^2 \\ 1 + (y^2 + y^3)^2 \end{bmatrix}.$$

Find formulae for the elementary differentials $F(t)$, for $t = [\tau]$, $[\tau^2]$ and $[\tau[\tau]]$.

31.2 For the Runge–Kutta method

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

find the elementary weights for the eight trees up to order 4. What is the order of this method?

31.3 For an arbitrary Runge–Kutta method, find the order condition corresponding to the tree

.

32   Low Order Explicit Methods

320   *Methods of orders less than 4*

It will be shown in Subsection 324 that, for an explicit method to have order p, at least $s = p$ stages are necessary. We derive methods up to $p = 3$, with exactly p stages, and then discuss briefly the advantages of using $s = p + 1$.

For $s = p = 1$ there is no choice beyond the Euler method with tableau

$$
\begin{array}{c|c}
0 & \\
\hline
& 1
\end{array}
$$

For $s = p = 2$, we have a one-parameter family of methods of the form

$$
\begin{array}{c|cc}
0 & & \\
c_2 & c_2 & \\
\hline
& 1 - \dfrac{1}{2c_2} & \dfrac{1}{2c_2}
\end{array}
$$

which satisfies the conditions $b_1 + b_2 = 1$ and $b_2 c_2 = \frac{1}{2}$, corresponding to the trees $\bullet$ and $\mathbf{\mathop{}}$.

For $s = p = 3$, we must satisfy four conditions, which are shown together with the corresponding trees as follows:

$$\bullet \qquad b_1 + b_2 + b_3 = 1, \tag{320a}$$

$$\qquad b_2 c_2 + b_3 c_3 = \frac{1}{2}, \tag{320b}$$

$$\qquad b_2 c_2^2 + b_3 c_3^2 = \frac{1}{3}, \tag{320c}$$

$$\qquad b_3 a_{32} c_2 = \frac{1}{6}. \tag{320d}$$

To solve these equations in the most straightforward manner, it is convenient to treat c_2 and c_3 as free parameters and to carry out three steps. First, solve for b_2 and b_3 from the linear system given by (320b) and (320c). Secondly, evaluate b_1 from (320a). Finally, solve for a_{32} from (320d). This plan will run into difficulties if the matrix of coefficients in (320b) and (320c) is singular; that is, if $c_2 c_3 (c_3 - c_2) = 0$. Assuming this does not occur, we have a further difficulty if the solution to (320b) and (320c) results in $b_3 = 0$. This anomaly, which occurs if $c_2 = \frac{2}{3}$, makes it impossible to solve (320d). A more careful analysis is necessary to resolve these difficulties, and it is possible to identify three cases where a solution can be found. These are

$$
\begin{array}{ll}
\text{I} & c_2 \neq 0 \neq c_3 \neq c_2 \neq \frac{2}{3}, \\
\text{II} & c_2 = c_3 = \frac{2}{3}, \ b_3 \neq 0, \\
\text{III} & c_2 = \frac{2}{3}, \ c_3 = 0, \ b_3 \neq 0.
\end{array}
$$

The coefficient tableaux for the three cases are summarized as follows, with the general form of the tableau given in each case: for case I we have

$$
\begin{array}{c|ccc}
0 \\
c_2 & c_2 \\
c_3 & \dfrac{c_3(3c_2 - 3c_2^2 - c_3)}{c_2(2 - 3c_2)} & \dfrac{c_3(c_3 - c_2)}{c_2(2 - 3c_2)} \\
\hline
& \dfrac{-3c_3 + 6c_2c_3 + 2 - 3c_2}{6c_2c_3} & \dfrac{3c_3 - 2}{6c_2(c_3 - c_2)} & \dfrac{2 - 3c_2}{6c_3(c_3 - c_2)}
\end{array} \; ;
$$

for case II,

$$
\begin{array}{c|ccc}
0 \\
\dfrac{2}{3} & \dfrac{2}{3} \\
\dfrac{2}{3} & \dfrac{2}{3} - \dfrac{1}{4b_3} & \dfrac{1}{4b_3} \\
\hline
& \dfrac{1}{4} & \dfrac{3}{4} - b_3 & b_3
\end{array} \; ;
$$

and for case III,

$$
\begin{array}{c|ccc}
0 \\
\dfrac{2}{3} & \dfrac{2}{3} \\
0 & -\dfrac{1}{4b_3} & \dfrac{1}{4b_3} \\
\hline
& \dfrac{1}{4} - b_3 & \dfrac{3}{4} & b_3
\end{array} \; .
$$

321 *Simplifying assumptions*

As the order being sought increases, the number of conditions rises rapidly and soon becomes unmanageable. For this reason, it is necessary to examine the relationships between the conditions corresponding to various trees. At the same time, we identify certain collections of order conditions which have some sort of central role. Since these special conditions will be of varying complexity, depending on the orders to which we apply them, they will be parameterized by one or more positive integers. For example, $E(\eta, \zeta)$ is a set of assumptions about a method that hold for all positive integers $k \leq \eta$ and $l \leq \zeta$.

The first of these conditions will be denoted by $B(\eta)$, and simply states that the conditions $\sum_{i=1}^{s} b_i c_i^{k-1} = k^{-1}$ hold for $k = 1, 2, \ldots, \eta$. For a method to be of order p, it is necessary that $B(p)$ holds, because this condition simply restates the order condition for the trees

$$
\bullet \quad \mathfrak{l} \quad \mathsf{V} \quad \mathsf{V} \quad \ldots .
$$

Table 321(I)   Order conditions corresponding to some pairs of related trees

t_1	$\Phi(t_1) = \dfrac{1}{\gamma(t_1)}$	t_2	$\tfrac{1}{2}\Phi(t_2) = \dfrac{1}{2\gamma(t_2)}$
	$\sum b_i a_{ij} c_j = \tfrac{1}{6}$		$\tfrac{1}{2}\sum b_i c_i^2 = \tfrac{1}{6}$
	$\sum b_i c_i a_{ij} c_j = \tfrac{1}{8}$		$\tfrac{1}{2}\sum b_i c_i^3 = \tfrac{1}{8}$
	$\sum b_k a_{ki} a_{ij} c_j = \tfrac{1}{24}$		$\tfrac{1}{2}\sum b_k a_{ki} c_i^2 = \tfrac{1}{24}$

To motivate condition $C(\eta)$, consider pairs of trees t_1 and t_2, with the same order, that differ in only one small respect. Suppose they are labelled with identical vertex sets and that the edge sets, say E_1 and E_2, respectively, differ only in that E_1 contains the edges $[i, j]$ and $[j, k]$, and that j and k do not occur in any of the other ordered pairs in E_1, whereas E_2 contains the edge $[i, k]$ instead of $[j, k]$. This will mean that the elementary weight corresponding to t_1 will have a factor $a_{ij} c_j$, whereas t_2 will have a corresponding factor c_i^2. Furthermore, the densities are also closely related in that $\gamma(t_1) = 2\gamma(t_2)$. Hence, the equations

$$\Phi(t_1) = \frac{1}{\gamma(t_1)} \qquad \text{and} \qquad \Phi(t_2) = \frac{1}{\gamma(t_2)}$$

will be equivalent if

$$\sum_{j=1}^{s} a_{ij} c_j = \frac{1}{2} c_i^2, \text{ for all } i = 1, 2, \ldots, s. \tag{321a}$$

We illustrate this by looking at some pairs of trees and noting the form of the equations

$$\Phi(t_1) = \frac{1}{\gamma(t_1)} \quad \text{and} \quad \frac{1}{2}\Phi(t_2) = \frac{1}{2\gamma(t_2)}.$$

These are displayed in Table 321(I).

It is clear that, if it were possible for (321a) to hold for *all* $i \in \{1, 2, \ldots, s\}$, then we could simply remove the order equations associated with the t_1 trees from consideration, because they will automatically be satisfied if the conditions $\Phi(t) = 1/\gamma(t)$ are satisfied for the t_2 trees. However, it is *not* possible in the case $i = 2$ because this gives the equation $\frac{1}{2}c_2^2 = 0$ which implies $c_2 = 0$. It will then follow in turn that $c_3 = 0$, $c_4 = 0$, $\ldots$ and all c components equal to zero will not be consistent even with the order condition $\sum b_i c_i = \frac{1}{2}$. While we cannot make use of the simplification of assuming
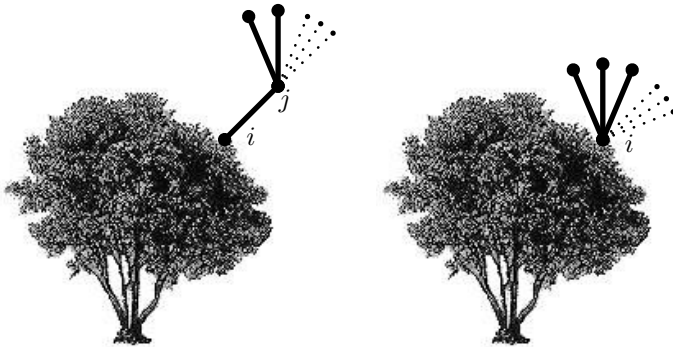
Figure 321(i) The $C(k)$ condition relating $\sum_j a_{ij} c_j^{k-1}$ (left-hand tree) to c_i^k (right-hand tree). The underlying tree is a pohutukawa (*Metrosideros excelsa*), also known as the 'New Zealand Christmas tree' because its bright red flowers bloom at Christmas-time.

(321a) in the case of explicit methods, we make extensive use of this and closely related conditions in the case of implicit methods. Furthermore, we can still use this sort of simplification applied to just *some* of the stages.

In addition to (321a), we can consider the possibility that conditions like

$$\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{1}{k} c_i^k, \quad i = 1, 2, \ldots, s, \tag{321b}$$

hold for $k = 1, 2, \ldots$. Assuming that these hold for $1 \le k \le \xi$, we denote this collection of conditions by $C(\xi)$. The consequences of $C(\xi)$ are that, for any pair of trees t_1 and t_2 for which $\Phi(t_1)$ contains a factor $a_{ij} c_j^{k-1}$, $\Phi(t_2)$ contains a factor $\frac{1}{k} c_i^k$ and the remaining factors are identical in the two expressions, then $\Phi(t_2) = 1/\gamma(t_2)$ implies $\Phi(t_1) = 1/\gamma(t_1)$. We illustrate this in Figure 321(i).

The $D(k)$ conditions interrelate three trees t_1, t_2 and t_3 for which the corresponding elementary weights differ only in that $\Phi(t_1)$ has a factor $b_i c_i^{k-1} a_{ij}$, $\Phi(t_2)$ has a factor b_j and $\Phi(t_3)$ has a factor $b_j c_j^k$. This means that these trees have forms like those shown in Figure 321(ii).

We illustrate this further, for the case $k = 1$, in Table 321(II). Note that if $D(1)$ holds, then the truth of $\Phi(t_1) = 1/\gamma(t_1)$ follows from $\Phi(t_2) = 1/\gamma(t_2)$ and $\Phi(t_3) = 1/\gamma(t_3)$. For explicit methods, $D(2)$ cannot hold, for similar reasons to the impossibility of $C(2)$. For implicit methods $D(s)$ is possible, as we shall see in Section 342.

Figure 321(ii)   The $D(k)$ condition relating $\sum_i b_i c_i^{k-1} a_{ij}$ (left-hand tree) to b_j (middle tree) and $b_j c_j^k$ (right-hand tree). The underlying tree is a kauri (*Agathis australis*). Although the immature tree shown is only a few metres tall, the most famous kauri tree, Tane Mahuta (Lord of the Forest), has a height of 40 m and a diameter, 1.5 m above ground level, of 5.21 m.

Table 321(II)   Sets of three related trees illustrating $D(1)$

t_1	$\Phi(t_1) = \dfrac{1}{\gamma(t_1)}$	t_2	$\Phi(t_2) = \dfrac{1}{\gamma(t_2)}$	t_3	$\Phi(t_3) = \dfrac{1}{\gamma(t_3)}$
	$\sum b_i a_{ij} c_j = \frac{1}{6}$		$\sum b_j c_j = \frac{1}{2}$		$\sum b_j c_j^2 = \frac{1}{3}$
	$\sum b_i a_{ij} c_j^2 = \frac{1}{12}$		$\sum b_j c_j^2 = \frac{1}{3}$		$\sum b_j c_j^3 = \frac{1}{4}$
	$\sum b_i a_{ij} a_{jk} c_k = \frac{1}{24}$		$\sum b_j a_{jk} c_k = \frac{1}{6}$		$\sum b_j c_j a_{jk} c_k = \frac{1}{8}$

Finally, the condition $E(\eta, \zeta)$ states that

$$\sum b_i c_i^{k-1} a_{ij} c_j^{l-1} = \frac{1}{l(k+l)}, \qquad k = 1, 2, \ldots, \eta, \quad l = 1, 2, \ldots, \zeta. \qquad (321c)$$

This simply expresses the fact that the order condition $\Phi(t) = 1/\gamma(t)$ is satisfied for trees $t = [\tau^{k-1}[\tau^{l-1}]]$ for $k \leq \eta$ and $l \leq \zeta$. This is a necessary condition for orders at least $\eta + \zeta$.

322   Methods of order 4

It is an interesting consequence of the fourth order conditions for a method with $s = 4$ stages, that $c_4 = 1$ and that $D(1)$ holds. This fact reduces significantly the number of conditions that remain to be solved; furthermore, it is possible to segment the derivation into two phases: the solution of the remaining order conditions and the evaluation of the elements in the final row of A to ensure that $D(1)$ is actually satisfied. Assuming that the method

$$
\begin{array}{c|cccc}
0 & & & & \\
c_2 & a_{21} & & & \\
c_3 & a_{31} & a_{32} & & \\
c_4 & a_{41} & a_{42} & a_{43} & \\
\hline
 & b_1 & b_2 & b_3 & b_4
\end{array}
$$

satisfies the fourth order conditions, then we can compute the values of

$$b_3(c_3 - c_4)(c_3 - c_2)c_3 = \sum b_i(c_i - c_4)(c_i - c_2)c_i = \tfrac{1}{4} - \tfrac{c_2+c_4}{3} + \tfrac{c_2 c_4}{2}, \quad (322a)$$

$$b_4 a_{43}(c_3 - c_2)c_3 = \sum b_i a_{ij}(c_j - c_2)c_j \qquad\qquad = \tfrac{1}{12} - \tfrac{c_2}{6}, \qquad\qquad (322b)$$

$$b_3(c_3 - c_4)a_{32}c_2 = \sum b_i(c_i - c_4)a_{ij}c_j \qquad\qquad = \tfrac{1}{8} - \tfrac{c_4}{6}, \qquad\qquad (322c)$$

$$b_4 a_{43} a_{32} c_2 = \sum b_i a_{ij} a_{jk} c_k \qquad\qquad\qquad = \tfrac{1}{24}. \qquad\qquad\quad (322d)$$

In each of these calculations, the first column is the only non-zero term in the middle column, while the final column is found by expanding the middle column into a linear combination of elementary weights and equating each of these to the right-hand sides of the corresponding order conditions. For example, (322a) is evaluated from the trees $\mathbf{V}$, $\mathbf{V}$ and $\mathbf{I}$ and uses the combination of order conditions

$$\Phi(\mathbf{V}) - (c_2 + c_4)\Phi(\mathbf{V}) + c_2 c_4 \Phi(\mathbf{I}) = \frac{1}{\gamma(\mathbf{V})} - \frac{c_2 + c_4}{\gamma(\mathbf{V})} + \frac{c_2 c_4}{\gamma(\mathbf{I})}.$$

From the first columns of (322a)–(322d), we observe that (322a)×(322d) = (322b)×(322c) so that, from the last columns, we find

$$\left(\frac{1}{4} - \frac{c_2 + c_4}{3} + \frac{c_2 c_4}{2}\right)\left(\frac{1}{24}\right) - \left(\frac{1}{12} - \frac{c_2}{6}\right)\left(\frac{1}{8} - \frac{c_4}{6}\right) = 0.$$

This relation simplifies to $c_2(c_4 - 1) = 0$ which, because $c_2 = 0$ is incompatible with (322d), implies $c_4 = 1$.

An alternative proof of this result, is found by using the following:

Lemma 322A *If P and Q are each 3×3 matrices such that their product has the form*

$$PQ = \begin{bmatrix} r_{11} & r_{12} & 0 \\ r_{21} & r_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where

$$\det\left(\begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}\right) \neq 0,$$

then either the last row of P is zero or the last column of Q is zero.

Proof. Because PQ is singular, either P is singular or Q is singular. In the first case, let $u^{\mathsf{T}} \neq 0$ be such that $u^{\mathsf{T}}P = 0$, and therefore $u^{\mathsf{T}}PQ = 0$; in the second case, let $v \neq 0$ be such that $Qv = 0$, and therefore $PQv = 0$. Because of the form of PQ, this implies that the first two components of u^{T} (or, respectively, the first two components of v) are zero.   □

To obtain the result that $D(1)$ necessarily holds if $s = p = 4$, we apply Lemma 322A with

$$P = \begin{bmatrix} b_2 & b_3 & b_4 \\ b_2 c_2 & b_3 c_3 & b_4 c_4 \\ \sum_{i=1}^{4} b_i a_{i2} - b_2(1-c_2) & \sum_{i=1}^{4} b_i a_{i3} - b_3(1-c_3) & \sum_{i=1}^{4} b_i a_{i4} - b_4(1-c_4) \end{bmatrix}$$

and

$$Q = \begin{bmatrix} c_2 & c_2^2 & \sum_{j=1}^{4} a_{2j}c_j - \frac{1}{2}c_2^2 \\ c_3 & c_3^2 & \sum_{j=1}^{4} a_{3j}c_j - \frac{1}{2}c_3^2 \\ c_4 & c_4^2 & \sum_{j=1}^{4} a_{4j}c_j - \frac{1}{2}c_4^2 \end{bmatrix}.$$

The value of the matrix PQ can be calculated from the order conditions. For example, the $(2,2)$ element is equal to

$$\begin{bmatrix} b_2 c_2 & b_3 c_3 & b_4 c_4 \end{bmatrix} \begin{bmatrix} c_2^2 & c_3^2 & c_4^2 \end{bmatrix}^{\mathsf{T}} = \sum_{i=1}^{4} b_i c_i^3 = \frac{1}{4}.$$

The elements in the last row and last column are a little more complicated to evaluate because they depend on linear combinations of elementary weights, but the relation of these elements in the product to the $C(2)$ and $D(1)$ conditions simplifies each of these elements to a zero value. In summary, the product of P and Q is

$$PQ = \begin{bmatrix} \frac{1}{2} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{4} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

so that the conditions of Lemma 322A are satisfied. The conclusion is that the last row of P or the last column of Q is zero. In particular, this means that either $\sum_{i=1}^{4} b_i a_{i4} - b_4(1 - c_4) = 0$ or $\sum_{j=1}^{4} a_{2j} c_j - \frac{1}{2} c_2^2 = 0$. These simplify to $b_4(1 - c_4) = 0$ or to $\frac{1}{2} c_2^2 = 0$, respectively. It is impossible that $c_2 = 0$ or that $b_4 = 1$, and hence $c_4 = 1$ and the $D(1)$ condition holds.

Since $D(1)$ holds, the set of additional equations we need to satisfy reduce to those associated with the trees $\bullet$, $\mathbf{!}$, $\mathbf{V}$ and $\mathbf{V}$ as well as with the tree $\mathbf{\sqrt{}}$. The order condition associated with the last of these is $\sum b_i c_i a_{ij} c_j = \frac{1}{8}$. It turns out to be more convenient to use, instead of this condition, the difference between this and with the condition associated with $\mathbf{!}$, that is, $\sum b_i a_{ij} c_j = \frac{1}{6}$, which is a consequence of other assumptions and of the $D(1)$ condition. Hence we assume $\sum b_i(1 - c_i) a_{ij} c_j = \frac{1}{24}$.

The steps we need to carry out to derive one of these methods are as follows:

(a) Choose c_2 and c_3, noting that $c_1 = 0$ and $c_4 = 1$.

(b) Choose b_1, b_2, b_3, b_4 to satisfy $\sum b_i c_i^{k-1} = 1/k$ for $k = 1, 2, 3, 4$.

(c) Choose a_{32} so that $b_3(1 - c_3) a_{32} c_2 = \frac{1}{24}$.

(d) Choose a_{41}, a_{42}, a_{43}, so that $\sum_i b_i a_{ij} = b_j(1 - c_j)$ for $j = 1, 2, 3$.

Carrying out this programme might present some difficulties. For example, if in step (a) the c_i are not distinct, then there might not exist a solution in step (b). It might also happen that the value of b_4, found in step (b), is zero, and this will make it impossible to carry out either step (c) or step (d). Even if a solution exists for the sub-problem that arises in each step, the solution might not be unique, and there could turn out to be a family of solutions. The general solution, which is valid except in these exceptional cases, is given by the following coefficients:

$$a_{21} = c_2,$$

$$a_{31} = \frac{c_3(c_3 + 4c_2^2 - 3c_2)}{2c_2(2c_2 - 1)},$$

$$a_{32} = -\frac{c_3(c_3 - c_2)}{2c_2(2c_2 - 1)},$$

$$a_{41} = \frac{-12c_3 c_2^2 + 12c_3^2 c_2^2 + 4c_2^2 - 6c_2 + 15c_2 c_3 - 12c_3^2 c_2 + 2 + 4c_3^2 - 5c_3}{2c_2 c_3(-4c_3 + 6c_3 c_2 + 3 - 4c_2)},$$

$$a_{42} = \frac{(c_2 - 1)(4c_3^2 - 5c_3 + 2 - c_2)}{2c_2(c_3 - c_2)(-4c_3 + 6c_3 c_2 + 3 - 4c_2)},$$

$$a_{43} = -\frac{(2c_2 - 1)(c_2 - 1)(c_3 - 1)}{c_3(c_3 - c_2)(-4c_3 + 6c_3 c_2 + 3 - 4c_2)},$$

$$b_1 = \frac{6c_3c_2 - 2c_3 - 2c_2 + 1}{12c_3c_2},$$

$$b_2 = -\frac{(2c_3 - 1)}{12c_2(c_2 - 1)(c_3 - c_2)},$$

$$b_3 = \frac{(2c_2 - 1)}{12c_3(c_2 - c_3c_2 + c_3^2 - c_3)},$$

$$b_4 = \frac{-4c_3 + 6c_3c_2 + 3 - 4c_2}{12(c_3 - 1)(c_2 - 1)}.$$

Kutta identified five special cases where a solution is certain to exist:

$$
\begin{array}{ll}
\text{I} & c_2 \notin \{0, \tfrac{1}{2}, \tfrac{1}{2} \pm \tfrac{\sqrt{3}}{6}, 1\}, \quad c_3 = 1 - c_2, \\
\text{II} & b_2 = 0, \quad c_2 \neq 0, \quad c_3 = \tfrac{1}{2}, \\
\text{III} & b_3 \neq 0, \quad c_2 = \tfrac{1}{2}, \quad c_3 = 0, \\
\text{IV} & b_4 \neq 0, \quad c_2 = 1, \quad c_3 = \tfrac{1}{2}, \\
\text{V} & b_3 \neq 0, \quad c_2 = c_3 = \tfrac{1}{2}.
\end{array}
$$

The coefficient tableaux are for case I,

$$
\begin{array}{c|cccc}
0 \\
1-c_3 & 1-c_3 \\
c_3 & \frac{c_3(1-2c_3)}{2(1-c_3)} & \frac{c_3}{2(1-c_3)} \\
1 & \frac{12c_3^3-24c_3^2+17c_3-4}{2(1-c_3)(6c_3-1-6c_3^2)} & \frac{c_3(1-2c_3)}{2(1-c_3)(6c_3-1-6c_3^2)} & \frac{1-c_3}{6c_3-1-6c_3^2} \\
\hline
& \frac{6c_3-1-6c_3^2}{12c_3(1-c_3)} & \frac{1}{12c_3(1-c_3)} & \frac{1}{12c_3(1-c_3)} & \frac{6c_3-1-6c_3^2}{12c_3(1-c_3)}
\end{array} \quad ;
$$

(322e)

for case II,

$$
\begin{array}{c|cccc}
0 \\
c_2 & c_2 \\
\frac{1}{2} & \frac{1}{2} - \frac{1}{8c_2} & \frac{1}{8c_2} \\
1 & \frac{1}{2c_2} - 1 & -\frac{1}{2c_2} & 2 \\
\hline
& \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6}
\end{array} \quad ;
$$

(322f)

for case III,

$$
\begin{array}{c|cccc}
0 \\
\frac{1}{2} & \frac{1}{2} \\
0 & -\frac{1}{12b_3} & \frac{1}{12b_3} \\
1 & -\frac{1}{2} - 6b_3 & \frac{3}{2} & 6b_3 \\
\hline
& \frac{1}{6} - b_3 & \frac{2}{3} & b_3 & \frac{1}{6}
\end{array} \quad ;
$$

(322g)

for case IV,

$$
\begin{array}{c|cccc}
0 \\
1 & 1 \\
\frac{1}{2} & \frac{3}{8} & \frac{1}{8} \\
1 & 1 - \frac{1}{4b_4} & -\frac{1}{12b_4} & \frac{1}{3b_4} \\
\hline
& \frac{1}{6} & \frac{1}{6} - b_4 & \frac{2}{3} & b_4
\end{array} \;; \tag{322h}
$$

and for case V,

$$
\begin{array}{c|cccc}
0 \\
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} - \frac{1}{6b_3} & \frac{1}{6b_3} \\
1 & 0 & 1 - 3b_3 & 3b_3 \\
\hline
& \frac{1}{6} & \frac{2}{3} - b_3 & b_3 & \frac{1}{6}
\end{array} \;. \tag{322i}
$$

Some interesting special choices within these cases are $c_3 = \frac{2}{3}$ in case I,

$$
\begin{array}{c|cccc}
0 \\
\frac{1}{3} & \frac{1}{3} \\
\frac{2}{3} & -\frac{1}{3} & 1 \\
1 & 1 & -1 & 1 \\
\hline
& \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8}
\end{array} \;,
$$

and $c_2 = \frac{1}{4}$ in case II,

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

A further, and somewhat eccentric, special choice in case II, is $c_2 = -\frac{1}{2}$:

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

The interest in this method, as for a similar method with $c_2 = -1$, is that it is possible to eliminate one stage of computation, by replacing F_2 by a quantity found in the *previous* step. The method contrived in this way is no longer a Runge–Kutta method, and has poorer stability, but it is more efficient in terms of order achieved per stages computed.

We also present the choices $b_3 = \frac{1}{12}$ in case III,

$$
\begin{array}{c|cccc}
0 \\
\frac{1}{2} & \frac{1}{2} \\
0 & -1 & 1 \\
1 & -1 & \frac{3}{2} & \frac{1}{2} \\
\hline
 & \frac{1}{12} & \frac{2}{3} & \frac{1}{12} & \frac{1}{6}
\end{array}
\quad,
$$

and $b_4 = \frac{1}{6}$ in case IV,

$$
\begin{array}{c|cccc}
0 \\
1 & 1 \\
\frac{1}{2} & \frac{3}{8} & \frac{1}{8} \\
1 & -\frac{1}{2} & -\frac{1}{2} & 2 \\
\hline
 & \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6}
\end{array}
\quad.
$$

Amongst the methods in case V, the 'classical Runge–Kutta method' is especially notable. The tableau is

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

Also in case V is a special method derived by Gill (1951), for the special purpose of reducing memory requirements for large problems. Gill found that by using a value $b_3 = \frac{1}{3} + \frac{\sqrt{2}}{6}$, or the conjugate of this which was rejected as having larger errors, it was possible to solve an N-dimensional system using only $3N$ stored numbers. For a general method with $s = p = 4$, the corresponding memory needs are $4N$. The tableau for Gill's method is

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

and is characterized by the condition

$$
\det \left(\begin{bmatrix} 1 & a_{31} & a_{32} \\ 1 & a_{41} & a_{42} \\ 1 & b_1 & b_2 \end{bmatrix} \right) = 0
$$

which, for a method in case V, imposes the constraint

$$18b_3^2 - 12b_3 + 1 = 0,$$

with solutions

$$b_3 = \frac{2 \pm \sqrt{2}}{6}.$$

323   New methods from old

As we seek explicit Runge–Kutta methods of higher and higher order, we observe relationships between methods of two adjacent orders. For example, fourth order methods are connected in a special way with certain methods with only three stages, but with a modified type of third order condition. Denote the fourth order method by

$$
\begin{array}{c|c}
c & A \\
\hline
 & b^\mathsf{T}
\end{array}
=
\begin{array}{c|cccc}
0 & & & & \\
c_2 & a_{21} & & & \\
c_3 & a_{31} & a_{32} & & \\
1 & a_{41} & a_{42} & a_{43} & \\
\hline
 & b_1 & b_2 & b_3 & b_4
\end{array}
\tag{323a}
$$

and consider the three-stage tableau

$$
\begin{array}{c|c}
\widetilde{c} & \widetilde{A} \\
\hline
 & \widetilde{b}^\mathsf{T}
\end{array}
=
\begin{array}{c|ccc}
0 & & & \\
c_2 & a_{21} & & \\
c_3 & a_{31} & a_{32} & \\
\hline
 & b_1 & b_2(1-c_2) & b_3(1-c_3)
\end{array}.
\tag{323b}
$$

If we denote the elementary weights for the new method (323b) by $\widetilde{\Phi}$, we find for the trees with order up to 3,

$$\widetilde{\Phi}(\bullet) = \frac{1}{2} = \frac{1}{(r(\bullet)+1)\gamma(\bullet)}, \tag{323c}$$

$$\widetilde{\Phi}(\mathord{\text{\textbf{!}}}) = \frac{1}{6} = \frac{1}{(r(\mathord{\text{\textbf{!}}})+1)\gamma(\mathord{\text{\textbf{!}}})}, \tag{323d}$$

$$\widetilde{\Phi}(\mathbf{V}) = \frac{1}{12} = \frac{1}{(r(\mathbf{V})+1)\gamma(\mathbf{V})}, \tag{323e}$$

$$\widetilde{\Phi}\left(\mathord{\text{\textbf{!}}}\right) = \frac{1}{24} = \frac{1}{\left(r\left(\mathord{\text{\textbf{!}}}\right)+1\right)\gamma\left(\mathord{\text{\textbf{!}}}\right)}. \tag{323f}$$

The conclusion that $\widetilde{\Phi}(t) = 1/\big((r(t)+1)\gamma(t)\big)$ is not in the least remarkable. In fact, such a conclusion will always hold if $\widetilde{b}^\mathsf{T} = b^\mathsf{T}A$, with obvious

adjustments made to c and A to form $\widetilde{c}$ and $\widetilde{A}$, but our interest here is in working in the opposite direction, from order 3 to order 4. If $\sum_{i=1}^{s} b_i = 1$ is satisfied for the four-stage method (323a), then the remainder of the order conditions are satisfied as a consequence of (323c)–(323f) and of the $D(1)$ assumption. We check these as follows, where the relevant trees are also shown:

$$
\begin{aligned}
\bullet \qquad & \sum_{i=1}^{s} b_i = 1, \\[4pt]
& \sum_{i=1}^{s} b_i c_i = \sum_{i=1}^{s} b_i - \sum_{i=1}^{s} b_i(1-c_i) && = 1 - \sum_{i=1}^{s} \widetilde{b}_i && = \tfrac{1}{2}, \\[4pt]
& \sum_{i=1}^{s} b_i c_i^2 = \sum_{i=1}^{s} b_i c_i - \sum_{i=1}^{s} b_i(1-c_i)c_i && = \tfrac{1}{2} - \sum_{i=1}^{s} \widetilde{b}_i c_i && = \tfrac{1}{3}, \\[4pt]
& \sum_{i,j=1}^{s} b_i a_{ij} c_j = \sum_{j=1}^{s} \widetilde{b}_j c_j && = \tfrac{1}{6}, \\[4pt]
& \sum_{i=1}^{s} b_i c_i^3 = \sum_{i=1}^{s} b_i c_i^2 - \sum_{i=1}^{s} b_i(1-c_i)c_i^2 && = \tfrac{1}{3} - \sum_{i=1}^{s} \widetilde{b}_i c_i^2 && = \tfrac{1}{4}, \\[4pt]
& \sum_{i,j=1}^{s} b_i c_i a_{ij} c_j = \sum_{i=1,j} b_i a_{ij} c_j - \sum_{i=1,j} \widetilde{b}_i a_{ij} c_j && = \tfrac{1}{8}, \\[4pt]
& \sum_{i,j=1}^{s} b_i a_{ij} c_j^2 = \sum_{j=1}^{s} \widetilde{b}_j c_j^2 && = \tfrac{1}{12}, \\[4pt]
& \sum_{i,j,k=1}^{s} b_i a_{ij} a_{jk} c_k = \sum_{j,k=1}^{s} \widetilde{b}_j a_{jk} c_k && = \tfrac{1}{24}.
\end{aligned}
$$

It is not possible to extend the principle illustrated in this result to higher orders without making some additional assumptions. We introduce the idea we need as follows:

Definition 323A *Consider a Runge–Kutta method given by the tableau*

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

For a tree t and stage i, let $\Phi_i(t)$ denote the elementary weight associated with t for the tableau

$$
\begin{array}{c|c}
c & A \\ \hline
 & e_i^{\mathsf{T}} A
\end{array}.
$$

Stage i has 'internal order q', if for all trees such that $r(t) \le q$,

$$
\Phi_i(t) = \frac{c_i^{r(t)}}{\gamma(t)}.
$$

The significance of this definition is that if stage i has internal order q, then, in any step with initial value $y_{n-1} = y(x_{n-1})$, the value computed in stage i satisfies $Y_i = y(x_{n-1} + hc_i) + O(h^{q+1})$. Note that the $C(q)$ condition is

necessary and sufficient for *every* stage to have internal order q, and this is possible only for implicit methods.

We are now in a position to generalize the remarks we have made about third and fourth order methods.

Theorem 323B *Let*

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

denote a Runge–Kutta method with $s-1$ stages and generalized order $p-1$, satisfying $\widetilde{c}_{s-1} \neq 1$. Let q be an integer such that $2q+2 \geq p$ and suppose that for any $i \in S \subset \{1, 2, \ldots, s-1\}$, the method has internal order q. If there exists $b \in \mathbb{R}^s$, with $b_s \neq 0$ such that

$$\sum_{i=1}^{s} b_i = 1, \tag{323g}$$

and such that $b_i \neq 0$ implies $i \in S$, $c_i \neq 1$ and $b_i(1 - c_i) = \widetilde{b}_i$, then the s-stage method

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

has order p, where $c^{\mathsf{T}} = [\ \widetilde{c}^{\mathsf{T}} \quad 1\ ]$ and the $s \times s$ matrix A is formed from $\widetilde{A}$ by adding an additional row with component $j \in \{1, 2, \ldots, s-1\}$ equal to $\left(\widetilde{b}_j - \sum_{i=1}^{s-1} b_i a_{ij}\right)/b_s$ and then adding an additional column of s zeros.

Proof. The case $p = 1$ follows from (323g), so we consider instead the case $p \geq 2$. Also, without loss of generality we assume that $1 \leq q \leq p - 1$, because internal order 1 is equivalent to $c_i = \sum_{j=1}^{s} a_{ij}$ and because $q \geq p$ implies internal order $p - 1$. We first prove that

$$\sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \qquad k = 1, 2, \ldots, p.$$

For $k = 1$ the result is equivalent to (323g). If the result has been proved for $k - 1 < p$, we verify it for k, thus completing an induction argument. We have

$$\sum_{i=1}^{s} b_i c_i^{k-1} = \sum_{i=1}^{s} b_i c_i^{k-2} - \sum_{i=1}^{s} \widetilde{b}_i c_i^{k-2} = \frac{1}{k-1} - \frac{1}{k(k-1)} = \frac{1}{k}.$$

The next step is to extend the internal order property to stage s. Write the

value of $\Phi_i(t)$ as $\sum_{j=1}^{s} a_{ij}\chi_j$. We then have

$$\frac{1}{\gamma(t)(r(t)+1)} = \sum_{j=1}^{s} \widetilde{b}_j\chi_j$$

$$= \sum_{i,j=1}^{s} b_i a_{ij}\chi_j$$

$$= b_s\left(\sum_{j=1}^{s} a_{sj}\chi_j - \frac{1}{\gamma(t)}\right) + \sum_{i=1}^{s} b_i\frac{c_i^{r(t)}}{\gamma(t)}$$

$$= b_s\left(\sum_{j=1}^{s} a_{sj}\chi_j - \frac{1}{\gamma(t)}\right) + \frac{1}{\gamma(t)(r(t)+1)},$$

implying that

$$\sum_{j=1}^{s} a_{sj}\chi_j = \frac{1}{\gamma(t)}.$$

Next we prove the order condition for a tree of the form $[\tau^{k-1}t_1]$ where $k + r(t_1) \le p$. We write $\Phi(t_1) = \sum_{i=1}^{s} b_i\chi_i$. For $k = 1$ we have

$$\Phi(t) = \sum_{i,j=1}^{s} b_i a_{ij}\chi_j = \sum_{j=1}^{s} \widetilde{b}_j\chi_j = \frac{1}{\gamma(t_1)(r(t_1)+1)} = \frac{1}{\gamma(t)}.$$

Now assume that $k > 1$ and that the result has been proved when k is replaced by $k - 1$. For the rest of this proof, we write $\Phi([t_1]) = \sum_{i=1}^{s} b_i\chi_i$. We have $b_i c_i^{k-1} = b_i c_i^{k-2} - \widetilde{b}_i c_i^{k-2}$ and hence

$$\Phi(t) = \Phi([\tau^{k-1}t_1])$$

$$= \sum_{i=1}^{s} b_i c_i^{k-1}\chi_i$$

$$= \sum_{i=1}^{s} b_i c_i^{k-2}\chi_i - \sum_{i=1}^{s} \widetilde{b}_i c_i^{k-2}\chi_i$$

$$= \frac{1}{\gamma(t_1)(r(t)-1)} - \frac{1}{\gamma(t_1)r(t)(r(t)-1)}$$

$$= \frac{1}{\gamma(t_1)r(t)}$$

$$= \frac{1}{\gamma(t)}.$$

Finally, we consider a tree of the form $t = [t_1 t_2 \cdots t_m]$, where $r(t_1) \geq r(t_2) \geq \cdots \geq r(t_m)$. Because $2q + 2 \geq p$, $r(t_k) \leq q$ for $k = 2, 3, \ldots, m$. We now have

$$
\begin{aligned}
\Phi(t) &= \Phi([t_1 t_2 \cdots t_m]) \\
&= \sum_{i=1}^{s} b_i \chi_i \prod_{k=2}^{m} \frac{c_i^{r(t_k)}}{\gamma(t_k)} \\
&= \sum_{i=1}^{s} b_i \chi_i c_i^{r(t)-r(t_1)-1} \frac{1}{\prod_{k=2}^{m} \gamma(t_k)} \\
&= \frac{1}{\prod_{k=2}^{m} \gamma(t_k)} \Phi([\tau^{r(t)-r(t_1)-1} t_1]) \\
&= \frac{1}{r(t)\gamma(t_1) \prod_{k=2}^{m} \gamma(t_k)} \\
&= \frac{1}{\gamma(t)}. \qquad\qquad\qquad\qquad\qquad\qquad\qquad \square
\end{aligned}
$$

Before we consider how to extend the benefits of Theorem 323B beyond the gain of a single order, we look again at the generalized order conditions

$$
\widetilde{\widetilde{\Phi}}(t) = \frac{1}{(r(t)+1)\gamma(t)}. \tag{323h}
$$

Because the series

$$
y(x_0) + \sum_{t \in T} \frac{\xi^{r(t)} h^{r(t)}}{\gamma(t)\sigma(t)} F(t)(y(x_0))
$$

represents the solution of

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

at $x = x_0 + \xi h$, we find by integrating term by term, from $\xi = 0$ to $\xi = 1$, that $h^{-1} \int_{x_0}^{x_0+h} y(x)dx$ has Taylor expansion

$$
y(x_0) + \sum_{t \in T} \frac{h^{r(t)}}{(r(t)+1)\gamma(t)\sigma(t)} F(t)(y(x_0)). \tag{323i}
$$

Hence a method satisfying (323h) for $r(t) \leq p$ agrees with (323i) to within $O(h^{p+1})$.

We can generalize the meaning of order further by replacing the single integral by the double integral

$$
\int_0^1 \int_0^{\overline{\xi}} y(x_0 + \xi h)d\xi d\overline{\xi},
$$

and we now find

$$h^{-2}\int_{x_0}^{x_0+h}\int_{x_0}^{\overline{x}} y(x)dxd\overline{x}=\tfrac{1}{2}y(x_0)+\sum_{t\in T}\frac{h^{r(t)}}{(r(t)+1)(r(t)+2)\gamma(t)\sigma(t)}F(t)(y(x_0)).$$

For a method with generalized order conditions, it might seem possible to carry out the process of reducing to one less stage and the second generalization of the order conditions, but this is of little value. When we have recovered the method with the first generalization, the last abscissa will have value 1, and it will not be possible to go further to recover a method satisfying the standard order conditions.

However, this difficulty can be overcome, to some extent, by setting the last component of the abscissa vector of the first generalized method to 0 rather than to 1, with appropriate modifications made to the method of recovery. To see how this works, consider the method with first level of generalized order equal to 3 whose tableau is

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

Note that this method was constructed to satisfy not only the four generalized order conditions

$$b^{\mathsf{T}}\mathbf{1}=\frac{1}{2},\quad b^{\mathsf{T}}c=\frac{1}{6},\quad b^{\mathsf{T}}c^2=\frac{1}{12},\quad b^{\mathsf{T}}Ac=\frac{1}{24},$$

but also the condition

$$\sum_{i=1}^{4}\frac{b_i}{1-c_i}=1,$$

which is imposed in anticipation of our intention to construct a fourth order method by adding an additional stage. The new method is

$$
\begin{array}{c|ccccc}
0 \\
\frac{1}{4} & \frac{1}{4} \\
\frac{1}{2} & \frac{1}{2} & 0 \\
\frac{3}{4} & 0 & \frac{1}{2} & \frac{1}{4} \\
0 & 0 & \frac{1}{6\beta} & -\frac{1}{3\beta} & \frac{1}{6\beta} \\
\hline
& -\beta & \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} & \beta
\end{array}
$$

and it is an easy matter to check that all the fourth order conditions are satisfied for any choice of the non-zero parameter β.

324   *Order barriers*

It is possible, as we have seen, to derive explicit methods with $s = p$ for $p = 1, 2, 3, 4$. These methods are optimal in the sense that $s < p$ is never possible and $p = 4$ is as high as it is possible to go with $s = p$. We now formalize these remarks.

Theorem 324A *If an explicit s-stage Runge–Kutta method has order p, then $s \geq p$.*

Proof. Let $t = [[\cdots [t] \cdots]]$ such that $r(t) = p > s$. The order condition associated with this tree is $\Phi(t) = 1/\gamma(t)$, where $\gamma(t) = p!$ and $\Phi(t) = b^{\mathsf{T}} A^{p-1} \mathbf{1}$. Because A is strictly lower triangular, $A^p = 0$. Hence, the order condition becomes $0 = 1/p!$, which has no solution. $\square$

Theorem 324B *If an explicit s-stage Runge–Kutta method has order $p \geq 5$, then $s > p$.*

Proof. Assume $s = p$. Evaluate the values of the following four expressions:

$$b^{\mathsf{T}} A^{p-4}(C - c_4 I)(C - c_2 I)c = \frac{6}{p!} - \frac{2(c_2 + c_4)}{(p-1)!} + \frac{c_2 c_4}{(p-2)!}, \qquad (324a)$$

$$b^{\mathsf{T}} A^{p-4}(C - c_4 I)Ac = \frac{3}{p!} - \frac{c_4}{(p-1)!}, \qquad (324b)$$

$$b^{\mathsf{T}} A^{p-4} A(C - c_2 I)c = \frac{2}{p!} - \frac{c_2}{(p-1)!}, \qquad (324c)$$

$$b^{\mathsf{T}} A^{p-4} A^2 c = \frac{1}{p!}. \qquad (324d)$$

From the left-hand sides of these expressions we observe that $(324a) \times (324d) = (324b) \times (324c)$. Evaluate the right-hand sides, and we find that

$$\left(\frac{6}{p!} - \frac{2(c_2 + c_4)}{(p-1)!} + \frac{c_2 c_4}{(p-2)!} \right) \left(\frac{1}{p!} \right) = \left(\frac{3}{p!} - \frac{c_4}{(p-1)!} \right) \left(\frac{2}{p!} - \frac{c_2}{(p-1)!} \right),$$

which simplifies to $c_2(c_4 - 1) = 0$.

Now consider the four expressions

$$b^{\mathsf{T}} A^{p-5}(C - c_5 I)A(C - c_2 I)c = \frac{8}{p!} - \frac{3c_2 + 2c_5}{(p-1)!} + \frac{c_2 c_5}{(p-2)!}, \qquad (324e)$$

$$b^{\mathsf{T}} A^{p-5}(C - c_5 I)A^2 c = \frac{4}{p!} - \frac{c_5}{(p-1)!}, \qquad (324f)$$

$$b^{\mathsf{T}} A^{p-5} A^2(C - c_2 I)c = \frac{2}{p!} - \frac{c_2}{(p-1)!}, \qquad (324g)$$

$$b^{\mathsf{T}} A^{p-5} A^3 c = \frac{1}{p!}. \qquad (324h)$$

Again we see that $(324e) \times (324h) = (324f) \times (324g)$, so that evaluating the right-hand sides, we find

$$\left(\frac{8}{p!} - \frac{3c_2 + 2c_5}{(p-1)!} + \frac{c_2 c_5}{(p-2)!} \right) \left(\frac{1}{p!} \right) = \left(\frac{4}{p!} - \frac{c_5}{(p-1)!} \right) \left(\frac{2}{p!} - \frac{c_2}{(p-1)!} \right),$$

leading to $c_2(c_5 - 1) = 0$. Since we cannot have $c_2 = 0$, it follows that $c_4 = c_5 = 1$. Now evaluate $b^{\mathsf{T}} A^{p-5}(C - e) A^2 c$. This equals $(4 - p)/p!$ by the order conditions but, in contradiction to this, it equals zero because component number i of $b^{\mathsf{T}} A^{p-5}$ vanishes unless $i \leq 5$. However, these components of $(C - e) A^2 c$ vanish. $\square$

The bound $s - p \geq 1$, which applies for $p \geq 5$, is superseded for $p \geq 7$ by $s - p \geq 2$. This is proved in Butcher (1965a). For $p \geq 8$ we have the stronger bound $s - p \geq 3$ (Butcher, 1985). It seems likely that the minimum value of $s - p$ rises steadily as p increases further, but there are no published results dealing with higher orders. On the other hand, it is known, because of the construction of a specific method (Hairer, 1978), that $p = 10$, $s = 17$ is possible.

That a sufficiently high s can be found to achieve order p follows immediately from Theorem 317A. We now derive an upper bound on the minimum value of such an s. This is done by constructing methods with odd orders, or methods satisfying the generalization of odd orders introduced in Subsection 323. In the latter case, we then use the results of that subsection to extend the result to the next even order higher.

Theorem 324C *For any positive integer p, an explicit Runge–Kutta method exists with order p and s stages, where*

$$s = \begin{cases} \frac{3p^2 - 10p + 24}{8}, & p \text{ even}, \\ \frac{3p^2 - 4p + 9}{8}, & p \text{ odd}. \end{cases}$$

Proof. We consider the case of p odd, but allow for generalized order conditions. If $p = 1 + 2m$, we construct first an implicit Runge–Kutta method with $1 + m$ stages, using (case I) standard order conditions and (case II) generalized order conditions. For case I, the order condition associated with the tree t is, as usual,

$$\Phi(t) = \frac{1}{\gamma(t)}.$$

In case II, this condition is replaced by

$$\Phi(t) = \frac{1}{(r(t) + 1)\gamma(t)}.$$

For the implicit method, the abscissae are at the zeros of the polynomial

$$\frac{d^m}{dx^m} x^{m+1}(x-1)^m, \text{ in case I,}$$

$$\frac{d^m}{dx^m} x^{m+1}(x-1)^{m+1}, \text{ in case II,}$$

with the zero $x = 1$ omitted in case II. It is clear that $x = 0$ is a zero in both cases and that the remaining zeros are distinct and lie in the interval $[0, 1)$. Denote the positive zeros by ξ_i, $i = 1, 2, \ldots, m$. We now construct methods with abscissae chosen from the successive rows of the following table:

row 0	0				
row 1	ξ_1				
row 2	ξ_1	ξ_2			
row 3	ξ_1	ξ_2	ξ_3		
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
row m	ξ_1	ξ_2	ξ_3	$\cdots$	ξ_m
row $m+1$	ξ_1	ξ_2	ξ_3	$\cdots$	ξ_m
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
row $2m$	ξ_1	ξ_2	ξ_3	$\cdots$	ξ_m

where there are exactly $m + 1$ repetitions of the rows with m members. The total number of stages will then be

$$s = 1 + \left(1 + 2 + \cdots + (m-1)\right) + (m+1)m = \frac{1}{2}(3m^2 + m + 2).$$

Having chosen $c = \begin{pmatrix} 0 & \xi_1 & \xi_1 & \xi_2 & \cdots & \xi_m \end{pmatrix}^{\mathsf{T}}$, we construct b^{T} with all components zero except the first component and the final m components. The non-zero components are chosen so that

$$b_1 + \sum_{i=1}^{m} b_{s-m+i} = \begin{cases} 1, & \text{case I} \\ \frac{1}{2}, & \text{case II} \end{cases}$$

$$\sum_{i=1}^{m} b_{s-m+i} \xi_i^{k-1} = \begin{cases} \frac{1}{k}, & \text{case I} \\ \frac{1}{k(k+1)}, & \text{case II} \end{cases}, \quad k = 1, 2, \ldots, 2m + 1.$$

The possibility that the non-zero b components can be found to satisfy these conditions follows from the theory of Gaussian quadrature. The final step in the construction of the method is choosing the elements of the matrix A. For i corresponding to a member of row k for $k = 1, 2, \ldots, m$, the only non-zero

a_{ij} are for $j = 1$ and for j corresponding to a member of row $k - 1$. Thus, the quadrature formula associated with this row has the form

$$\int_0^{c_i} \phi(x)dx \approx w_0\phi(0) + \sum_{j=1}^{k-1} w_j\phi(\xi_j)$$

and the coefficients are chosen to make this exact for ϕ a polynomial of degree $k - 1$. For i a member of row $k = m + 1, m + 2, \ldots, 2m$, the non-zero a_{ij} are found in a similar way based on the quadrature formula

$$\int_0^{c_i} \phi(x)dx \approx w_0\phi(0) + \sum_{j=1}^{m} w_j\phi(\xi_j).$$

The method constructed in this way has order, or generalized order, respectively, equal to $p = 2m+1$. To see this, let $\widehat{Y}_i$ denote the approximation to $y(x_{n-1} + h\xi_i)$ in stage $1 + i$ of the order $2m + 1$ Radau I method (in case I) or the order $2m + 2$ Lobatto method (in case II). It is easy to see that the stages corresponding to row k approximate the $\widehat{Y}$ quantities to within $O(h^{k+1})$. Thus the full method has order $2m + 1$ in case I and generalized order $2m + 1$ in case II. Add one more stage to the case II methods, as in Theorem 323B, and we obtain order $p = 2m + 2$ with $s = \frac{1}{2}(3m^2 + m + 4)$ stages compared with $p = 2m + 1$ and $s = \frac{1}{2}(3m^2 + m + 2)$ stages in case I. This gives the result of the theorem.   □

325   Methods of order 5

We saw in Theorem 324B that for orders greater than 4, $s = p$ is impossible. Hence, we assume that $s = 6$. We assume the $D(1)$ condition and the $C(2)$ condition applied to all stages except the second. We also need to assume the subsidiary conditions $b_2 = \sum_{i=3}^{5} b_i(1 - c_i)a_{i2} = 0$. These conditions dispose of all conditions except those associated with the trees
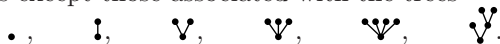
The second and third of these turn out to be consequences of the $D(1)$ and $C(2)$ conditions, and we find that some of the elements in the final row can be evaluated in two different but consistent ways. The condition associated with can be replaced by the difference of this condition and the automatically satisfied condition associated with ; see (325h) below. This last modification of the order conditions we actually solve has the advantage that it removes the last row of the A matrix from the calculation until, at the end, we compute this row using the $D(1)$ condition.

Collecting these comments together, we summarize the defining equations for a fifth order method. Where we write 'choose' one of the coefficients, we mean that it can be set to an arbitrary value, excluding only a finite set of

possibilities. We do not state in detail what constitute the exceptional cases, but these can be identified with little difficulty:

$$c_6 = 1, \tag{325a}$$

$$\text{Choose } c_2, c_3, c_4, c_5, \tag{325b}$$

$$\sum_{i=1}^{6} b_i c_i^{k-1} = \frac{1}{k}, \qquad k = 1, 2, \ldots, 5, \tag{325c}$$

$$\text{Choose } a_{42}, \tag{325d}$$

$$\sum_{j=2}^{i-1} a_{ij} c_j = \frac{1}{2} c_i^2, \qquad i = 3, 4, 5, \tag{325e}$$

$$\sum_{j=1}^{i-1} a_{ij} = c_i, \qquad i = 2, 3, 4, 5, \tag{325f}$$

$$\sum_{i=3}^{5} b_i (1 - c_i) a_{i2} = 0, \tag{325g}$$

$$b_5 (1 - c_5) a_{54} c_4 (c_4 - c_3) = \frac{1}{60} - \frac{c_3}{24}, \tag{325h}$$

$$\sum_{i=j+1}^{6} b_i a_{ij} = b_j (1 - c_j), \qquad j = 1, 2, 3, 4, 5. \tag{325i}$$

The following schema shows which of these various defining equations are used in the choice of particular coefficients of the method:

$$
\begin{array}{c|cccccc}
0 & & & & & & \\
(325b) & (325f) & & & & & \\
(325b) & (325f) & (325e) & & & & \\
(325b) & (325f) & (325d) & (325e) & & & \\
(325b) & (325f) & (325g) & (325e) & (325h) & & \\
(325a) & (325i) & (325i) & (325i) & (325i) & (325i) & \\
\hline
 & (325c) & 0 & (325c) & (325c) & (325c) & (325c)
\end{array}
$$

We give a single example of a method derived in this manner:

$$
\begin{array}{c|cccccc}
0 & & & & & & \\
\frac{1}{4} & \frac{1}{4} & & & & & \\
\frac{1}{4} & \frac{1}{8} & \frac{1}{8} & & & & \\
\frac{1}{2} & 0 & 0 & \frac{1}{2} & & & \\
\frac{3}{4} & \frac{3}{16} & -\frac{3}{8} & \frac{3}{8} & \frac{9}{16} & & \\
1 & -\frac{3}{7} & \frac{8}{7} & \frac{6}{7} & -\frac{12}{7} & \frac{8}{7} & \\
\hline
 & \frac{7}{90} & 0 & \frac{16}{45} & \frac{2}{15} & \frac{16}{45} & \frac{7}{90}
\end{array}.
$$

The first methods of this order, derived by Kutta (1901), have a different structure. One of these, as corrected by Nyström (1925), is

$$
\begin{array}{c|ccccccc}
0 \\
\frac{1}{3} & \frac{1}{3} \\
\frac{2}{5} & \frac{4}{25} & \frac{6}{25} \\
1 & \frac{1}{4} & -3 & \frac{15}{4} \\
\frac{2}{3} & \frac{2}{27} & \frac{10}{9} & -\frac{50}{81} & \frac{8}{81} \\
\frac{4}{5} & \frac{2}{25} & \frac{12}{25} & \frac{2}{15} & \frac{8}{75} & 0 \\
\hline
& \frac{23}{192} & 0 & \frac{125}{192} & 0 & -\frac{27}{64} & \frac{125}{192}
\end{array}
$$

As we have pointed out in Subsection 316, the order conditions for a scalar first order differential equation are less restrictive than for the general vector case, if orders of 5 or more are under consideration. This suggests the existence of methods whose orders, when applied to a single first order differential equation, may be 5, whereas it is only 4 when applied to a higher-dimensional system. An example of such a method is given in Butcher (1995).

326   Methods of order 6

The first methods of order 6 were derived by Huťa (1956, 1957). Although his methods used $s = 8$ stages, it is possible to find methods of this order with $s = 7$. Just as for order 5, we assume the modified $C(2)$ condition and the $D(1)$ condition. We also assume the quadrature conditions so that the only order conditions that remain are $\Phi(t) = 1/\gamma(t)$ for the trees

$$
t = \qquad \mathbf{Y}, \qquad \mathbf{Y}, \qquad \mathbf{Y} \qquad \text{and} \qquad \mathbf{Y}.
$$

Linear combinations of these with other order conditions whose truth is automatic appear in (326h)–(326k) below, where we have listed all the conditions we need to specify a method:

$$
b_2 = 0, \tag{326a}
$$

$$
\sum_{i=1}^{7} b_i(1-c_i)(c_i-c_6)(c_i-c_3)(c_i-c_4)c_i =
\frac{1}{30} - \frac{c_3+c_4+c_6}{20} + \frac{c_3c_4+c_3c_6+c_4c_6}{12} - \frac{c_3c_4c_6}{6}, \tag{326b}
$$

$$
\sum_{i=1}^{7} b_i(1-c_i)(c_i-c_6)(c_i-c_4)c_i = \frac{1}{20} - \frac{c_4+c_6}{12} + \frac{c_4c_6}{6}, \tag{326c}
$$

$$
\sum_{i=1}^{7} b_i(1-c_i)(c_i-c_6)c_i = \frac{1}{12} - \frac{c_6}{6}, \tag{326d}
$$

$$\sum_{i=1}^{7} b_i(1 - c_i)c_i = \frac{1}{6}, \tag{326e}$$

$$\sum_{i=1}^{7} b_i c_i = \frac{1}{2}, \tag{326f}$$

$$\sum_{i=1}^{7} b_i = 1, \tag{326g}$$

$$\sum_{i,j=1}^{7} b_i(1 - c_i)a_{ij}(c_j - c_3)c_j = \frac{1}{60} - \frac{c_3}{24}, \tag{326h}$$

$$\sum_{i,j=1}^{7} b_i(1 - c_i)(c_i - c_6)a_{ij}(c_j - c_3)c_j = \frac{1}{90} - \frac{c_3}{40} - \frac{c_6}{60} + \frac{c_3 c_6}{24}, \tag{326i}$$

$$\sum_{i,j=1}^{7} b_i(1 - c_i)a_{ij}(c_j - c_4)(c_j - c_3)c_j = \frac{1}{120} - \frac{c_3 + c_4}{60} + \frac{c_3 c_4}{24}, \tag{326j}$$

$$\sum_{i,j,k=1}^{7} b_i(1 - c_i)a_{ij}a_{jk}(c_k - c_3)c_k = \frac{1}{360} - \frac{c_3}{120}, \tag{326k}$$

$$\sum_{j=1}^{7} a_{ij}c_j = \frac{1}{2}c_i^2, \qquad i \neq 2, \tag{326l}$$

$$\sum_{j=1}^{7} a_{ij} = c_i, \qquad i = 1, 2, \ldots, 7, \tag{326m}$$

$$\sum_{i=1}^{7} b_i a_{ij} = b_i(1 - c_j), \; j = 1, 2, \ldots, 7, \tag{326n}$$

$$\sum_{i=1}^{7} b_i(1 - c_i)a_{i2} = 0, \tag{326o}$$

$$\sum_{i=1}^{7} b_i(1 - c_i)c_i a_{i2} = 0, \tag{326p}$$

$$\sum_{i,j=1}^{7} b_i(1 - c_i)a_{ij}a_{j2} = 0. \tag{326q}$$

This rather formidable set of equations can be solved in a systematic and straightforward manner except for one detail: there are three equations, (326i), (326j) and (326k), each involving a_{54} and a_{65} and no other elements of A. Hence, we need to ensure, by restricting the choice of c, that these equations are consistent. To find the consistency condition, note that the left-hand sides of these equations are related by (326i)×(326j) = (326b)×(326k). The consistency condition, found from the right-hand sides, simplifies to

$$(c_6 - 1)\left(c_4(2 - 10c_3 + 15c_3^2) - c_3\right) = 0. \tag{326r}$$

We can eliminate the factor $c_6 - 1$ because, if it were zero, then it would follow that $c_3 = \frac{1}{3}$ and that $c_4 = 1$, which are consistent with the vanishing of the second factor, which leads to

$$c_4 = \frac{c_3}{2 - 10c_3 + 15c_3^2}. \tag{326s}$$

Having chosen c_3, and therefore c_4, together with arbitrary c_2, c_5 and c_6 and the known value $c_7 = 1$, excluding some impossible cases, we can solve for the components of b^T from (326a)–(326g). We can then solve for a_{54}, a_{64} and a_{65} from the consistent equations (326h)–(326k). We then solve for a_{32} from (326l) and then for a_{42}, a_{43}, a_{52}, a_{53}, a_{62} and a_{63} from (326l) with $i = 4, 5, 6$ and from (326o), (326p) and (326q). It remains to compute the first column of A from (326m) and the last row from (326n).

The following example is of a method derived from these equations:

$$
\begin{array}{c|ccccccc}
0 \\
\frac{1}{3} & \frac{1}{3} \\
\frac{2}{3} & 0 & \frac{2}{3} \\
\frac{1}{3} & \frac{1}{12} & \frac{1}{3} & -\frac{1}{12} \\
\frac{5}{6} & \frac{25}{48} & -\frac{55}{24} & \frac{35}{48} & \frac{15}{8} \\
\frac{1}{6} & \frac{3}{20} & -\frac{11}{24} & -\frac{1}{8} & \frac{1}{2} & \frac{1}{10} \\
1 & -\frac{261}{260} & \frac{33}{13} & \frac{43}{156} & -\frac{118}{39} & \frac{32}{195} & \frac{80}{39} \\
\hline
 & \frac{13}{200} & 0 & \frac{11}{40} & \frac{11}{40} & \frac{4}{25} & \frac{4}{25} & \frac{13}{200}
\end{array}
$$

It is possible to derive sixth order methods in other ways. For example, Huťa used the $C(3)$ with subsidiary conditions for stages 2 and 3. However, he used $s = 8$, and this gave him more freedom in the choice of c.

The alternative example of a method of this order that we give uses $C(2)$ and $D(2)$ with subsidiary conditions to repair the gaps in the order conditions caused by $C(2)$ not applying to stage 2 and $D(2)$ not holding for stage 6. It is necessary to choose $b_2 = 0$, and to require that c_3, c_4 and c_5 are related so that the right-hand side vanishes in the equations

$$\sum_{i,j=1}^{7} b_i(1 - c_i)(c_i - c_5)a_{ij}c_j(c_j - c_3) = \frac{1}{90} - \frac{c_3}{40} - \frac{c_5}{60} + \frac{c_3c_5}{24},$$

$$\sum_{i=1}^{7} b_i(1 - c_i)(c_i - c_3)(c_i - c_4)(c_i - c_5)c_i =$$

$$\frac{1}{30} - \frac{c_3 + c_4 + c_5}{20} + \frac{c_3c_4 + c_3c_5 + c_4c_5}{12} - \frac{c_3c_4c_5}{6},$$

because the left-hand sides are identically zero. A method derived along these lines is as follows:

$$
\begin{array}{c|ccccccc}
0 \\
\frac{2}{5} & \frac{2}{5} \\
\frac{4}{5} & 0 & \frac{4}{5} \\
\frac{2}{9} & \frac{169}{1458} & \frac{110}{729} & -\frac{65}{1458} \\
\frac{8}{15} & -\frac{44}{675} & -\frac{88}{135} & \frac{76}{351} & \frac{336}{325} \\
0 & \frac{21}{106} & 0 & -\frac{105}{689} & -\frac{324}{689} & \frac{45}{106} \\
1 & -\frac{2517}{4864} & -\frac{55}{38} & \frac{10615}{31616} & \frac{567}{7904} & \frac{7245}{4864} & \frac{2597}{2432} \\
\hline
& 0 & 0 & \frac{1375}{4992} & \frac{6561}{20384} & \frac{3375}{12544} & \frac{53}{768} & \frac{19}{294}
\end{array}
$$

327 Methods of orders greater than 6

Methods with order 7 must have at least nine stages. It is possible to construct such a method using the principles of Subsection 323, extending the approach used in Subsection 326. The abscissa vector is chosen as

$$
c = \begin{bmatrix} 0 & \tfrac{1}{3}c_4 & \tfrac{2}{3}c_4 & c_4 & c_5 & c_6 & c_7 & 0 & 1 \end{bmatrix}^{\mathsf{T}},
$$

and the orders of stages numbered $4, 5, \ldots, 9$ are forced to be 3. To achieve consistency of the conditions

$$
\sum b_i(1 - c_i)a_{ij}a_{jk}c_k(c_k - c_4)(c_k - c_5) =
$$
$$
\frac{1}{4\cdot5\cdot6\cdot7} - \frac{c_4 + c_5}{3\cdot4\cdot5\cdot6} + \frac{c_4 c_5}{2\cdot3\cdot4\cdot5},
$$
$$
\sum b_i(1 - c_i)a_{ij}c_j(c_j - c_4)(c_j - c_5)(c_j - c_6) =
$$
$$
\frac{1}{5\cdot6\cdot7} - \frac{c_4 + c_5 + c_6}{4\cdot5\cdot6} + \frac{c_4 c_5 + c_4 c_6 + c_5 c_6}{3\cdot4\cdot5} - \frac{c_4 c_5 c_6}{2\cdot3\cdot4},
$$
$$
\sum b_i(1 - c_i)c_i a_{ij}c_j(c_j - c_4)(c_j - c_5) = \frac{1}{4\cdot6\cdot7} - \frac{c_4 + c_5}{3\cdot5\cdot6} + \frac{c_4 c_5}{2\cdot4\cdot5},
$$

it is found that

$$
c_6 = \frac{u - 12v + 7uv}{3 - 12u + 24v + 14u^2 - 70uv + 105v^2},
$$

where $u = c_4 + c_5$ and $v = c_4 c_5$. The value of c_7 is selected to ensure that

$$
\int_0^1 x(1 - x)(x - c_4)(x - c_5)(x - c_6)(x - c_7)dx = 0.
$$

The tableau for a possible method derived along these lines is

$$
\begin{array}{c|ccccccccc}
0 \\
\frac{1}{6} & \frac{1}{6} \\
\frac{1}{3} & 0 & \frac{1}{3} \\
\frac{1}{2} & \frac{1}{8} & 0 & \frac{3}{8} \\
\frac{2}{11} & \frac{148}{1331} & 0 & \frac{150}{1331} & -\frac{56}{1331} \\
\frac{2}{3} & -\frac{404}{243} & 0 & -\frac{170}{27} & \frac{4024}{1701} & \frac{10648}{1701} \\
\frac{6}{7} & \frac{2466}{2401} & 0 & \frac{1242}{343} & -\frac{19176}{16807} & -\frac{51909}{16807} & \frac{1053}{2401} \\
0 & \frac{5}{154} & 0 & 0 & \frac{96}{539} & -\frac{1815}{20384} & -\frac{405}{2464} & \frac{49}{1144} \\
1 & -\frac{113}{32} & 0 & -\frac{195}{22} & \frac{32}{7} & \frac{29403}{3584} & -\frac{729}{512} & \frac{1029}{1408} & \frac{21}{16} \\
\hline
 & 0 & 0 & 0 & \frac{32}{105} & \frac{1771561}{6289920} & \frac{243}{2560} & \frac{16807}{74880} & \frac{77}{1440} & \frac{11}{270}
\end{array}
$$

.

Order 8 requires 11 stages, and methods of this order were derived by Curtis (1970)and Cooper and Verner (1972). In each case the abscissae were based on the Lobatto quadrature formula with three internal points. We quote the method of Cooper and Verner in Table 327(I).

Although order 9 has not attracted much interest, and it is unknown how many stages are required to achieve this order, order 10 has posed a challenge. In Curtis (1975) a method of order 10 was presented with 18 stages. However, using an ingenious combination of various simplifying assumptions, Hairer (1978) accomplished this feat in 17 stages. It is still not known if fewer stages are possible.

Exercises 32

32.1 Find a method with $s = p = 3$ such that $c = [0, \frac{1}{2}, 1]$.

32.2 Find a method with $s = p = 3$ such that $c = [0, \frac{1}{3}, 1]$.

32.3 Find a method with $s = p = 4$ such that $b_1 = 0$ and $c_2 = \frac{1}{5}$.

32.4 Find a method with $s = p = 4$ such that $b_2 = 0$ and $c_2 = \frac{1}{4}$.

32.5 Find a method with $s = p = 4$ such that $b_1 = 0$ and $c_3 = 0$.

32.6 Show that Lemma 322A can be used to prove that $c_4 = 1$, if $s = p \geq 4$.

32.7 Show that Lemma 322A can be used to prove that $c_5 = 1$, if $s = p \geq 5$ leading to an alternative proof of Theorem 324B.

Table 327(I)   Tableau for Cooper–Verner eighth order method

c	1	2	3	4	5	6	7	8	9	10	11
0											
$\frac{1}{2}$	$\frac{1}{2}$										
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$									
$\frac{7+\sqrt{21}}{14}$	$\frac{1}{7}$	$\frac{-7-3\sqrt{21}}{98}$	$\frac{21+5\sqrt{21}}{49}$								
$\frac{7+\sqrt{21}}{14}$	$\frac{11+\sqrt{21}}{84}$	0	$\frac{18+4\sqrt{21}}{63}$	$\frac{21-\sqrt{21}}{252}$							
$\frac{1}{2}$	$\frac{5+\sqrt{21}}{48}$	0	$\frac{9+\sqrt{21}}{36}$	$\frac{-231+14\sqrt{21}}{360}$	$\frac{63-7\sqrt{21}}{80}$						
$\frac{7-\sqrt{21}}{14}$	$\frac{10-\sqrt{21}}{42}$	0	$\frac{-432+92\sqrt{21}}{315}$	$\frac{633-145\sqrt{21}}{90}$	$\frac{-504+115\sqrt{21}}{70}$	$\frac{63-13\sqrt{21}}{35}$					
$\frac{7-\sqrt{21}}{14}$	$\frac{1}{14}$	0	0	0	$\frac{14-3\sqrt{21}}{126}$	$\frac{13-3\sqrt{21}}{63}$	$\frac{1}{9}$				
$\frac{1}{2}$	$\frac{1}{32}$	0	0	0	$\frac{91-21\sqrt{21}}{576}$	$\frac{11}{72}$	$\frac{-385-75\sqrt{21}}{1152}$	$\frac{63+13\sqrt{21}}{128}$			
$\frac{7+\sqrt{21}}{14}$	$\frac{1}{14}$	0	0	0	$\frac{1}{9}$	$\frac{-733-147\sqrt{21}}{2205}$	$\frac{515+111\sqrt{21}}{504}$	$\frac{-51-11\sqrt{21}}{56}$	$\frac{132+28\sqrt{21}}{245}$		
1	0	0	0	0	$\frac{-42+7\sqrt{21}}{18}$	$\frac{-18+28\sqrt{21}}{45}$	$\frac{-273-53\sqrt{21}}{72}$	$\frac{301+53\sqrt{21}}{72}$	$\frac{28-28\sqrt{21}}{45}$	$\frac{49-7\sqrt{21}}{18}$	
	$\frac{1}{20}$	0	0	0	0	0	0	$\frac{49}{180}$	$\frac{16}{45}$	$\frac{49}{180}$	$\frac{1}{20}$

33   Runge–Kutta Methods with Error Estimates

330   Introduction

Practical computations with Runge–Kutta methods usually require a means of local error estimation. This is because stepsizes are easy to adjust so as to follow the behaviour of the solution, but the optimal sequence of stepsizes depends on the local truncation error. Of course, the exact truncation error cannot realistically be found, but asymptotically correct approximations to it can be computed as the integration proceeds. One way of looking at this is that *two* separate approximations to the solution at a step value x_n are found. Assuming that the solution value at the previous point is regarded as exact, because it is the *local* error that is being approximated, denote the two solutions found at the current point by y_n and $\widehat{y}_n$. Suppose the two approximations have orders p and q, respectively, so that

$$y_n = y(x_n) + O(h^{p+1}), \qquad \widehat{y}_n = y(x_n) + O(h^{q+1}).$$

Then, if $q > p$,
$$\widehat{y}_n - y_n = y(x_n) - y_n + O(h^{p+2}),$$

which can be used as an approximation to the error committed in the step. Furthermore, the approximation becomes increasingly accurate as h becomes small. Thus $\widehat{y}_n - y_n$ is used as the error estimator.

Even though we emphasize the construction of method pairs for which $q = p+1$, and for which it is y_n (rather than the asymptotically more accurate approximation $\widehat{y}_n$) that is propagated as the numerical approximation at x_n, customary practice is to use the *higher* order as the propagated value. This is sometimes interpreted as 'local extrapolation', in the sense that the error estimate is added to the approximate solution as a correction. While the estimator is still used as a stepsize controller, it is now no longer related asymptotically to the local truncation error.

We review the 'deferred approach to the limit' of Richardson (1927) and then consider specially constructed Runge–Kutta tableaux, which combine two methods, with orders one apart, built into one. The classical method of this type is due to Merson (1957), but we also consider built-in estimators due to Fehlberg (1968, 1969), Verner (1978) and Dormand and Prince (1980). Some of the methods derived for the author's previous book (Butcher, 1987) will also be recalled.

331   Richardson error estimates

Richardson extrapolation consists of calculating a result in a manner that depends on a small parameter, and for which the error in the calculation varies systematically as the parameter varies. By using a sequence of values of the parameter, much of the effect of the errors can be eliminated so that

improved accuracy results. In numerical quadrature, for example, the method of Romberg (1955) is based on calculating an integral $I = \int_a^b \phi(x)dx$ using the trapezoidal rule with a stepsize h equal to an integer divisor of $b - a$. For a single choice of h, the result computed can be expanded by an asymptotic formula of the form

$$T(h) = I + C_1 h^2 + C_2 h^4 + \cdots ,$$

so that, using a sequence $h = H, 2^{-1}H, 2^{-2}H, \ldots$, we arrive at the approximations $T_0 = T(H)$, $T_1 = T(\frac{1}{2}H)$, $T_2 = T(\frac{1}{4}H)$, ... with expansions

$$T_0 = I + C_1 H^2 + C_2 H^4 + \cdots ,$$
$$T_1 = I + \tfrac{1}{4}C_1 H^2 + \tfrac{1}{16}C_2 H^4 + \cdots ,$$
$$T_2 = I + \tfrac{1}{16}C_1 H^2 + \tfrac{1}{256}C_2 H^4 + \cdots ,$$
$$\vdots \qquad\qquad \vdots$$

By forming

$$T_{01} = \frac{4}{3}T_1 - \frac{1}{3}T_0,$$
$$T_{12} = \frac{4}{3}T_2 - \frac{1}{3}T_1,$$
$$\vdots \qquad\qquad \vdots$$

we obtain an 'improved' sequence in which the $C_1 H^2$ terms are eliminated from the asymptotic expansions so that convergence towards the exact result I is more rapid as terms in the sequence are calculated. Similarly, a second sequence of improved approximations can be found from

$$T_{012} = \frac{16}{15}T_{12} - \frac{1}{15}T_{01},$$
$$T_{123} = \frac{16}{15}T_{23} - \frac{1}{15}T_{12},$$
$$\vdots \qquad\qquad \vdots$$

This idea has an application to Runge–Kutta methods for ordinary differential equations on the small scale of a single step, repeated with two steps and half the original value of h. Let y_{n-1} denote an incoming approximation for $y(x_{n-1})$ and y_n the solution computed as an approximation to $y(x_n) = y(x_{n-1} + h)$ using a Runge–Kutta method with tableau

$$
\begin{array}{c|ccccc}
0 \\
c_2 & a_{21} \\
c_3 & a_{31} & a_{32} \\
\vdots & \vdots & \vdots & \ddots \\
c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} \\
\hline
 & b_1 & b_2 & \cdots & b_{s-1} & b_s
\end{array}
\;.
$$

Repeating the calculation with h replaced by $\frac{1}{2}h$ but carrying out two steps, rather than only one, is equivalent to taking a single step with the original h, but using the tableau

$$
\begin{array}{c|cccccccccc}
0 \\
\frac{1}{2}c_2 & \frac{1}{2}a_{21} \\
\frac{1}{2}c_3 & \frac{1}{2}a_{31} & \frac{1}{2}a_{32} \\
\vdots & \vdots & \vdots & \ddots \\
\frac{1}{2}c_s & \frac{1}{2}a_{s1} & \frac{1}{2}a_{s2} & \cdots & \frac{1}{2}a_{s,s-1} \\
\frac{1}{2} & \frac{1}{2}b_1 & \frac{1}{2}b_2 & \cdots & \frac{1}{2}b_{s-1} & \frac{1}{2}b_s \\
\frac{1}{2}+\frac{1}{2}c_2 & \frac{1}{2}b_1 & \frac{1}{2}b_2 & \cdots & \frac{1}{2}b_{s-1} & \frac{1}{2}b_s & \frac{1}{2}a_{21} \\
\frac{1}{2}+\frac{1}{2}c_3 & \frac{1}{2}b_1 & \frac{1}{2}b_2 & \cdots & \frac{1}{2}b_{s-1} & \frac{1}{2}b_s & \frac{1}{2}a_{31} & \frac{1}{2}a_{32} \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \ddots \\
\frac{1}{2}+\frac{1}{2}c_s & \frac{1}{2}b_1 & \frac{1}{2}b_2 & \cdots & \frac{1}{2}b_{s-1} & \frac{1}{2}b_s & \frac{1}{2}a_{s1} & \frac{1}{2}a_{s2} & \cdots & \frac{1}{2}a_{s,s-1} \\
\hline
 & \frac{1}{2}b_1 & \frac{1}{2}b_2 & \cdots & \frac{1}{2}b_{s-1} & \frac{1}{2}b_s & \frac{1}{2}b_1 & \frac{1}{2}b_2 & \cdots & \frac{1}{2}b_{s-1} & \frac{1}{2}b_s
\end{array}
$$

Denote the result computed by this $2s$-stage method by $\widehat{y}_n$, and note that if the local truncation error in y_n is $C(x_n)h^{p+1} + O(h^{p+2})$, so that

$$y_n = y(x_n) - C(x_n)h^{p+1} + O(h^{p+2}), \tag{331a}$$

then

$$\widehat{y}_n = y(x_n) - 2^{-p}C(x_n)h^{p+1} + O(h^{p+2}), \tag{331b}$$

because the error in computing $\widehat{y}_n$ is $2^{-p-1}C(x_n)h^{p+1} + O(h^{p+2})$ contributed from each of two steps.

From the difference of (331a) and (331b) we find

$$\widehat{y}_n - y_n = (1 - 2^{-p})C(x_n)h^{p+1} + O(h^{p+2}),$$

so that the local truncation error in y_n can be approximated by

$$(1 - 2^{-p})^{-1}(\widehat{y}_n - y_n). \tag{331c}$$

This seems like an expensive way of computing the error in the result computed using an s-stage method, because the additional computations required for the estimation take twice as long as the result itself. However, the additional cost becomes more reasonable when we realize that it is not y_n but $\widehat{y}_n$ that should be propagated. The additional cost on this basis is something like 50%. Actually, it is slightly less than this because the calculation of the derivative of y_{n-1} is shared by each of the two methods, and needs to be carried out only once.

332   Methods with built-in estimates

Instead of using the Richardson technique it is possible to combine two methods into one by constructing a tableau with common stages but two alternative output coefficient vectors. The following method, due to Merson (1957), seems to have been the first attempt at constructing this type of stepsize control mechanism:

$$
\begin{array}{c|ccccc}
0 \\
\frac{1}{3} & \frac{1}{3} \\
\frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\
\frac{1}{2} & \frac{1}{8} & 0 & \frac{3}{8} \\
1 & \frac{1}{2} & 0 & -\frac{3}{2} & 2 \\
\hline
& \frac{1}{6} & 0 & 0 & \frac{2}{3} & \frac{1}{6} \\
\hline
& \frac{1}{10} & 0 & \frac{3}{10} & \frac{2}{5} & \frac{1}{5}
\end{array}\ .
$$

The interpretation of this tableau, which contains two b^{T} vectors, is that it combines two methods given by

$$
\begin{array}{c|ccccc}
0 \\
\frac{1}{3} & \frac{1}{3} \\
\frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\
\frac{1}{2} & \frac{1}{8} & 0 & \frac{3}{8} \\
1 & \frac{1}{2} & 0 & -\frac{3}{2} & 2 \\
\hline
& \frac{1}{6} & 0 & 0 & \frac{2}{3} & \frac{1}{6}
\end{array}
\tag{332a}
$$

and by

$$
\begin{array}{c|ccccc}
0 \\
\frac{1}{3} & \frac{1}{3} \\
\frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\
\frac{1}{2} & \frac{1}{8} & 0 & \frac{3}{8} \\
1 & \frac{1}{2} & 0 & -\frac{3}{2} & 2 \\
\hline
& \frac{1}{10} & 0 & \frac{3}{10} & \frac{2}{5} & \frac{1}{5}
\end{array}\ .
\tag{332b}
$$
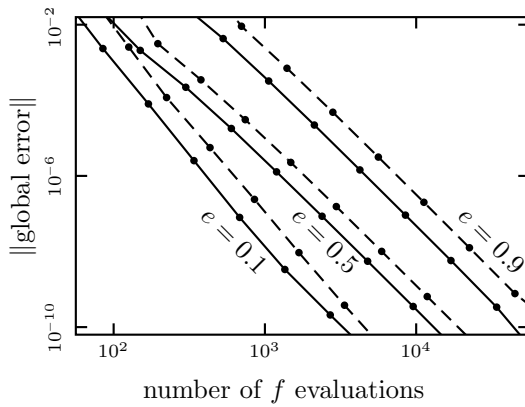
Figure 332(i)   Two alternative stepsize control mechanisms based on
Richardson (dashed line) and built-in (solid line) error estimates

In Merson's derivation of this method, (332a) was shown to be of order 4. Although (332b) has order only 3, it becomes effectively of order 5 if used to solve linear problems with constant coefficients. The difference between the results computed by the two methods can, it is suggested, be used as a local error estimator. To show how well the method works in practice, an experiment using this technique has been carried out and the results summarized in Figure 332(i). The three problems attempted are the Kepler orbit problem with eccentricities $e = 0.1$, $e = 0.5$ and $e = 0.9$, respectively.

333   A class of error-estimating methods

In the search for efficient step-control mechanisms, we consider $(s + 1)$-stage methods of the form

$$
\begin{array}{c|ccccc}
0 & & & & & \\
c_2 & a_{21} & & & & \\
c_3 & a_{31} & a_{32} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & \\
1 & a_{s+1,1} & a_{s+1,2} & \cdots & a_{s+1,s-1} & a_{s+1,s} \\
\hline
 & b_1 & b_2 & \cdots & b_{s-1} & b_s & b_{s+1}
\end{array}
\tag{333a}
$$

with order $p + 1$, with the coefficients chosen so that the embedded method

$$
\begin{array}{c|ccccc}
0 \\
c_2 & a_{21} \\
c_3 & a_{31} & a_{32} \\
\vdots & \vdots & \vdots & \ddots \\
c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} \\
\hline
& a_{s+1,1} & a_{s+1,2} & \cdots & a_{s+1,s-1} & a_{s+1,s}
\end{array}
\qquad (333\mathrm{b})
$$

has order p.

Even though this method *formally* has $s + 1$ stages, in terms of computational cost it can be regarded as having only s, because the derivative calculation needed for stage $s+1$ is identical to the first derivative calculation in the succeeding step. It is convenient to write order conditions for the embedded method pair in terms of the number $B = b_{s+1}$ and the artificial tableau

$$
\begin{array}{c|ccccc}
0 \\
c_2 & a_{21} \\
c_3 & a_{31} & a_{32} \\
\vdots & \vdots & \vdots & \ddots \\
c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} \\
\hline
& b_1 & b_2 & \cdots & b_{s-1} & b_s
\end{array}
\qquad (333\mathrm{c})
$$

An elementary weight, computed using this tableau, differs from that of the full method by a single term. This additional term is formed by multiplying B by the derivative of the order p result found by the method represented by (333b). This enables us to form modified order conditions for (333c), which will ensure that both (333a) and (333b) satisfy the correct conditions. We denote the elementary weights for (333c) by $\Phi(t)$.

Theorem 333A *If* (333b) *has order* p *and* (333a) *has order* $p + 1$ *and* $B = b_{s+1}$, *then*

$$
\Phi(t) = \frac{1 - Br(t)}{\gamma(t)}, \qquad r(t) \le p + 1. \qquad (333\mathrm{d})
$$

Conversely, if (333d) *holds with* $c_s \ne 1$ *and* $B \ne 0$ *and, in addition,*

$$
b_{s+1} = B, \qquad (333\mathrm{e})
$$

$$
a_{s+1,s} = B^{-1} b_s (1 - c_s), \qquad (333\mathrm{f})
$$

$$
a_{s+1,j} = B^{-1} \left(b_j (1 - c_j) - \sum_{i=1}^{s} b_i a_{ij} \right), \qquad j = 1, 2, \ldots, s - 1, \qquad (333\mathrm{g})
$$

then (333b) *has order* p *and* (333a) *has order* $p + 1$.

Proof. For a given tree t, let $\widehat{\Phi}(t)$ denote the elementary weight for (333a) and $\overline{\Phi}(t)$ the elementary weight for (333b). Because the latter method has order p, it follows that for a tree $t = [t_1 t_2 \cdots t_m]$, with order not exceeding $p+1$, we have $\overline{\Phi}(t_i) = 1/\gamma(t_i)$, for $i = 1, 2, \ldots, m$. Hence, for a method identical with (333a) except for b^{T} replaced by the basis vector e_{s+1}^{T}, the elementary weight corresponding to t will be

$$\prod_{i=1}^{m} \frac{1}{\gamma(t_i)} = \frac{r(t)}{\gamma(t)}.$$

Adding B multiplied by this quantity to $\Phi(t)$ gives the result

$$\Phi(t) + B\frac{r(t)}{\gamma(t)} = \widehat{\Phi}(t) = \frac{1}{\gamma(t)},$$

which is equivalent to (333d).

To prove the converse, we first note that, because $B \neq 0$, the previous argument can be reversed. That is, if (333b) has order p then (333d) implies that (333a) has order $p + 1$. Hence, it is only necessary to prove that (333b) has order p. We calculate $\overline{\Phi}(t)$, for $r(t) \leq p$ as follows, where we have written $\chi_i(t)$ for the coefficient of b_i in $\Phi(t)$

$$
\begin{aligned}
\overline{\Phi}(t) &= B^{-1}\sum_{j=1}^{s} b_j (1 - c_j)\chi_j(t) - B^{-1}\sum_{i=1}^{s}\sum_{j=1}^{s-1} b_i a_{ij}\chi_j(t) \\
&= B^{-1}(\Phi(t) - \Phi(t\tau) - \Phi(\tau t)) \\
&= B^{-1}\left(\frac{1 - Br(t)}{\gamma(t)} - \frac{r(t)(1 - B(1 + r(t)))}{(1 + r(t))\gamma(t)} - \frac{1 - B(1 + r(t))}{(1 + r(t))\gamma(t)}\right) \\
&= \frac{1}{\gamma(t)}. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \Box
\end{aligned}
$$

Although the derivation is carried out from a modified version of the order conditions, it is convenient to display a particular method in the format

$$
\begin{array}{c|ccccc}
0 & & & & & \\
c_2 & a_{21} & & & & \\
c_3 & a_{31} & a_{32} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
\hline
c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & \\
\hline
& b_1 & b_2 & \cdots & b_{s-1} & b_s \\
& d_1 & d_2 & \cdots & d_{s-1} & d_s
\end{array} \; ,
$$

where

$$
[\,d_1 \quad d_2 \quad \cdots \quad d_{s-1} \quad d_s\,] = [\,b_1 - a_{s1} \quad b_2 - a_{s2} \quad \cdots \quad b_{s-1} - a_{s,s-1} \quad b_s\,]
$$

is the vector of coefficients in the proposed error estimator. That is, $h \sum_{i=1}^{s} d_i f(Y_i)$ is used to evaluate the difference between the order p approximation $y_{n-1} + h \sum_{i=1}^{s} a_{s+1,i} f(Y_i)$ and the supposedly more accurate approximation of order $p+1$ given by $y_{n-1} + h \sum_{i=1}^{s} b_i f(Y_i)$. The dashed line above row number s of the tableau is intended to indicate that the row below it is the approximation to be propagated and, of course, the dashed line below the b^{T} vector separates the order $p+1$ approximation from the error estimator.

Now let us look at some example of these embedded methods. Methods of orders 1 and 2 are easy to derive and examples of each of these are as follows:

$$
\begin{array}{c|cc}
0 \\
\hline
1 & 1 \\
\hline
& \frac{1}{2} & \frac{1}{2} \\
\hline
& -\frac{1}{2} & \frac{1}{2}
\end{array}
$$

and

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

Observe that for the second order method, the third order method in which it is embedded is actually the classical fourth order method.

Order 3 embedded in order 4 requires $s = 4$ stages. From the modified order conditions we find that

$$
b_3(c_3 - c_4)c_3(c_3 - c_2) = \left(\tfrac{1}{4} - B\right) - (c_2 + c_4)\left(\tfrac{1}{3} - B\right) + c_2 c_4\left(\tfrac{1}{2} - B\right), \quad \text{(333h)}
$$

$$
b_4 a_{43} c_3(c_3 - c_2) = \left(\tfrac{1}{12} - \tfrac{B}{3}\right) - c_2\left(\tfrac{1}{6} - \tfrac{B}{2}\right), \quad \text{(333i)}
$$

$$
b_3(c_3 - c_4)a_{32} c_2 = \left(\tfrac{1}{8} - \tfrac{B}{2}\right) - c_4\left(\tfrac{1}{6} - \tfrac{B}{2}\right), \quad \text{(333j)}
$$

$$
b_4 a_{43} a_{32} c_2 = \left(\tfrac{1}{24} - \tfrac{B}{6}\right), \quad \text{(333k)}
$$

so that, equating the products (333h)×(333k) and (333i)×(333j) and simplifying, we find the consistency condition

$$
c_4 = \frac{1 - 7B + 12B^2}{1 - 6B + 12B^2}.
$$

For example, choosing $B = \tfrac{1}{12}$ to give $c_4 = \tfrac{6}{7}$, together with $c_2 = \tfrac{2}{7}$ and $c_3 = \tfrac{4}{7}$, yields the tableau

$$
\begin{array}{c|cccccc}
0 \\
\frac{2}{7} & \frac{2}{7} \\
\frac{4}{7} & -\frac{8}{35} & \frac{4}{5} \\
\frac{6}{7} & \frac{29}{42} & -\frac{2}{3} & \frac{5}{6} \\
1 & \frac{1}{6} & \frac{1}{6} & \frac{5}{12} & \frac{1}{4} \\
\hline
& \frac{11}{96} & \frac{7}{24} & \frac{35}{96} & \frac{7}{48} & \frac{1}{12} \\
& -\frac{5}{96} & \frac{1}{8} & -\frac{5}{96} & -\frac{5}{48} & \frac{1}{12}
\end{array} \;.
$$

Order 4 embedded in order 5 requires $s = 6$. That is, there are seven stages overall, but the last stage derivative is identical to the first stage derivative for the following step. To derive a method of this type, make the simplifying assumption

$$
\sum_{j=1}^{6} a_{ij}c_j = \frac{1}{2}c_i^2, \qquad i \neq 2,
$$

together with the subsidiary conditions

$$
b_2 = \sum_{i=3}^{6} b_i a_{i2} = \sum_{i=3}^{6} b_i c_i a_{i2} = \sum_{i=4}^{6}\sum_{j=3}^{i-1} b_i a_{ij} a_{j2} = 0.
$$

Also, impose order conditions for the trees $\;\bullet\;\; {\Large\mathsf{!}}\;\; \mathsf{V}\;\; \mathsf{V}\;$ but instead of the corresponding conditions for the trees $\;\mathsf{Y}\;\; \mathsf{w}\;\; \mathsf{Y}\;\; \mathsf{Y}\;\; \mathsf{Y}\;$, use linear combinations as follows:

$$
\sum_{6 \geq i > j \geq 4} b_i a_{ij} c_j (c_j - c_3) = \left(\tfrac{1}{12} - \tfrac{1}{3}B\right) - c_3\left(\tfrac{1}{6} - \tfrac{1}{2}B\right), \tag{333l}
$$

$$
\begin{aligned}
\sum_{5 \geq i \geq 5} b_i c_i (c_i - c_6)(c_i - c_4)(c_i - c_3) = &\left(\tfrac{1}{5} - B\right) - (c_6 + c_4 + c_3)\left(\tfrac{1}{4} - B\right) \\
&+ (c_6 c_4 + c_6 c_3 + c_4 c_3)\left(\tfrac{1}{3} - B\right) \\
&- c_6 c_4 c_3\left(\tfrac{1}{2} - B\right),
\end{aligned} \tag{333m}
$$

$$
\begin{aligned}
\sum_{5 \geq i > j \geq 4} b_i(c_i - c_6) a_{ij} c_j (c_j - c_3) = &\left(\tfrac{1}{15} - \tfrac{1}{3}B\right) - c_6\left(\tfrac{1}{12} - \tfrac{1}{3}B\right) \\
&- c_3\left(\tfrac{1}{8} - \tfrac{1}{2}B\right) + c_6 c_3\left(\tfrac{1}{6} - \tfrac{1}{2}B\right),
\end{aligned} \tag{333n}
$$

$$
\begin{aligned}
\sum_{6 \geq i > j \geq 5} b_i a_{ij} c_j (c_i - c_4)(c_j - c_3) = &\left(\tfrac{1}{20} - \tfrac{1}{4}B\right) - (c_4 + c_3)\left(\tfrac{1}{12} - \tfrac{1}{3}B\right) \\
&+ c_4 c_3\left(\tfrac{1}{6} - \tfrac{1}{2}B\right),
\end{aligned} \tag{333o}
$$

$$
\sum_{6 \geq i > j > k \geq 4} b_i a_{ij} a_{jk} c_k (c_k - c_3) = \left(\tfrac{1}{60} - \tfrac{1}{12}B\right) - c_3\left(\tfrac{1}{24} - \tfrac{1}{6}B\right). \tag{333p}
$$

The left-hand sides of (333m)–(333p) consist of only a single term and we see that the product of (333m) and (333p) is equal to the product of (333n) and (333o). Thus we obtain consistency conditions for the values of a_{65} and a_{54} by comparing the products of the corresponding right-hand sides. After considerable manipulation and simplification, we find that this consistency condition reduces to

$$c_6 = 1 - \frac{q_0 B}{q_0 - q_1 B + q_2 B^2}, \tag{333q}$$

with

$$q_0 = 10c_3^2 c_4 + 2c_4 - 8c_3 c_4 - c_3,$$
$$q_1 = 60c_3^2 c_4 - 56c_3 c_4 + 16c_4 - 8c_3,$$
$$q_2 = 120c_3^2 c_4 - 120c_3 c_4 + 40c_4 - 20c_3.$$

Construction of the method consists of selecting c_2, c_3, c_4, c_5 and B; choosing c_6 in accordance with (333q); evaluating a_{65} and a_{54} from the consistent equations (333n), (333o) and (333p); and then evaluating a_{64} from (333l). The remaining coefficients are then evaluated using the remaining conditions that have been stated.

An example of a method in this family is

0							
$\frac{1}{4}$	$\frac{1}{4}$						
$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$					
$\frac{1}{2}$	0	$-\frac{1}{2}$	1				
$\frac{13}{20}$	$\frac{13}{200}$	$-\frac{299}{1000}$	$\frac{78}{125}$	$\frac{13}{50}$			
$\frac{4}{5}$	$\frac{548}{7475}$	$\frac{688}{2875}$	$\frac{572}{2875}$	$-\frac{88}{575}$	$\frac{132}{299}$		
1	$\frac{37}{312}$	0	$\frac{4}{33}$	$\frac{8}{9}$	$-\frac{100}{117}$	$\frac{575}{792}$	
	$\frac{41}{520}$	0	$\frac{58}{165}$	$\frac{16}{135}$	$\frac{50}{351}$	$\frac{575}{2376}$	$\frac{1}{15}$
	$-\frac{31}{780}$	0	$\frac{38}{165}$	$-\frac{104}{135}$	$\frac{350}{351}$	$-\frac{575}{1188}$	$\frac{1}{15}$

For $p = 5$, that is, a fifth order method embedded within a sixth order method, $s = 8$ seems to be necessary. We present a single example of a method satisfying these requirements. For all stages except the second, the stage order is at least 2, and for stages after the third, the stage order is at least 3. Under these assumptions, together with subsidiary conditions, it is found that for consistency, a relation between c_4, c_5, c_6, c_8 and B must hold. Given that these are satisfied, the derivation is straightforward but lengthy and will not be presented here. The example of a method pair constructed in this way is shown in the following tableau:

$$
\begin{array}{c|ccccccccc}
0 \\
\frac{1}{9} & \frac{1}{9} \\
\frac{1}{9} & \frac{1}{18} & \frac{1}{18} \\
\frac{1}{6} & \frac{1}{24} & 0 & \frac{1}{8} \\
\frac{1}{3} & \frac{1}{6} & 0 & -\frac{1}{2} & \frac{2}{3} \\
\frac{1}{2} & \frac{15}{8} & 0 & -\frac{63}{8} & 7 & -\frac{1}{2} \\
\frac{3}{4} & -\frac{93}{22} & 0 & \frac{24921}{1408} & -\frac{10059}{704} & \frac{735}{1408} & \frac{735}{704} \\
\frac{17}{19} & \frac{86547313055}{10295610642} & 0 & -\frac{96707067}{2867062} & \frac{15526951598}{571978869} & \frac{27949088}{81711267} & -\frac{452648800}{245133801} & \frac{270189568}{467982711} \\
\hline
1 & \frac{98}{765} & 0 & 0 & -\frac{9}{83} & \frac{1071}{1600} & -\frac{11}{75} & \frac{64}{225} & \frac{390963}{2257600} \\
\hline
& \frac{188}{3315} & 0 & 0 & \frac{1593}{7553} & \frac{2943}{20800} & \frac{197}{975} & \frac{576}{2275} & \frac{2476099}{29348800} & \frac{2}{39} \\
\hline
& -\frac{142}{1989} & 0 & 0 & \frac{2412}{7553} & -\frac{549}{1040} & \frac{68}{195} & -\frac{128}{4095} & -\frac{130321}{1467440} & \frac{2}{39}
\end{array}
$$

334   The methods of Fehlberg

Early attempts to incorporate error estimators into Runge–Kutta methods are exemplified by the work of Fehlberg (1968, 1969). In writing the coefficients of methods from this paper, a tabular form is used as follows:

$$
\begin{array}{c|c}
c & A \\
\hline
& b^{\mathsf{T}} \\
\hline
& \widehat{b}^{\mathsf{T}} \\
\hline
& d^{\mathsf{T}}
\end{array}\;.
$$

The significance of this augmented tableau is that

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

is a Runge–Kutta method of order p, while

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

is a Runge–Kutta method of order $p+1$. The additional vector $d^{\mathsf{T}} = \widehat{b}^{\mathsf{T}} - b^{\mathsf{T}}$ is used for error estimation. The fifth order method, with additional sixth order output for error estimation, recommended by Fehlberg, is

$$
\begin{array}{c|cccccccc}
0 \\
\frac{1}{6} & \frac{1}{6} \\
\frac{4}{15} & \frac{4}{75} & \frac{16}{75} \\
\frac{2}{3} & \frac{5}{6} & -\frac{8}{3} & \frac{5}{2} \\
\frac{4}{5} & -\frac{8}{5} & \frac{144}{25} & -4 & \frac{16}{25} \\
1 & \frac{361}{320} & -\frac{18}{5} & \frac{407}{128} & -\frac{11}{80} & \frac{55}{128} \\
0 & -\frac{11}{640} & 0 & \frac{11}{256} & -\frac{11}{160} & \frac{11}{256} & 0 \\
1 & \frac{93}{640} & -\frac{18}{5} & \frac{803}{256} & -\frac{11}{160} & \frac{99}{256} & 0 & 1 \\
\hline
& \frac{31}{384} & 0 & \frac{1125}{2816} & \frac{9}{32} & \frac{125}{768} & \frac{5}{66} & 0 & 0 \\
\hline
& \frac{7}{1408} & 0 & \frac{1125}{2816} & \frac{9}{32} & \frac{125}{768} & 0 & \frac{5}{66} & \frac{5}{66} \\
\hline
& -\frac{5}{66} & 0 & 0 & 0 & 0 & -\frac{5}{66} & \frac{5}{66} & \frac{5}{66}
\end{array}\,.
$$

We also present a similar method with $p = 7$. This also comes from Fehlberg's paper, subject to the correction of some minor misprints. The augmented tableau is

$$
\begin{array}{c|cccccccccccc}
0 \\
\frac{2}{27} & \frac{2}{27} \\
\frac{1}{9} & \frac{1}{36} & \frac{1}{12} \\
\frac{1}{6} & \frac{1}{24} & 0 & \frac{1}{8} \\
\frac{5}{12} & \frac{5}{12} & 0 & -\frac{25}{16} & \frac{25}{16} \\
\frac{1}{2} & \frac{1}{20} & 0 & 0 & \frac{1}{4} & \frac{1}{5} \\
\frac{5}{6} & -\frac{25}{108} & 0 & 0 & \frac{125}{108} & -\frac{65}{27} & \frac{125}{54} \\
\frac{1}{6} & \frac{31}{300} & 0 & 0 & 0 & \frac{61}{225} & -\frac{2}{9} & \frac{13}{900} \\
\frac{2}{3} & 2 & 0 & 0 & -\frac{53}{6} & \frac{704}{45} & -\frac{107}{9} & \frac{67}{90} & 3 \\
\frac{1}{3} & -\frac{91}{108} & 0 & 0 & \frac{23}{108} & -\frac{976}{135} & \frac{311}{54} & -\frac{19}{60} & \frac{17}{6} & -\frac{1}{12} \\
1 & \frac{2383}{4100} & 0 & 0 & -\frac{341}{164} & \frac{4496}{1025} & -\frac{301}{82} & \frac{2133}{4100} & \frac{45}{82} & \frac{45}{164} & \frac{18}{41} \\
0 & \frac{3}{205} & 0 & 0 & 0 & 0 & -\frac{6}{41} & -\frac{3}{205} & -\frac{3}{41} & \frac{3}{41} & \frac{6}{41} & 0 \\
1 & -\frac{1777}{4100} & 0 & 0 & -\frac{341}{164} & \frac{4496}{1025} & -\frac{289}{82} & \frac{2193}{4100} & \frac{51}{82} & \frac{33}{164} & \frac{12}{41} & 0 & 1 \\
\hline
& \frac{41}{840} & 0 & 0 & 0 & 0 & \frac{34}{105} & \frac{9}{35} & \frac{9}{35} & \frac{9}{280} & \frac{9}{280} & \frac{41}{840} & 0 & 0 \\
\hline
& 0 & 0 & 0 & 0 & 0 & \frac{34}{105} & \frac{9}{35} & \frac{9}{35} & \frac{9}{280} & \frac{9}{280} & 0 & \frac{41}{840} & \frac{41}{840} \\
\hline
& -\frac{41}{840} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{41}{840} & \frac{41}{840} & \frac{41}{840}
\end{array}\,.
$$

The two methods presented here, along with some of the other Runge–Kutta pairs derived by Fehlberg, have been criticized for a reason associated with computational robustness. This is that the two quadrature formulae characterized by the vectors b^{T} and $\hat{b}^{\mathsf{T}}$ are identical. Hence, if the differential equation being solved is approximately equal to a pure quadrature problem, then error estimates will be too optimistic.

Although the methods were intended by Fehlberg to be used as order p schemes together with asymptotically correct error estimators, such methods are commonly implemented in a slightly different way. Many numerical analysts argue that it is wasteful to propagate a low order approximation when a higher order approximation is available. This means that the method $(A, \hat{b}^\mathsf{T}, c)$, rather than (A, b^T, c), would be used to produce output values. The order $p+1$ method will have a different stability region than that of the order p method, and this needs to be taken into account. Also there is no longer an asymptotically correct error estimator available. Many practical codes have no trouble using the difference of the order p and order $p+1$ approximations to control stepsize, even though it is the higher order result that is propagated.

335 The methods of Verner

The methods of Verner overcome the fault inherent in many of the Fehlberg methods, that the two embedded methods both have the same underlying quadrature formula. The following method from Verner (1978) consists of a fifth order method which uses just the first six stages together with a sixth order method based on all of the eight stages. Denote the two output coefficient vectors by b^T and $\hat{b}^\mathsf{T}$, respectively. As usual we give the difference $\hat{b}^\mathsf{T} - b^\mathsf{T}$ which is used for error estimation purposes:

$$
\begin{array}{c|cccccccc}
0 \\
\frac{1}{18} & \frac{1}{18} \\
\frac{1}{6} & -\frac{1}{12} & \frac{1}{4} \\
\frac{2}{9} & -\frac{2}{81} & \frac{4}{27} & \frac{8}{81} \\
\frac{2}{3} & \frac{40}{33} & -\frac{4}{11} & -\frac{56}{11} & \frac{54}{11} \\
1 & -\frac{369}{73} & \frac{72}{73} & \frac{5380}{219} & -\frac{12285}{584} & \frac{2695}{1752} \\
\frac{8}{9} & -\frac{8716}{891} & \frac{656}{297} & \frac{39520}{891} & -\frac{416}{11} & \frac{52}{27} & 0 \\
1 & \frac{3015}{256} & -\frac{9}{4} & -\frac{4219}{78} & \frac{5985}{128} & -\frac{539}{384} & 0 & \frac{693}{3328} \\
\hline
& \frac{3}{80} & 0 & \frac{4}{25} & \frac{243}{1120} & \frac{77}{160} & \frac{73}{700} & 0 & 0 \\
& \frac{57}{640} & 0 & -\frac{16}{65} & \frac{1377}{2240} & \frac{121}{320} & 0 & \frac{891}{8320} & \frac{2}{35} \\
\hline
& \frac{33}{640} & 0 & -\frac{132}{325} & \frac{891}{2240} & -\frac{33}{320} & -\frac{73}{700} & \frac{891}{8320} & \frac{2}{35}
\end{array}
$$

As for the Fehlberg methods, we have a choice as to whether we use the fifth or sixth order approximation as output for propagation purposes. Even though the sixth order choice leaves us without an asymptotically correct local error estimator, the use of this more accurate approximation has definite advantages. In Figure 335(i) the stability regions for the two approximations are plotted. It is clear that stability considerations favour the higher order method.
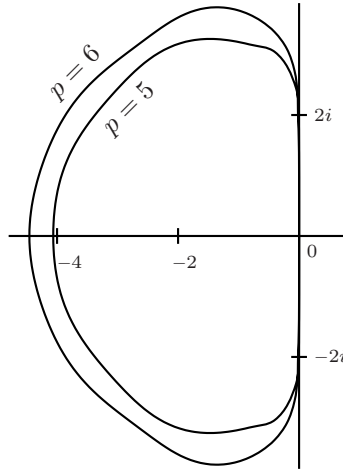
Figure 335(i) Stability regions of embedded Verner method with orders 5 and 6

336 The methods of Dormand and Prince

If it is accepted that in using a Runge–Kutta pair, comprising methods of orders p and $p + 1$, it is the *higher* order member of the pair that is going to be propagated, then it is logical to take some care over the properties of this order $p+1$ method. In the methods introduced in Dormand and Prince (1980), this point of view is adopted. The first of these method pairs, referred to by the authors as 'RK5(4)7M', is designed to have a low value of the 2-norm of the vector of sixth order error coefficients. This method has the tableau

$$
\begin{array}{c|ccccccc}
0 \\
\frac{1}{5} & \frac{1}{5} \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} \\
\frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} \\
\frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} \\
1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} \\
1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} \\
\hline
& \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
& \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40} \\
\hline
& -\frac{71}{57600} & 0 & \frac{71}{16695} & -\frac{71}{1920} & \frac{17253}{339200} & -\frac{22}{525} & \frac{1}{40}
\end{array}
$$

(336a)

It is emphasized that the first of the output approximations has order $p+1 = 5$ and is the result propagated. This method, like those derived in Subsection 333, have the so-called FSAL ('first same as last') property in which the

vector b^{T}, corresponding to the output approximation, has its last component zero and is in fact identical to the last row of A. This means that, while this particular method has seven stages, it operates as though it only had six because the evaluation of the seventh and last stage derivative can be retained to serve as the first stage derivative for the subsequent step.

An alternative choice of free parameters leads to the following method:

$$
\begin{array}{c|ccccccc}
0 \\
\frac{2}{9} & \frac{2}{9} \\
\frac{1}{3} & \frac{1}{12} & \frac{1}{4} \\
\frac{5}{9} & \frac{55}{324} & -\frac{25}{108} & \frac{50}{81} \\
\frac{2}{3} & \frac{83}{330} & -\frac{13}{22} & \frac{61}{66} & \frac{9}{110} \\
1 & -\frac{19}{28} & \frac{9}{4} & \frac{1}{7} & -\frac{27}{7} & \frac{22}{7} \\
1 & \frac{19}{200} & 0 & \frac{3}{5} & -\frac{243}{400} & \frac{33}{40} & \frac{7}{80} \\
\hline
 & \frac{19}{200} & 0 & \frac{3}{5} & -\frac{243}{400} & \frac{33}{40} & \frac{7}{80} & 0 \\
 & \frac{431}{5000} & 0 & \frac{333}{500} & -\frac{7857}{10000} & \frac{957}{1000} & \frac{193}{2000} & -\frac{1}{50} \\
\hline
 & -\frac{11}{1250} & 0 & \frac{33}{500} & -\frac{891}{5000} & \frac{33}{250} & \frac{9}{1000} & -\frac{1}{50}
\end{array}
\tag{336b}
$$

Although this has larger error constants overall (as measured by the 2-norm of the sixth order error vector), it has the advantage of a longer stability interval than that of (336a).

For comparison, a method pair with exactly six stages (but of course without the FSAL property) was also presented in the Dormand and Prince paper. This method, given by

$$
\begin{array}{c|cccccc}
0 \\
\frac{1}{5} & \frac{1}{5} \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} \\
\frac{3}{5} & \frac{3}{10} & -\frac{9}{10} & \frac{6}{5} \\
\frac{2}{3} & \frac{226}{729} & -\frac{25}{27} & \frac{880}{729} & \frac{55}{729} \\
1 & -\frac{181}{270} & \frac{5}{2} & -\frac{266}{297} & -\frac{91}{27} & \frac{189}{55} \\
\hline
 & \frac{19}{216} & 0 & \frac{1000}{2079} & -\frac{125}{216} & \frac{81}{88} & \frac{5}{56} \\
 & \frac{31}{540} & 0 & \frac{190}{297} & -\frac{145}{108} & \frac{351}{220} & \frac{1}{20} \\
\hline
 & -\frac{11}{360} & 0 & \frac{10}{63} & -\frac{55}{72} & \frac{27}{40} & -\frac{11}{280}
\end{array}
,
$$

seems to be less efficient than the FSAL method.

In the derivation of these method pairs, some attention is devoted to the properties of the approximation which is *not* propagated. In particular, care is taken to ensure that this approximation has an acceptable stability region. In

any implementation of these methods, $\widehat{b}^{\mathsf{T}}$ does not play a direct role because stepsize is controlled using the vector of coefficients $d^{\mathsf{T}} = \widehat{b}^{\mathsf{T}} - b^{\mathsf{T}}$. Rescaling this vector by a non-zero factor is then equivalent to rescaling the user-imposed tolerance. From this point of view, the restriction of methods to those for which the non-propagated approximation has good stability properties is unnecessary.

Exercises 33

33.1 To overcome the perceived disadvantage of using Richardson extrapolation as in Subsection 331, is it feasible to modify the method so that a proportion of the estimated error (331c) is subtracted from the result $\widehat{y}_n$?

33.2 Find a problem for which the Merson method gives reasonable error estimating performance.

33.3 Find a problem which exposes the error estimating deficiencies of the Merson method.

33.4 Find a method of order 3 embedded in order 4, based on equations (333h)–(333k) with $B = \frac{1}{6}$, $c_2 = \frac{2}{3}$, $c_3 = \frac{1}{3}$.

33.5 Find an example of a differential equation system for which the methods given in Subsection 334 are likely to have misleading error estimates.

34   Implicit Runge–Kutta Methods

340   Introduction

The possibility that the coefficient matrix A in a Runge–Kutta method might not be strictly lower triangular has very important consequences. These more general methods, known as 'implicit Runge–Kutta methods', are difficult to actually use, because the explicit stage-by-stage implementation scheme enjoyed by explicit methods is no longer available and needs to be replaced by an iterative computation. However, there are several very good reasons, both theoretical and practical, for moving these methods into the centre of our attention. Perhaps the most important theoretical reason for regarding implicit methods as the standard examples of Runge–Kutta methods is the fact that implicit methods have a group structure. We explore this in detail in Section 38. In the explicit case, methods do not have explicit methods as inverses, and thus explicit methods possess only a semi-group structure. Stiff problems cannot be solved efficiently using explicit methods: this fact is the most important practical reason for paying special attention to implicit methods. However, there are other problem classes, such as differential-algebraic equations, for which implicit Runge–Kutta methods also have a vital role.

341 Solvability of implicit equations

As we have remarked, explicit evaluation of the stages is not, in general, possible for an implicit Runge–Kutta method. However, under mild assumptions on the smoothness of the function f it is easy to see that, for sufficiently small h, the values of Y_1, Y_2, ..., Y_s, and hence the output from a step, exist and are unique. Suppose that f satisfies a Lipschitz condition

$$\|f(\eta) - f(\overline{\eta})\| \le L\|\eta - \overline{\eta}\|$$

and consider the stages in a step with size h from initial value y_0. We can identify the values of Y_i, $i = 1, 2, \ldots, s$, as comprising the components of a vector in $\mathbb{R}^{sN}$ which is a fixed point of the mapping

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix} \mapsto \phi(Y) = \begin{bmatrix} y_0 + h\sum_{i=1}^{s} a_{1i}f(Y_i) \\ y_0 + h\sum_{i=1}^{s} a_{2i}f(Y_i) \\ \vdots \\ y_0 + h\sum_{i=1}^{s} a_{si}f(Y_i) \end{bmatrix}.$$

Define a metric on $\mathbb{R}^{sN}$ by the formula

$$\rho(Y, \overline{Y}) = \max_{i=1}^{s} \|Y_i - \overline{Y}_i\|,$$

and estimate $\rho(\phi(Y), \phi(\overline{Y}))$ as follows:

$$\rho(\phi(Y), \phi(\overline{Y})) = \max_{i=1}^{s} \|\sum_{j=1}^{s} ha_{ij}\|f(Y_j) - \overline{f}(Y_j)\|$$

$$\le |h| \max_{i=1}^{s} \sum_{j=1}^{s} |a_{ij}|L\|Y_j - \overline{Y}_j\|$$

$$\le |h|L\|A\|_\infty \max_{j=1}^{s} \|Y_j - \overline{Y}_j\|$$

$$\le |h|L\|A\|_\infty \rho(Y, \overline{Y}),$$

so that the conditions for the contraction mapping principle are satisfied as long as

$$|h| \le (L\|A\|_\infty)^{-1}.$$

In practice, this result is of little value, because implicit Runge–Kutta methods are usually used only for stiff problems for which L is typically unreasonably large. In this case it is usually more efficient to use some variant of the Newton method. We discuss this question further in Subsection 360.

342   Methods based on Gaussian quadrature

We recall the Legendre polynomials on the interval $[0, 1]$

$$P_0^*(x) = 1,$$
$$P_1^*(x) = 2x - 1,$$
$$P_2^*(x) = 6x^2 - 6x + 1,$$
$$P_3^*(x) = 20x^3 - 30x^2 + 12x - 1,$$

$$\vdots \qquad \vdots$$

where we use the notation P_n^* for the member of the sequence with degree n. Note that P_n^* is related to P_n, the Legendre polynomials on the standard symmetric interval $[-1, 1]$, by $P_n^*(x) = P_n(2x-1)$. Amongst the rich collection of properties of this polynomial sequence, we state:

Lemma 342A *There exist polynomials $P_n^* : [0, 1] \rightarrow \mathbb{R}$, of degrees n, for $n = 0, 1, 2, \ldots$ with the properties that*

$$\int_0^1 P_m^*(x) P_n^*(x) dx = 0, \qquad m \neq n, \tag{342a}$$

$$P_n^*(1) = 1, \qquad n = 0, 1, 2, \ldots. \tag{342b}$$

Furthermore, the polynomials defined by (342a) and (342b) have the following additional properties:

$$P_n^*(1-x) = (-1)^n P_n^*(x), \qquad n = 0, 1, 2, \ldots, \tag{342c}$$

$$\int_0^1 P_n^*(x)^2 dx = \frac{1}{2n+1}, \qquad n = 0, 1, 2, \ldots, \tag{342d}$$

$$P_n^*(x) = \frac{1}{n!} \left(\frac{d}{dx} \right)^n (x^2 - x)^n, \qquad n = 0, 1, 2, \ldots, \tag{342e}$$

$$nP_n^*(x) = (2x-1)(2n-1)P_{n-1}^*(x) - (n-1)P_{n-2}^*(x), \quad n = 2, 3, 4, \ldots, \tag{342f}$$

P_n^* *has n distinct real zeros in the interval $(0, 1)$,* $\quad n = 0, 1, 2, \ldots.$ (342g)

Proof. We give only outline proofs of these well-known results. The orthogonality property (342a), of the polynomials defined by (342e), follows by repeated integration by parts. The value at $x = 1$ follows by substituting $x = 1 + \xi$ in (342e) and evaluating the coefficient of the lowest degree term. The fact that P_n^* is an even or odd polynomial in $2x - 1$, as stated in (342c), follows from (342e). The highest degree coefficients in P_n^* and P_{n-1}^* can be compared so that $nP_n^*(x) - (2x - 1)(2n - 1)P_{n-1}^*(x)$ is a polynomial, Q say, of degree less than n. Because Q has the same parity as n, it is of degree

less than $n - 1$. A simple calculation shows that Q is orthogonal to P_k^* for $k < n - 2$. Hence, (342f) follows except for the value of the P_{n-2}^* coefficient, which is resolved by substituting $x = 1$. The final result (342g) is proved by supposing, on the contrary, that $P_n^*(x) = Q(x)R(x)$, where the polynomial factors Q and R have degrees $m < n$ and $n - m$, respectively, and where R has no zeros in $(0, 1)$. We now find that $\int_0^1 P_n^*(x)Q(x)dx = 0$, even though the integrand is not zero and has a constant sign. □

In preparation for constructing a Runge–Kutta method based on the zeros c_i, $i = 1, 2, \ldots, s$ of P_s^*, we look at the associated quadrature formula.

Lemma 342B *Let c_1, c_2, . . . denote the zeros of P_s^*. Then there exist positive numbers b_1, b_2, . . . , b_s such that*

$$\int_0^1 \phi(x)dx = \sum_{i=1}^s b_i\phi(c_i), \qquad (342\text{h})$$

for any polynomial of degree less than $2s$. The b_i are unique.

Proof. Choose b_i, $i = 1, 2, \ldots, s$, so that (342h) holds for any ϕ of degree less than s. Because the c_i are distinct the choice of the b_i is unique. To prove that (342h) holds for degree up to $2s - 1$, write

$$\phi(x) = P_s^*(x)Q(x) + R(x),$$

where the quotient Q and the remainder R have degrees not exceeding $s - 1$. We now have

$$\int_0^1 \phi(x)dx = \int_0^1 P_s^*(x)Q(x)dx + \int_0^1 R(x)dx = 0 + \sum_{i=1}^s b_iR(c_i) = \sum_{i=1}^s b_i\phi(c_i).$$

To prove the b_i are positive, let $\phi(x)$ denote the square of the polynomial formed by dividing $P_s^*(x)$ by $x - c_i$. Substitute into (342h), and the result follows. □

We note that the choice of the c_i as the zeros of P_s^* is the only one possible for (342h) to hold for ϕ of degree as high as $2s - 1$. If this were not the case, let

$$S(x) = \prod_{i=1}^s (x - c_i)$$

and substitute $\phi(x) = S(x)Q(x)$ for any polynomial Q of degree less than s. It is found that S is orthogonal to all polynomials of lower degree and hence, apart from a scale factor, is identical to P_s^*.

We now consider the possibility of constructing an s-stage implicit Runge–Kutta method with order $2s$. If such a method exists, then the values of the

vectors c and b^{T} are known. In the case $s = 2$ we can explore the possibility of choosing the only free parameters that remain, to satisfy four additional order conditions. Surprisingly, this can be done. Write the tableau in the form

$$
\begin{array}{c|cc}
\frac{1}{2} - \frac{\sqrt{3}}{6} & a_{11} & \frac{1}{2} - \frac{\sqrt{3}}{6} - a_{11} \\
\frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{2} + \frac{\sqrt{3}}{6} - a_{22} & a_{22} \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}
\qquad (342\text{i})
$$

For the trees $\bullet$, $\mathbf{!}$, V, V, the order conditions are satisfied. These are just the $B(4)$ conditions introduced in Subsection 321. The remaining trees and the conditions that result from substituting the values from (342i) and simplifying are:

$$
a_{11} = a_{22},
$$
$$
(1 - \sqrt{3})a_{11} + (1 + \sqrt{3})a_{22} = \frac{1}{2},
$$
$$
a_{11} = a_{22},
$$
$$
(1 + \sqrt{3})a_{11} + (1 - \sqrt{3})a_{22} + 2\sqrt{3}(a_{11}^2 - a_{22}^2) = \frac{1}{2}.
$$

These are all satisfied by $a_{11} = a_{22} = \frac{1}{4}$.

We also notice that $C(2)$ and $D(2)$ are satisfied by these values, and it is natural to ask if it is possible, in general, to satisfy both $C(s)$ and $D(s)$ assuming that the b^{T} and c vectors have been chosen to satisfy the quadrature conditions. A crucial link in the chain connecting these conditions is $E(s, s)$, given by (321c), and we present a result which expresses the essential connections between them. It will be convenient to write $G(\eta)$ to represent the fact that a given Runge–Kutta method has order η.

Theorem 342C

$$
G(2s) \Rightarrow B(2s), \qquad (342\text{j})
$$
$$
G(2s) \Rightarrow E(s, s), \qquad (342\text{k})
$$
$$
B(2s) \wedge C(s) \wedge D(s) \Rightarrow G(2s), \qquad (342\text{l})
$$
$$
B(2s) \wedge C(s) \Rightarrow E(s, s), \qquad (342\text{m})
$$
$$
B(2s) \wedge E(s, s) \Rightarrow C(s), \qquad (342\text{n})
$$
$$
B(2s) \wedge D(s) \Rightarrow E(s, s), \qquad (342\text{o})
$$
$$
B(2s) \wedge E(s, s) \Rightarrow D(s). \qquad (342\text{p})
$$

Proof. The first two results (342j), (342k) are consequences of the order conditions. Given that $C(s)$ is true, all order conditions based on trees containing the structure $\cdots [\tau^{k-1}] \cdots$, with $k \leq s$, can be removed, as we
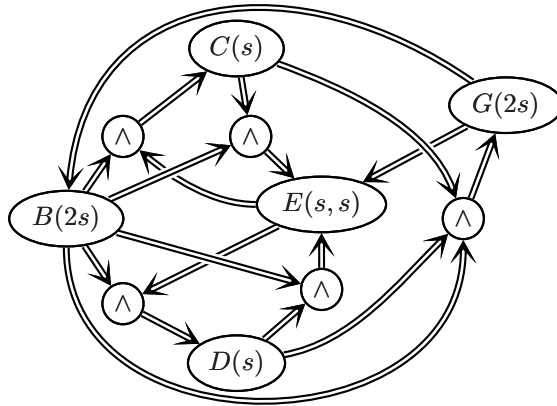
Figure 342(i)   Schema representing Theorem 342C

saw in Subsection 321. Similarly, the condition $D(s)$ enables us to remove from consideration all trees of the form $[\tau^{k-1}[\cdots]]$. Hence, if both $C(s)$ and $D(s)$ are true, the only trees remaining are those associated with the trees covered by $B(2s)$. Hence, (342l) follows. Multiply the matrix of quantities that must be zero according to the $C(s)$ condition

$$
\begin{bmatrix}
\sum_j a_{1j} - c_1 & \sum_j a_{1j}c_j - \frac{1}{2}c_1^2 & \cdots & \sum_j a_{1j}c_j^{s-1} - \frac{1}{s}c_1^s \\
\sum_j a_{2j} - c_2 & \sum_j a_{2j}c_j - \frac{1}{2}c_2^2 & \cdots & \sum_j a_{2j}c_j^{s-1} - \frac{1}{s}c_2^s \\
\vdots & \vdots & & \vdots \\
\sum_j a_{sj} - c_s & \sum_j a_{sj}c_j - \frac{1}{2}c_s^2 & \cdots & \sum_j a_{sj}c_j^{s-1} - \frac{1}{s}c_s^s
\end{bmatrix}
$$

by the non-singular matrix

$$
\begin{bmatrix}
b_1 & b_2 & \cdots & b_s \\
b_1 c_1 & b_2 c_2 & \cdots & b_s c_s \\
\vdots & \vdots & & \vdots \\
b_1 c_1^{s-1} & b_2 c_2^{s-1} & \cdots & b_s c_s^{s-1}
\end{bmatrix}
$$

and the result is the matrix of $E(s,s)$ conditions. Hence, (342m) follows and, because the matrix multiplier is non-singular, (342n) also follows. The final results (342o) and (342p) are proved in similar way.   □

A schema summarizing Theorem 342C is shown in Figure 342(i). To turn this result into a recipe for constructing methods of order $2s$ we have:

Corollary 342D *A Runge–Kutta method has order $2s$ if and only if its coefficients are chosen as follows:*

(i) *Choose c_1, c_2, $\ldots$, c_s as the zeros of P_s^*.*

(ii) *Choose b_1, b_2, $\ldots$, b_s to satisfy the $B(s)$ condition.*

(iii) *Choose a_{ij}, $i, j = 1, 2, \ldots, s$, to satisfy the $C(s)$ condition.*

Proof. If the method has order $2s$ then $B(2s)$ is satisfied. This implies (i) and (ii). Because the order is $2s$, $E(s, s)$ is satisfied and this, together with $B(2s)$, implies (iii). Conversely, if (i) and (ii) are satisfied, then $B(2s)$ holds and this in turn implies $E(s, s)$. This fact, together with $B(2s)$, implies $D(s)$. Finally, use (342l) to complete the proof. $\square$

We conclude this introduction to the Gauss methods by listing the tableaux for $s = 1, 2, 3$ and orders $2, 4, 6$, respectively:

$$s = 1, \quad p = 2,$$

$$
\begin{array}{c|c}
\frac{1}{2} & \frac{1}{2} \\
\hline
& 1
\end{array} \; ;
$$

$$s = 2, \quad p = 4,$$

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

$$s = 3, \quad p = 6,$$

$$
\begin{array}{c|ccc}
\frac{1}{2} - \frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9} - \frac{\sqrt{15}}{15} & \frac{5}{36} - \frac{\sqrt{15}}{30} \\
\frac{1}{2} & \frac{5}{36} + \frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36} - \frac{\sqrt{15}}{24} \\
\frac{1}{2} + \frac{\sqrt{15}}{10} & \frac{5}{36} + \frac{\sqrt{15}}{30} & \frac{2}{9} + \frac{\sqrt{15}}{15} & \frac{5}{36} \\
\hline
& \frac{5}{18} & \frac{4}{9} & \frac{5}{18}
\end{array} \; .
$$

343   Reflected methods

Given a Runge–Kutta method,

$$
\begin{array}{c|cccc}
c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\
c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\
\vdots & \vdots & \vdots & & \vdots \\
c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\
\hline
& b_1 & b_2 & \cdots & b_s
\end{array}
\tag{343a}
$$

we construct a method which exactly undoes the work of the given method. If the signs of the coefficients are then all reversed, the resulting method is known as the 'reflection' (Scherer, 1977, 1978) of the original method. Because the exact solution is its own reflection, it is natural to consider whether Runge–Kutta methods that have this property have any advantage over other methods. In particular, the Gauss methods are their own reflections, as we will see. Reflected methods are now commonly known as 'adjoint methods'; for references to modern applications and research, see Hairer, Lubich and Wanner (2006).

For method (343a), the stages and the final output at the end of step n are given by

$$Y_i = y_{n-1} + h \sum_{j=1}^{s} a_{ij} f(Y_j), \quad i = 1, 2, \ldots, s, \tag{343b}$$

$$y_n = y_{n-1} + h \sum_{j=1}^{s} b_j f(Y_j). \tag{343c}$$

Subtract (343c) from (343b) so that the stage values are written in terms of the result found at the *end* of the step. Also rearrange (343c) so that it gives y_{n-1} in terms of y_n. Thus, the result that works in the reverse direction is given by the equations

$$Y_i = y_n + h \sum_{j=1}^{s} (a_{ij} - b_j) f(Y_j), \quad i = 1, 2, \ldots, s,$$

$$y_{n-1} = y_n + h \sum_{j=1}^{s} (-b_j) f(Y_j).$$

This reversed method has tableau

$$
\begin{array}{c|cccc}
c_1 - \sum_{j=1}^{s} b_j & a_{11} - b_1 & a_{12} - b_2 & \cdots & a_{1s} - b_s \\
c_2 - \sum_{j=1}^{s} b_j & a_{21} - b_1 & a_{22} - b_2 & \cdots & a_{2s} - b_s \\
\vdots & \vdots & \vdots & & \vdots \\
c_s - \sum_{j=1}^{s} b_j & a_{s1} - b_1 & a_{s2} - b_2 & \cdots & a_{ss} - b_s \\
\hline
& -b_1 & -b_2 & \cdots & -b_s
\end{array} \; .
$$

Reverse the signs and we have the tableau for the reflection of (343a)

$$
\begin{array}{c|cccc}
\sum_{j=1}^{s} b_j - c_1 & b_1 - a_{11} & b_2 - a_{12} & \cdots & b_s - a_{1s} \\
\sum_{j=1}^{s} b_j - c_2 & b_1 - a_{21} & b_2 - a_{22} & \cdots & b_s - a_{2s} \\
\vdots & \vdots & \vdots & & \vdots \\
\sum_{j=1}^{s} b_j - c_s & b_1 - a_{s1} & b_2 - a_{s2} & \cdots & b_s - a_{ss} \\
\hline
& b_1 & b_2 & \cdots & b_s
\end{array} \; .
$$

It is easy to verify the following result, which we present without proof.

Theorem 343A *The reflection of the reflection of a Runge–Kutta method is the original method.*

If a method satisfies some of the simplifying assumptions introduced in Subsection 321, then we consider the possibility that the reflection of the method satisfies corresponding conditions. To enable us to express these connections conveniently, we write $\widetilde{B}(\eta)$, $\widetilde{C}(\eta)$, $\widetilde{D}(\eta)$ and $\widetilde{E}(\eta, \zeta)$ to represent $B(\eta)$, $C(\eta)$, $D(\eta)$ and $E(\eta, \zeta)$, respectively, but with reference to the reflected method. We then have:

Theorem 343B *If η and ζ are positive integers, then*

$$B(\eta) \Rightarrow \widetilde{B}(\eta), \tag{343d}$$

$$B(\eta) \wedge C(\eta) \Rightarrow \widetilde{C}(\eta), \tag{343e}$$

$$B(\eta) \wedge D(\eta) \Rightarrow \widetilde{D}(\eta), \tag{343f}$$

$$B(\eta + \zeta) \wedge E(\eta, \zeta) \Rightarrow \widetilde{E}(\eta, \zeta). \tag{343g}$$

Proof. Let P and Q be arbitrary polynomials of degrees less than η and less than ζ, respectively. By using the standard polynomial basis, we see that $B(\eta)$, $C(\eta)$, $D(\eta)$ and $E(\eta, \zeta)$ are equivalent respectively to the statements

$$\sum_{j=1}^{s} b_j P(c_j) = \int_0^1 P(x)dx, \tag{343h}$$

$$\sum_{j=1}^{s} a_{ij} P(c_j) = \int_0^{c_i} P(x)dx, \qquad i = 1, 2, \ldots, s, \tag{343i}$$

$$\sum_{i=1}^{s} b_i P(c_i) a_{ij} = b_j \int_{c_j}^1 P(x)dx, \qquad j = 1, 2, \ldots, s, \tag{343j}$$

$$\sum_{i,j=1}^{s} b_i P(c_i) a_{ij} Q(c_j) = \int_0^1 P(x) \left(\int_0^x Q(x)dx \right) dx. \tag{343k}$$

In each part of the result $B(\eta)$ holds with $\eta \geq 1$, and hence we can assume that $\sum_{i=1}^{s} b_i = 1$. Hence the reflected tableau can be assumed to be

$1 - c_1$	$b_1 - a_{11}$	$b_2 - a_{12}$	$\cdots$	$b_s - a_{1s}$
$1 - c_2$	$b_1 - a_{21}$	$b_2 - a_{22}$	$\cdots$	$b_s - a_{2s}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$1 - c_s$	$b_1 - a_{s1}$	$b_2 - a_{s2}$	$\cdots$	$b_s - a_{ss}$
	b_1	b_2	$\cdots$	b_s

To prove (343d) we have, using (343h),

$$\sum_{j=1}^{s} b_j P(1 - c_j) = \int_0^1 P(1 - x)dx = \int_0^1 P(x)dx.$$

To prove (343e) we use (343i) to obtain

$$\sum_{j=1}^{s}(b_j - a_{ij})P(1 - c_j) = \int_0^1 P(x)dx - \int_0^{c_i} P(1 - x)dx$$

$$= \int_0^1 P(x)dx - \int_{1-c_i}^1 P(x)dx$$

$$= \int_0^{1-c_i} P(x)dx.$$

Similarly, we prove (343f) using (343j):

$$\sum_{i=1}^{s} b_i P(1 - c_i)(b_j - a_{ij}) = b_j \int_0^1 P(x)dx - b_j \int_{c_j}^1 P(1 - x)dx$$

$$= b_j \left(\int_0^1 P(x)dx - \int_0^{1-c_j} P(x)dx \right)$$

$$= b_j \int_{1-c_j}^1 P(x)dx.$$

Finally, use (343k) to prove (343g):

$$\sum_{i,j=1}^{s} b_i P(1 - c_i)(b_j - a_{ij})Q(1 - c_j)$$

$$= \int_0^1 P(x)dx \int_0^1 Q(x)dx - \int_0^1 P(1 - x) \left(\int_0^x Q(1 - x)dx \right) dx$$

$$= \int_0^1 P(x)dx \int_0^1 Q(x)dx - \int_0^1 P(1 - x) \left(\int_{1-x}^1 Q(x)dx \right) dx$$

$$= \int_0^1 P(x)dx \int_0^1 Q(x)dx - \int_0^1 P(x) \left(\int_x^1 Q(x)dx \right) dx$$

$$= \int_0^1 P(x) \left(\int_0^x Q(x)dx \right) dx. \qquad \square$$

344   Methods based on Radau and Lobatto quadrature

It will be shown in Subsection 353 that the Gauss methods have stability regions equal to exactly the left half-plane, and they are therefore A-stable.

For many stiff problems, it is desirable to sacrifice order to gain L-stability, so that the stability function satisfies the property $\lim_{|z| \to \infty} |R(z)| = 0$. We explore methods based on quadrature formulae of orders $2s - 1$ or $2s - 2$. Instead of choosing c_1, c_2, ..., c_s to obtain as high a degree as possible for polynomials ϕ such that

$$\int_0^1 \phi(x)dx = \sum_{i=1}^s b_i\phi(c_i), \tag{344a}$$

we choose either (i) $c_1 = 0$, (ii) $c_s = 1$ or (iii) $c_1 = 0$ *and* $c_s = 1$. The remaining unspecified c_i are then chosen to make (344a) true for a polynomial of degree as high as is still possible.

A 'Radau I quadrature formula' is an interpolational quadrature formula on $[0, 1]$ where the abscissae are chosen as the zeros of $P_s^*(x) + P_{s-1}^*(x)$; a 'Radau II quadrature formula' is an interpolational quadrature formula on $[0, 1]$ where the abscissae are chosen as the zeros of $P_s^*(x) - P_{s-1}^*(x)$ and a 'Lobatto quadrature formula' is an interpolational quadrature formula on $[0, 1]$ where the abscissae are chosen as the zeros of $P_s^*(x) - P_{s-2}^*(x)$. Note that 'Lobatto' is sometimes referred to as 'Lobatto III', to bring the naming of these formulae into a consistent pattern. These three quadrature formulae are the ones sought. We have:

Theorem 344A *Let $c_1 < c_2 < \cdots < c_s$ be chosen as abscissae of the Radau I, the Radau II or the Lobatto quadrature formula, respectively. Then:*

I *For the Radau I formula, $c_1 = 0$. This formula is exact for polynomials of degree up to $2s - 2$.*

II *For the Radau II formula, $c_s = 1$. This formula is exact for polynomials of degree up to $2s - 2$.*

III *For the Lobatto formula, $c_1 = 0, c_s = 1$. This formula is exact for polynomials of degree up to $2s - 3$.*

Furthermore, for each of the three quadrature formulae, $c_i \in [0, 1]$, for $i = 1, 2, \ldots, s$, and $b_i > 0$, for $i = 1, 2, \ldots, s$.

Proof. The fact that $x = 1$ is a zero of $P_s^*(x) - P_{s-1}^*(x)$ and of $P_s^*(x) - P_{s-2}^*(x)$ follows from (342b). The fact that $x = 0$ is a zero of $P_s^*(x) + P_{s-1}^*(x)$ and of $P_s^*(x) - P_{s-2}^*(x)$ follows from (342b) and (342c), with $x = 1$. Let ϕ denote an arbitrary polynomial of degree not exceeding $2s - 2$ in the Radau cases or $2s - 3$ in the Lobatto case. Divide this by the polynomial satisfied by the abscissae and write Q for the quotient and R for the remainder. We have in the three cases,

$$\phi(x) = Q(x)(P_s^*(x) + P_{s-1}^*(x)) + R(x), \qquad \text{Radau I case,}$$
$$\phi(x) = Q(x)(P_s^*(x) - P_{s-1}^*(x)) + R(x), \qquad \text{Radau II case,}$$
$$\phi(x) = Q(x)(P_s^*(x) - P_{s-2}^*(x)) + R(x), \qquad \text{Lobatto case.}$$

Table 344(I)   Methods in the Radau and Lobatto families

Name	Choice of b^{T} and c	Choice of A
Radau I	Radau I quadrature	$C(s)$
Radau IA	Radau I quadrature	The reflections of Radau II
Radau II	Radau II quadrature	$D(s)$
Radau IIA	Radau II quadrature	The reflections of Radau I
Lobatto III	Lobatto quadrature	$C(s-1)$, $a_{1s} = a_{2s} = \cdots = a_{ss} = 0$
Lobatto IIIA	Lobatto quadrature	$C(s)$
Lobatto IIIB	Lobatto quadrature	$D(s)$
Lobatto IIIC	Lobatto quadrature	The reflections of Lobatto III

Evaluate the approximate integral of ϕ written in this form, and the terms involving Q are zero because of orthogonality, and the terms involving R are exact because of the interpolational nature of the quadrature.

In the Radau cases, to prove that the abscissae are always in $[0,1]$ and that the weights are positive, use a homotopy $t \mapsto P_s^* \pm t P_{s-1}^*$, where the upper sign is used for Radau I and the lower sign for Radau II. If any of the weights becomes zero, then for this value of t, the quadrature formula has a greater order than is possible. Furthermore, no abscissae can move outside $[0,1]$, until t reaches a value $t = 1$. The proof is slightly more complicated in the Lobatto case, where we use the homotopy $t \mapsto P_s^* - t P_{s-2}^*$. Because of the symmetry of the quadrature formula for all t, $c_1 = 0$ and $c_s = 1$ both occur at the same time and this is when $t = 1$. If a weight passes through zero, then we again obtain a contradiction to the optimality of Gaussian quadrature because two weights vanish simultaneously. The one case not covered by this argument is when s is odd and the weight corresponding to $c_{(s+1)/2} = \frac{1}{2}$ vanishes. However, it is impossible that as t moves from 0 to 1, it passes through a point for which this happens because in this case the remaining abscissae would have to be the zeros of P_{s-1}^*. By (342f), this occurs only for $t = -(n-1)/n$, and this has the wrong sign.   □

Given the choice of c and b^{T} in accordance with the requirements of Radau I, Radau II or Lobatto quadrature, the choice of A to yield a Runge–Kutta of the same order as for the underlying quadrature formula remains. The most obvious choice, of making the methods as close to explicit as possible, is inappropriate for stiff problems, but makes the method more efficient for non-stiff problems. Other choices can be made in terms of the C and D conditions, and in terms of specific choices of specific elements of A. To distinguish these from the simple (closest to explicit) choices, a letter A, B or C is added to the designation for the method. A summary of many of the methods in the Radau and Lobatto families is given in Table 344(I).

Selected examples of these methods are as follows, where we note that Lobatto IIIB with $s = 2$ does not exist:

Radau I $\qquad (s = 2, p = 3),$

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

Radau IA $\qquad (s = 2, p = 3),$

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

Radau II $\qquad (s = 2, p = 3),$

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

Radau IIA $\qquad (s = 2, p = 3),$

$$
\begin{array}{c|cc}
\frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\
1 & \frac{3}{4} & \frac{1}{4} \\
\hline
 & \frac{3}{4} & \frac{1}{4}
\end{array}
$$

Radau I $\qquad (s = 3, p = 5),$

$$
\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
\frac{6-\sqrt{6}}{10} & \frac{9+\sqrt{6}}{75} & \frac{24+\sqrt{6}}{120} & \frac{168-73\sqrt{6}}{600} \\
\frac{6+\sqrt{6}}{10} & \frac{9-\sqrt{6}}{75} & \frac{168+73\sqrt{6}}{600} & \frac{24-\sqrt{6}}{120} \\
\hline
 & \frac{1}{9} & \frac{16+\sqrt{6}}{36} & \frac{16-\sqrt{6}}{36}
\end{array}
$$

Radau IA $\qquad (s = 3, p = 5),$

$$
\begin{array}{c|ccc}
0 & \frac{1}{9} & \frac{-1-\sqrt{6}}{18} & \frac{-1+\sqrt{6}}{18} \\
\frac{6-\sqrt{6}}{10} & \frac{1}{9} & \frac{88+7\sqrt{6}}{360} & \frac{88-43\sqrt{6}}{360} \\
\frac{6+\sqrt{6}}{10} & \frac{1}{9} & \frac{88+43\sqrt{6}}{360} & \frac{88-7\sqrt{6}}{360} \\
\hline
 & \frac{1}{9} & \frac{16+\sqrt{6}}{36} & \frac{16-\sqrt{6}}{36}
\end{array}
$$

Radau II $(s = 3, p = 5)$,

$$
\begin{array}{c|ccc}
\frac{4-\sqrt{6}}{10} & \frac{24-\sqrt{6}}{120} & \frac{24-11\sqrt{6}}{120} & 0 \\
\frac{4+\sqrt{6}}{10} & \frac{24+11\sqrt{6}}{120} & \frac{24+\sqrt{6}}{120} & 0 \\
1 & \frac{6-\sqrt{6}}{12} & \frac{6+\sqrt{6}}{12} & 0 \\
\hline
 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9}
\end{array}
$$

Radau IIA $(s = 3, p = 5)$,

$$
\begin{array}{c|ccc}
\frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\
\frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\
1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\
\hline
 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9}
\end{array}
$$

Lobatto III $(s = 2, p = 2)$,

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

Lobatto IIIA $(s = 2, p = 2)$,

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

Lobatto IIIC $(s = 2, p = 2)$,

$$
\begin{array}{c|cc}
0 & \frac{1}{2} & -\frac{1}{2} \\
1 & \frac{1}{2} & \frac{1}{2} \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}
$$

Lobatto III $(s = 3, p = 4)$,

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

Lobatto IIIA $(s = 3, p = 4)$,

$$
\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\
1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\
\hline
 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6}
\end{array}
$$

Lobatto IIIB $(s = 3, p = 4)$,

$$
\begin{array}{c|ccc}
0 & \frac{1}{6} & -\frac{1}{6} & 0 \\
\frac{1}{2} & \frac{1}{6} & \frac{1}{3} & 0 \\
1 & \frac{1}{6} & \frac{5}{6} & 0 \\
\hline
 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6}
\end{array}
$$

Lobatto IIIC $(s = 3, p = 4)$,

$$
\begin{array}{c|ccc}
0 & \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} \\
\frac{1}{2} & \frac{1}{6} & \frac{5}{12} & -\frac{1}{12} \\
1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\
\hline
 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6}
\end{array}
$$

Lobatto III $(s = 4, p = 6)$,

$$
\begin{array}{c|cccc}
0 & 0 & 0 & 0 & 0 \\
\frac{5-\sqrt{5}}{10} & \frac{5+\sqrt{5}}{60} & \frac{1}{6} & \frac{15-7\sqrt{5}}{60} & 0 \\
\frac{5+\sqrt{5}}{10} & \frac{5-\sqrt{5}}{60} & \frac{15+7\sqrt{5}}{60} & \frac{1}{6} & 0 \\
1 & \frac{1}{6} & \frac{5-\sqrt{5}}{12} & \frac{5+\sqrt{5}}{12} & 0 \\
\hline
 & \frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12}
\end{array}
$$

Lobatto IIIA $(s = 4, p = 6)$,

$$
\begin{array}{c|cccc}
0 & 0 & 0 & 0 & 0 \\
\frac{5-\sqrt{5}}{10} & \frac{11+\sqrt{5}}{120} & \frac{25-\sqrt{5}}{120} & \frac{25-13\sqrt{5}}{120} & \frac{-1+\sqrt{5}}{120} \\
\frac{5+\sqrt{5}}{10} & \frac{11-\sqrt{5}}{120} & \frac{25+13\sqrt{5}}{120} & \frac{25+\sqrt{5}}{120} & \frac{-1-\sqrt{5}}{120} \\
1 & \frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12} \\
\hline
 & \frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12}
\end{array}
$$

Lobatto IIIB $(s = 4, p = 6)$,

$$
\begin{array}{c|cccc}
0 & \frac{1}{12} & \frac{-1-\sqrt{5}}{24} & \frac{-1+\sqrt{5}}{24} & 0 \\
\frac{5-\sqrt{5}}{10} & \frac{1}{12} & \frac{25+\sqrt{5}}{120} & \frac{25-13\sqrt{5}}{120} & 0 \\
\frac{5+\sqrt{5}}{10} & \frac{1}{12} & \frac{25+13\sqrt{5}}{120} & \frac{25-\sqrt{5}}{120} & 0 \\
1 & \frac{1}{12} & \frac{11-\sqrt{5}}{120} & \frac{11+\sqrt{5}}{120} & 0 \\
\hline
& \frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12}
\end{array}
$$

Lobatto IIIC $(s = 4, p = 6)$,

$$
\begin{array}{c|cccc}
0 & \frac{1}{12} & -\frac{\sqrt{5}}{12} & \frac{\sqrt{5}}{12} & -\frac{1}{12} \\
\frac{5-\sqrt{5}}{10} & \frac{1}{12} & \frac{1}{4} & \frac{10-7\sqrt{5}}{60} & \frac{\sqrt{5}}{60} \\
\frac{5+\sqrt{5}}{10} & \frac{1}{12} & \frac{10+7\sqrt{5}}{60} & \frac{1}{4} & -\frac{\sqrt{5}}{60} \\
1 & \frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12} \\
\hline
& \frac{1}{12} & \frac{5}{12} & \frac{5}{12} & \frac{1}{12}
\end{array}
$$

Lobatto III $(s = 5, p = 8)$,

$$
\begin{array}{c|ccccc}
0 & 0 & 0 & 0 & 0 & 0 \\
\frac{7-\sqrt{21}}{14} & \frac{1}{14} & \frac{1}{9} & \frac{13-3\sqrt{21}}{63} & \frac{14-3\sqrt{21}}{126} & 0 \\
\frac{1}{2} & \frac{1}{32} & \frac{91+21\sqrt{21}}{576} & \frac{11}{72} & \frac{91-21\sqrt{21}}{576} & 0 \\
\frac{7+\sqrt{21}}{14} & \frac{1}{14} & \frac{14+3\sqrt{21}}{126} & \frac{13+3\sqrt{21}}{63} & \frac{1}{9} & 0 \\
1 & 0 & \frac{7}{18} & \frac{2}{9} & \frac{7}{18} & 0 \\
\hline
& \frac{1}{20} & \frac{49}{180} & \frac{16}{45} & \frac{49}{180} & \frac{1}{20}
\end{array}
$$

Lobatto IIIA $(s = 5, p = 8)$,

$$
\begin{array}{c|ccccc}
0 & 0 & 0 & 0 & 0 & 0 \\
\frac{7-\sqrt{21}}{14} & \frac{119+3\sqrt{21}}{1960} & \frac{343-9\sqrt{21}}{2520} & \frac{392-96\sqrt{21}}{2205} & \frac{343-69\sqrt{21}}{2520} & \frac{-21+3\sqrt{21}}{1960} \\
\frac{1}{2} & \frac{13}{320} & \frac{392+105\sqrt{21}}{2880} & \frac{8}{45} & \frac{392-105\sqrt{21}}{2880} & \frac{3}{320} \\
\frac{7+\sqrt{21}}{14} & \frac{119-3\sqrt{21}}{1960} & \frac{343+69\sqrt{21}}{2520} & \frac{392+96\sqrt{21}}{2205} & \frac{343+9\sqrt{21}}{2520} & \frac{-21-3\sqrt{21}}{1960} \\
1 & \frac{1}{20} & \frac{49}{180} & \frac{16}{45} & \frac{49}{180} & \frac{1}{20} \\
\hline
& \frac{1}{20} & \frac{49}{180} & \frac{16}{45} & \frac{49}{180} & \frac{1}{20}
\end{array}
$$

RUNGE–KUTTA METHODS   229

Lobatto IIIB $(s=5, p=8)$,

$$
\begin{array}{c|ccccc}
0 & \frac{1}{20} & \frac{-7-\sqrt{21}}{120} & \frac{1}{15} & \frac{-7+\sqrt{21}}{120} & 0 \\
\frac{7-\sqrt{21}}{14} & \frac{1}{20} & \frac{343+9\sqrt{21}}{2520} & \frac{56-15\sqrt{21}}{315} & \frac{343-69\sqrt{21}}{2520} & 0 \\
\frac{1}{2} & \frac{1}{20} & \frac{49+12\sqrt{12}}{360} & \frac{8}{45} & \frac{49-12\sqrt{12}}{360} & 0 \\
\frac{7+\sqrt{21}}{14} & \frac{1}{20} & \frac{343+69\sqrt{21}}{2520} & \frac{56+15\sqrt{21}}{315} & \frac{343-9\sqrt{21}}{2520} & 0 \\
1 & \frac{1}{20} & \frac{119-3\sqrt{21}}{360} & \frac{13}{45} & \frac{119+3\sqrt{21}}{360} & 0 \\
\hline
 & \frac{1}{20} & \frac{49}{180} & \frac{16}{45} & \frac{49}{180} & \frac{1}{20}
\end{array}
$$

Lobatto IIIC $(s=5, p=8)$,

$$
\begin{array}{c|ccccc}
0 & \frac{1}{20} & -\frac{7}{60} & \frac{2}{15} & -\frac{7}{60} & \frac{1}{20} \\
\frac{7-\sqrt{21}}{14} & \frac{1}{20} & \frac{29}{180} & \frac{47-15\sqrt{21}}{315} & \frac{203-30\sqrt{21}}{1260} & -\frac{3}{140} \\
\frac{1}{2} & \frac{1}{20} & \frac{329+105\sqrt{21}}{2880} & \frac{73}{360} & \frac{329-105\sqrt{21}}{2880} & \frac{3}{160} \\
\frac{7+\sqrt{21}}{14} & \frac{1}{20} & \frac{203+30\sqrt{21}}{1260} & \frac{47+15\sqrt{21}}{315} & \frac{29}{180} & -\frac{3}{140} \\
1 & \frac{1}{20} & \frac{49}{180} & \frac{16}{45} & \frac{49}{180} & \frac{1}{20} \\
\hline
 & \frac{1}{20} & \frac{49}{180} & \frac{16}{45} & \frac{49}{180} & \frac{1}{20}
\end{array}
$$

Exercises 34

34.1 Show that there is a unique Runge–Kutta method of order 4 with $s = 3$ for which A is lower triangular with $a_{11} = a_{33} = 0$. Find the tableau for this method.

34.2 Show that the implicit Runge–Kutta given by the tableau

$$
\begin{array}{c|cccc}
0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & \frac{1}{8} & \frac{1}{8} & 0 & 0 \\
\frac{7}{10} & -\frac{1}{100} & \frac{14}{25} & \frac{3}{20} & 0 \\
1 & \frac{2}{7} & 0 & \frac{5}{7} & 0 \\
\hline
 & \frac{1}{14} & \frac{32}{81} & \frac{250}{567} & \frac{5}{54}
\end{array}
$$

has order 5.

34.3 Find the tableau for the Gauss method with $s = 4$ and $p = 8$.

34.4 Show that Gauss methods are invariant under reflection.

35   Stability of Implicit Runge–Kutta Methods

350   A-stability, $A(\alpha)$-stability and L-stability

We recall that the stability function for a Runge–Kutta method (238b) is the rational function

$$R(z) = 1 + zb^{\mathsf{T}}(I - zA)^{-1}\mathbf{1}, \tag{350a}$$

and that a method is A-stable if

$$|R(z)| \leq 1, \quad \text{whenever } \operatorname{Re}(z) \leq 0.$$

For the solution of stiff problems, A-stability is a desirable property, and there is sometimes a preference for methods to be L-stable; this means that the method is A-stable and that, in addition,

$$R(\infty) = 0. \tag{350b}$$

Where A-stability is impossible or difficult to achieve, a weaker property is acceptable for the solution of many problems.

Definition 350A *Let α denote an angle satisfying $\alpha \in (0, \pi)$ and let $S(\alpha)$ denote the set of points $x + iy$ in the complex plane such that $x \leq 0$ and $-\tan(\alpha)|x| \leq y \leq \tan(\alpha)|x|$. A Runge–Kutta method with stability function $R(z)$ is $A(\alpha)$-stable if $|R(z)| \leq 1$ for all $z \in S(\alpha)$.*

The region $S(\alpha)$ is illustrated in Figure 350(i) in the case of the Runge–Kutta method

$$\begin{array}{c|ccc}
\lambda & \lambda & 0 & 0 \\
\frac{1+\lambda}{2} & \frac{1-\lambda}{2} & \lambda & 0 \\
1 & -\frac{(1-\lambda)(1-9\lambda+6\lambda^2)}{1-3\lambda+6\lambda^2} & \frac{2(1-\lambda)(1-6\lambda+6\lambda^2)}{1-3\lambda+6\lambda^2} & \lambda \\
\hline
 & \frac{1+3\lambda}{6(1-\lambda)^2} & \frac{2(1-3\lambda)}{3(1-\lambda)^2} & \frac{1-3\lambda+6\lambda^2}{6(1-\lambda)^2}
\end{array} \,, \tag{350c}$$

where $\lambda \approx 0.158984$ is a zero of $6\lambda^3 - 18\lambda^2 + 9\lambda - 1$. This value of λ was chosen to ensure that (350b) holds, even though the method is not A-stable. It is, in fact, $A(\alpha)$-stable with $\alpha \approx 1.31946 \approx 75.5996°$.

351   Criteria for A-stability

We first find an alternative expression for the rational function (350a).

Lemma 351A *Let (A, b, c) denote a Runge–Kutta method. Then its stability function is given by*

$$R(z) = \frac{\det\left(I + z(\mathbf{1}b^{\mathsf{T}} - A)\right)}{\det(I - zA)}.$$
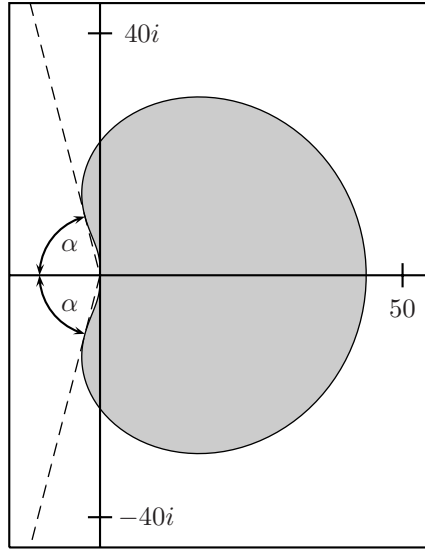
Figure 350(i)   A(α) stability region for the method (350c)

Proof. Because a rank 1 $s \times s$ matrix uv^T has characteristic polynomial $\det(Iw - uv^\mathsf{T}) = w^{s-1}(w - v^\mathsf{T}u)$, a matrix of the form $I + uv^\mathsf{T}$ has characteristic polynomial $(w-1)^{s-1}(w - 1 - v^\mathsf{T}u)$ and determinant of the form $1 + v^\mathsf{T}u$. Hence,

$$\det\left(I + z\mathbf{1}b^\mathsf{T}(I - zA)^{-1}\right) = 1 + zb^\mathsf{T}(I - zA)^{-1}\mathbf{1} = R(z).$$

We now note that

$$I + z(\mathbf{1}b^\mathsf{T} - A) = \left(I + z\mathbf{1}b^\mathsf{T}(I - zA)^{-1}\right)(I - zA),$$

so that

$$\det\left(I + z(\mathbf{1}b^\mathsf{T} - A)\right) = R(z)\det(I - zA). \qquad \square$$

Now write the stability function of a Runge–Kutta method as the ratio of two polynomials

$$R(z) = \frac{N(z)}{D(z)}$$

and define the E-polynomial by

$$E(y) = D(iy)D(-iy) - N(iy)N(-iy).$$

Theorem 351B *A Runge–Kutta method with stability function $R(z) = N(z)/D(z)$ is A-stable if and only if (a) all poles of R (that is, all zeros of D) are in the right half-plane and (b) $E(y) \geq 0$, for all real y.*

Proof. The necessity of (a) follows from the fact that if z^* is a pole then $\lim_{z \to z^*} |R(z)| = \infty$, and hence $|R(z)| > 1$, for z close enough to z^*. The necessity of (b) follows from the fact that $E(y) < 0$ implies that $|R(iy)| > 1$, so that $|R(z)| > 1$ for some $z = -\epsilon + iy$ in the left half-plane. Sufficiency of these conditions follows from the fact that (a) implies that R is analytic in the left half-plane so that, by the maximum modulus principle, $|R(z)| > 1$ in this region implies $|R(z)| > 1$ on the imaginary axis, which contradicts (b). $\square$

352   *Padé approximations to the exponential function*

Given a function f, assumed to be analytic at zero, with $f(0) \neq 0$, and given non-negative integers l and m, it is sometimes possible to approximate f by a rational function

$$f(z) \approx \frac{N(z)}{D(z)},$$

with N of degree l and D of degree m and with the error in the approximation equal to $O(z^{l+m+1})$. In the special case $m = 0$, this is exactly the Taylor expansion of f about $z = 0$, and when $l = 0$, $D(z)/N(z)$ is the Taylor expansion of $1/f(z)$.

For some specially contrived functions and particular choices of the degrees l and m, the approximation will not exist. An example of this is

$$f(z) = 1 + \sin(z) \approx 1 + z - \frac{1}{6}z^3 + \cdots, \tag{352a}$$

with $l = 2$, $m = 1$ because it is impossible to choose a to make the coefficient of z^3 equal to zero in the Taylor expansion of $(1 + az)f(z)$.

When an approximation

$$f(z) = \frac{N_{lm}(z)}{D_{lm}(z)} + O(z^{l+m+1})$$

exists, it is known as the '(l, m) Padé approximation' to f. The array of Padé approximations for $l, m = 0, 1, 2, \ldots$ is referred to as 'the Padé table' for the function f.

Padé approximations to the exponential function are especially interesting to us, because some of them are equal to the rational functions of some important Gauss, Radau and Lobatto methods. We show that the full Padé table exists for this function and, at the same time, we find explicit values for the coefficients in N and D and for the next two terms in the Taylor series for $N(z) - \exp(z)D(z)$. Because it is possible to rescale both N and D by an arbitrary factor, we specifically choose a normalization for which $N(0) = D(0) = 1$.

Theorem 352A *Let $l, m \geq 0$ be integers and define polynomials N_{lm} and D_{lm} by*

$$N_{lm}(z) = \frac{l!}{(l+m)!} \sum_{i=0}^{l} \frac{(l+m-i)!}{i!(l-i)!} z^i, \qquad (352b)$$

$$D_{lm}(z) = \frac{m!}{(l+m)!} \sum_{i=0}^{m} \frac{(l+m-i)!}{i!(m-i)!} (-z)^i. \qquad (352c)$$

Also define

$$C_{lm} = (-1)^m \frac{l!m!}{(l+m)!(l+m+1)!}.$$

Then

$$N_{lm}(z) - \exp(z)D_{lm}(z) + C_{lm} z^{l+m+1} + \frac{m+1}{l+m+2} C_{lm} z^{l+m+2} = O(z^{l+m+3}). \quad (352d)$$

Proof. In the case $m = 0$, the result is equivalent to the Taylor series for $\exp(z)$; by multiplying both sides of (352d) by $\exp(-z)$ we find that the result is also equivalent to the Taylor series for $\exp(-z)$ in the case $l = 0$. We now suppose that $l \geq 1$ and $m \geq 1$, and that (352d) has been proved if l is replaced by $l-1$ or m replaced is by $m-1$. We deduce the result for the given values of l and m so that the theorem follows by induction.

Because the result holds with l replaced by $l-1$ or with m replaced by $m-1$, we have

$$N_{l-1,m}(z) - \exp(z)D_{l-1,m}(z) + \left(1 + \frac{m+1}{l+m+1}z\right) C_{l-1,m} z^{l+m} = O(z^{l+m+2}), \quad (352e)$$

$$N_{l,m-1}(z) - \exp(z)D_{l,m-1}(z) + \left(1 + \frac{m}{l+m+1}z\right) C_{l,m-1} z^{l+m} = O(z^{l+m+2}). \quad (352f)$$

Multiply (352e) by $l/(l+m)$ and (352f) by $m/(l+m)$, and we find that the coefficient of z^{l+m} has the value

$$\frac{l}{l+m} C_{l-1,m} + \frac{m}{l+m} C_{l,m-1} = 0.$$

The coefficient of z^{l+m+1} is found to be equal to C_{lm}. Next we verify that

$$\frac{l}{l+m} N_{l-1,m}(z) + \frac{m}{l+m} N_{l,m-1}(z) - N_{lm}(z) = 0 \qquad (352g)$$

and that

$$\frac{l}{l+m} D_{l-1,m}(z) + \frac{m}{l+m} D_{l,m-1}(z) - D_{lm}(z) = 0. \qquad (352h)$$

Table 352(I)   Padé approximations N_{lm}/D_{lm} for $l, m = 0, 1, 2, 3$

m\\l	0	1	2	3
0	1	$1+z$	$1+z+\frac{1}{2}z^2$	$1+z+\frac{1}{2}z^2+\frac{1}{6}z^3$
1	$\frac{1}{1-z}$	$\frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$	$\frac{1+\frac{2}{3}z+\frac{1}{6}z^2}{1-\frac{1}{3}z}$	$\frac{1+\frac{3}{4}z+\frac{1}{4}z^2+\frac{1}{24}z^3}{1-\frac{1}{4}z}$
2	$\frac{1}{1-z+\frac{1}{2}z^2}$	$\frac{1+\frac{1}{3}z}{1-\frac{2}{3}z+\frac{1}{6}z^2}$	$\frac{1+\frac{1}{2}z+\frac{1}{12}z^2}{1-\frac{1}{2}z+\frac{1}{12}z^2}$	$\frac{1+\frac{3}{5}z+\frac{3}{20}z^2+\frac{1}{60}z^3}{1-\frac{2}{5}z+\frac{1}{20}z^2}$
3	$\frac{1}{1-z+\frac{1}{2}z^2-\frac{1}{6}z^3}$	$\frac{1+\frac{1}{4}z}{1-\frac{3}{4}z+\frac{1}{4}z^2-\frac{1}{24}z^3}$	$\frac{1+\frac{2}{5}z+\frac{1}{20}z^2}{1-\frac{3}{5}z+\frac{3}{20}z^2-\frac{1}{60}z^3}$	$\frac{1+\frac{1}{2}z+\frac{1}{10}z^2+\frac{1}{120}z^3}{1-\frac{1}{2}z+\frac{1}{10}z^2-\frac{1}{120}z^3}$

The coefficient of z^i in (352g) is

$$\frac{(l-1)!(l+m-i-1)!}{(l+m)!i!(l-i)!}\left(l(l-i)+ml-l(l+m-i)\right)=0,$$

so that (352g) follows. The verification of (352h) is similar and will be omitted. It now follows that

$$N_{lm}(z)-\exp(z)D_{lm}(z)+C_{lm}z^{l+m+1}+\tfrac{m+1}{l+m+2}\widetilde{C}_{lm}z^{l+m+2}=O(z^{l+m+3}),\quad (352i)$$

and we finally need to prove that $\widetilde{C}_{lm}=C_{lm}$. Operate on both sides of (352i) with the operator $(d/dz)^{l+1}$ and multiply the result by $\exp(-z)$. This gives

$$P(z)+\left(\tfrac{m+1}{l+m+2}\tfrac{(l+m+2)!}{(m+1)!}\widetilde{C}_{lm}-\tfrac{(l+m+1)!}{m!}C_{lm}\right)z^{m+1}=O(z^{m+2}),\qquad (352j)$$

where P is the polynomial of degree m given by

$$P(z)=\frac{(l+m+1)!}{m!}C_{lm}z^m-\left(1+\frac{d}{dz}\right)^{l+1}D_{lm}(z).$$

It follows from (352j) that $\widetilde{C}_{lm}=C_{lm}$. $\square$

The formula we have found for a possible (l,m) Padé approximation to $\exp(z)$ is unique. This is not the case for an arbitrary function f, as the example of the function given by (352a) shows; the $(2,1)$ approximation is not unique. The case of the exponential function is covered by the following result:

Theorem 352B *The function N_{lm}/D_{lm}, where the numerator and denominator are given by (352b) and (352c), is the unique (l,m) Padé approximation to the exponential function.*

Proof. If $\widehat{N}_{lm}/\widehat{D}_{lm}$ is a second such approximation then, because these functions differ by $O(z^{l+m+1})$,

$$N_{lm}\widehat{D}_{lm}-\widehat{N}_{lm}D_{lm}=0,$$

Table 352(II)   Diagonal members of the Padé table N_{mm}/D_{mm} for $m = 0, 1, 2, \ldots, 7$

m	$\dfrac{N_{mm}}{D_{mm}}$
0	1
1	$\dfrac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}$
2	$\dfrac{1 + \frac{1}{2}z + \frac{1}{12}z^2}{1 - \frac{1}{2}z + \frac{1}{12}z^2}$
3	$\dfrac{1 + \frac{1}{2}z + \frac{1}{10}z^2 + \frac{1}{120}z^3}{1 - \frac{1}{2}z + \frac{1}{10}z^2 - \frac{1}{120}z^3}$
4	$\dfrac{1 + \frac{1}{2}z + \frac{3}{28}z^2 + \frac{1}{84}z^3 + \frac{1}{1680}z^4}{1 - \frac{1}{2}z + \frac{3}{28}z^2 - \frac{1}{84}z^3 + \frac{1}{1680}z^4}$
5	$\dfrac{1 + \frac{1}{2}z + \frac{1}{9}z^2 + \frac{1}{72}z^3 + \frac{1}{1008}z^4 + \frac{1}{30240}z^5}{1 - \frac{1}{2}z + \frac{1}{9}z^2 - \frac{1}{72}z^3 + \frac{1}{1008}z^4 - \frac{1}{30240}z^5}$
6	$\dfrac{1 + \frac{1}{2}z + \frac{5}{44}z^2 + \frac{1}{66}z^3 + \frac{1}{792}z^4 + \frac{1}{15840}z^5 + \frac{1}{665280}z^6}{1 - \frac{1}{2}z + \frac{5}{44}z^2 - \frac{1}{66}z^3 + \frac{1}{792}z^4 - \frac{1}{15840}z^5 + \frac{1}{665280}z^6}$
7	$\dfrac{1 + \frac{1}{2}z + \frac{3}{26}z^2 + \frac{5}{312}z^3 + \frac{5}{3432}z^4 + \frac{1}{11440}z^5 + \frac{1}{308880}z^6 + \frac{1}{17297280}z^7}{1 - \frac{1}{2}z + \frac{3}{26}z^2 - \frac{5}{312}z^3 + \frac{5}{3432}z^4 - \frac{1}{11440}z^5 + \frac{1}{308880}z^6 - \frac{1}{17297280}z^7}$

because the expression on the left-hand side is $O(z^{l+m+1})$, and is at the same time a polynomial of degree not exceeding $l+m$. Hence, the only way that two distinct approximations can exist is when they can be cancelled to a rational function of lower degrees. This means that for some (l, m) pair, there exists a Padé approximation for which the error coefficient is zero. However, since $\exp(z)$ is not *equal* to a rational function, there is some higher exponent k and a non-zero constant C such that

$$N_{lm}(z) - \exp(z)D_{lm}(z) = Cz^k + O(z^{k+1}), \tag{352k}$$

with $k \geq l + m + 2$. Differentiate (352k) $k - m - 1$ times, multiply the result by $\exp(-z)$ and then differentiate a further $m + 1$ times. This leads to the contradictory conclusion that $C = 0$. $\square$

Expressions for the (l, m) Padé approximations are given in Table 352(I) for $l, m = 0, 1, 2, 3$. To extend the information further, Table 352(II) is presented to give the values for $l = m = 0, 1, 2, \ldots, 7$. Similar tables are also given for the first and second sub-diagonals in Tables 352(III) and 352(IV), respectively, and error constants corresponding to entries in each of these three tables are presented in Table 352(V).

Table 352(III)   First sub-diagonal members of the Padé table $N_{m-1,m}/D_{m-1,m}$
for $m = 1, 2, \ldots, 7$

m	$\dfrac{N_{m-1,m}}{D_{m-1,m}}$
1	$\dfrac{1}{1-z}$
2	$\dfrac{1 + \frac{1}{3}z}{1 - \frac{2}{3}z + \frac{1}{6}z^2}$
3	$\dfrac{1 + \frac{2}{5}z + \frac{1}{20}z^2}{1 - \frac{3}{5}z + \frac{3}{20}z^2 - \frac{1}{60}z^3}$
4	$\dfrac{1 + \frac{3}{7}z + \frac{1}{14}z^2 + \frac{1}{210}z^3}{1 - \frac{4}{7}z + \frac{1}{7}z^2 - \frac{2}{105}z^3 + \frac{1}{840}z^4}$
5	$\dfrac{1 + \frac{4}{9}z + \frac{1}{12}z^2 + \frac{1}{126}z^3 + \frac{1}{3024}z^4}{1 - \frac{5}{9}z + \frac{5}{36}z^2 - \frac{5}{252}z^3 + \frac{5}{3024}z^4 - \frac{1}{15120}z^5}$
6	$\dfrac{1 + \frac{5}{11}z + \frac{1}{11}z^2 + \frac{1}{99}z^3 + \frac{1}{1584}z^4 + \frac{1}{55440}z^5}{1 - \frac{6}{11}z + \frac{3}{22}z^2 - \frac{2}{99}z^3 + \frac{1}{528}z^4 - \frac{1}{9240}z^5 + \frac{1}{332640}z^6}$
7	$\dfrac{1 + \frac{6}{13}z + \frac{5}{52}z^2 + \frac{5}{429}z^3 + \frac{1}{1144}z^4 + \frac{1}{25740}z^5 + \frac{1}{1235520}z^6}{1 - \frac{7}{13}z + \frac{7}{52}z^2 - \frac{35}{1716}z^3 + \frac{7}{3432}z^4 - \frac{7}{51480}z^5 + \frac{7}{1235520}z^6 - \frac{1}{8648640}z^7}$

For convenience, we write $V_{mn}(z)$ for the two-dimensional vector whose first component is $N_{lm}(z)$ and whose second component is $D_{lm}(z)$. From the proof of Theorem 352A, it can be seen that the three such vectors $V_{l-1,m}(z)$, $V_{l,m-1}(z)$ and $V_{l,m}(z)$ are related by

$$lV_{l-1,m}(z) + mV_{l,m-1}(z) = (l+m)V_{l,m}(z).$$

Many similar relations between neighbouring members of a Padé table exist, and we present three of them. In each case the relation is between three Padé vectors of successive denominator degrees.

Theorem 352C *If $l, m \geq 2$ then*

$$V_{lm}(z) = \left(1 + \frac{m-l}{(l+m)(l+m-2)}z\right)V_{l-1,m-1}(z)$$

$$+ \frac{(l-1)(m-1)}{(l+m-1)(l+m-2)^2(l+m-3)}z^2 V_{l-2,m-2}(z).$$

Table 352(IV) Second sub-diagonal members of the Padé table
$N_{m-2,m}/D_{m-2,m}$ for $m = 2, 3, \ldots, 7$

m	$\dfrac{N_{m-2,m}}{D_{m-2,m}}$
2	$\dfrac{1}{1 - z + \frac{1}{2}z^2}$
3	$\dfrac{1 + \frac{1}{4}z}{1 - \frac{3}{4}z + \frac{1}{4}z^2 - \frac{1}{24}z^3}$
4	$\dfrac{1 + \frac{1}{3}z + \frac{1}{30}z^2}{1 - \frac{2}{3}z + \frac{1}{5}z^2 - \frac{1}{30}z^3 + \frac{1}{360}z^4}$
5	$\dfrac{1 + \frac{3}{8}z + \frac{3}{56}z^2 + \frac{1}{336}z^3}{1 - \frac{5}{8}z + \frac{5}{28}z^2 - \frac{5}{168}z^3 + \frac{1}{336}z^4 - \frac{1}{6720}z^5}$
6	$\dfrac{1 + \frac{2}{5}z + \frac{1}{15}z^2 + \frac{1}{180}z^3 + \frac{1}{5040}z^4}{1 - \frac{3}{5}z + \frac{1}{6}z^2 - \frac{1}{36}z^3 + \frac{1}{336}z^4 - \frac{1}{5040}z^5 + \frac{1}{151200}z^6}$
7	$\dfrac{1 + \frac{5}{12}z + \frac{5}{66}z^2 + \frac{1}{132}z^3 + \frac{1}{2376}z^4 + \frac{1}{95040}z^5}{1 - \frac{7}{12}z + \frac{7}{44}z^2 - \frac{7}{264}z^3 + \frac{7}{2376}z^4 - \frac{7}{31680}z^5 + \frac{1}{95040}z^6 - \frac{1}{3991680}z^7}$

Proof. Let

$$V(z) = V_{lm}(z) - \left(1 + \frac{m - l}{(l + m)(l + m - 2)}z\right)V_{l-1,m-1}(z)$$
$$- \frac{(l - 1)(m - 1)}{(l + m - 1)(l + m - 2)^2(l + m - 3)}z^2 V_{l-2,m-2}(z).$$

It is easy to verify that the coefficients of z^0, z^1 and z^2 vanish in both components of $V(z)$. We also find that

$$[1 \quad - \exp(z)]V(z) = O(z^{l+m-1}).$$

If $V(z)$ is not the zero vector, we find that

$$z^{-2}[1 \quad - \exp(z)]V(z) = O(z^{l+m-3}),$$

contradicting the uniqueness of Padé approximations of degrees $(l - 2, m - 2)$.
□

Theorems 352D and 352E which follow are proved in the same way as Theorem 352C and the details are omitted.

Table 352(V)   Error constants for diagonal and first two sub-diagonals

m	$C_{m-2,m}$	$C_{m-1,m}$	C_{mm}
0			1
1		$-\frac{1}{2}$	$-\frac{1}{12}$
2	$\frac{1}{6}$	$\frac{1}{72}$	$\frac{1}{720}$
3	$-\frac{1}{480}$	$-\frac{1}{7200}$	$-\frac{1}{100800}$
4	$\frac{1}{75600}$	$\frac{1}{1411200}$	$\frac{1}{25401600}$
5	$-\frac{1}{20321280}$	$-\frac{1}{457228800}$	$-\frac{1}{10059033600}$
6	$\frac{1}{8382528000}$	$\frac{1}{221298739200}$	$\frac{1}{5753767219200}$
7	$-\frac{1}{4931800473600}$	$-\frac{1}{149597947699200}$	$-\frac{1}{4487938430976000}$

Theorem 352D *If $l \geq 1$ and $m \geq 2$ then*

$$V_{lm}(z) = \left(1 - \frac{l}{(l+m)(l+m-1)}z\right) V_{l,m-1}(z)$$
$$+ \frac{l(m-1)}{(l+m)(l+m-1)^2(l+m-2)}z^2 V_{l-1,m-2}(z).$$

Theorem 352E *If $l \geq 0$ and $m \geq 2$ then*

$$V_{lm}(z) = \left(1 - \frac{1}{l+m}z\right)V_{l+1,m-1}(z) + \frac{m-1}{(l+m)^2(l+m-1)}z^2 V_{l,m-2}(z).$$

353   A-stability of Gauss and related methods

We consider the possible A-stability of methods whose stability functions correspond to members on the diagonal and first two sub-diagonals of the Padé table for the exponential function. These include the Gauss methods and the Radau IA and IIA methods as well as the Lobatto IIIC methods. A corollary is that the Radau IA and IIA methods and the Lobatto IIIC methods are L-stable.

Theorem 353A *Let s be a positive integer and let*

$$R(z) = \frac{N(z)}{D(z)}$$

denote the $(s-d, s)$ member of the Padé table for the exponential function, where $d = 0, 1$ or 2. Then

$$|R(z)| \leq 1,$$

for all complex z satisfying $\mathrm{Re}\, z \leq 0$.

Proof. We use the E-polynomial. Because $N(z) = \exp(z)D(z) + O(z^{2s-d+1})$, we have

$$
\begin{aligned}
E(y) &= D(iy)D(-iy) - N(iy)N(-iy) \\
&= D(iy)D(-iy) - \exp(iy)D(iy)\exp(-iy)D(-iy) + O(y^{2s-d+1}) \\
&= O(y^{2s-d+1}).
\end{aligned}
$$

Because $E(y)$ has degree not exceeding $2s$ and is an even function, either $E(y) = 0$, in the case $d = 0$, or $E(y) = Cy^{2s}$ with $C > 0$, in the cases $d = 1$ and $d = 2$. In all cases, $E(y) \geq 0$ for all real y.

To complete the proof, we must show that the denominator of R has no zeros in the left half-plane. Without loss of generality, we assume that $\mathrm{Re}\, z < 0$ and we prove that $D(z) \neq 0$. Write D_0, D_1, ..., D_s for the denominators of the sequence of Padé approximations given by

$$
V_{00}, V_{11}, \ldots, V_{s-1,s-1}, V_{s-d,s},
$$

so that $D(z) = D_s(z)$. From Theorems 352C, 352D and 352E, we have

$$
D_k(z) = D_{k-1}(z) + \frac{1}{4(2k-1)(2k-3)}z^2 D_{k-2}, \quad k = 2, 3, \ldots, s-1, \quad (353a)
$$

and

$$
D_s(z) = (1 - \alpha z)D_{s-1} + \beta z^2 D_{s-2}, \quad (353b)
$$

where the constants α and β will depend on the value of d and s. However, $\alpha = 0$ if $d = 0$ and $\alpha > 0$ for $d = 1$ and $d = 2$. In all cases, $\beta > 0$.

Consider the sequence of complex numbers, ζ_k, for $k = 1, 2, \ldots, s$, defined by

$$
\zeta_1 = 1 - \frac{1}{2}z,
$$

$$
\zeta_k = 1 + \frac{1}{4(2k-1)(2k-3)}z^2\zeta_{k-1}^{-1}, \quad k = 2, 3, \ldots, s-1,
$$

$$
\zeta_s = (1 - \alpha z) + \beta z^2 \zeta_{s-1}^{-1}.
$$

This means that $\zeta_1/z = -1/2 + 1/z$ has negative real part. We prove by induction that ζ_k/z also has negative real part for $k = 2, 3, \ldots, s$. We see this by noting that

$$
\frac{\zeta_k}{z} = \frac{1}{z} + \frac{1}{4(2k-1)(2k-3)}\left(\frac{\zeta_{k-1}}{z}\right)^{-1}, \quad k = 2, 3, \ldots, s-1,
$$

$$
\frac{\zeta_s}{z} = \frac{1}{z} - \alpha + \beta\left(\frac{\zeta_{s-1}}{z}\right)^{-1}.
$$

The fact that $D_s(z)$ cannot vanish now follows by observing that

$$D_s(z) = \zeta_1 \zeta_2 \zeta_3 \cdots \zeta_s.$$

Hence, $D = D_s$ does not have a zero in the left half-plane.   □

Alternative proofs of this and related results have been given by Axelsson (1969, 1972), Butcher (1977), Ehle (1973), Ehle and Picel (1975), Watts and Shampine (1972) and Wright (1970).

354   Order stars

We have identified some members of the Padé table for the exponential function for which the corresponding numerical methods are A-stable. We now ask: are there other members of the table with this property? It will be seen that everything hinges on the value of $m-l$, the degree of the denominator minus the degree of the numerator. It is clear that if $m-l < 0$, A-stability is impossible, because in this case

$$|R(z)| \to \infty,$$

as $z \to \infty$, and hence, for some z satisfying $\mathrm{Re}\, z < 0$, $|R(z)|$ is greater than 1. For $m-l \in \{0, 1, 2\}$, A-stability follows from Theorem 353A. Special cases with $m-l > 2$ suggest that these members of the Padé table are not A-stable. For the third sub-diagonal, this was proved by Ehle (1969), and for the fourth and fifth sub-diagonals by Nørsett (1974). Based on these observations, Ehle (1973) conjectured that no case with $m-l > 2$ can be A-stable. This result was eventually proved by Wanner, Hairer and Nørsett (1978), and we devote this subsection to introducing the approximations considered in that paper and to proving the Ehle conjecture.

In Subsection 216, we touched on the idea of an order star. Associated with the stability function $R(z)$ for a Runge–Kutta method, we consider the set of points in the complex plane such that

$$|\exp(-z)R(z)| > 1.$$

This is known as the 'order star' of the method, and the set of points such that

$$|\exp(-z)R(z)| < 1$$

is the 'dual order star'. The common boundary of these two sets traces out an interesting path, as we see illustrated in Figure 354(i), for the case of the $(1, 3)$ Padé approximation given by

$$R(z) = \frac{1 + \frac{1}{4}z}{1 - \frac{3}{4}z + \frac{1}{4}z^2 - \frac{1}{24}z^3}.$$
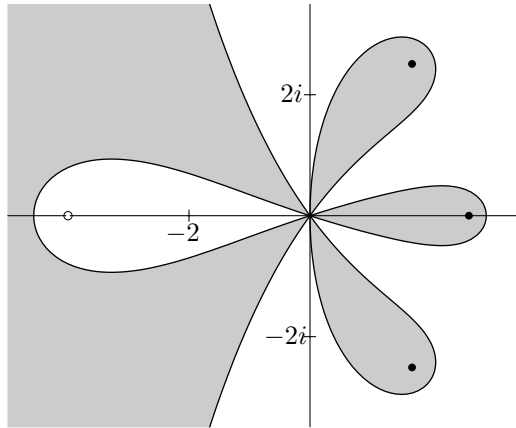
Figure 354(i)   Order star for the $(1, 3)$ Padé approximation to exp

In this diagram, the dual order star, which can also be described as the 'relative stability region', is the interior of the unshaded region. The order star is the interior of the shaded region.

In Butcher (1987) an attempt was made to present an informal survey of order stars leading to a proof of the Ehle result. In the present volume, the discussion of order stars will be even more brief, but will serve as an introduction to an alternative approach to achieve similar results. In addition to Wanner, Hairer and Nørsett (1978), the reader is referred to Iserles and Nørsett (1991) for fuller information and applications of order stars.

The 'order star', for a particular rational approximation to the exponential function, disconnects into 'fingers' emanating from the origin, which may be bounded or not, and similar remarks apply to 'dual fingers' which are the connected components of the dual star. The following statements summarize the key properties of order stars for applications of the type we are considering. Because we are including only hints of the proofs, we refer to them as remarks rather than as lemmas or theorems. Note that S denotes the order star for a specific 'method' and I denotes the imaginary axis.

Remark 354A *A method is A-stable if and only if S has no poles in the negative half-plane and $S \cup I = \emptyset$, because the inclusion of the exponential factor does not alter the set of poles and does not change the magnitude of the stability function on I.*

Remark 354B *There exists $\rho_0 > 0$ such that, for all $\rho \geq \rho_0$, functions $\theta_1(\rho)$ and $\theta_2(\rho)$ exist such that the intersection of S with the circle $|z| = \rho$ is the set $\{\rho \exp(i\theta) : \theta_1 < \theta < \theta_2\}$ and where $\lim_{\rho \to \infty} \theta_1(\rho) = \pi/2$ and $\lim_{\rho \to \infty} \theta_2(\rho) = 3\pi/2$, because at a great distance from the origin, the*
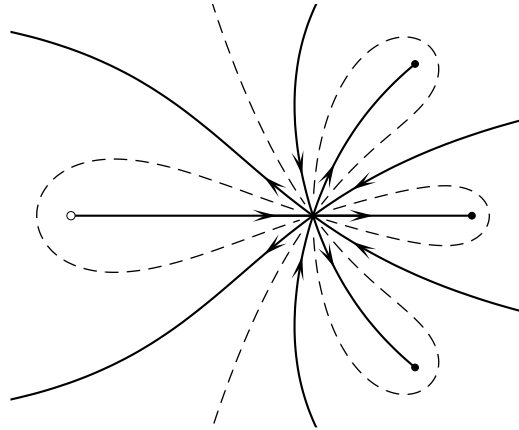
Figure 354(ii)   Relation between order arrows and order stars

behaviour of the exponential function multiplied by the rational function on which the order star is based is dominated by the exponential factor.

Remark 354C *For a method of order p, the arcs $\{r\exp(i(j+\frac{1}{2})\pi/(p+1)) : 0 \leq r\}$, where $j = 0, 1, \ldots, 2p+1$, are tangential to the boundary of S at 0, because $\exp(-z)R(z) = 1 + Cz^{p+1} + O(|z|^{p+2})$, so that $|\exp(-z)R(z)| = 1 + \mathrm{Re}(Cz^{p+1}) + O(|z|^{p+2})$.*

It is possible that m bounded fingers can join together to make up a finger of multiplicity m. Similarly, m dual fingers in $\overline{S}$ can combine to form a dual finger with multiplicity m.

Remark 354D *Each bounded finger of S, with multiplicity m, contains at least m poles, counted with their multiplicities, because, by the Cauchy–Riemann conditions, the argument of $\exp(-z)R(z)$ increases monotonically as the boundary of the order star is traced out in a counter-clockwise direction.*

In the following subsection, we introduce a slightly different tool for studying stability questions. The basic idea is to use, rather than the fingers and dual fingers as in order star theory, the lines of steepest ascent and descent from the origin. Since these lines correspond to values for which $R(z)\exp(-z)$ is real and positive, we are, in reality, looking at the set of points in the complex plane where this is the case.

We illustrate this by presenting, in Figure 354(ii), a modified version of Figure 354(i), in which the boundary of the order star is shown as a dashed line and the 'order arrows', as we call them, are shown with arrow heads showing the direction of ascent.

355   Order arrows and the Ehle barrier

For a stability function $R(z)$ of order p, define two types of 'order arrows' as follows:

Definition 355A *The locus of points in the complex plane for which $\phi(z) = R(z)\exp(-z)$ is real and positive is said to be the 'order web' for the rational function R. The part of the order web connected to 0 is the 'principal order web'. The rays emanating from 0 with increasing value of ϕ are 'up arrows' and those emanating from 0 with decreasing ϕ are 'down arrows'.*

The up and down arrows leave the origin in a systematic pattern:

Theorem 355B *Let R be a rational approximation to* exp *of exact order p, so that*
$$R(z) = \exp(z) - Cz^{p+1} + O(z^{p+2}),$$
where the error constant C is non-zero. If $C < 0$ $(C > 0)$ there are up (down) arrows tangential at 0 to the rays with arguments $k2\pi i/(p+1)$, $k = 0, 1, \ldots, p$, and down (up) arrows tangential at 0 to the rays with arguments $(2k+1)\pi i/(p+1)$, $k = 0, 1, \ldots, p$.

Proof. If, for example, $C < 0$, consider the set $\{r\exp(i\theta) : r > 0, \theta \in [k2\pi i/(p+1) - \epsilon, k2\pi i/(p+1) + \epsilon\}$, where ϵ and r are both small and $k \in \{0, 1, 2, \ldots, p\}$. We have
$$R(z)\exp(-z) = 1 + (-C)r^{p+1}\exp((p+1)\theta) + O(r^{p+2}).$$

For r sufficiently small, the last term is negligible and, for ϵ sufficiently small, the real part of $(-C)r^{p+1}\exp((p+1)\theta))$ is positive. The imaginary part changes sign so that an up arrow lies in this wedge. The cases of the down arrows and for $C > 0$ are proved in a similar manner.   □

Where the arrows leaving the origin terminate is of crucial importance.

Theorem 355C *The up arrows terminate either at poles of R or at $-\infty$. The down arrows terminate either at zeros of R or at $+\infty$.*

Proof. Consider a point on an up arrow for which $|z|$ is sufficiently large to ensure that it is not possible that z is a pole or that z is real with $(d/dz)(R(z)\exp(-z)) = 0$. In this case we can assume without loss of generality that $\mathrm{Im}(z) \geq 0$. Write $R(z) = Kz^n + O(|z|^{n-1})$ and assume that $K > 0$ (if $K < 0$, a slight change is required in the details which follow). If $z = x + iy = r\exp(i\theta)$, then
$$\begin{aligned} w(z) &= R(z)\exp(-z) \\ &= Kr^n\exp(-x)\big(1 + O(r^{-1})\big)\exp\big(i(n\theta - y + O(r^{-1}))\big). \end{aligned}$$

Because θ cannot leave the interval $[0, \pi]$, then for w to remain real, y is bounded as $z \to \infty$. Furthermore, $w \to \infty$ implies that $x \to -\infty$.

The result for the down arrows is proved in a similar way. $\qquad\square$

We can obtain more details about the fate of the arrows from the following result.

Theorem 355D *Let R be a rational approximation to* $\exp$ *of order p with numerator degree n and denominator degree d. Let $\widehat{n}$ denote the number of down arrows terminating at zeros and $\widehat{d}$ the number of up arrows terminating at poles of R. Then*

$$\widehat{n} + \widehat{d} \geq p.$$

Proof. There are $p + 1 - \widehat{n}$ down arrows and $p + 1 - \widehat{d}$ up arrows terminating at $+\infty$ and $-\infty$, respectively. Let θ and ϕ be the minimum angles with the properties that all the down arrows which terminate at $+\infty$ lie within θ on either side of the positive real axis and all the up arrows which terminate at $-\infty$ lie within an angle ϕ on either side of the negative real axis. Hence

$$2\theta \geq \frac{(p - \widehat{n})2\pi}{p + 1}, \quad 2\phi \geq \frac{(p - \widehat{d})2\pi}{p + 1}.$$

Because up arrows and down arrows cannot cross and, because there is a wedge with angle equal to at least $\pi/(p+1)$ between the last down arrow and the first up arrow, it follows that $2\theta + 2\phi + 2\pi/(p+1) \leq 2\pi$. Hence we obtain the inequality

$$\frac{2p + 1 - \widehat{n} - \widehat{d}}{p + 1} 2\pi \leq 2\pi,$$

and the result follows. $\qquad\square$

For Padé approximations we can obtain precise values of $\widehat{n}$ and $\widehat{d}$.

Theorem 355E *Let $R(z)$ denote a Padé approximation to $\exp(z)$, with degrees n (numerator) and d (denominator). Then n of the down arrows terminate at zeros and d of the up arrows terminate at poles.*

Proof. Because $p = n + d$, $n \geq \widetilde{n}$ and $d \geq \widetilde{d}$, it follows from Theorem 355D that

$$p = n + d \geq \widetilde{n} + \widetilde{d} \geq p$$

and hence that $(n - \widetilde{n}) + (d - \widetilde{d}) = 0$. Since both terms are non-negative they must be zero and the result follows. $\qquad\square$

Before proving the 'Ehle barrier', we establish a criterion for A-stability based on the up arrows that terminate at poles.

Theorem 355F *A Runge–Kutta method is A-stable only if all poles of the stability function $R(z)$ lie in the right half-plane and no up arrow of the order web intersects with or is tangential to the imaginary axis.*

Proof. The requirement on the poles is obvious. If an up arrow intersects or is tangential to the imaginary axis then there exists y such that

$$|R(iy)\exp(-iy)| > 1.$$

Because $|\exp(-iy)| = 1$, it follows that $|R(iy)| > 1$ and the method is not A-stable.   □

We are now in a position to prove the result formerly known as the Ehle conjecture (Ehle, 1973),but which we will also refer to as the 'Ehle barrier'.

Theorem 355G *Let $R(z)$ denote the stability function of a Runge–Kutta method. If $R(z)$ is an (n, d) Padé approximation to $\exp(z)$ then the Runge–Kutta is not A-stable unless $d \leq n + 2$.*

Proof. If $d \geq n + 3$ and $p = n + d$, it follows that $d \geq \frac{1}{2}(p + 3)$. By Theorem 355E, at least d up arrows terminate at poles. Suppose these leave zero in directions between $-\theta$ and $+\theta$ from the positive real axis. Then

$$2\theta \geq \frac{2\pi(d - 1)}{p + 1} \geq \pi,$$

and at least one up arrow, which terminates at a pole, is tangential to the imaginary axis or passes into the left half-plane. If the pole is in the left half-plane, then the stability function is unbounded in this half-plane. On the other hand, if the pole is in the right half-plane, then the up arrow must cross the imaginary axis. In either case, the method cannot be A-stable, by Theorem 355F.   □

356   AN-stability

Linear stability analysis is based on the linear test problem

$$y'(x) = qy(x),$$

so that

$$y_n = R(z)y_{n-1},$$

where $z = hq$. Even though this analysis provides useful information about the behaviour of a numerical method when applied to a stiff problem, even more is learned from generalizing this analysis in two possible ways. The first

of these generalizations allows the linear factor q to be time-dependent so that the test problem becomes

$$y'(x) = q(x)y(x). \tag{356a}$$

A second generalization, which we explore in Subsection 357, allows the differential equation to be non-linear.

When (356a) is numerically solved using an implicit Runge–Kutta method (A, b^{T}, c), the stage values satisfy the equations

$$Y_i = y_{n-1} + \sum_{j=1}^{s} a_{ij} hq(x_{n-1} + hc_j)Y_j, \qquad i = 1, 2, \ldots, s,$$

and the output result is

$$y_n = y_{n-1} + \sum_{i=1}^{s} b_i hq(x_{n-1} + hc_i)Y_i.$$

Let Z denote the diagonal matrix given by

$$Z = \begin{bmatrix} hq(x_{n-1} + hc_1) & 0 & \cdots & 0 \\ 0 & hq(x_{n-1} + hc_2) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & hq(x_{n-1} + hc_s) \end{bmatrix}$$

$$= \mathrm{diag}\left(\begin{bmatrix} hq(x_{n-1} + hc_1) & hq(x_{n-1} + hc_2) & \cdots & hq(x_{n-1} + hc_s) \end{bmatrix} \right).$$

This makes it possible to write the vector of stage values in the form

$$Y = y_{n-1}\mathbf{1} + AZY,$$

so that

$$Y = (I - AZ)^{-1}\mathbf{1}y_{n-1}.$$

The output value is given by

$$y_n = y_{n-1} + b^{\mathsf{T}}ZY = \left(1 + b^{\mathsf{T}}Z(I - AZ)^{-1}\mathbf{1}\right)y_{n-1} = R(Z)y_{n-1}.$$

The function $R(Z)$ introduced here is the non-autonomous generalization of the linear stability function.

We are mainly concerned with situations in which the stage abscissae are distinct and where they do not interfere with the stages of adjoining steps. This means that we can regard the diagonal elements of Z as different from each other and independent of the values in the steps that come before or after the current step. With this in mind, we define a non-autonomous counterpart of A-stability that will guarantee that we obtain stable behaviour as long as the real part of $q(x)$ is never positive. This is appropriate because the exact solution to (356a) is never increasing under this assumption, and we want to guarantee that this property carries over to the computed solution.

Definition 356A *A Runge–Kutta method* (A, b^T, c) *is 'AN-stable' if the function*

$$R(Z) = 1 + b^\mathsf{T} Z (I - AZ)^{-1} \mathbf{1},$$

where $Z = \operatorname{diag}\left(\begin{bmatrix} z_1 & z_2 & \cdots & z_s \end{bmatrix}\right)$ *is bounded in magnitude by 1 whenever* $z_1, z_2, \ldots, z_s$ *are in the left half-plane.*

It is interesting that a simple necessary and sufficient condition exists for AN-stability. In Theorem 356C we state this criterion and prove it only in terms of necessity. Matters become complicated if the method can be reduced to a method with fewer stages that gives exactly the same computed result. This can happen, for example, if there exists $j \in \{1, 2, \ldots, s\}$ such that $b_j = 0$, and furthermore, $a_{ij} = 0$ for all $i = 1, 2, \ldots, s$, except perhaps for $i = j$. Deleting stage j has no effect on the numerical result computed in a step. We make a detailed study of reducibility in Subsection 381, but in the meantime we identify 'irreducibility in the sense of Dahlquist and Jeltsch', or 'DJ-irreducibility', (Dahlquist and Jeltsch, 1979) as the property that a tableau cannot be reduced in the sense of Definition 356B.

Definition 356B *A Runge–Kutta method is 'DJ-reducible' if there exists a partition of the stages*

$$\{1, 2, \ldots, s\} = S \cup S_0,$$

with S_0 *non-empty, such that if* $i \in S$ *and* $j \in S_0$,

$$b_j = 0 \text{ and } a_{ij} = 0.$$

The 'reduced method' is the method formed by deleting all stages numbered by members of the set S_0.

The necessary condition to be given in Theorem 356C will be strengthened under DJ-irreducibility in Corollary 356D.

Theorem 356C *Let* (A, b^T, c) *be an implicit Runge–Kutta method. Then the method is AN-stable only if*

$$b_j \geq 0, \qquad j = 1, 2, \ldots, s,$$

and the matrix

$$M = \operatorname{diag}(b) A + A^\mathsf{T} \operatorname{diag}(b) - b b^\mathsf{T}$$

is positive semi-definite.

Proof. If $b_j < 0$ then choose $Z = -t \operatorname{diag}(e_j)$, for t positive. The value of $R(Z)$ becomes

$$R(Z) = 1 - t b_j + O(t^2),$$

which is greater than 1 for t sufficiently small. Now consider Z chosen with purely imaginary components

$$Z = i\operatorname{diag}(vt),$$

where v has real components and t is a small positive real. We have

$$
\begin{aligned}
R(Z) &= 1 + itb^\mathsf{T}\operatorname{diag}(v)\mathbf{1} - t^2 b^\mathsf{T}\operatorname{diag}(v)A\operatorname{diag}(v)\mathbf{1} + O(t^3) \\
&= 1 + itb^\mathsf{T}v - t^2 v^\mathsf{T}\operatorname{diag}(b)Av + O(t^3),
\end{aligned}
$$

so that

$$|R(Z)|^2 = 1 - t^2 v^\mathsf{T}Mv + O(t^3).$$

Since this cannot exceed 1 for t small and any choice of v, M is positive semi-definite. □

Since there is no practical interest in reducible methods, we might look at the consequences of assuming a method is irreducible. This result was published in Dahlquist and Jeltsch (1979):

Corollary 356D *Under the same conditions of Theorem 356C, with the additional assumption that the method is DJ-irreducible,*

$$b_j > 0, \qquad j = 1, 2, \ldots, s.$$

Proof. Suppose that for $i \le \overline{s}$, $b_i > 0$, but that for $i > \overline{s}$, $b_i = 0$. In this case, M can be written in partitioned form as

$$
M = \begin{bmatrix} \overline{M} & N \\ N^\mathsf{T} & 0 \end{bmatrix}
$$

and this cannot be positive semi-definite unless $N = 0$. This implies that

$$a_{ij} = 0, \quad \text{whenever } i \le \overline{s} < j,$$

implying that the method is reducible to a method with only $\overline{s}$ stages. □

357   Non-linear stability

The second generalization of A-stability we consider is the assumption that, even though the function f is non-linear, it satisfies the condition that

$$\langle f(u) - f(v), u - v \rangle \le 0, \tag{357a}$$

where $\langle\cdot\rangle$ denotes a semi-inner product, with corresponding semi-norm defined by

$$|u| = \langle u, u \rangle^{1/2}.$$

The reason for our interest in the assumption (357a) is that if there are two solutions y and z to the same differential equations, but with possibly different initial values, then the norm difference of y and z satisfies the bound

$$|y(x) - z(x)| \leq |y(x_0) - z(x_0)|,$$

because

$$\frac{d}{dx}|y(x) - z(x)|^2 = 2\langle f(y(x)) - f(z(x)), y(x) - z(x)\rangle \leq 0.$$

The corresponding property for a Runge–Kutta method would be that the sequences of computed solutions satisfy

$$|y_n - z_n| \leq |y_{n-1} - z_{n-1}|. \qquad (357b)$$

It would equally be possible to use a simpler type of test problem, such as $Y'(x) = F(Y(x))$, where

$$\langle\!\langle g(U), U\rangle\!\rangle \leq 0, \qquad (357c)$$

because (357a) can be expressed using (357c). If $\langle \cdot \rangle$ is the semi-inner product on $\mathbb{R}^N$ used in (357a), with $|\cdot|$ the corresponding semi-norm, then we can define a quasi-inner product $\langle\!\langle \cdot \rangle\!\rangle$ on $\mathbb{R}^{2N}$, with corresponding norm $\|\cdot\|$, by the formula

$$\left\langle\!\!\!\left\langle \begin{bmatrix} u \\ v \end{bmatrix}, \begin{bmatrix} \widetilde{u} \\ \widetilde{v} \end{bmatrix} \right\rangle\!\!\!\right\rangle = \langle u, \widetilde{u}\rangle - \langle u, \widetilde{v}\rangle - \langle v, \widetilde{u}\rangle + \langle v, \widetilde{v}\rangle.$$

The semi-norms defined from these quasi-inner products are related by

$$\left\|\begin{bmatrix} u \\ v \end{bmatrix}\right\| = \langle\!\langle u - v, u - v\rangle\!\rangle = |u - v|^2,$$

and we can write the condition (357a) in the form

$$\left\langle\!\!\!\left\langle G\left(\begin{bmatrix} u \\ v \end{bmatrix}\right), \begin{bmatrix} u \\ v \end{bmatrix} \right\rangle\!\!\!\right\rangle \leq 0,$$

where G is defined by

$$G\left(\begin{bmatrix} u \\ v \end{bmatrix}\right) = \begin{bmatrix} f(u) \\ f(v) \end{bmatrix}.$$

Furthermore, the requirement on a numerical method (357b) can be written in the form

$$\|Y_n\| \leq \|Y_{n-1}\|.$$

Hence we lose no generality in using a test problem which satisfies (357c) rather than the formally more complicated condition (357a). We therefore adopt this requirement, but revert to the more conventional notation of using $\langle \cdot \rangle$ for a standard semi-inner product with $\| \cdot \|$ the corresponding norm.

Even though we have simplified the notation in one way, it is appropriate to generalize it in another. We really need to avoid the use of autonomous problems because of the intimate relationship that will be found between AN-stability and the type of non-linear stability we are discussing here. When Definition 357A was first introduced, it was referred to as 'B-stability', because it is one step more stringent than A-stability. In the non-autonomous form in which it seems to be a more useful concept, a more natural name is BN-stability.

Definition 357A *A Runge–Kutta (A, b^{T}, c) is 'BN-stable' if for any initial value problem*

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

satisfying the condition

$$\langle f(x, u), u \rangle \leq 0,$$

the sequence of computed solutions satisfies

$$\|y_n\| \leq \|y_{n-1}\|.$$

The crucial result is that for an irreducible non-confluent method, AN-stability and BN-stability are equivalent. Because of the fundamental importance of the necessary and sufficient conditions for a Runge–Kutta method to have either, and therefore both, of these properties, we formalize these conditions:

Definition 357B *A Runge–Kutta method (A, b^{T}, c) is 'algebraically stable' if $b_i > 0$, for $i = 1, 2, \ldots, s$, and if the matrix M, given by*

$$M = \mathrm{diag}(b)A + A^{\mathsf{T}}\mathrm{diag}(b) - bb^{\mathsf{T}}, \tag{357d}$$

is positive semi-definite.

We now show the sufficiency of this property.

Theorem 357C *If a Runge–Kutta method is algebraically stable then it is BN-stable.*

Proof. Let $F_i = f(x_{n-1} + hc_i, Y_i)$. We note that if M given by (357d) is positive semi-definite, then there exist vectors $v_l \in \mathbb{R}^s$, $l = 1, 2, \ldots, \overline{s} \leq s$, such that

$$M = \sum_{l=1}^{\overline{s}} \mu_l \mu_l^{\mathsf{T}}.$$

This means that a quadratic form can be written as the sum of squares as follows:

$$\xi^{\mathsf{T}} M \xi = \sum_{l=1}^{\bar{s}} (\mu_l^{\mathsf{T}} \xi)^2 .$$

Furthermore, a quadratic form of inner products

$$\sum_{i,j=1}^{s} m_{ij} \langle U_i, U_j \rangle$$

is equal to

$$\sum_{l=1}^{\bar{s}} \left\| \sum_{i=1}^{s} \mu_{li} U_i \right\|^2 ,$$

and cannot be negative. We show that

$$\|y_n\| - \|y_{n-1}\|^2 = 2h \sum_{i=1}^{s} b_i \langle Y_i, F_i \rangle - h^2 \sum_{i,j=1}^{s} m_{ij} \langle F_i, F_j \rangle, \tag{357e}$$

so that the result will follow. To prove (357e), we use the equations

$$Y_i = y_{n-1} + h \sum_{j=1}^{s} a_{ij} F_j, \tag{357f}$$

$$Y_i = y_n + h \sum_{j=1}^{s} (a_{ij} - b_j) F_j, \tag{357g}$$

which hold for $i = 1, 2, \ldots, s$. In each case, form the quasi-inner product with F_i, and we find

$$\langle Y_i, F_i \rangle = \langle y_{n-1}, F_i \rangle + h \sum_{j=1}^{s} a_{ij} \langle F_i, F_j \rangle,$$

$$\langle Y_i, F_i \rangle = \langle y_n, F_i \rangle + h \sum_{j=1}^{s} (a_{ij} - b_j) \langle F_i, F_j \rangle.$$

Hence,

$$2h \sum_{i=1}^{s} b_i \langle Y_i, F_i \rangle = \left\langle y_n + y_{n-1}, h \sum_{i=1}^{s} b_i F_i \right\rangle$$

$$= h^2 \sum_{i,j=1}^{s} (2 b_i a_{ij} - b_i b_j) \langle F_i, F_j \rangle.$$

Substitute y_n and y_{n-1} from (357f) and (357g) and rearrange to deduce (357e). □

Our final aim in this discussion of non-autonomous and non-linear stability is to show that BN-stability implies AN-stability. This will give the satisfactory conclusion that algebraic stability is equivalent to each of these concepts.

Because we have formulated BN-stability in terms of a quasi-inner product over the real numbers, we first need to see how (356a) can be expressed in a suitable form. Write the real and imaginary parts of $q(x)$ as $\alpha(x)$ and $\beta(x)$, respectively. Also write $y(x) = \xi(x) + i\eta(x)$ and write $\zeta(x)$ for the function with values in $\mathbb{R}^2$ whose components are $\xi(x)$ and $\eta(x)$, respectively.

Thus, because

$$
\begin{aligned}
y'(x) &= (\alpha(x) + i\beta(x))(\xi(x) + i\eta(x)) \\
&= (\alpha(x)\xi(x) - \beta(x)\eta(x)) + i(\beta(x)\xi(x) + \alpha(x)\eta(x)),
\end{aligned}
$$

we can write

$$
\zeta'(x) = Q\zeta,
$$

where

$$
Q = \left[\begin{array}{cc} \alpha(x) & -\beta(x) \\ \beta(x) & \alpha(x) \end{array} \right].
$$

Using the usual inner product we now have the dissipativity property

$$
\langle Qv, v \rangle = \alpha \|v\|^2 \leq 0,
$$

if $\alpha \leq 0$.

What we have found is that the test problem for AN-stability is an instance of the test problem for BN-stability. This means that we can complete the chain of equivalences interconnecting AN-stability, BN-stability and algebraic stability. The formal statement of the final step is as follows:

Theorem 357D *If an irreducible non-confluent Runge–Kutta method is BN-stable, then it is AN-stable.*

358   BN-stability of collocation methods

In the case of methods satisfying the collocation conditions

$$
\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{1}{k} c_i^k, \qquad\qquad i, k = 1, 2, \ldots, s,
$$

$$
\sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \qquad\qquad k = 1, 2, \ldots, s,
$$

a congruence transformation of M, using the Vandermonde matrix

$$
V = \left[\begin{array}{cccc} \mathbf{1} & c & c^2 & \cdots & c^{s-1} \end{array} \right],
$$

where powers of c are interpreted in a componentwise manner, leads to considerable simplification. Denote

$$\epsilon_k = \sum_{i=1}^{s} b_i c_i^{k-1} - \frac{1}{k}, \qquad k = 1, 2, \ldots, 2s,$$

so that $\epsilon_1 = \epsilon_2 = \cdots = \epsilon_s = 0$. Calculate the (k, l) element of $V^{\mathsf{T}} M V$. This has the value

$$\sum_{i=1}^{s} c_i^{k-1} \sum_{j=1}^{s} c_j^{l-1} (b_i a_{ij} + b_j a_{ji} - b_i b_j)$$

$$= \sum_{i=1}^{s} \frac{1}{l} b_i c_i^{k+l-1} + \sum_{j=1}^{s} \frac{1}{k} b_j c_j^{k+l-1} - \frac{1}{kl}$$

$$= \frac{1}{l(k+l)} + \frac{1}{l} \epsilon_{k+l} + \frac{1}{k(k+l)} + \frac{1}{k} \epsilon_{k+l} - \frac{1}{kl}$$

$$= \frac{k+l}{kl} \epsilon_{k+l}.$$

Thus,

$$V^{\mathsf{T}} M V = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & \frac{s+1}{s} \epsilon_{s+1} \\ 0 & 0 & 0 & \cdots & \frac{s+1}{2(s-1)} \epsilon_{s+1} & \frac{s+2}{2s} \epsilon_{s+2} \\ 0 & 0 & 0 & \cdots & \frac{s+2}{3(s-1)} \epsilon_{s+2} & \frac{s+3}{3s} \epsilon_{s+3} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & \frac{s+1}{2(s-1)} \epsilon_{s+1} & \frac{s+2}{3(s-1)} \epsilon_{s+2} & \cdots & \frac{2s-2}{(s-1)^2} \epsilon_{2s-2} & \frac{2s-1}{s(s-1)} \epsilon_{2s-1} \\ \frac{s+1}{s} \epsilon_{s+1} & \frac{s+2}{2s} \epsilon_{s+2} & \frac{s+3}{3s} \epsilon_{s+3} & \cdots & \frac{2s-1}{s(s-1)} \epsilon_{2s-1} & \frac{2s}{s^2} \epsilon_{2s} \end{bmatrix}.$$

A symmetric positive semi-definite matrix cannot have a zero diagonal element unless all the elements on the same row and column are also zero. Hence, we deduce that $\epsilon_i = 0$ for $i = s+1, s+2, \ldots, 2s-1$. Thus, the only way for M to be positive semi-definite is that

$$V^{\mathsf{T}} M V = \frac{2s}{s^2} \epsilon_{2s} e_s e_s^{\mathsf{T}}$$

and that

$$\epsilon_{2s} \geq 0. \tag{358a}$$

Combining these remarks with a criterion for (358a), we state:

Theorem 358A *A collocation Runge–Kutta method is algebraically stable if and only if the abscissae are zeros of a polynomial of the form*

$$P_s^* - \theta P_{s-1}^*, \tag{358b}$$

where $\theta \geq 0$.

Proof. Because $\epsilon_i = 0$ for $i = 1, 2, \ldots, 2s - 1$, it follows that

$$\int_0^1 P(x)\phi(x)dx = 0, \tag{358c}$$

where $\phi(x)$ is a polynomial of degree s, with positive leading coefficient and zeros c_1, c_2, $\ldots$, c_s and P is any polynomial of degree not exceeding $s - 2$. Furthermore, if P is a polynomial of degree $s - 1$ and positive leading coefficient, the integral in (358c) has the same sign as $-\epsilon_{2s}$. Because of the orthogonality of ϕ and polynomials of degree less than $s - 1$, ϕ is a positive constant multiple of (358b). Apart from a positive factor, we can now evaluate the integral in (358c), with $P(x) = P_{s-1}^*(x)$,

$$\int_0^1 P_{s-1}^*(x)(P_s^*(x) - \theta P_{s-1}^*(x))dx = -\theta \int_0^1 P_{s-1}^*(x)^2 dx,$$

which has the opposite sign to θ. $\qquad\qquad\square$

A consequence of this result is that both Gauss and Radau IIA methods are algebraically stable. Many other methods used for the solution of stiff problems have stage order *lower* than s and are therefore not collocation methods. A general characterization of algebraic stable methods is found by using a transformation based not on the Vandermonde matrix V, but on a generalized Vandermonde matrix based on the polynomials that are essentially the same as P_i^*, for $i = 0, 1, 2, \ldots, s - 1$.

359   The V and W transformations

We refer to the transformation of M using the Vandermonde matrix V to form $V^{\mathsf{T}}MV$, as the 'V transformation'. We now introduce the more sophisticated W transformation.

We recall Corollary 356D, which enables us to confine our attention to irreducible methods in which b^{T} has only positive elements. Construct a sequence of polynomials P_0, P_1, $\ldots$, P_{s-1} with degrees 0, 1, $\ldots$, $s - 1$, respectively, which are orthonormal in the sense that

$$\sum_{i=1}^{s} b_i P_{k-1}(c_i) P_{l-1}(c_i) = \delta_{kl}, \qquad k, l = 1, 2, \ldots, s. \qquad (359a)$$

We can assume that the leading coefficients are all positive. Define W as the generalized Vandermonde matrix

$$W = \begin{bmatrix} P_0(c) & P_1(c) & \cdots & P_{s-1}(c) \end{bmatrix}$$

$$= \begin{bmatrix} P_0(c_1) & P_1(c_1) & \cdots & P_{s-1}(c_1) \\ P_0(c_2) & P_1(c_2) & \cdots & P_{s-1}(c_2) \\ \vdots & \vdots & & \vdots \\ P_0(c_s) & P_1(c_s) & \cdots & P_{s-1}(c_s) \end{bmatrix}. \qquad (359b)$$

This matrix can be constructed using the Gram–Schmidt process, or what is algebraically equivalent, from a QR factorization

$$B^{1/2} V = (B^{1/2} W) R,$$

where $B^{1/2} = \operatorname{diag}(\sqrt{b_1}, \sqrt{b_2}, \cdots, \sqrt{b_s})$ and R is upper triangular with positive elements on the diagonal. Note that the coefficients in P_0, P_1, $\ldots$, P_{s-1} can be read off from the columns of R.

If b^{T} and c are weight and abscissa vectors for a Runge–Kutta method of order p, then as long as $k + l \leq p + 1$, (359a) implies that

$$\int_0^1 P_{k-1}(x) P_{l-1}(x) dx = \sum_{i=1}^{s} b_i P_{k-1}(c_i) P_{l-1}(c_i) = \delta_{kl},$$

implying that P_0, P_1, $\ldots$, $P_{[(p-1)/2]}$ are orthonormal with respect to integration on $[0, 1]$. This means that they are necessarily the normalized Legendre polynomials on this interval, given by

$$P_k(z) = \sqrt{2k+1} \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} \binom{k+i}{i} z^i.$$

In particular, $P_0(x) = 1$ and the first column of W is $\mathbf{1}$. Because of orthonormality, it follows that $\mathbf{1}^{\mathsf{T}} B W = e_1^{\mathsf{T}}$.

We now focus our attention on the matrix $X = W^{\mathsf{T}} B A W$. This is significant because

$$W^{\mathsf{T}} M W = X + X^{\mathsf{T}} - (W^{\mathsf{T}} B \mathbf{1})(\mathbf{1}^{\mathsf{T}} B W) = (X - \tfrac{1}{2} e_1 e_1^{\mathsf{T}}) + (X - \tfrac{1}{2} e_1 e_1^{\mathsf{T}})^{\mathsf{T}}.$$

Because M, and therefore $W^{\mathsf{T}} M W$, is the zero matrix for the Gauss method, it follows that $X - \tfrac{1}{2} e_1 e_1^{\mathsf{T}}$ is skew-symmetric. Denote X by X_{G} in this special case. We now evaluate X_{G} in full.

Lemma 359A *Let*
$$X_G = W^\mathsf{T} BAW,$$
where A and $B = \mathrm{diag}(b)$ are as for the Gauss method of order $2s$. Also let
$$\xi_k = \frac{1}{2\sqrt{4k^2 - 1}}, \qquad k = 1, 2, \ldots, s - 1.$$

Then
$$X_G = \begin{bmatrix}
\frac{1}{2} & -\xi_1 & 0 & 0 & \cdots & 0 & 0 \\
\xi_1 & 0 & -\xi_2 & 0 & \cdots & 0 & 0 \\
0 & \xi_2 & 0 & -\xi_3 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 0 & -\xi_{s-1} \\
0 & 0 & 0 & 0 & \cdots & \xi_{s-1} & 0
\end{bmatrix}.$$

Proof. From linear combinations of identities included in the condition $E(s, s)$, given by (321c), we have
$$\sum_{i=1}^{s} \sum_{j=1}^{s} b_i \phi(c_i) a_{ij} \psi(c_j) = \int_0^1 \phi(u) \int_0^u \psi(v) dv du,$$

for polynomials ϕ and ψ each with degree less than s. Use the polynomials $\phi = P_{k-1}$, $\psi = P_{l-1}$ and we have a formula for the (k, l) element of X_G. Add to this the result for k and l interchanged and use integration by parts. We have
$$(X_G)_{kl} + (X_G)_{lk} = \int_0^1 P_{k-1}(u) du \int_0^1 P_{l-1}(v) dv = \delta_{k1}\delta_{l1}.$$

This result determines the diagonal elements of X_G, and also implies the skew-symmetric form of $X_G - \frac{1}{2}e_1 e_1^\mathsf{T}$. We now determine the form of the lower triangular elements. If $k > l + 1$, the integral $\int_0^u P_{l-1}(v) dv$ has lower degree than P_{k-1} and is therefore orthogonal to it. Thus, in this case, $(X_G)_{kl} = 0$. It remains to evaluate $(X_G)_{k,k-1}$ for $k = 1, 2, \ldots, s - 1$. The integral $\int_0^u P_{k-1}(v) dv$ is a polynomial in u of degree k and can be written in the form $\theta P_k(u)$ added to a polynomial of degree less than k. The integral of $P_k(u)$ multiplied by the polynomial of degree less than k is zero, by orthogonality, and the integral reduces to
$$\int_0^1 \theta P_k(u)^2 du = \theta.$$

The value of θ can be found by noting that the coefficient of v^{k-1} in $P_{k-1}(v)$ is $\sqrt{2k-1}\binom{2k-2}{k-1}$, with a similar formula for the leading coefficient of $P_k(u)$.

Hence,

$$(X_{\mathrm{G}})_{k,k-1} = \theta = \frac{\frac{1}{k}\sqrt{2k-1}\binom{2k-2}{k-1}}{\sqrt{2k+1}\binom{2k}{k}} = \frac{1}{2\sqrt{4k^2-1}}. \qquad \Box$$

The computation of elements of $X = W^{\mathsf{T}}BAW$ for any Runge–Kutta method, for which W makes sense, will lead to the same (k,l) elements as in X_{G} as long as $k+l \le p+1$. We state this formally.

Corollary 359B *Let (A,b,c) denote a Runge–Kutta method for which $B = \mathrm{diag}(b)$ is positive definite and for which the abscissae are distinct. Define W by (359b) and X by $X = W^{\mathsf{T}}BAW$. Then $X_{kl} = (X_G)_{kl}$, as long as $k+l \le p+1$.*

The W transformation is related in an interesting way to the $C(m)$ and $D(m)$ conditions, which can be written in the equivalent forms

$$C(m): \quad \sum_{j=1}^{s} a_{ij}P_{k-1}(c_j) = \int_0^{c_i} P_{k-1}(x)dx, \qquad k \le m, \ i=1,2,\dots,s,$$

$$D(m): \quad \sum_{i=1}^{s} b_i P_{k-1}(c_i)a_{ij} = b_j \int_{c_j}^{1} P_{k-1}(x)dx, \quad k \le m, \ j=1,2,\dots,s.$$

It follows from these observations that, if $B(m)$ and $C(m)$ are true, then the first m columns of X will be the same as for X_{G}. Similarly, if $B(m)$ and $D(m)$, then the first m rows of X and X_{G} will agree.

Amongst the methods known to be algebraically stable, we have already encountered the Gauss and Radau IIA methods. We can extend this list to include further methods.

Theorem 359C *The Gauss, Radau IA, Radau IIA and Lobatto IIIC methods are algebraically stable.*

Proof. We have already settled the Gauss and Radau IIA cases, using the V transformation, making use of the $C(s)$ and $B(p)$ conditions, as in Theorem 358A.

To prove the result for Radau IA methods, use the $D(s)$ and $B(2s-1)$ conditions:

$$\sum_{i,j=1}^{s} c_i^{k-1}b_i a_{ij}c_j^{l-1} + \sum_{i,j=1}^{s} c_i^{k-1}b_j a_{ji}c_j^{l-1}$$

$$= \frac{1}{k}\sum_{j=1}^{s} b_j(1-c_j^k)c_j^{l-1} + \frac{1}{l}\sum_{i=1}^{s} b_i(1-c_i^l)c_i^{k-1} - \frac{1}{kl}$$

$$= \frac{1}{kl} - \frac{k+l}{kl}\sum_{i=1}^{s} b_i c_i^{k+l-1}.$$

The value of this expression is zero if $k+l \leq 2s-1$. Although it can be verified directly that the value is positive in the remaining case $k = l = s$, it is enough to show that the $(1,1)$ element of M is positive, because this will have the same sign as the only non-zero eigenvalue of the rank 1 matrix $V^{\mathsf{T}}MV$. We note that all values in the first column of A are equal to b_1 because these give the unique solution to the $D(s)$ condition applied to the first column. Hence, we calculate the $(1,1)$ element of M to be

$$2b_1 a_{11} - b_1^2 = b_1^2 > 0.$$

In the case of the Lobatto IIIC methods, we can use a combination of the $C(s-1)$ and $D(s-1)$ conditions to evaluate the (k,l) and (l,k) elements of M, where $k \leq s-1$ and $l \leq s$. The value of these elements is

$$\sum_{i,j=1}^{s} c_i^{k-1} b_i a_{ij} c_j^{l-1} + \sum_{i,j=1}^{s} c_i^{k-1} b_j a_{ji} c_j^{l-1}$$

$$= \frac{1}{k} \sum_{j=1}^{s} (1 - c_j^k) c_j^{l-1} + \frac{1}{k} \sum_{i=1}^{s} b_i c_i^{k+l-1} - \frac{1}{kl}$$

$$= \frac{1}{k} \sum_{j=1}^{s} b_j c_j^{l-1} - \frac{1}{kl}$$

$$= 0.$$

The final step of the proof is the same as for the Radau IA case, because again $a_{i1} = b_1$, for $i = 1, 2, \ldots, s$. $\square$

The V transformation was used to simplify questions concerning algebraic stability in Butcher (1975) and Burrage (1978). The W transformation was introduced in Hairer and Wanner (1981, 1982). Recent results on the W transformation, and especially application to symplectic methods, were presented in Hairer and Leone (2000) .

Exercises 35

35.1 Show that a Runge–Kutta method with $R(z) = N(z)/D(z)$, where N and D have no common factors, cannot be A-stable unless the coefficients in $D(z)$ alternate in sign.

35.2 Show that the error constant for the $(s-d, s)$ Padé approximation to the exponential function has sign $(-1)^s$. Deduce that, if $d > 0$ and $d = 3$ (mod 4) or $d = 0$ (mod 4), then $|R(iy)| > 0$, for sufficiently small real y.

35.3 Show that the implicit Runge–Kutta method with tableau

$$
\begin{array}{c|cc}
\frac{1}{4} & \frac{7}{24} & -\frac{1}{24} \\[4pt]
\frac{3}{4} & \frac{13}{24} & \frac{5}{24} \\[4pt]
\hline
& \frac{1}{2} & \frac{1}{2}
\end{array}
$$

is A-stable but not AN-stable.

35.4 For the $(0, m)$ Padé approximation $N(z)/D(z)$, show that the denominator D has zeros with negative real part, for $m \geq 5$.

35.5 For the $(1, m)$ Padé approximation $N(z)/D(z)$, show that the denominator D has zeros with negative real part, for $m \geq 7$.

36 Implementable Implicit Runge–Kutta Methods

360 Implementation of implicit Runge–Kutta methods

Because of the implicit nature of these methods, every step requires the solution of an algebraic system. For an s-stage method with an N-dimensional problem, there are sN unknowns to evaluate and these satisfy sN equations. If f is nonlinear, then the large system of equations to be solved is also non-linear. However, there are linear parts of it, and it may be possible to exploit this in their numerical solution. Let A denote the coefficient matrix; then the stage values need to be computed as solutions to the system of equations

$$
Y_1 = y_{n-1} + h \sum_{j=1}^{s} a_{1j} f(Y_j),
$$

$$
Y_2 = y_{n-1} + h \sum_{j=1}^{s} a_{2j} f(Y_j),
$$

$$
\vdots \qquad\qquad \vdots
$$

$$
Y_s = y_{n-1} + h \sum_{j=1}^{s} a_{sj} f(Y_j).
$$

For an N-dimensional differential equation system, this amounts to a system of sN non-linear equations.

We consider how to solve these equations using a full Newton method. This requires going through the following steps:

1. Compute approximations to Y_1, Y_2, ..., Y_s using information available at the start of the step. Denote these 'predicted' values by $Y_i^{[0]}$, $i = 1, 2, \ldots, s$.

2. Carry out a sequence of iterations leading to approximations $Y_i^{[k]}$, for $k = 1, 2, \ldots$, $i = 1, 2, \ldots, s$. These are given by the formulae

$$Y_i^{[k]} = Y_i^{[k-1]} - \Delta_i,$$

where

$$\sum_{j=1}^{s} m_{ij} \Delta_j = \phi_i, \qquad i = 1, 2, \ldots, s, \tag{360a}$$

with

$$\phi_i = Y_i^{[k-1]} - y_{n-1} - h \sum_{j=1}^{s} a_{ij} f(Y_j^{[k-1]})$$

and

$$m_{ij} = \delta_{ij} I - h a_{ij} f'(Y_j^{[k-1]}).$$

3. Test for convergence and terminate when each of $\|\Delta_1\|$, $\|\Delta_2\|$, ..., $\|\Delta_s\|$ are sufficiently small. Suppose that this happens in the computation of iteration k.

4. Assign $Y_i^{[k]}$ to Y_i, for each $i = 1, 2, \ldots, s$.

In a practical calculation, it is usual to simplify this computation in various ways. Most importantly, the solution of (360a), preceded by the evaluation of the elements of m_{ij} which depend on f' evaluated at each stage and in each iteration, requires a large number of algebraic operations; these are to be avoided whenever possible.

A typical simplification is to replace the value of $f'(Y_j^{[k-1]})$ by a constant approximation to this Jacobian matrix. This approximation is maintained at a fixed value over every iteration and over each stage, and possibly over many steps. This means that the $sN \times sN$ matrix with elements built up from the submatrices m_{ij} can be replaced by a matrix of the form

$$I_s \otimes I_N - hA \otimes J, \tag{360b}$$

where J is the Jacobian approximation. The cost, measured solely in terms of linear algebra costs, divides into two components. First, the factorization of the matrix (360b), carried out from time to time during the computation, costs a small multiple of $s^3 N^3$ floating point operations. Secondly, the solution of (360a) costs a small multiple of $s^2 N^2$ arithmetic operations per iteration.

It is the aim of the study of implementable methods to lower the factors s^3 in the occasional part of the cost and to lower the factor s^2 in the 'per iteration' part of the cost.

361   Diagonally implicit Runge–Kutta methods

Because of the excessive cost in evaluating the stages in a fully implicit Runge–Kutta method, we consider the so-called 'diagonally implicit Runge–Kutta' or DIRK methods (Alexander, 1977). For these methods, the coefficient matrix A has a lower triangular structure with equal elements on the diagonal. Note that sometimes these methods are referred to as 'singly diagonally implicit' or SDIRK, with DIRK methods not necessarily having equal diagonals. Earlier names for methods in this general class are semi-implicit Runge–Kutta methods (Butcher, 1965)and semi-explicit (Nørsett, 1974).

The advantage of these methods is that the stages can be evaluated sequentially rather than as one great implicit system. We consider here the derivation of some low order members of this class with a brief analysis of their stability regions.

To obtain order 2 with two stages, consider the tableau

$$\begin{array}{c|cc} \lambda & \lambda & 0 \\ c_2 & c_2 - \lambda & \lambda \\ \hline & b_1 & b_2 \end{array}.$$

The order conditions are

$$b_1 + b_2 = 1, \tag{361a}$$

$$b_1 \lambda + b_2 c_2 = \tfrac{1}{2}, \tag{361b}$$

with solution $b_1 = \frac{2c_2 - 1}{2(c_2 - \lambda)}$, $b_2 = \frac{1 - 2\lambda}{2(c_2 - \lambda)}$. The method is A-stable if $\lambda \geq \frac{1}{4}$ and L-stable if $\lambda = 1 \pm \frac{1}{2}\sqrt{2}$. A particularly attractive choice is $c_2 = 1$, $\lambda = 1 - \frac{1}{2}\sqrt{2}$, for which the tableau is

$$\begin{array}{c|cc} 1 - \tfrac{1}{2}\sqrt{2} & 1 - \tfrac{1}{2}\sqrt{2} & 0 \\ 1 & \tfrac{1}{2}\sqrt{2} & 1 - \tfrac{1}{2}\sqrt{2} \\ \hline & \tfrac{1}{2}\sqrt{2} & 1 - \tfrac{1}{2}\sqrt{2} \end{array}.$$

For $s = p = 3$, the stability function is given by

$$R(z) = \frac{1 + (1 - 3\lambda)z + (\tfrac{1}{2} - 3\lambda + 3\lambda^2)z^2 + (\tfrac{1}{6} - \tfrac{3}{2}\lambda + 3\lambda^2 - \lambda^3)z^3}{(1 - \lambda z)^3}$$

and the E-polynomial is found to be

$$E(y) = \left(\frac{1}{12} - \lambda + 3\lambda^2 - 2\lambda^3\right)y^4 + \left(-\frac{1}{36} + \frac{\lambda}{2} - \frac{13\lambda^2}{4} + \frac{28\lambda^3}{3} - 12\lambda^4 + 6\lambda^5\right)y^6.$$

For $E(y) \geq 0$, for all $y > 0$, it is necessary and sufficient for A-stability that $\lambda \in [\frac{1}{3}, \widetilde{\lambda}]$, where $\widetilde{\lambda} \approx 1.0685790213$ is a zero of the coefficient of y^4 in $E(y)$. For

L-stability there is only one possible choice in this interval: $\lambda \approx 0.4358665215$, a zero of the coefficient of z^3 in the numerator of $R(z)$. Assuming λ is chosen as this value, a possible choice for the remaining coefficients is given by the tableau

$$
\begin{array}{c|ccc}
\lambda & \lambda & 0 & 0 \\
\frac{1}{2}(1+\lambda) & \frac{1}{2}(1-\lambda) & \lambda & 0 \\
1 & \frac{1}{4}(-6\lambda^2+16\lambda-1) & \frac{1}{4}(6\lambda^2-20\lambda+5) & \lambda \\
\hline
& \frac{1}{4}(-6\lambda^2+16\lambda-1) & \frac{1}{4}(6\lambda^2-20\lambda+5) & \lambda
\end{array}
.
$$

362   The importance of high stage order

The asymptotic error behaviour of a numerical method underlines the importance of the order p in ensuring high accuracy at minimal computing cost, as long as sufficient accuracy is required. If, for two methods, the asymptotic local truncation errors are, respectively, $C_1 h^{p_1+1}$ and $C_2 h^{p_2+1}$, where $p_2 > p_1$, then the second method will always be more efficient as long as h is taken to be sufficiently small. This argument ignores the fact that the methods might have differing costs per step, and therefore the stepsizes that make the work done by the methods comparable might be vastly different. It also ignores the fact that C_1 and C_2 can have such values that, for moderate stepsizes, the first method may be more efficient. This argument also ignores the fact that it is not just local errors that matter, but rather the accumulated global error after many steps; from the global error point of view it is also true that high orders will always eventually win over low orders. This ignores the case of special problems where there might be a cancellation of errors, so that in effect the order is greater than it would be for a general problem.

   If the stage order is significantly lower than the order, then the final result computed will have depended for its value on much less accurate answers evaluated along the way. For non-stiff problems this is not a serious difficulty, because the order conditions take into account the need for the effect of these internal errors to cancel each other out. Asymptotically this also happens for stiff problems, but the magnitude of the stepsize required to enjoy the benefits of this asymptotic behaviour may depend drastically on the nature of the problem and on some quantitative measure of its stiffness.

   To investigate this question, Prothero and Robinson (1974) considered a special family of problems of the form

$$
y'(x) = L\big(y(x) - g(x)\big) + g'(x), \qquad y(x_0) = g(x_0),
$$

where L is a negative constant and g is a smooth function that varies at a moderate rate. We first look at the extreme 'non-stiff' case $L = 0$. In this case the Prothero and Robinson problem becomes

$$
y'(x) = g'(x),
$$

and the defining equations for the solution computed by the Runge–Kutta method are

$$Y = y_{n-1}\mathbf{1} + hAG', \tag{362a}$$

$$y_n = y_0 + hb^\mathsf{T}G', \tag{362b}$$

where G' is the subvector made up from the values of $g'(x)$ evaluated at the stage values. We also write G for the corresponding vector of $G(x)$ values. Thus

$$G = \begin{bmatrix} g(x_{n-1} + hc_1) \\ g(x_{n-1} + hc_2) \\ \vdots \\ g(x_{n-1} + hc_s) \end{bmatrix}, \qquad G' = \begin{bmatrix} g'(x_{n-1} + hc_1) \\ g'(x_{n-1} + hc_2) \\ \vdots \\ g'(x_{n-1} + hc_s) \end{bmatrix}.$$

We see that the accuracy of the computation of y_n, as an approximation to $y(x_n)$, is independent of the A matrix and is determined by the accuracy of the quadrature formula

$$\sum_{i=1}^{s} b_i \phi'(c_i) \approx \int_0^1 \phi'(\xi)d\xi, \tag{362c}$$

which we assume to be of order p. This means that (362c) is exact for ϕ a polynomial of degree up to p, and the error will be approximately

$$\frac{1}{p!}\left(\frac{1}{p+1} - \sum_{i=1}^{s} b_i c_i^p\right)\phi^{(p+1)}(0)$$

and the error in the Runge–Kutta method for this problem will be

$$\frac{h^{p+1}}{p!}\left(\frac{1}{p+1} - \sum_{i=1}^{s} b_i c_i^p\right)g^{(p+1)}(x_{n-1}) + O(h^{p+2}). \tag{362d}$$

Now return to the full Prothero and Robinson problem

$$y'(x) = L\Big(y(x) - g(x)\Big) + g'(x),$$

for which the computed results satisfy

$$Y = y_{n-1}\mathbf{1} + hA\left(L(Y - G) + G'\right),$$

$$y_n = y_{n-1} + hb^\mathsf{T}\left(L(Y - G) + G'\right).$$

Eliminate Y, and we find

$$y_n = \Big(1 + hLb^\mathsf{T}(I - hLA)^{-1}\mathbf{1}\Big)y_{n-1} + hb^\mathsf{T}(I - hLA)^{-1}(G' - LG),$$

where the coefficient of y_{n-1} is seen to be the stability function value

$$R(hL) = 1 + hLb^\mathsf{T}(I - hLA)^{-1}\mathbf{1}.$$

By rearranging this expression we see that

$$
\begin{aligned}
y_n &= R(hL)\Big(y_{n-1} - g(x_{n-1})\Big) + g(x_{n-1}) + hb^\mathsf{T}G' \\
&\quad + hLb^\mathsf{T}(I - hLA)^{-1}\Big(hAG' - (G - g(x_{n-1}))\Big) \\
&= R(hL)\Big(y_{n-1} - g(x_{n-1})\Big) + g(x_n) - \epsilon_0 - hLb^\mathsf{T}(I - hLA)^{-1}\epsilon,
\end{aligned}
$$

where

$$\epsilon_0 = h\int_0^1 g'(x_{n-1} + h\xi)d\xi - h\sum_{i=1}^s b_i g'(x_{n-1} + hc_i)$$

is the non-stiff error term given approximately by (362d) and ϵ is the vector of errors in the individual stages with component i given by

$$h\int_0^{c_i} g'(x_{n-1} + h\xi)d\xi - h\sum_{j=1}^s a_{ij}g'(x_{n-1} + hc_j).$$

If L has a moderate size, then $hLb^\mathsf{T}(I - hLA)^{-1}\epsilon$ can be expanded in the form

$$hLb^\mathsf{T}(I + hLA + h^2L^2A^2 + \cdots)\epsilon$$

and error behaviour of order p can be verified term by term.

On the other hand, if hL is large, a more realistic idea of the error is found using the expansion

$$(I - hLA)^{-1} = -\frac{1}{hL}A^{-1} - \frac{1}{h^2L^2}A^{-2} - \cdots,$$

and we obtain an approximation to the error, $g(x_n) - y_n$, given by

$$
\begin{aligned}
g(x_n) - y_n &= R(hL)\Big(g(x_{n-1}) - y_{n-1}\Big) + \epsilon_0 \\
&\quad - b^\mathsf{T}A^{-1}\epsilon - h^{-1}L^{-1}b^\mathsf{T}A^{-2}\epsilon - h^{-2}L^{-2}b^\mathsf{T}A^{-3}\epsilon - \cdots.
\end{aligned}
$$

Even though the stage order may be low, the *final* stage may have order p. This will happen, for example, if the final row of A is identical to the vector b^T. In this special case, the term $b^\mathsf{T}A^{-1}\epsilon$ will cancel ϵ_0.

In other cases, the contributions from $b^\mathsf{T}A^{-1}\epsilon$ might dominate ϵ_0, if the stage order is less than the order.

Define

$$\eta_n = \epsilon_0 + hLb^\mathsf{T}(I - hLA)^{-1}\epsilon, \qquad n > 0,$$

with η_0 defined as the initial error $g(x_0) - y_0$. The accumulated truncation error after n steps is equal to

$$\sum_{i=0}^{n} R(hL)^{n-i} \eta_i \approx \sum_{i=0}^{n} R(\infty)^{n-i} \eta_i.$$

There are three important cases which arise in a number of widely use methods. If $R(\infty) = 0$, as in the Radau IA, Radau IIA and Lobatto IIIC methods, or for that matter in any L-stable method, then we can regard the global truncation error as being just the error in the final step. Thus, if the local error is $O(h^{q+1})$ then the global error would also be $O(h^{q+1})$. On the other hand, for the Gauss method with s stages, $R(\infty) = (-1)^s$. For the methods for which $R(\infty) = 1$, then we can further approximate the global error as the integral of the local truncation error multiplied by h^{-1}. Hence, a local error $O(h^{q+1})$ would imply a global error of $O(h^q)$. In the cases for which $R(\infty) = -1$ we would expect the global error to be $O(h^{q+1})$, because of cancellation of η_i over alternate steps.

We explore a number of example methods to see what can be expected for both local and global error behaviour.

For the Gauss methods, for which $p = 2s$, we can approximate ϵ_0 by

$$\frac{h^{2s+1}}{(2s)!} \left(\frac{1}{2s+1} - \sum_{i=1}^{s} b_i c_i^{2s} \right) g^{(2s+1)}(x_{n-1}) + O(h^{2s+2}),$$

which equals

$$\frac{h^{2s+1} s!^4}{(2s)!^3 (2s+1)} g^{(2s+1)}(x_{n-1}) + O(h^{2s+2}). \tag{362e}$$

Now consider the term $-b^{\mathsf{T}} A^{-1} \epsilon$. This is found to equal

$$\frac{h^{s+1} s!}{(2s)!(s+1)} g^{(s+1)}(x_{n-1}) + O(h^{s+2}),$$

which, if $|hL|$ is large, dominates (362e).

We also consider the important case of the Radau IIA methods. In this case ϵ_0 is approximately

$$\frac{h^{2s}}{(2s-1)!} \left(\frac{1}{2s} - \sum_{i=1}^{s} b_i c_i^{2s-1} \right) g^{(2s)}(x_{n-1}) + O(h^{2s+1})$$

$$= -\frac{h^{2s} s!(s-1)!^3}{2(2s-1)!^3} g^{(2s)}(x_{n-1}) + O(h^{2s+1}).$$

As we have remarked, for $|hL|$ large, this term is cancelled by $-b^\mathsf{T}A^{-1}\epsilon$. Hence, the local truncation error can be approximated in this case by $-(hL)^{-1}b^\mathsf{T}A^{-2}\epsilon$. The value of this is

$$\frac{s!}{(s+1)(2s-1)!}\frac{1}{hL}g^{(s)}(x_{n-1})h^s + O(L^{-1}h^s).$$

To summarize: for very stiff problems and moderate stepsizes, a combination modelled for the Prothero–Robinson problem by a high value of hL, the stage order, rather than the classical order, plays a crucial role in determining the error behaviour. For this reason, we consider criteria other than super-convergence as important criteria in the identification of suitable methods for the solution of stiff problems. In particular, we look for methods that are capable of cheap implementation.

363   Singly implicit methods

We consider methods for which the stage order q and the order are related by $p = q = s$. To make the methods cheaply implementable, we also assume that

$$\sigma(A) = \{\lambda\}. \tag{363a}$$

The detailed study of methods for which A has a one-point spectrum and for which $q \geq p-1$ began with Burrage (1978). The special case $q = p$ was further developed in Butcher (1979), and this led to the implementation of STRIDE described in Burrage, Butcher and Chipman (1980).

Given $q = p$ and (363a), there will be a constraint on the abscissae of the method. To explore this, write down the $C(s)$ conditions

$$\sum_{j=1}^{s} a_{ij}c_j^{k-1} = \frac{1}{k}c_i^k, \qquad i, k = 1, 2, \ldots, s,$$

or, more compactly,

$$Ac^{k-1} = \frac{1}{k}c^k, \qquad k = 1, 2, \ldots, s, \tag{363b}$$

where c^k denotes the component-by-component power.

We can now evaluate $A^{k-1}\mathbf{1}$ by induction. In fact,

$$A^k\mathbf{1} = \frac{1}{k!}c^k, \qquad k = 1, 2, \ldots, s, \tag{363c}$$

because the case $k = 1$ is just (363b), also with $k = 1$; and the case $k > 1$ follows from (363c) with k replaced by $k - 1$ and from (363b).

Because of (363a) and the Cayley–Hamilton theorem, we have

$$(A - \lambda I)^s = 0.$$

Table 363(I)   Laguerre polynomials L_s for degrees $s = 1, 2, \ldots, 8$

s	$L_s(\xi)$
1	$1 - \xi$
2	$1 - 2\xi + \frac{1}{2}\xi^2$
3	$1 - 3\xi + \frac{3}{2}\xi^2 - \frac{1}{6}\xi^3$
4	$1 - 4\xi + 3\xi^2 - \frac{2}{3}\xi^3 + \frac{1}{24}\xi^4$
5	$1 - 5\xi + 5\xi^2 - \frac{5}{3}\xi^3 + \frac{5}{24}\xi^4 - \frac{1}{120}\xi^5$
6	$1 - 6\xi + \frac{15}{2}\xi^2 - \frac{10}{3}\xi^3 + \frac{5}{8}\xi^4 - \frac{1}{20}\xi^5 + \frac{1}{720}\xi^6$
7	$1 - 7\xi + \frac{21}{2}\xi^2 - \frac{35}{6}\xi^3 + \frac{35}{24}\xi^4 - \frac{7}{40}\xi^5 + \frac{7}{720}\xi^6 - \frac{1}{5040}\xi^7$
8	$1 - 8\xi + 14\xi^2 - \frac{28}{3}\xi^3 + \frac{35}{12}\xi^4 - \frac{7}{15}\xi^5 + \frac{7}{180}\xi^6 - \frac{1}{630}\xi^7 + \frac{1}{40320}\xi^8$

Post-multiply by $\mathbf{1}$ and expand using the binomial theorem, and we find

$$\sum_{i=0}^{s} \binom{s}{i} (-\lambda)^{s-i} A^i \mathbf{1} = 0.$$

Using (363c), we find that

$$\sum_{i=0}^{s} \binom{s}{i} (-\lambda)^{s-i} \frac{1}{i!} c^i = 0.$$

This must hold for each component separately so that, for $i = 1, 2, \ldots, s$, c_i/λ is a zero of

$$\sum_{i=0}^{s} \binom{s}{i} (-1)^i \frac{(-\xi)^i}{i!}.$$

However, this is just the Laguerre polynomial of degree s, usually denoted by $L_s(\xi)$, and it is known that all its zeros are real and positive. For convenience, expressions for these polynomials, up to degree 8, are listed in Table 363(I) and approximations to the zeros are listed in Table 363(II). We saw in Subsection 361 that for $\lambda = \xi^{-1}$ for the case of three doubly underlined zeros of orders 2 and 3, L-stability is achieved. Double underlining to show similar choices for other orders is continued in the table and these are the only possibilities that exist (Wanner, Hairer and Nørsett, 1978). This means that there are no L-stable methods – and in fact there is not even an A-stable method – with $s = p = 7$ or with $s = p > 8$. Even though fully L-stable methods are confined to the eight cases indicated in this table, there are other choices of $\lambda = \xi^{-1}$ that give stability which is acceptable for many problems. In each of the values of ξ for which there is a *single* underline, the method is A(α)-stable with $\alpha \geq 1.55 \approx 89°$.

Table 363(II)   Zeros of Laguerre polynomials for degrees $s = 1, 2, \ldots, 8$

s	$\xi_1, \ldots, \xi_s$			
1	1.0000000000			
2	0.5857864376	3.4142135624		
3	0.4157745568	2.2942803603	6.2899450829	
4	0.3225476896	1.7457611012	4.5366202969	9.3950709123
5	0.2635603197	1.4134030591	3.5964257710	7.0858100059
	12.6408008443			
6	0.2228466042	1.1889321017	2.9927363261	5.7751435691
	9.8374674184	15.9828739806		
7	0.1930436766	1.0266648953	2.5678767450	4.9003530845
	8.1821534446	12.7341802918	19.3957278623	
8	0.1702796323	0.9037017768	2.2510866299	4.2667001703
	7.0459054024	10.7585160102	15.7406786413	22.8631317369

The key to the efficient implementation of singly implicit methods is the similarity transformation matrix that transforms the coefficient matrix to lower triangular form. Let T denote the matrix with (i, j) element

$$t_{ij} = L_{j-1}(\xi_i), \qquad i, j = 1, 2, \ldots, s.$$

The principal properties of T and its relationship to A are as follows:

Theorem 363A *The (i, j) element of T^{-1} is equal to*

$$\frac{\xi_j}{s^2 L_{s-1}(\xi_j)^2} L_{i-1}(\xi_j). \tag{363d}$$

Let $\widetilde{A}$ denote $T^{-1}AT$; then

$$\widetilde{A} = \lambda \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}. \tag{363e}$$

Proof. To prove (363d), use the Christoffel–Darboux formula for Laguerre polynomials in the form

$$\sum_{k=0}^{s-1} L_k(x) L_k(y) = \frac{s}{x-y} \big(L_s(y) L_{s-1}(x) - L_s(x) L_{s-1}(y) \big).$$

For $i \neq j$, substitute $x = \xi_i, y = \xi_j$ to find that rows i and j of T are orthogonal. To evaluate the inner product of row i with itself, substitute $y = \xi_i$ and take the limit as $x \to \xi_i$. It is found that

$$\sum_{k=0}^{s-1} L_k(\xi_k)^2 = -s L'_s(\xi_i) L_{s-1}(\xi_i) = \frac{s^2 L_{s-1}(\xi_i)^2}{\xi_i}. \qquad (363f)$$

The value of TT^{T} as a diagonal matrix with (i,i) element given by (363f) is equivalent to (363d).

The formula for $\widetilde{A}$ is verified by evaluating

$$\sum_{j=1}^{s} a_{ij} L_{k-1}(\xi_j) = \sum_{j=1}^{s} a_{ij} L_{k-1}(c_j/\lambda)$$

$$= \int_0^{\lambda \xi_i} L_{k-1}(c_j/\lambda) dt$$

$$= \lambda \int_0^{\xi_i} L_{k-1}(t) dt$$

$$= \lambda \int_0^{\xi_i} (L'_{k-1}(t) - L'_k(t)) dt$$

$$= \lambda (L_{k-1}(\xi_i) - L_k(\xi_i)) dt,$$

where we have used known properties of Laguerre polynomials. The value of this sum is equivalent to (363e). $\qquad \square$

For convenience we sometimes write

$$J = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix},$$

so that (363e) can be written

$$\widetilde{A} = \lambda(I - J).$$

We now consider the possible A-stability or L-stability of singly implicit methods. This hinges on the behaviour of the rational functions

$$R(z) = \frac{N(z)}{(1 - \lambda z)^s},$$

where the degree of the polynomial $N(z)$ is no more than s, and where

$$N(z) = \exp(z)(1 - \lambda z)^s + O(z^{s+1}).$$

We can obtain a formula for $N(z)$ as follows:

$$N(z) = \sum_{i=0}^{s-i} (-\lambda)^i L_s^{(s-i)} \left(\frac{1}{\lambda}\right) z^i,$$

where $L_n^{(m)}$ denotes the m-fold derivative of L_n, rather than a generalized Laguerre polynomial. To verify the L-stability of particular choices of s and λ, we note that all poles of $N(z)/(1 - \lambda z)^s$ are in the right half-plane. Hence, it is necessary only to test that $|D(z)|^2 - |(1 - \lambda z)^s|^2 \geq 0$, whenever z is on the imaginary axis. Write $z = iy$ and we find the 'E-polynomial' defined in this case as

$$E(y) = (1 + \lambda^2 y^2)^s - N(iy)N(-iy),$$

with $E(y) \geq 0$ for all real y as the condition for A-stability. Although A-stability for $s = p$ is confined to the cases indicated in Table 363(II), it will be seen in the next subsection that higher values of s can lead to additional possibilities.

We conclude this subsection by constructing the two-stage L-stable singly implicit method of order 2. From the formulae for the first few Laguerre polynomials,

$$L_0(x) = 1, \quad L_1(x) = 1 - x, \quad L_2(x) = 1 - 2x + \frac{1}{2}x^2,$$

we find the values of ξ_1 and ξ_2, and evaluate the matrices T and T^{-1}. We have

$$\xi_1 = 2 - \sqrt{2}, \qquad \xi_2 = 2 + \sqrt{2}$$

and

$$T = \begin{bmatrix} L_0(\xi_1) & L_1(\xi_1) \\ L_0(\xi_2) & L_1(\xi_2) \end{bmatrix} = \begin{bmatrix} 1 & -1 + \sqrt{2} \\ 1 & -1 - \sqrt{2} \end{bmatrix}, \quad T^{-1} = \begin{bmatrix} \frac{1}{2} + \frac{\sqrt{2}}{4} & \frac{1}{2} - \frac{\sqrt{2}}{4} \\ \frac{\sqrt{2}}{4} & -\frac{\sqrt{2}}{4} \end{bmatrix}.$$

For L-stability, choose $\lambda = \xi_2^{-1} = 1 - \frac{1}{2}\sqrt{2}$, and we evaluate $A = \lambda T(I - J)T^{-1}$ to give the tableau

$$\begin{array}{c|cc} 3 - 2\sqrt{2} & \frac{5}{4} - \frac{3}{4}\sqrt{2} & \frac{7}{4} - \frac{5}{4}\sqrt{2} \\ 1 & \frac{1}{4} + \frac{1}{4}\sqrt{2} & \frac{3}{4} - \frac{1}{4}\sqrt{2} \\ \hline & \frac{1}{4} + \frac{1}{4}\sqrt{2} & \frac{3}{4} - \frac{1}{4}\sqrt{2} \end{array} \qquad (363\text{g})$$

In the implementation of this, or any other, singly implicit method, the actual entries in this tableau are not explicitly used. To emphasize this point, we look in detail at a single Newton iteration for this method. Let $M = I - h\lambda f'(y_{n-1})$. Here the Jacobian matrix f' is supposed to have been evaluated at the start of the current step. In practice, a Jacobian evaluated at an earlier time value might give satisfactory performance, but we do not dwell on this point here. If the method were to be implemented with no special use made of its singly implicit structure, then we would need, instead of the $N \times N$ matrix M, a $2N \times 2N$ matrix $\widetilde{M}$ given by

$$\widetilde{M} = \begin{bmatrix} I - ha_{11}f'(y_{n-1}) & -ha_{12}f'(y_{n-1}) \\ -ha_{21}f'(y_{n-1}) & I - ha_{22}f'(y_{n-1}) \end{bmatrix}.$$

In this 'fully implicit' situation, a single iteration would start with the input approximation y_{n-1} and existing approximations to the stage values and stage derivatives Y_1, Y_2, hF_1 and hF_2. It will be assumed that these are consistent with the requirements that

$$Y_1 = y_{n-1} + a_{11}hF_1 + a_{12}hF_2, \quad Y_2 = y_{n-1} + a_{21}hF_1 + a_{22}hF_2,$$

and the iteration process will always leave these conditions intact.

364 *Generalizations of singly implicit methods*

In an attempt to improve the performance of existing singly implicit methods, Butcher and Cash (1990) considered the possibility of adding additional diagonally implicit stages. For example, if $s = p + 1$ is chosen, then the coefficient matrix has the form

$$A = \begin{bmatrix} \lambda\widehat{A} & 0 \\ b^{\mathsf{T}} & \lambda \end{bmatrix},$$

where $\widehat{A}$ is the matrix

$$\widehat{A} = T(I - J)T^{-1}.$$

An appropriate choice of λ is made by balancing various considerations. The first of these is good stability, and the second is a low error constant. Minor considerations would be convenience, the avoidance of coefficients with abnormally large magnitudes or with negative signs, where possible, and a preference for methods in which the c_i lie in $[0, 1]$. We illustrate these ideas for the case $p = 2$ and $s = 3$, for which the general form for a method would be

$$\begin{array}{c|ccc}
\lambda(2 - \sqrt{2}) & \lambda(1 - \frac{1}{4}\sqrt{2}) & \lambda(1 - \frac{3}{4}\sqrt{2}) & 0 \\
\lambda(2 + \sqrt{2}) & \lambda(1 + \frac{3}{4}\sqrt{2}) & \lambda(1 + \frac{1}{4}\sqrt{2}) & 0 \\
1 & \frac{2+3\sqrt{2}}{4} - \frac{\lambda(1+\sqrt{2})}{2} - \frac{\sqrt{2}}{8\lambda} & \frac{2-3\sqrt{2}}{4} - \frac{\lambda(1-\sqrt{2})}{2} + \frac{\sqrt{2}}{8\lambda} & \lambda \\
\hline
 & \frac{2+3\sqrt{2}}{4} - \frac{\lambda(1+\sqrt{2})}{2} - \frac{\sqrt{2}}{8\lambda} & \frac{2-3\sqrt{2}}{4} - \frac{\lambda(1-\sqrt{2})}{2} + \frac{\sqrt{2}}{8\lambda} & \lambda
\end{array}.$$
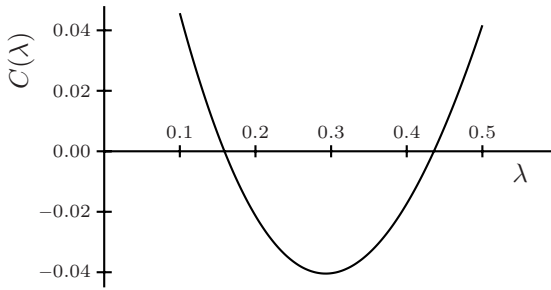
Figure 364(i)   Error constant $C(\lambda)$ for $\lambda \in [0.1, 0.5]$

The only choice available is the value of λ, and we consider the consequence of making various choices for this number. The first criterion is that the method should be A-stable, and we analyse this by calculating the stability function

$$R(z) = \frac{N(z)}{D(z)} = \frac{1 + (1 - 3\lambda)z + (\frac{1}{2} - 3\lambda + 3\lambda^2)z^2}{(1 - \lambda z)^3}$$

and the E-polynomial

$$E(y) = |D(iy)|^2 - |N(iy)|^2 = \left(3\lambda^4 - \left(\tfrac{1}{2} - 3\lambda + 3\lambda^2\right)^2\right)y^4 + \lambda^6 y^6.$$

For A-stability, the coefficient of y^4 must be non-negative. The condition for this is that

$$\frac{3 - \sqrt{3 + 2\sqrt{3}}}{2(3 - \sqrt{3})} \leq \lambda \leq \frac{3 + \sqrt{3 + 2\sqrt{3}}}{2(3 - \sqrt{3})},$$

or that λ lies in the interval $[0.180425, 2.185600]$. The error constant $C(\lambda)$, defined by $\exp(z) - R(z) = C(\lambda)z^3 + O(z^4)$, is found to be

$$C(\lambda) = \frac{1}{6} - \frac{3}{2}\lambda + 3\lambda^2 - \lambda^3,$$

and takes on values for $\lambda \in [0.1, 0.5]$, as shown in Figure 364(i).

The value of b_1 is positive for $\lambda > 0.125441$. Furthermore b_2 is positive for $\lambda < 0.364335$. Since $b_1 + b_2 + \lambda = 1$, we obtain moderately sized values of all components of b^T if $\lambda \in [0.125441, 0.364335]$. The requirement that c_1 and c_2 lie in $(0, 1)$ is satisfied if $\lambda < (2 - \sqrt{2})^{-1} \approx 0.292893$. Leaving aside the question of convenience, we should perhaps choose $\lambda \approx 0.180425$ so that the error constant is small, the method is A-stable, and the other minor considerations are all satisfied. Convenience might suggest an alternative value $\lambda = \frac{1}{5}$.

365 Effective order and DESIRE methods

An alternative way of forcing singly implicit methods to be more appropriate for practical computation is to generalize the order conditions. This has to be done without lowering achievable accuracy, and the use of effective order is indicated. Effective order is discussed in a general setting in Subsection 389 but, for methods with high stage order, a simpler analysis is possible.

Suppose that the quantities passed from one step to the next are not necessarily intended to be highly accurate approximations to the exact solution, but rather to modified quantities related to the exact result by weighted Taylor series. For example, the input to step n might be an approximation to

$$y(x_{n-1}) + \alpha_1 h y'(x_{n-1}) + \alpha_2 h^2 y''(x_{n-1}) + \cdots + \alpha_p h^p y^{(p)}(y_{n-1}).$$

We could regard a numerical method, which produces an output equal to

$$y_n = y(x_n) + \alpha_1 h y'(x_n) + \alpha_2 h^2 y''(x_n) + \cdots + \alpha_p h^p y^{(p)}(y_n) + O(h^{p+1}),$$

as a satisfactory alternative to a method of classical order p.

We explore this idea through the example of the effective order generalization of the L-stable order 2 singly implicit method with the tableau (363g). For this method, the abscissae are necessarily equal to $3 - 2\sqrt{2}$ and 1, which are quite satisfactory for computation. However, we consider other choices, because in the more complicated cases with $s = p > 2$, at least one of the abscissae is outside the interval $[0, 1]$, for A-stability.

If the method is required to have only *effective* order 2, then we can assume that the incoming and outgoing approximations are equal to

$$y_{n-1} = y(x_{n-1}) + h\alpha_1 y'(x_{n-1}) + h^2 \alpha_2 y''(x_{n-1}) + O(h^{p+1}),$$
$$y_n = y(x_n) + h\alpha_1 y'(x_n) + h^2 \alpha_2 y''(x_n) + O(h^{p+1}),$$

respectively. Suppose that the stage values are required to satisfy

$$Y_1 = y(x_{n-1} + hc_1) + O(h^3), \qquad Y_2 = y(x_{n-1} + hc_2) + O(h^3),$$

with corresponding approximations for the stage derivatives. In deriving the order conditions, it can be assumed, without loss of generality, that $n = 1$. The order conditions for the two stages and for the output approximation $y_n = y_1$ are

$$y(x_0 + hc_1) = y(x_0) + h\alpha_1 y'(x_0) + h^2 \alpha_2 y''(x_0)$$
$$+ ha_{11} y'(x_0 + hc_1) + ha_{12} y'(x_0 + hc_2) + O(h^3),$$
$$y(x_0 + hc_2) = y(x_0) + h\alpha_1 y'(x_0) + h^2 \alpha_2 y''(x_0)$$
$$+ ha_{21} y'(x_0 + hc_1) + ha_{22} y'(x_0 + hc_2) + O(h^3),$$
$$y(x_1) + h\alpha_1 y'(x_1) + h^2 \alpha_2 y''(x_1)$$
$$= y(x_0) + h\alpha_1 y'(x_0) + h^2 \alpha_2 y''(x_0)$$
$$+ hb_1 y'(x_0 + hc_1) + hb_2 y'(x_0 + hc_2) + O(h^3).$$

These can be converted into algebraic relations on the various free parameters by expanding by Taylor series about x_0 and equating coefficients of $hy'(x_0)$ and $h^2 y''(x_0)$. This gives the conditions

$$c_1 = \alpha_1 + a_{11} + a_{12},$$
$$\tfrac{1}{2} c_1^2 = \alpha_2 + a_{11} c_1 + a_{12} c_2,$$
$$c_2 = \alpha_1 + a_{21} + a_{22},$$
$$\tfrac{1}{2} c_2^2 = \alpha_2 + a_{21} c_1 + a_{22} c_2,$$
$$1 + \alpha_1 = \alpha_1 + b_1 + b_2,$$
$$\tfrac{1}{2} + \alpha_1 + \alpha_2 = \alpha_2 + b_1 c_1 + b_2 c_2.$$

Because of the single-implicitness condition $\sigma(A) = \{\lambda\}$, we also have

$$a_{11} + a_{22} = 2\lambda,$$
$$a_{11} a_{22} - a_{21} a_{12} = \lambda^2.$$

Assuming that c_1 and c_2 are distinct, a solution to these equations always exists, and it leads to the values

$$\alpha_1 = \tfrac{1}{2}(c_1 + c_2) - 2\lambda, \quad \alpha_2 = \tfrac{1}{2} c_1 c_2 - \lambda(c_1 + c_2) + \lambda^2,$$

together with the tableau

$$
\begin{array}{c|cc}
c_1 & -\frac{c_2 - c_1}{2} + \lambda + \frac{\lambda^2}{c_2 - c_1} & \lambda - \frac{\lambda^2}{c_2 - c_1} \\
c_2 & \lambda + \frac{\lambda^2}{c_2 - c_1} & \frac{c_2 - c_1}{2} + \lambda - \frac{\lambda^2}{c_2 - c_1} \\
\hline
 & \frac{1}{2} + \frac{2\lambda - \frac{1}{2}}{c_2 - c_1} & \frac{1}{2} - \frac{2\lambda - \frac{1}{2}}{c_2 - c_1}
\end{array}
$$

In the special case $c^{\mathsf{T}} = [0, 1]$, with $\lambda = 1 - \tfrac{1}{2}\sqrt{2}$ for L-stability, we find $\alpha_1 = \sqrt{2} - \tfrac{3}{2}$ and $\alpha_2 = \tfrac{1}{2}(1 - \sqrt{2})$ and the tableau

$$
\begin{array}{c|cc}
0 & \frac{1}{2}(4 - 3\sqrt{2}) & \frac{1}{2}(\sqrt{2} - 1) \\
1 & \frac{1}{2}(5 - 3\sqrt{2}) & \frac{1}{2}\sqrt{2} \\
\hline
 & 2 - \sqrt{2} & \sqrt{2} - 1
\end{array}
$$

Combine the effective order idea with the diagonal extensions introduced in Subsection 364, and we obtain 'DESIRE' methods (diagonally extended implicit Runge–Kutta methods using effective order). These are exemplified by the example with $p = 2$, $s = 3$ and $\lambda = \frac{1}{5}$. For this method, $\alpha_1 = -\frac{3}{20}$, $\alpha_2 = \frac{1}{400}$ and the coefficient tableau is

$$
\begin{array}{c|ccc}
0 & \frac{31}{200} & -\frac{1}{200} & 0 \\
\frac{1}{2} & \frac{81}{200} & \frac{49}{200} & 0 \\
1 & \frac{71}{200} & \frac{119}{200} & \frac{1}{5} \\
\hline
 & \frac{103}{250} & \frac{119}{250} & \frac{14}{125}
\end{array}
\;.
$$

Exercises 36

36.1 Derive the tableau for the two-stage order 2 diagonally implicit method satisfying (361a), (361b) with $\lambda = 1 - \frac{1}{2}\sqrt{2}$ and $c_2 = 3\lambda$.

36.2 Rewrite the method in Exercise 36.1 so that the value of Y_1 in step n is the input and the value of Y_1 in step $n + 1$ is the output.

36.3 Show that the method derived in Exercise 36.2 has stage order 2.

36.4 Derive a diagonally implicit method with $s = p = 3$ and with $\lambda = c_2 = \frac{1}{3}$, $c_2 = \frac{2}{3}$, $c_3 = 1$.

36.5 Derive a diagonally implicit method with $s = p = 3$, $\lambda = 1$, $c_2 = \frac{1}{3}$, $c_3 = 1$, $b_1 = 0$.

36.6 Show that for an L-stable method of the type described in Subsection 364 with $p = 3$, $s = 4$, the minimum possible value of λ is approximately 0.2278955169, a zero of the polynomial

$$185976\lambda^{12} - 1490400\lambda^{11} + 4601448\lambda^{10} - 7257168\lambda^9 + 6842853\lambda^8$$
$$-4181760\lambda^7 + 1724256\lambda^6 - 487296\lambda^5 + 94176\lambda^4 - 12192\lambda^3 + 1008\lambda^2 - 48\lambda + 1.$$

37 Symplectic Runge–Kutta Methods

370 Maintaining quadratic invariants

We recall Definition 357B in which the matrix M plays a role, where the elements of M are

$$m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j. \tag{370a}$$

Now consider a problem for which

$$y^\mathsf{T} Q f(y) = 0, \tag{370b}$$

for all y. It is assumed that Q is a symmetric matrix so that (370b) is equivalent to the statement that $y(x)^\mathsf{T} Q y(x)$ is invariant.

We want to characterize Runge–Kutta methods with the property that $y_n^\mathsf{T} Q y_n$ is invariant with n so that the the numerical solution preserves the conservation law possessed by the problem. If the input to step 1 is y_0, then the output will be

$$y_1 = y_0 + h \sum_{i=1}^{s} b_i F_i, \tag{370c}$$

where the stage derivatives are $F_i = f(Y_i)$, with

$$Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} F_j.$$

From (370b) it follows that

$$F_i^\mathsf{T} Q y_0 = -h \sum_{j=1}^{s} a_{ij} F_i^\mathsf{T} Q F_j. \tag{370d}$$

Use (370c) to calculate $y_1^\mathsf{T} Q y_1$ and substitute from (370d) to obtain the result

$$y_1^\mathsf{T} Q y_1 = y_0^\mathsf{T} Q y_0 - h^2 \sum_{i,j=1}^{s} m_{ij} F_i^\mathsf{T} Q F_j,$$

with m_{ij} given by (370a).

Thus $M = 0$ implies that quadratic invariants are preserved and, in particular, that symplectic behaviour is maintained. Accordingly, we have the following definition:

Definition 370A *A Runge–Kutta method (A, b^T, c) is symplectic if*

$$M = \operatorname{diag}(b) A + A^\mathsf{T} \operatorname{diag}(b) - b b^\mathsf{T}$$

is the zero matrix.

The property expressed by Definition 370A was first found by Cooper (1987) and, as a characteristic of symplectic methods, by Lasagni (1988), Sanz-Serna (1988) and Suris (1988).

371   Examples of symplectic methods

A method with a single stage is symplectic only if $2 b_1 a_{11} - b_1^2 = 0$. For consistency, that is order at least 1, $b_1 = 1$ and hence $c_1 = a_{11} = \frac{1}{2}$; this is just the implicit mid-point rule. We can extend this in two ways: by either looking at methods where A is lower triangular or looking at the methods with stage order s.

For lower triangular methods we will assume that none of the b_i is zero. The diagonals can be found from $2b_i a_{ii} = b_i^2$ to be $a_{ii} = \frac{1}{2}b_i$. For the elements of A below the diagonal we have $b_i a_{ij} = b_i b_j$ so that $a_{ij} = b_j$. This gives a tableau

$$
\begin{array}{c|cccccc}
\frac{1}{2}b_1 & \frac{1}{2}b_1 & & & & & \\
b_1 + \frac{1}{2}b_2 & b_1 & \frac{1}{2}b_2 & & & & \\
b_1 + b_2 + \frac{1}{2}b_3 & b_1 & b_2 & \frac{1}{2}b_3 & & & \\
\vdots & \vdots & \vdots & \vdots & \ddots & & \\
b_1 + \cdots + b_{s-1} + \frac{1}{2}b_s & b_1 & b_2 & b_3 & \cdots & \frac{1}{2}b_s \\
\hline
 & b_1 & b_2 & b_3 & \cdots & b_s
\end{array} \ .
$$

This method is identical with s steps of the mid-point rule with stepsizes $b_1 h$, $b_2 h$, ..., $b_s h$.

For methods with order and stage order equal to s, we have, in the notation of Subsection 358, $\epsilon_i = 0$ for $i = s+1$, $s+2$, ..., $2s$. This follows from the observation that $V^{\mathsf{T}} M V = 0$. Thus, in addition to $B(s)$, $B(2s)$ holds. Hence, the abscissae of the method are the zeros of P_s^* and the method is the s-stage Gauss method.

372   Order conditions

Given rooted trees t, u and a symplectic Runge–Kutta method, we consider the relationship between the elementary weights $\phi(tu)$, $\phi(ut)$, $\phi(t)$, $\phi(u)$. Write

$$
\Phi(t) = \sum_{i=1} b_i \phi_i, \quad \Phi(u) = \sum_{i=1} b_i \psi_i.
$$

Then we find

$$
\Phi(tu) = \sum_{i,j=1}^{s} b_i \phi_i a_{ij} \psi_j,
$$

$$
\Phi(ut) = \sum_{i,j=1}^{s} b_j \psi_j a_{ji} \phi_i,
$$

so that

$$
\Phi(tu) + \Phi(ut) = \sum_{i,j=1}^{s} (b_i a_{ij} + b_j a_{ji}) \phi_i \psi_j
$$

$$
= \sum_{i,j=1}^{s} (b_i b_j) \phi_i \psi_j
$$

$$
= \Phi(t)\Phi(u).
$$

Assuming the order conditions $\Phi(t) = 1/\gamma(t)$ and $\Phi(u) = 1/\gamma(u)$ are satisfied, then

$$\Phi(tu) - \frac{1}{\gamma(tu)} + \Phi(ut) - \frac{1}{\gamma(ut)} = 0. \tag{372a}$$

Using this fact, we can prove the following theorem:

Theorem 372A *Let (A, b^{T}, c) be a symplectic Runge–Kutta method. The method has order p if and only if for each non-superfluous tree and any vertex in this tree as root, $\Phi(t) = 1/\gamma(t)$, where t is the rooted tree with this vertex.*

Proof.   We need only to prove the sufficiency of this criterion. If two rooted trees belong to the same tree but have vertices v_0, $\widehat{v}$ say, then there is a sequence of vertices $v_0, v_1, \ldots, v_m = \widehat{v}$, such that v_{i-1} and v_i are adjacent for $i = 1, 2, \ldots, m$. This mean that rooted trees t, u exist such that tu is the rooted tree with root v_{i-1} and ut is the rooted tree with root v_i. We are implicitly using induction on the order of trees and hence we can assume that $\Phi(t) = 1/\gamma(t)$ and $\Phi(u) = 1/\gamma(u)$. Hence, if one of the order conditions for the trees tu and ut is satisfied, then the other is. By working along the chain of possible roots $v_0, v_1, \ldots, v_m$, we see that the order condition associated with the root v_0 is equivalent to the condition for $\widehat{v}$. In the case of superfluous trees, one choice of adjacent vertices would imply that $t = u$. Hence, (372a) is equivalent to $2\Phi(tt) = 2/\gamma(tt)$ so that the order condition associated with tt is satisfied and all rooted trees belonging to the same tree are also satisfied.
□

373   Experiments with symplectic methods

The first experiment uses the simple pendulum based on the Hamiltonian $H(p, q) = p^2/2 - \cos(q)$ and initial value $(p, q) = (1, 0)$. The amplitude is found to be $\pi/3 \approx 1.047198$ and the period to be approximately 6.743001. Numerical solutions, displayed in Figure 373(i), were found using the Euler, implicit Euler and the implicit mid-point rule methods. Only the last of these is symplectic and its behaviour reflects this. That is, like the exact solution which is also shown, the area of the initial set remains unchanged, even though its shape is distorted.

The second experiment is based on problem (122c), which evolves on the unit sphere $y_1^2 + y_2^2 + y_3^2 = 1$. The value of $y_1^2 + y_2^2 + y_3^2$ is calculated by the Euler method, the implicit Euler method and the implicit mid-point rule method. Only the last of these is symplectic. The computed results are shown in Figure 373(ii). In each case a stepsize $h = 0.1$ was used. Although results are shown for only 500 time steps, the actual experiment was extended much further. There is no perceptible deviation from $y_1^2 + y_2^2 + y_3^2 = 1$ for the first million steps.
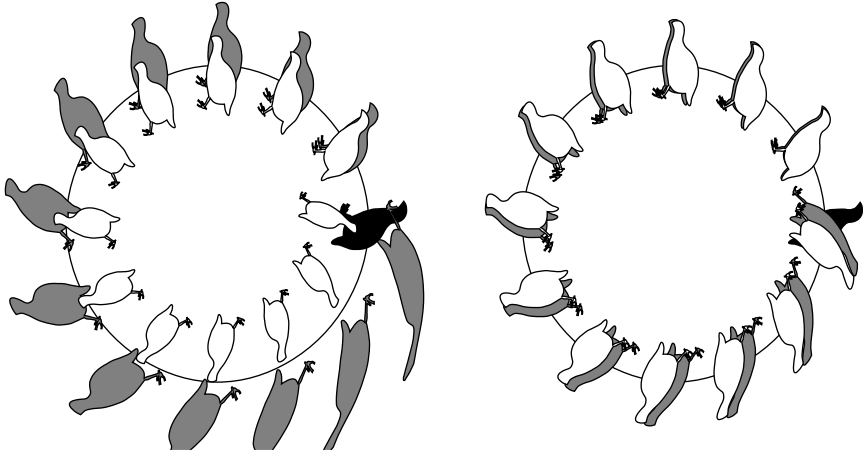
Figure 373(i) Solutions of the Hamiltonian problem $H(p,q) = p^2/2 - \cos(q)$. Left: Euler method (grey) and implicit Euler method (white). Right: exact solution (grey) and implicit mid-point method (white). The underlying image depicts the takahe *Porphyrio hochstetteri*, rediscovered in 1948 after many years of presumed extinction.
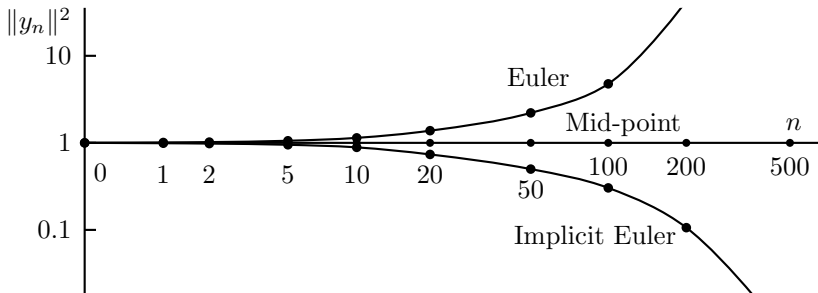
Figure 373(ii) Experiments for problem (122c). The computed value of $\|y_n\|^2$ is shown after $n = 1, 2, \ldots$, steps.

Exercises 37

37.1 Do two-stage symplectic Runge–Kutta methods exist which have order 3 but not order 4?

37.2 Do three-stage order 3 symplectic Runge–Kutta methods exist for which A is lower triangular?

38 Algebraic Properties of Runge–Kutta Methods

380 Motivation

For any specific N-dimensional initial value problem, Runge–Kutta methods can be viewed as mappings from $\mathbb{R}^N$ to $\mathbb{R}^N$. However, the semi-group generated by such mappings has a significance independent of the particular initial value problem, or indeed of the vector space in which solution values lie. If a method with s_1 stages is composed with a second method with s_2 stages, then the combined method with $s_1 + s_2$ stages can be thought of as the product of the original methods. It turns out that this is not quite the best way of formulating this product, and we need to work with equivalence classes of Runge–Kutta methods. This will also enable us to construct a group, rather than a mere semi-group.

It will be shown that the composition group of Runge–Kutta equivalent classes is homomorphic to a group on mappings from trees to real numbers. In fact the mapping that corresponds to a specific Runge–Kutta method is just the function that takes each tree to the associated elementary weight.

There are several reasons for introducing and studying these groups. For Runge–Kutta methods themselves, it is possible to gain a better understanding of the order conditions by looking at them in this way. Furthermore, methods satisfying certain simplifying assumptions, notably the C and D conditions, reappear as normal subgroups of the main group. An early application of this theory is the introduction of the concept of 'effective order'. This is a natural generalization from this point of view, but makes very little sense from a purely computational point of view. While effective order was not widely accepted at the time of its discovery, it has been rediscovered (López-Marcos, Sanz-Serna and Skeel, 1996) and has now been seen to have further ramifications.

The final claim that is made for this theory is that it has applications to the analysis of the order of general linear methods. In this guise a richer structure, incorporating an additive as well as a multiplicative operation, needs to be used; the present section also examines this more elaborate algebra.

The primary source for this theory is Butcher (1972), but it is also widely known through the work of Hairer and Wanner (1974). Recently the algebraic structures described here have been rediscovered through applications in theoretical physics. For a review of these developments, see Brouder (2000).

Before proceeding with this programme, we remark that the mappings from trees to real numbers, which appear as members of the algebraic systems introduced in this section, are associated with formal Taylor series of the form

$$a(\emptyset)y(x) + \sum_{t \in T} \frac{a(t)}{\sigma(t)} h^{r(t)} F(t)(y(x)). \tag{380a}$$

Such expressions as this were given the name B-series by Hairer and Wanner

(1974) and written
$$B(a, y(x)),$$
where $a : T^\# \to \mathbb{R}$, with $T^\#$ denoting the set of rooted trees T together with an additional empty tree $\emptyset$. Because of the central role of the exact solution series, in which $a(\emptyset) = 1$ and $a(t) = 1/\gamma(t)$, Hairer and Wanner scale the terms in the series slightly differently, and write

$$
\begin{aligned}
B(a, y(x)) &= a(\emptyset)y(x) + \sum_{t \in T} \frac{\alpha(t)a(t)}{r(t)!} h^{r(t)} F(t)(y(x)) \\
&= a(\emptyset)y(x) + \sum_{t \in T} \frac{a(t)}{\gamma(t)\sigma(t)!} h^{r(t)} F(t)(y(x)),
\end{aligned}
\tag{380b}
$$

where $\alpha(t)$ is the function introduced in Subsection 302. This means that the B-series representing a Runge–Kutta method with order p will have $a(t) = 1$ whenever $r(t) \leq p$. In this book we concentrate on the coefficients themselves, rather than on the series, but it will be the interpretation as coefficients in (380a), and not as coefficients in (380b), that will always be intended.

381   Equivalence classes of Runge–Kutta methods

We consider three apparently distinct ways in which two Runge–Kutta methods may be considered equivalent. Our aim will be to define these three equivalence relations and then show that they are actually *equivalent* equivalence relations. By this we mean that if two methods are equivalent in one of the three senses then they are equivalent also in each of the other senses. We temporarily refer to these three equivalence relations as 'equivalence', 'Φ-equivalence' and 'P-equivalence', respectively.

Definition 381A *Two Runge–Kutta methods are 'equivalent' if, for any initial value problem defined by an autonomous function f satisfying a Lipschitz condition, and an initial value y_0, there exists $h_0 > 0$ such that the result computed by the first method is identical with the result computed by the second method, if $h \leq h_0$.*

Definition 381B *Two Runge–Kutta methods are 'Φ-equivalent' if, for any $t \in T$, the elementary weight $\Phi(t)$ corresponding to the first method is equal to $\Phi(t)$ corresponding to the second method.*

In introducing P-equivalence, we need to make use of the concept of reducibility of a method. By this we mean that the method can be replaced by a method with fewer stages formed by eliminating stages that do not contribute in any way to the final result, and combining stages that are essentially the same into a single stage. We now formalize these two types of reducibility.

Definition 381C *A Runge–Kutta method (A, b^{T}, c) is '0-reducible' if the stage index set can be partitioned into two subsets $\{1, 2, \ldots, s\} = P_0 \cup P_1$ such that $b_i = 0$ for all $i \in P_0$ and such that $a_{ij} = 0$ if $i \in P_1$ and $j \in P_0$. The method formed by deleting all stages indexed by members of P_0 is known as the '0-reduced method'.*

Definition 381D *A Runge–Kutta method (A, b^{T}, c) is 'P-reducible' if the stage index set can be partitioned into $\{1, 2, \ldots, s\} = P_1 \cup P_2 \cup \cdots \cup P_{\overline{s}}$ and if, for all $I, J = 1, 2, \ldots, \overline{s}$, $\sum_{j \in P_J} a_{ij}$ is constant for all $i \in P_I$. The method $(\overline{A}, \overline{b}^{\mathsf{T}}, \overline{c})$, with $\overline{s}$ stages with $\overline{a}_{IJ} = \sum_{j \in P_J} a_{ij}$, for $i \in P_I$, $\overline{b}_I = \sum_{i \in P_I} b_i$ and $\overline{c}_I = c_i$, for $i \in P_I$, is known as the P-reduced method.*

Definition 381E *A Runge–Kutta method is 'irreducible' if it is neither 0-reducible nor P-reducible. The method formed from a method by first carrying out a P-reduction and then carrying out a 0-reduction is said to be the 'reduced method'.*

Definition 381F *Two Runge–Kutta methods are 'P-equivalent' if each of them reduces to the same reduced method.*

Theorem 381G *Let (A, b^{T}, c) be an irreducible s-stage Runge–Kutta method. Then, for any two stage indices $i, j \in \{1, 2, \ldots, s\}$, there exists a Lipschitz-continuous differential equation system such that $Y_i \neq Y_j$. Furthermore, there exists $t \in T$, such that $\Phi_i(t) \neq \Phi_j(t)$.*

Proof. If i, j exist such that

$$\Phi_i(t) = \Phi_j(t) \quad \text{for all } t \in T, \tag{381a}$$

then define a partition $P = \{P_1, P_2, \ldots, P_{\overline{s}}\}$ of $\{1, 2, \ldots, s\}$ such that i and j are in the same component of the partition if and only if (381a) holds. Let $\mathcal{A}$ denote the algebra of vectors in $\mathbb{R}^s$ such that, if i and j are in the same component of P, then the i and j components of $v \in \mathcal{A}$ are identical. The algebra is closed under vector space operations and under component-by-component multiplication. Note that the vector with every component equal to 1 is also in $\mathcal{A}$. Let $\widehat{\mathcal{A}}$ denote the subalgebra generated by the vectors made up from the values of the elementary weights for the stages for all trees. That is, if $t \in T$, then $v \in \mathbb{R}^s$ defined by $v_i = \Phi_i(t)$, $i = 1, 2, \ldots, s$, is in $\widehat{\mathcal{A}}$, as are the component-by-component products of the vectors corresponding to any finite set of trees. In particular, by using the empty set, we can regard the vector defined by $v_i = 1$ as also being a member of $\widehat{\mathcal{A}}$. Because of the way in which elementary weights are constructed, $v \in \widehat{\mathcal{A}}$ implies $Av \in \widehat{\mathcal{A}}$. We now show that $\widehat{\mathcal{A}} = \mathcal{A}$. Let I and J be two distinct members of P. Then because $t \in T$ exists so that $\Phi_i(t) \neq \Phi_j(t)$ for $i \in I$ and $j \in J$, we can find $v \in \widehat{\mathcal{A}}$ so that $v_i \neq v_j$. Hence, if $w = (v_i - v_j)^{-1}(v - v_j 1)$, where 1 in this

context represents the vector in $\mathbb{R}^s$ with every component equal to 1, then $w_i = 1$ and $w_j = 0$. Form the product of all such members of the algebra for $J \neq I$ and we deduce that the characteristic function of I is a member of $\mathcal{A}$. Since the S such vectors constitute a basis for this algebra, it follows that $\widehat{\mathcal{A}} = \mathcal{A}$. Multiply the characteristic function of J by A and note that, for all $i \in I \in P$, the corresponding component in the product is the same. This contradicts the assumption that the method is irreducible. Suppose it were possible that two stages, Y_i and Y_j, say, give identical results for any Lipschitz continuous differential equation, provided $h > 0$ is sufficiently small. We now prove the contradictory result that $\Phi_i(t) = \Phi_j(t)$ for all $t \in T$. If there were a $t \in T$ for which this does not hold, then write U for a finite subset of T containing t as in Subsection 314. Construct the corresponding differential equation as in that subsection and consider a numerical solution using the Runge–Kutta method (A, b^{T}, c) and suppose that t corresponds to component k of the differential equation. The value of component k of Y_i is $\Phi_i(t)$ and the value of component k of Y_j is $\Phi_j(t)$. $\qquad\square$

Now the key result interrelating the three equivalence concepts.

Theorem 381H *Two Runge–Kutta methods are equivalent if and only if they are P-equivalent and if and only if they are Φ-equivalent.*

Proof.
P-equivalence $\Rightarrow$ equivalence. It will enough to prove that if $i, j \in P_I$, in any P-reducible Runge–Kutta method, where we have used the notation of Definition 381D, then for any initial value problem, as in Definition 381A, $Y_i = Y_j$, for $h < h_0$. Calculate the stages by iteration starting with $Y_i^{[0]} = \eta$, for every $i \in \{1, 2, \ldots, s\}$. The value of $Y_i^{[k]}$ in iteration k will be identical for all i in the same partitioned component.

P-equivalence $\Rightarrow$ Φ-equivalence. Let the stages be partitioned according to $\{1, 2, \ldots, s\} = P_1 \cup P_2 \cup \cdots \cup P_{\overline{s}}$ and assume that a Runge–Kutta method is reducible with respect to this partition. It will be enough to prove that, for all $t \in T$, $\Phi_i(t) = \Phi_j(t)$ if i and j belong to the same component. This follows by induction on the order of t. It is true for $t = \tau$ because $\Phi_i(t) = c_i$ is constant for all i in the same component. For $t = [t_1 t_2 \cdots t_m]$,

$$\Phi_i([t_1 t_2 \cdots t_m]) = \sum_{j=1}^{s} a_{ij} \prod_{k=1}^{m} \Phi_j(t_k)$$

and this also is constant for all i in the same component.

Φ-equivalence $\Rightarrow$ P-equivalence. Suppose two methods are Φ-equivalent but not P-equivalent. Combine the s stages of method 1 and the $\widehat{s}$ stages of method 2, together with the output approximations, into a single method and

replace this by a reduced method. Because the original methods are not P-equivalent, the output approximations in the combined method are not in the same partition. Hence, by Theorem 381G, there exists $t \in T$ such that $\Phi_i(t)$ takes on different values for these two approximations.

   Equivalence $\Rightarrow$ P-equivalence. Suppose two methods are equivalent but not P-equivalent. Carry out the same construction as in the immediately previous part of the proof. By Theorem 381G, there is an initial value problem satisfying the requirements of Definition 381A such that Y_i takes on different values for the two output approximations. This contradicts the assumption that the original methods are equivalent. $\square$

382   The group of Runge–Kutta methods

Consider two equivalence classes of Runge–Kutta methods and choose a representative member of each of these classes. Because of the results of the previous subsection, equivalence is the same as Φ-equivalence and the same as P-equivalence. To see how to construct the composition product for the classes, form a tableau

$$
\begin{array}{c|cccccccc}
c_1 & a_{11} & a_{12} & \cdots & a_{1s} & 0 & 0 & \cdots & 0 \\
c_2 & a_{21} & a_{22} & \cdots & a_{2s} & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
c_s & a_{s1} & a_{s2} & \cdots & a_{ss} & 0 & 0 & \cdots & 0 \\
\\
\sum_{i=1}^{s} b_i + \widehat{c}_1 & b_1 & b_2 & \cdots & b_s & \widehat{a}_{11} & \widehat{a}_{12} & \cdots & \widehat{a}_{1\widehat{s}} \\
\sum_{i=1}^{s} b_i + \widehat{c}_2 & b_1 & b_2 & \cdots & b_s & \widehat{a}_{21} & \widehat{a}_{22} & \cdots & \widehat{a}_{2\widehat{s}} \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
\sum_{i=1}^{s} b_i + \widehat{c}_{\widehat{s}} & b_1 & b_2 & \cdots & b_s & \widehat{a}_{\widehat{s}1} & \widehat{a}_{\widehat{s}2} & \cdots & \widehat{a}_{\widehat{s}\widehat{s}} \\
\hline
 & b_1 & b_2 & \cdots & b_s & \widehat{b}_1 & \widehat{b}_2 & \cdots & \widehat{b}_{\widehat{s}}
\end{array}
\tag{382a}
$$

from the elements of the tableaux for the two methods (A, b^{T}, c) and $(\widehat{A}, \widehat{b}^{\mathsf{T}}, \widehat{c})$, respectively. We have written s and $\widehat{s}$ for the numbers of stages in the first and second method, respectively.

   By writing y_0 for the initial value for the first method and y_1 for the value computed in a step and then writing y_2 for the result computed by the second method *using y_1 for its initial value*, we see that y_2 is the result computed by the product method defined by (382a). To see why this is the case, denote the stage values by Y_i, $i = 1, 2, \ldots, s$, for the first method and by $\widehat{Y}_i$, $i = 1, 2, \ldots, \widehat{s}$, for the second method. The variables F_i and $\widehat{F}_i$ will denote the values of $f(Y_i)$ and $f(\widehat{Y}_i)$.

The values of the stages and of the final results computed within the first and second steps are

$$Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} F_j, \qquad i = 1, 2, \ldots, s, \tag{382b}$$

$$y_1 = y_0 + h \sum_{j=1}^{s} b_j F_j, \tag{382c}$$

$$\widehat{Y}_i = y_1 + h \sum_{j=1}^{\widehat{s}} \widehat{a}_{ij} \widehat{F}_j, \qquad i = 1, 2, \ldots, \widehat{s}, \tag{382d}$$

$$y_2 = y_1 + h \sum_{j=1}^{\widehat{s}} \widehat{b}_j \widehat{F}_j. \tag{382e}$$

Substitute y_1 from (382c) into (382d) and (382e), and we see that the coefficients for the stages in the second step and for the final output value y_2 are given as in the tableau (382a).

If m_1 and m_2 denote the methods (A, b^{T}, c) and $(\widehat{A}, \widehat{b}^{\mathsf{T}}, \widehat{c})$, respectively, write $m_1 \cdot m_2$ for the method defined by (382a). Also, for a given method m, we write $[m]$ for the equivalence class containing m. The notation $m \equiv \overline{m}$ will signify that m and $\overline{m}$ are equivalent methods.

We are interested in multiplication of equivalent classes, rather than of particular methods within these classes. Hence, we attempt to use the method given by (382a) as defining a new class of equivalent methods, which we can use as the product of the original two classes. The only possible difficulty could be that the result might depend on the particular choice of representative member for the two original classes. That no such difficulty arises follows from the following theorem:

Theorem 382A *Let m_1, m_2, $\overline{m}_1$, $\overline{m}_2$ denote Runge–Kutta methods, such that*

$$m_1 \equiv \overline{m}_1 \quad and \quad m_2 \equiv \overline{m}_2. \tag{382f}$$

Then

$$[m_1 \cdot m_2] = [\overline{m}_1 \cdot \overline{m}_2].$$

Proof. We note that an equivalent statement is

$$m_1 \cdot m_2 \equiv \overline{m}_1 \cdot \overline{m}_2. \tag{382g}$$

Let y_1 and y_2 denote the output values over the two steps for the sequence of steps constituting $m_1 \cdot m_2$, and $\overline{y}_1$ and $\overline{y}_2$ denote the corresponding output values for $\overline{m}_1 \cdot \overline{m}_2$. If f satisfies a Lipschitz condition and if h is sufficiently

small, then $y_1 = \overline{y}_1$ because $m_1 \equiv \overline{m}_1$, and $y_2 = \overline{y}_2$ because $m_2 \equiv \overline{m}_2$. Hence, (382g) and therefore (382f) follows.   □

Having constructed a multiplicative operation, we now construct an identity element and an inverse for equivalence classes of Runge–Kutta methods. For the identity element we consider the class containing any method m_0 that maps an initial value to an equal value, for a problem defined by a Lipschitz continuous function, provided that h is sufficiently small. It is clear that $[m_0 \cdot m] = [m \cdot m_0] = [m]$ for any Runge–Kutta method m. It will be convenient to denote the identity equivalence class by the symbol 1, where it will be clear from the context that this meaning is intended.

To define the inverse of an equivalence class, start with a particular representative $m = (A, b^{\mathsf{T}}, c)$, with s stages, and consider the tableau

$$
\begin{array}{c|cccc}
c_1 - \sum_{j=1}^{s} b_j & a_{11} - b_1 & a_{12} - b_2 & \cdots & a_{1s} - b_s \\
c_2 - \sum_{j=1}^{s} b_j & a_{21} - b_1 & a_{22} - b_2 & \cdots & a_{2s} - b_s \\
\vdots & \vdots & \vdots & & \vdots \\
c_s - \sum_{j=1}^{s} b_j & a_{s1} - b_1 & a_{s2} - b_2 & \cdots & a_{ss} - b_s \\
\hline
& -b_1 & -b_2 & \cdots & -b_s
\end{array}
$$

As we saw in Subsection 343, this method exactly undoes the work of m. Denote this new method by m^{-1}, and we prove the following result:

Theorem 382B *Let m denote a Runge–Kutta method. Then*

$$
[m \cdot m^{-1}] = [m^{-1} \cdot m] = 1.
$$

Proof. The tableaux for the two composite methods $m \cdot m^{-1}$ and $m^{-1} \cdot m$ are, respectively,

$$
\begin{array}{c|cccc cccc}
c_1 & a_{11} & a_{12} & \cdots & a_{1s} & 0 & 0 & \cdots & 0 \\
c_2 & a_{21} & a_{22} & \cdots & a_{2s} & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
c_s & a_{s1} & a_{s2} & \cdots & a_{ss} & 0 & 0 & \cdots & 0 \\
c_1 & b_1 & b_2 & \cdots & b_s & a_{11} - b_1 & a_{12} - b_2 & \cdots & a_{1s} - b_s \\
c_2 & b_1 & b_2 & \cdots & b_s & a_{21} - b_1 & a_{22} - b_2 & \cdots & a_{2s} - b_s \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
c_s & b_1 & b_2 & \cdots & b_s & a_{s1} - b_1 & a_{s2} - b_2 & \cdots & a_{ss} - b_s \\
\hline
& b_1 & b_2 & \cdots & b_s & -b_1 & -b_2 & \cdots & -b_s
\end{array}
$$

and

$$
\left[
\begin{array}{c|cccccccc}
c_1 - \sum_{j=1}^{s} b_j & a_{11} - b_1 & a_{12} - b_2 & \cdots & a_{1s} - b_s & 0 & 0 & \cdots & 0 \\
c_2 - \sum_{j=1}^{s} b_j & a_{21} - b_1 & a_{22} - b_2 & \cdots & a_{2s} - b_s & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
c_s - \sum_{j=1}^{s} b_j & a_{s1} - b_1 & a_{s2} - b_2 & \cdots & a_{ss} - b_s & 0 & 0 & \cdots & 0 \\
c_1 - \sum_{j=1}^{s} b_j & -b_1 & -b_2 & \cdots & -b_s & a_{11} & a_{12} & \cdots & a_{1s} \\
c_2 - \sum_{j=1}^{s} b_j & -b_1 & -b_2 & \cdots & -b_s & a_{21} & a_{22} & \cdots & a_{2s} \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
c_s - \sum_{j=1}^{s} b_j & -b_1 & -b_2 & \cdots & -b_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\
\hline
& -b_1 & -b_2 & \cdots & -b_s & b_1 & b_2 & \cdots & b_s
\end{array}
\right].
$$

Each of these methods is P-reducible to the methods m and m^{-1}, respectively, but in each case with b^{T} replaced by the zero vector, so that each lies in the equivalence class 1. $\qquad\square$

383   The Runge–Kutta group

While the group of equivalent classes of Runge–Kutta methods is conceptually very simple, it is difficult to use for detailed manipulations. We turn to a second group that is closely related to it, but which has a more convenient representation.

Let G_1 denote the set of functions on T, the rooted trees, to the real numbers. We define a binary relation on G_1 that makes it a group. It is convenient to widen the scope of our discussion by making use of forests. By a 'forest', we mean a set of vertices V and a set of edges E such that each edge is an ordered pair of members of V under the restrictions that each vertex appears as the second member of at most one edge. If $[v_1, v_2]$, $[v_2, v_3]$, $\ldots$, $[v_{n-1}, v_n]$ are edges, we write $v_1 < v_n$. We will require this relation to be a partial ordering.

Suppose that V and E can be partitioned as $V = V_1 \cup V_2 \cup \cdots \cup V_k$, $E = E_1 \cup E_2 \cup \cdots \cup E_k$, where each of (V_i, E_i), $i = 1, 2, \ldots, k$, is connected and is therefore a rooted tree. A function $\alpha : T \to \mathbb{R}$ can be extended multiplicatively to a function on the set of all forests by defining

$$
\alpha\big((V, E)\big) = \prod_{i=1}^{k} \alpha\big((V_i, E_i)\big).
$$

If (V, E) is a forest and $\widehat{V}$ is a subset of V, then the sub-forest induced by $\widehat{V}$ is the forest $(\widehat{V}, \widehat{E})$, where $\widehat{E}$ is the intersection of $\widehat{V} \times \widehat{V}$ and E. A special

case is when a sub-forest $(\widehat{V}, \widehat{E})$ satisfies the requirement that for any two vertices u, v of E such that $u < v$ and $v \in \widehat{E}$, u is also a member of $\widehat{E}$. In this case we write

$$(\widehat{V}, \widehat{E}) \triangleleft (V, E).$$

From now on we write forests by single letters Q, R, S, and interpret $R \triangleleft S$ accordingly. If $R \triangleleft S$ then $S \setminus R$ will denote the forest induced by the difference of the vertex sets of S and R, respectively.

We can now define a product of two multiplicative mappings of forests to real numbers. If α and β are two such mappings, then we write

$$(\alpha\beta)(S) = \sum_{R \triangleleft S} \alpha(S \setminus R)\beta(R). \tag{383a}$$

We need to verify that $\alpha\beta$ is multiplicative if the same is true for α and β.

Lemma 383A *Let α and β be multiplicative mappings from the forests to the real numbers. Then $\alpha\beta$ is multiplicative.*

Proof. It will be sufficient to consider the value of $(\alpha\beta)(S)$, where $S = S_1 \cup S_2$. Each $R \triangleleft S$ can be written as $R = R_1 \cup R_2$, where $R_1 \triangleleft S_1$ and $R_2 \triangleleft S_2$. We now have

$$(\alpha\beta)(S) = \sum_{R \triangleleft S} \alpha(S \setminus R)\beta(R)$$

$$= \sum_{R_1 \triangleleft S_1} \alpha(S_1 \setminus R_1)\beta(R_1) \sum_{R_2 \triangleleft S_2} \alpha(S_2 \setminus R_2)\beta(R_2)$$

$$= (\alpha\beta)(S_1)(\alpha\beta)(S_2). \qquad \square$$

We next show that the product we have defined is associative.

Lemma 383B *Let α, β and γ be multiplicative mappings from forests to reals. Then*

$$(\alpha\beta)\gamma = \alpha(\beta\gamma).$$

Proof. If $Q \triangleleft R \triangleleft S$ then $(R \setminus Q) \triangleleft (S \setminus Q)$. Hence, we find

$$((\alpha\beta)\gamma)(S) = \sum_{Q \triangleleft S} (\alpha\beta)(S \setminus Q)\gamma(Q)$$

$$= \sum_{Q \triangleleft S} \sum_{(R \setminus Q) \triangleleft (S \setminus Q)} \alpha((S \setminus Q) \setminus (R \setminus Q))\beta(R \setminus Q)\gamma(Q)$$

$$= \sum_{Q \triangleleft R} \sum_{R \triangleleft S} \alpha(S \setminus R)\beta(R \setminus Q)\gamma(Q)$$

$$= \sum_{R \triangleleft S} \alpha(S \setminus R)(\beta\gamma)(R)$$

$$= (\alpha(\beta\gamma))(S). \qquad \square$$

We can now restrict multiplication to trees, and we note that associativity still remains. The semi-group that has been constructed on the set G_1 is actually a group because we can construct both left and right inverses, $\alpha_{\text{left}}^{-1}$ and $\alpha_{\text{right}}^{-1}$ say, for any $\alpha \in G_1$, which must be equal because

$$\alpha_{\text{left}}^{-1} = \alpha_{\text{left}}^{-1}\left(\alpha\alpha_{\text{right}}^{-1}\right) = \left(\alpha_{\text{left}}^{-1}\alpha\right)\alpha_{\text{right}}^{-1} = \alpha_{\text{right}}^{-1}.$$

Lemma 383C *Given* $\alpha \in G_1$, *there exist a left inverse and a right inverse.*

Proof. We show, by induction on the order of t, that it is possible to construct β such that $(\alpha\beta)(t) = 0$ or $(\beta\alpha)(t) = 0$, for all $t \in T$. Because $(\alpha\beta)(\tau) = (\beta\alpha)(\tau) = \alpha(\tau) + \beta(\tau)$, the result is clear for order 1. Suppose the result has been proved for all trees of order less than that of $t \neq \tau$; then we note that

$$(\alpha\beta)(t) = \alpha(t) + \beta(t) + \phi(t, \alpha, \beta)$$

and

$$(\beta\alpha)(t) = \alpha(t) + \beta(t) + \phi(t, \beta, \alpha),$$

where $\phi(t, \alpha, \beta)$ involves the values of α and β only for trees with orders less than $r(t)$. Hence, it is possible to assign a value to $\beta(t)$ so that $(\alpha\beta)(t) = 0$ or that $(\beta\alpha)(t) = 0$, respectively. Thus it is possible to construct β as a left inverse or right inverse of α. $\qquad\square$

Having established the existence of an inverse for any $\alpha \in G_1$, we find a convenient formula for α^{-1}. We write S for a tree t, written in the form (V, E), and $\mathcal{P}(S)$ for the set of all partitions of S. This means that if $P \in \mathcal{P}(S)$, then P is a forest formed by possibly removing some of the edges from E. Another way of expressing this is that the components of P are trees (V_i, E_i), for $i = 1, 2, \ldots, n$, where V is the union of V_1, V_2, $\ldots$, V_n and each E_i is a subset of E. The integer n, denoting the number of components of P, will be written as $\#P$. We write t_i as the tree represented by (V_i, E_i).

Lemma 383D *Given* $\alpha \in G_1$ *and* $t \in T$, *written in the form* (V, E), *then*

$$\alpha^{-1}(t) = \sum_{P \in \mathcal{P}(S)} \prod_{i=1}^{\#P} (-\alpha(t_i)). \qquad (383b)$$

Proof. Construct a mapping $\beta \in G_1$ equal to the right-hand side of (383b). We show that for any $t \in T$, $(\alpha\beta)(t) = 0$ so that $\alpha\beta = 1$. Let $t = (V, E)$. For any partition P with components (V_i, E_i), for $i = 1, 2, \ldots, n$, we consider the set of possible combinations of $\{1, 2, \ldots, n\}$, with the restriction that if C is such a combination, then no edge $(v_1, v_2) \in E$ exists with $v_1 \in V_i$ and $v_2 \in V_j$, with i and j distinct members of C. Let $\mathcal{C}(P)$ denote the set of all

such combinations of $P \in \mathcal{P}(t)$. Given $C \in P$, denote by $\overline{C}$ the complement of C in P.

The value of $(\alpha\beta)(t)$ can be written in the form

$$\sum_{P \in \mathcal{P}(t)} \sum_{C \in \mathcal{C}(P)} \prod_{i \in C} \alpha(t_i)(-1)^{\#\overline{C}} \prod_{j \in \overline{C}} \alpha(t_j).$$

For any particular partition P, the total contribution is

$$\sum_{C \in \mathcal{C}(P)} (-1)^{n-\#C} \prod_{i=1}^{\#P} \alpha(t_i).$$

This is zero because $\sum_{C \in \mathcal{C}(P)} (-1)^{n-\#C} = 0$. $\square$

384   A homomorphism between two groups

We show that the groups introduced in Subsections 382 and 383 are related in such a way that the former is isomorphic to a subgroup of the latter. The mapping between elements of the group that provides this homomorphism maps an equivalence class of Runge–Kutta methods to the function on T to $\mathbb{R}$ defined by the elementary weights associated with a representative member of the class. We need to establish that products in the first group are preserved in the second. This means that if m and $\widehat{m}$ are Runge–Kutta methods and $\Phi : T \to \mathbb{R}$ and $\widehat{\Phi} : T \to \mathbb{R}$ are the elementary weight functions for m and $\widehat{m}$, respectively, then $\Phi\widehat{\Phi}$ is the elementary weight function associated with $m\widehat{m}$.

Theorem 384A *Let* $\Phi : T \to \mathbb{R}$ *be the elementary weight function associated with* (A, b^{T}, c) *and* $\widehat{\Phi} : T \to \mathbb{R}$ *the elementary weight function associated with* $(\widehat{A}, \widehat{b}^{\mathsf{T}}, \widehat{c})$. *Let* $\widetilde{\Phi} : T \to \mathbb{R}$ *denote the elementary weight function for the product method as represented by* (382a). *Then*

$$\widetilde{\Phi} = \Phi\widehat{\Phi}.$$

Proof. Denote the $(s + \widehat{s})$-stage composite coefficient matrices by $(\widetilde{A}, \widetilde{b}^{\mathsf{T}}, \widetilde{c})$ with the elements of $\widetilde{A}$ and $\widetilde{b}^{\mathsf{T}}$ given by

$$\widetilde{a}_{ij} = \begin{cases} a_{ij}, & i \le s, \quad j \le s, \\ 0, & i \le s, \quad j > s, \\ b_j, & i > s, \quad j \le s, \\ \widehat{a}_{i-s,j-s}, & i > s, \quad j > s. \end{cases}$$

$$\widetilde{b}_i = \begin{cases} b_i, & i \le s, \\ \widehat{b}_{i-s}, & i > s. \end{cases}$$

For a tree t, such that $r(t) = n$, represented by the vertex–edge pair (V, E), with root $\rho \in V$, write the elementary weight $\widetilde{\Phi}(t)$ in the form

$$\widetilde{\Phi}(t) = \sum_{i \in I} \widetilde{b}_{i(\rho)} \prod_{(v,w) \in E} \widetilde{a}_{i(v),i(w)}. \tag{384a}$$

In this expression, I is the set of all mappings from V to the set $\{1, 2, \ldots, \widetilde{s}\}$ and, for $i \in I$ and $v \in V$, $i(v)$ denotes the value to which the vertex v maps.

If $v < w$ and $i(v) \leq s < i(w)$ then the corresponding term in (384a) is zero. Hence, we sum only over I' defined as the subset of I from which such i are omitted. For any $i \in I'$, define $R \lhd S = (V, E)$ such that all the vertices associated with R map into $\{s+1, s+2, \ldots, s+\widehat{s}\}$. Collect together all $i \in I'$ which share a common R so that (384a) can be written in the form

$$\widetilde{\Phi}(t) = \sum_{R \lhd S} \sum_{i \in I_R} \widetilde{b}_{i(\rho)} \prod_{(v,w) \in E} \widetilde{a}_{i(v),i(w)}.$$

For each R, the terms in the sum have total value $\Phi(S \setminus R)\widehat{\Phi}(R)$, and the result follows. $\qquad\square$

385   A generalization of G_1

It will be convenient to build an algebraic system similar to G_1, but possessing, in addition to the group structure, a vector space structure. We cannot exactly achieve all of this, but we can achieve almost all of it. The way we go about this is to add to T an additional member, known as the 'empty tree' and denoted by $\emptyset$. The augmented set of trees will be denoted by $T^{\#}$. We write G for the set of mappings $T^{\#} \to \mathbb{R}$ and G_1 for the set of those members of G for which $\emptyset$ maps to 1. We define the operation $G_1 \times G \to G$ just as for the group operation except that the coefficient of $\alpha(t)$ in the formula for $(\alpha\beta)(t)$ is $\beta(\emptyset)$. With this understanding we retain the associativity property, in cases where it makes sense. That is, if $\alpha, \beta \in G_1$ and $\gamma \in G$, then

$$(\alpha\beta)\gamma = \alpha(\beta\gamma).$$

Furthermore, left-multiplication by an element of G_1 is linear in the sense that

$$\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma,$$

whenever $\alpha \in G_1$ and $\beta, \gamma \in G$. Furthermore,

$$\alpha(c\beta) = c\alpha\beta,$$

where, for a scalar c, $c\beta$ is the mapping that takes t to $c\beta(t)$ for all $t \in T^{\#}$.

The generalization we have introduced has a simple significance in terms of Runge–Kutta tableaux and methods. Instead of computing the output value from a step of computation by the formula

$$y_0 + h \sum_{i=1}^{s} b_i F_i, \tag{385a}$$

where y_0 is the input value and F_1, F_2, ..., F_s are stage derivatives, we can replace (385a) by

$$b_0 y_0 + h \sum_{i=1}^{s} b_i F_i.$$

To express this in a tableau, we place the coefficient b_0 in the spare space at the left of the last line. Thus, the tableau would have the form

$$
\begin{array}{c|cccc}
c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\
c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\
\vdots & \vdots & \vdots & & \vdots \\
c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\
\hline
b_0 & b_1 & b_2 & \cdots & b_s
\end{array} \, .
$$

As a Runge–Kutta method, to be used in the conventional manner as a one-step method for advancing the solution of a differential equation, this makes no sense at all, if $b_0 \neq 1$. Indeed, the method would not even be covariant with respect to shifts of origin. However, the process of computing with a single step of this contrived method may play an important role as part of a more sophisticated computation. An important example of a generalized Runge–Kutta method is given by the one-stage tableau

$$
\begin{array}{c|c}
0 & 0 \\
\hline
0 & 1
\end{array} \, . \tag{385b}
$$

This method does nothing other than computing h multiplied by the derivative of the input value. Combined with linear operations, all Runge–Kutta methods can be built up from this basic method. The elementary weights associated with this method are given by

$$
\Phi(t) = \begin{cases} 1, & t = \tau, \\ 0, & t \neq \tau. \end{cases}
$$

386 Recursive formula for the product

We consider a formalism for the product on $G_1 \times G \to G$, based on the second of the recursive constructions of trees defined in Subsection 300. That is, for

two trees t, u, we define tu as the tree formed by joining the roots of t and u with the root of t regarded as the root of the product. Corresponding to $t \in T^\#$, we define $\widehat{t} : G_1 \to \mathbb{R}$ by the formula

$$\widehat{t}(\alpha) = \alpha(t), \qquad \alpha \in G_1.$$

The set of all $\widehat{t}$, for $t \in T$, will be denoted by $\widehat{T}$. We extend the dot-product notation from $T \times T \to T$ to $\widehat{T} \times \widehat{T} \to \widehat{T}$ by the formula

$$\widehat{t} \cdot \widehat{u} = \widehat{tu}.$$

Since $\widehat{T}^\#$ denotes a set of linear functionals on G, it is natural to consider also the vector space spanned by such functionals and extend the dot-product notation to make the product of two functionals bilinear. We denote this set of functionals by G^*.

We can now define a special function, $\lambda : G_1 \times T \to G^*$, by the recursion

$$\lambda(\alpha, \tau) = \widehat{\tau},$$
$$\lambda(\alpha, tu) = \lambda(\alpha, t)\lambda(\alpha, u) + \alpha(u)\lambda(\alpha, t).$$

This enables us to generate expressions for $\alpha\beta$ for all trees.

Theorem 386A *For $\alpha \in G_1$ and $\beta \in G$,*

$$(\alpha\beta)(\emptyset) = \beta(\emptyset),$$
$$(\alpha\beta)(t) = \lambda(\alpha, t)(\beta) + \alpha(t)\beta(\emptyset).$$

Proof. In this proof only, we introduce the notation $R \dot{\vartriangleleft} S$ to denote $R \vartriangleleft S$, with $R \neq \emptyset$. If a tree t is represented by the set S of vertices, with an implied set of edges, then the notation t_R, where $R \vartriangleleft S$, will denote the tree formed from the elements of R, with the induced set of edges. With this terminology, we can write (383a) in the form

$$(\alpha\beta)(t) = \sum_{R \dot{\vartriangleleft} S} \alpha(S \setminus R)\beta(R) + \alpha(t)\beta(\emptyset).$$

Hence, we need to show that

$$\lambda(\alpha, t) = \sum_{R \dot{\vartriangleleft} S} \alpha(S \setminus R)\widehat{t}_R.$$

This is obvious in the case $t = \tau$. We now consider a tree tu with t represented by S and u represented by Q. This means that tu can be represented by the graph (V, E), where V is the union of the vertex sets associated with S and

Table 386(I)   The function λ for trees of orders 1 to 5

t		$r(t)$	$\lambda(\alpha, t)$
$\bullet$	τ	1	$\widehat{\tau}$
	$\tau\tau$	2	$\widehat{\widehat{\tau}\tau} + \alpha(\tau)\widehat{\tau}$
	$\tau\tau\cdot\tau$	3	$\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau} + 2\alpha(\tau)\widehat{\widehat{\tau}\tau} + \alpha(\tau)^2\widehat{\tau}$
	$\tau\cdot\tau\tau$	3	$\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau} + \alpha(\tau)\widehat{\widehat{\tau}\tau} + \alpha(\tau\tau)\widehat{\tau}$
	$(\tau\tau\cdot\tau)\tau$	4	$(\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau})\widehat{\tau} + 3\alpha(\tau)\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau} + 3\alpha(\tau)^2\widehat{\widehat{\tau}\tau} + \alpha(\tau)^3\widehat{\tau}$
	$\tau\tau\cdot\tau\tau$	4	$\widehat{\widehat{\tau}\tau}\cdot\widehat{\widehat{\tau}\tau} + \alpha(\tau)\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau} + \alpha(\tau)\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau} +$ $(\alpha(\tau)^2 + \alpha(\tau\tau))\widehat{\widehat{\tau}\tau} + \alpha(\tau)\alpha(\tau\tau)\widehat{\tau}$
	$\tau(\tau\tau\cdot\tau)$	4	$\widehat{\tau}(\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau}) + 2\alpha(\tau)\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau} + \alpha(\tau)^2\widehat{\widehat{\tau}\tau} + \alpha(\tau\tau\cdot\tau)\widehat{\tau}$
	$\tau(\tau\cdot\tau\tau)$	4	$\widehat{\tau}(\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau}) + \alpha(\tau)\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau} + \alpha(\tau\tau)\widehat{\widehat{\tau}\tau} + \alpha(\tau\cdot\tau\tau)\widehat{\tau}$
	$(\tau\tau\cdot\tau)\tau\cdot\tau$	5	$(\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau})\widehat{\tau}\cdot\widehat{\tau} + 4\alpha(\tau)(\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau})\widehat{\tau} + 6\alpha(\tau)^2\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau} +$ $4\alpha(\tau)^3\widehat{\widehat{\tau}\tau} + \alpha(\tau)^4\widehat{\tau}$
	$(\tau\tau\cdot\tau)\cdot\tau\tau$	5	$(\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau})\cdot\widehat{\widehat{\tau}\tau} + 2\alpha(\tau)\widehat{\widehat{\tau}\tau}\cdot\widehat{\widehat{\tau}\tau} + \alpha(\tau)(\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau})\widehat{\tau} +$ $2\alpha(\tau)^2\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau} + (\alpha(\tau)^2 + \alpha(\tau\tau))\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau} +$ $(\alpha(\tau)^3 + 2\alpha(\tau)\alpha(\tau\tau))\widehat{\widehat{\tau}\tau} + \alpha(\tau)^2\alpha(\tau\tau)\widehat{\tau}$
	$\tau\tau\cdot(\tau\tau\cdot\tau)$	5	$\widehat{\widehat{\tau}\tau}\cdot(\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau}) + 2\alpha(\tau)\widehat{\widehat{\tau}\tau}\cdot\widehat{\widehat{\tau}\tau} + \alpha(\tau)\widehat{\tau}(\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau}) +$ $\alpha(\tau)^2\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau} + 2\alpha(\tau)^2\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau} +$ $(\alpha(\tau)^3 + \alpha(\tau\tau\cdot\tau))\widehat{\widehat{\tau}\tau} + \alpha(\tau)\alpha(\tau\tau\cdot\tau)\widehat{\tau}$
	$\tau\tau\cdot(\tau\cdot\tau\tau)$	5	$\widehat{\widehat{\tau}\tau}\cdot(\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau}) + \alpha(\tau)\widehat{\widehat{\tau}\tau}\cdot\widehat{\widehat{\tau}\tau} + \alpha(\tau)\widehat{\tau}(\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau}) +$ $\alpha(\tau\tau)\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau} + \alpha(\tau)^2\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau} +$ $(\alpha(\tau)\alpha(\tau\tau) + \alpha(\tau\cdot\tau\tau))\widehat{\widehat{\tau}\tau} + \alpha(\tau)\alpha(\tau\cdot\tau\tau)\widehat{\tau}$
	$(\tau\cdot\tau\tau)\cdot\tau\tau$	5	$(\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau})\cdot\widehat{\widehat{\tau}\tau} + 2\alpha(\tau)\widehat{\widehat{\tau}\tau}\cdot\widehat{\widehat{\tau}\tau} + \alpha(\tau)^2\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau} +$ $2\alpha(\tau\tau)\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau} + 2\alpha(\tau)\alpha(\tau\tau)\widehat{\widehat{\tau}\tau} + \alpha(\tau\tau)^2\widehat{\tau}$
	$\tau\cdot(\tau\tau\cdot\tau)\tau$	5	$\widehat{\tau}\cdot(\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau})\widehat{\tau} + 3\alpha(\tau)\widehat{\tau}(\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau}) + 3\alpha(\tau)^2\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau} +$ $\alpha(\tau)^3\widehat{\widehat{\tau}\tau} + \alpha((\tau\tau\cdot\tau)\tau)\widehat{\tau}$
	$\tau(\tau\tau\cdot\tau\tau)$	5	$\widehat{\tau}(\widehat{\widehat{\tau}\tau}\cdot\widehat{\widehat{\tau}\tau}) + \alpha(\tau)\widehat{\tau}(\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau}) + \alpha(\tau)\widehat{\tau}(\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau}) +$ $(\alpha(\tau)^2 + \alpha(\tau\tau))\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau} + \alpha(\tau)\alpha(\tau\tau)\widehat{\widehat{\tau}\tau} + \alpha(\tau\tau\cdot\tau\tau)\widehat{\tau}$
	$\tau\cdot\tau(\tau\tau\cdot\tau)$	5	$\widehat{\tau}\cdot\widehat{\tau}(\widehat{\widehat{\tau}\tau}\cdot\widehat{\tau}) + 2\alpha(\tau)\widehat{\tau}(\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau}) + \alpha(\tau)^2\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau} +$ $\alpha(\tau\tau\cdot\tau)\widehat{\widehat{\tau}\tau} + \alpha(\tau(\tau\tau\cdot\tau))\widehat{\tau}$
	$\tau\cdot\tau(\tau\cdot\tau\tau)$	5	$\widehat{\tau}\cdot\widehat{\tau}(\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau}) + \alpha(\tau)\widehat{\tau}(\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau}) + \alpha(\tau\tau)\widehat{\tau}\cdot\widehat{\widehat{\tau}\tau} +$ $\alpha(\tau\cdot\tau\tau)\widehat{\widehat{\tau}\tau} + \alpha(\tau(\tau\cdot\tau\tau))\widehat{\tau}$

Q, and E is the union of the corresponding edge sets together with additional edge connecting the two roots. Temporarily we write $(V, E) = SQ$. If $R \trianglelefteq S$ and $P \trianglelefteq Q$ then the set of subgraphs related to SQ by the relation $X \trianglelefteq SQ$ are of the form $X = RP$ or of the form $X = R$. Hence,

Table 386(II) Formulae for $(\alpha\beta)(t_i)$ up to trees of order 5

i	$r(t_i)$	t_i	$(\alpha\beta)(t_i)$
0	0	$\emptyset$	β_0
1	1		$\beta_1 + \alpha_1\beta_0$
2	2		$\beta_2 + \alpha_1\beta_1 + \alpha_2\beta_0$
3	3		$\beta_3 + 2\alpha_1\beta_2 + \alpha_1^2\beta_1 + \alpha_3\beta_0$
4	3		$\beta_4 + \alpha_1\beta_2 + \alpha_2\beta_1 + \alpha_4\beta_0$
5	4		$\beta_5 + 3\alpha_1\beta_3 + 3\alpha_1^2\beta_2 + \alpha_1^3\beta_1 + \alpha_5\beta_0$
6	4		$\beta_6 + \alpha_1\beta_4 + \alpha_1\beta_3 + (\alpha_1^2 + \alpha_2)\beta_2 + \alpha_1\alpha_2\beta_1 + \alpha_6\beta_0$
7	4		$\beta_7 + 2\alpha_1\beta_4 + \alpha_1^2\beta_2 + \alpha_3\beta_1 + \alpha_7\beta_0$
8	4		$\beta_8 + \alpha_1\beta_4 + \alpha_2\beta_2 + \alpha_4\beta_1 + \alpha_8\beta_0$
9	5		$\beta_9 + 4\alpha_1\beta_5 + 6\alpha_1^2\beta_3 + 4\alpha_1^3\beta_2 + \alpha_1^4\beta_1 + \alpha_9\beta_0$
10	5		$\beta_{10} + 2\alpha_1\beta_6 + \alpha_1\beta_5 + \alpha_1^2\beta_4 + (2\alpha_1^2 + \alpha_2)\beta_3 + \\ (2\alpha_1\alpha_2 + \alpha_1^3)\beta_2 + \alpha_1^2\alpha_2\beta_1 + \alpha_{10}\beta_0$
11	5		$\beta_{11} + \alpha_1\beta_7 + 2\alpha_1\beta_6 + 2\alpha_1^2\beta_4 + \alpha_1^2\beta_3 + (\alpha_1^3 + \alpha_3)\beta_2 + \\ \alpha_1\alpha_3\beta_1 + \alpha_{11}\beta_0$
12	5		$\beta_{12} + \alpha_1\beta_8 + \alpha_1\beta_6 + \alpha_1^2\beta_4 + \alpha_2\beta_3 + (\alpha_1\alpha_2 + \alpha_4)\beta_2 + \\ \alpha_1\alpha_4\beta_1 + \alpha_{12}\beta_0$
13	5		$\beta_{13} + 2\alpha_1\beta_6 + 2\alpha_2\beta_4 + \alpha_1^2\beta_3 + 2\alpha_1\alpha_2\beta_2 + \alpha_2^2\beta_1 + \alpha_{13}\beta_0$
14	5		$\beta_{14} + 3\alpha_1\beta_7 + 3\alpha_1^2\beta_4 + \alpha_1^3\beta_2 + \alpha_5\beta_1 + \alpha_{14}\beta_0$
15	5		$\beta_{15} + \alpha_1\beta_8 + \alpha_1\beta_7 + (\alpha_1^2 + \alpha_2)\beta_4 + \alpha_1\alpha_2\beta_2 + \alpha_6\beta_1 + \alpha_{15}\beta_0$
16	5		$\beta_{16} + 2\alpha_1\beta_8 + \alpha_1^2\beta_4 + \alpha_3\beta_2 + \alpha_7\beta_1 + \alpha_{16}\beta_0$
17	5		$\beta_{17} + \alpha_1\beta_8 + \alpha_2\beta_4 + \alpha_4\beta_2 + \alpha_8\beta_1 + \alpha_{17}\beta_0$

$$\sum_{X \trianglelefteq SQ} \alpha(SQ \setminus X)\widehat{t}_X = \sum_{P \trianglelefteq Q} \sum_{R \trianglelefteq S} \alpha(SQ \setminus PR)\widehat{t}_{PR} + \sum_{R \trianglelefteq S} \alpha(SQ \setminus R)\widehat{t}_R$$

$$= \sum_{P \trianglelefteq Q} \alpha(Q \setminus P)\widehat{t}_P \sum_{R \trianglelefteq S} \alpha(S \setminus R)\widehat{t}_R + \alpha((S \setminus R)Q) \sum_{R \trianglelefteq S} \widehat{t}_R$$

$$= \lambda(\alpha, t)\lambda(\alpha, u) + \alpha(u)\lambda(\alpha, t)$$

$$= \lambda(\alpha, tu). \qquad \square$$

Table 386(III)   Formulae for $(\alpha^{-1})(t_i)$ up to trees of order 5

i	$r(t_i)$	t_i	$(\alpha^{-1})(t_i)$
1	1	$\bullet$	$-\alpha_1$
2	2		$\alpha_1^2 - \alpha_2$
3	3		$2\alpha_1\alpha_2 - \alpha_1^3 - \alpha_3$
4	3		$2\alpha_1\alpha_2 - \alpha_1^3 - \alpha_4$
5	4		$3\alpha_1\alpha_3 - 3\alpha_2\alpha_1^2 + \alpha_1^4 - \alpha_5$
6	4		$\alpha_1\alpha_3 + \alpha_1\alpha_4 + \alpha_2^2 - 3\alpha_2\alpha_1^2 + \alpha_1^4 - \alpha_6$
7	4		$2\alpha_1\alpha_4 + \alpha_1\alpha_3 - 3\alpha_1^2\alpha_2 + \alpha_1^4 - \alpha_7$
8	4		$2\alpha_1\alpha_4 + \alpha_2^2 - 3\alpha_1^2\alpha_2 + \alpha_1^4 - \alpha_8$
9	5		$4\alpha_1\alpha_5 - 6\alpha_1^2\alpha_3 + 4\alpha_1^3\alpha_2 - \alpha_1^5 - \alpha_9$
10	5		$2\alpha_1\alpha_6 + \alpha_1\alpha_5 + \alpha_2\alpha_3 - \alpha_1^2\alpha_4 - 3\alpha_1^2\alpha_3 + 4\alpha_1\alpha_2 - \alpha_1^5 - \alpha_{10}$
11	5		$\alpha_1\alpha_7 + 2\alpha_1\alpha_6 + \alpha_2\alpha_3 - 2\alpha_1\alpha_2^2 - \alpha_1^2\alpha_3 - 2\alpha_1^2\alpha_4 +$ $4\alpha_1^3\alpha_2 - \alpha_1^5 - \alpha_{11}$
12	5		$\alpha_1\alpha_8 + \alpha_1\alpha_6 + \alpha_2\alpha_3 + \alpha_2\alpha_4 - 3\alpha_1\alpha_2^2 - \alpha_1^2\alpha_3 - 2\alpha_1^2\alpha_4 +$ $4\alpha_1^3\alpha_2 - \alpha_1^5 - \alpha_{12}$
13	5		$2\alpha_1\alpha_6 + 2\alpha_2\alpha_4 - \alpha_1^2\alpha_3 - 2\alpha_1^2\alpha_4 - 3\alpha_1\alpha_2^2 + 4\alpha_1^3\alpha_2 - \alpha_1^5 - \alpha_{13}$
14	5		$3\alpha_1\alpha_7 + \alpha_1\alpha_5 - 3\alpha_1^2\alpha_4 - 3\alpha_1^2\alpha_3 + 4\alpha_1^3\alpha_2 - \alpha_1^5 - \alpha_{14}$
15	5		$\alpha_1\alpha_8 + \alpha_1\alpha_7 + \alpha_1\alpha_6 + \alpha_2\alpha_4 - 2\alpha_1\alpha_2^2 - \alpha_1^2\alpha_3 - 3\alpha_1^2\alpha_4 +$ $4\alpha_1^3\alpha_2 - \alpha_1^5 - \alpha_{15}$
16	5		$2\alpha_1\alpha_8 + \alpha_1\alpha_7 + \alpha_2\alpha_3 - 2\alpha_1\alpha_2^2 - \alpha_1^2\alpha_3 - 3\alpha_1^2\alpha_4 +$ $4\alpha_1^3\alpha_2 - \alpha_1^5 - \alpha_{16}$
17	5		$2\alpha_1\alpha_8 + 2\alpha_2\alpha_4 - 3\alpha_1\alpha_2^2 + 4\alpha_1^3\alpha_2 - \alpha_1^5 - \alpha_{17}$

As examples of the use of the algorithm for evaluating λ, and thence values of the product on $G_1 \times G$, we find

$$\lambda(\alpha, \tau) = \widehat{\tau}, \tag{386a}$$

$$\lambda(\alpha, \tau\tau) = \widehat{\tau}\widehat{\tau} + \alpha(\tau)\widehat{\tau}, \tag{386b}$$

$$\lambda(\alpha, \tau\tau\cdot\tau) = (\widehat{\tau}\widehat{\tau} + \alpha(\tau)\widehat{\tau})\cdot\widehat{\tau} + \alpha(t)(\widehat{\tau}\widehat{\tau} + \alpha(\tau)\widehat{\tau})$$

$$= \widehat{\tau}\widehat{\tau}\cdot\widehat{\tau} + 2\alpha(\tau)\widehat{\tau}\widehat{\tau} + \alpha(\tau)^2\widehat{\tau}, \tag{386c}$$

$$\lambda(\alpha, \tau\cdot\tau\tau) = \widehat{\tau}\cdot(\widehat{\tau}\widehat{\tau} + \alpha(\tau)\widehat{\tau}) + \alpha(\tau\tau)\widehat{\tau}$$

$$= \widehat{\tau}\cdot\widehat{\tau}\widehat{\tau} + \alpha(\tau)\widehat{\tau}\widehat{\tau} + \alpha(\tau\tau)\widehat{\tau}. \tag{386d}$$

The values of $\lambda(\alpha, t)$ are continued in Table 386(I) up to trees of order 5. For convenience, each tree is given in product form as well as in pictorial form.

From (386a)–(386d), we find

$$(\alpha\beta)(\tau) = \beta(\tau) + \alpha(\tau)\beta(\emptyset),$$
$$(\alpha\beta)(\tau\tau) = \beta(\tau\tau) + \alpha(\tau)\beta(\tau) + \alpha(\tau\tau)\beta(\emptyset),$$
$$(\alpha\beta)(\tau\tau \cdot \tau) = \beta(\tau\tau \cdot \tau) + 2\alpha(\tau)\beta(\tau\tau) + \alpha(\tau)^2\beta(\tau) + \alpha(\tau\tau \cdot \tau)\beta(\emptyset),$$
$$(\alpha\beta)(\tau \cdot \tau\tau) = \beta(\tau \cdot \tau\tau) + \alpha(\tau)\beta(\tau\tau) + \alpha(\tau\tau)\beta(\tau) + \alpha(\tau \cdot \tau\tau)\beta(\emptyset).$$

It will be convenient to extend these formulae up to trees of order 5, and we present this in Table 386(II). For convenience, we denote the empty tree by t_0 and the trees of order 1 to 5 by t_i, $i = 1, 2, \ldots, 17$. We also write α_i and β_i for $\alpha(t_i)$ and $\beta(t_i)$, respectively. Note that α_0 does not appear in this table because it always has the value $\alpha(\emptyset) = 1$.

Because Table 386(II) has reference value, we supplement the information it contains with Table 386(III), which gives the formulae for $(\alpha^{-1})(t)$ where $r(t) \leq 5$ and $\alpha \in G_1$.

387   Some special elements of G

As we have remarked, $D \in G$ represents the differentiation operation, scaled by the unit stepsize h. If ξ denotes the element in G_1 corresponding to a generalized Runge–Kutta tableau

$$
\frac{c \;\big|\; A}{1 \;\big|\; b^{\mathsf{T}}} =
\begin{array}{c|ccccc}
c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\
c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\
\vdots & \vdots & \vdots & & \vdots \\
c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\
\hline
1 & b_1 & b_2 & \cdots & b_s
\end{array}
\tag{387a}
$$

then ξD will correspond to the s-stage tableau

$$
\begin{array}{c|ccccc}
c_1 & a_{11} & a_{12} & \cdots & a_{1s} & 0 \\
c_2 & a_{21} & a_{22} & \cdots & a_{2s} & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
c_s & a_{s1} & a_{s2} & \cdots & a_{ss} & 0 \\
\sum_{i=1}^{s} b_i & b_1 & b_2 & \cdots & b_s & 0 \\
\hline
0 & 0 & 0 & \cdots & 0 & 1
\end{array}
\tag{387b}
$$

The result computed by (387b) is just $hf(\widehat{y})$, where $\widehat{y}$ is the result computed by (387a). With this understanding, we have an alternative means of defining the group element corresponding to each of the stages, as well as the final

result, of a Runge–Kutta method. Denote the members of G_1 corresponding to the stages Y_i, $i = 1, 2, \ldots, s$, of (387a) by η_i and the output result by ξ; then

$$\eta_i = 1 + \sum_{j=1}^{s} a_{ij} \eta_j D,$$

$$\xi = 1 + \sum_{i=1}^{s} b_i \eta_i D. \tag{387c}$$

In the case of a generalized method, where b_0 is the coefficient of y_{n-1} in the formula for y_n, (387c) is replaced by

$$\xi = b_0 1 + \sum_{i=1}^{s} b_i \eta_i D,$$

where, in this context, 1 is the group-theoretic identity in G.

In addition to D, it is convenient to introduce an element $E \in G$, defined by

$$E(\emptyset) = 1,$$

$$E(t) = \frac{1}{\gamma(t)}, \qquad t \in T.$$

This means that E corresponds to the exact solution of the differential equation as represented by the Picard iteration scheme introduced in Section 311. The conditions for order p for the Runge–Kutta method (387a) can now be written in the form

$$\xi(t) = E(t), \qquad r(t) \le p.$$

Finally, we define a sequence of members of G which correspond to the computation of the Taylor coefficients at the initial point, scaled in terms of powers of h. If T_k corresponds to the method which, on input $y_0 = y(x_0)$, computes $h^k y^{(k)}(x_0)$, then we require that

$$T_k(\emptyset) = 0,$$

$$T_k(t) = \begin{cases} \alpha(t), & r(t) = k, \\ 0, & r(t) \ne k. \end{cases}$$

Obviously, $T_1 = D$, but D^n is not defined for $n \ge 2$.

We can relate T_1, T_2, $\ldots$ with E by writing

$$E = 1 + \sum_{k=1}^{\infty} \frac{1}{k!} T_k, \tag{387d}$$

where the result is interpreted as meaning that

$$E(t) = 1(t) + \sum_{k=1}^{\infty} \frac{1}{k!} T_k(t),$$

for any $t \in T$.

Since E takes the exact solution to a differential equation through one unit step h, it is natural to ask how we would represent the solution at a general point θh advanced from the initial point. We write this as $E^{(\theta)}$, and we note that

$$E^{(\theta)}(t) = \theta^{r(t)} E(t),$$

for all $t \in T$. We can generalize (387d) in the form

$$E^{(\theta)} = 1 + \sum_{k=1}^{\infty} \frac{\theta^k}{k!} T_k,$$

and note that, for θ an integer n, we have

$$E^{(n)} = E^n.$$

This property is, to some extent, characteristic of E, and we have:

Theorem 387A *If $\alpha \in G_1$ such that $\alpha(\tau) = 1$, and m is an integer with $m \notin \{0, 1, -1\}$, then $\alpha^{(m)} = \alpha^m$ implies that $\alpha = E$.*

Proof. For any tree $t \neq \tau$, we have $\alpha^{(m)}(t) = r(t)^m \alpha(t) + Q_1$ and $\alpha^m(t) = m\alpha(t) + Q_2$, where Q_1 and Q_2 are expressions involving $\alpha(u)$ for $r(u) < r(t)$. Suppose that $\alpha(u)$ has been proved equal to $E(u)$ for all such trees. Then

$$\alpha^{(m)}(t) = r(t)^m \alpha(t) + Q_1,$$
$$\alpha^m(t) = m\alpha(t) + Q_2,$$
$$E^{(m)}(t) = r(t)^m E(t) + Q_1,$$
$$E^m(t) = mE(t) + Q_2,$$

so that $\alpha^{(m)}(t) = \alpha^m(t)$ implies that

$$(r(t)^m - m)(\alpha(t) - E(t)) = 0,$$

implying that $\alpha(t) = E(t)$, because $r(t)^m \neq m$ whenever $r(t) > 1$ and $m \notin \{0, 1, -1\}$. $\qquad\square$

Of the three excluded values of m in Theorem 387A, only $m = -1$ is interesting. Methods for which $\alpha^{(-1)} = \alpha^{-1}$ have a special property which makes them of potential value as the source of efficient extrapolation

procedures. Consider the solution of an initial value problem over an interval $[x_0, \overline{x}]$ using n steps of a Runge–Kutta method with stepsize $h = (\overline{x} - x_0)/n$. Suppose the computed solution can be expanded in an asymptotic series in h,

$$y(\overline{x}) + \sum_{i=1}^{\infty} C_i h^i. \tag{387e}$$

If the elementary weight function for the method is α, then the method corresponding to $(\alpha^{(-1)})^{-1}$ exactly undoes the work of the method but with h reversed. This means that the asymptotic error expansion for this reversed method would correspond to changing the sign of h in (387e). If $\alpha = (\alpha^{(-1)})^{-1}$, this would give exactly the same expansion, so that (387e) is an *even* function. It then becomes possible to extend the applicability of the method by extrapolation in even powers only.

388   Some subgroups and quotient groups

Let H_p denote the linear subspace of G defined by

$$H_p = \{\alpha \in G : \alpha(t) = 0, \text{ whenever } r(t) \le p\}.$$

If $\alpha, \beta \in G$ then $\alpha = \beta + H_p$ will mean that $\alpha - \beta$ is a member of H_p. The subspace is an ideal of G in the sense of the following result:

Theorem 388A *Let* $\alpha \in G_1$, $\beta \in G_1$, $\gamma \in G$ *and* $\delta \in G$ *be such that* $\alpha = \beta + H_p$ *and* $\gamma = \delta + H_p$. *Then* $\alpha\gamma = \beta\delta + H_p$.

Proof. Two members of G differ by a member of H_p if and only if they take identical values for any t such that $r(t) \le p$. For any such t, the formula for $(\alpha\gamma)(t)$ involves only values of $\alpha(u)$ and $\gamma(u)$ for $r(u) < r(t)$. Hence, $(\alpha\gamma)(t) = (\beta\delta)(t)$. $\qquad\square$

An alternative interpretation of H_p is to use instead $1 + H_p \in G_1$ as a subgroup of G_1. We have:

Theorem 388B *Let* $\alpha, \beta \in G_1$; *then*

$$\alpha = \beta + H_p \tag{388a}$$

if and only if

$$\alpha = \beta(1 + H_p). \tag{388b}$$

Proof. Both (388a) and (388b) are equivalent to the statement $\alpha(t) = \beta(t)$ for all t such that $r(t) \le p$. $\qquad\square$

Furthermore, we have:

Theorem 388C *The subgroup* $1 + H_p$ *is a normal subgroup of* G_1.

Proof. Theorem 388B is equally true if (388b) is replaced by $\alpha = (1 + H_p)\beta$. Hence, for any $\beta \in G_1$, $(1 + H_p)\beta = \beta(1 + H_p)$.   □

Quotient groups of the form $G_1/(1 + H_p)$ can be formed, and we consider their significance in the description of numerical methods. Suppose that m and $\overline{m}$ are Runge–Kutta methods with corresponding elementary weight functions α and $\overline{\alpha}$. If m and $\overline{m}$ are related by the requirement that for any smooth problem the results computed by these methods in a single step differ by $O(h^{p+1})$, then this means that $\alpha(t) = \overline{\alpha}(t)$, whenever $r(t) \leq p$. However, this is identical to the statement that

$$\overline{\alpha} \in (1 + H_p)\alpha,$$

which means that α and $\overline{\alpha}$ map canonically into the same member of the quotient group $G_1/(1 + H_p)$.

Because we also have the ideal H_p at our disposal, this interpretation of equivalent computations modulo $O(h^{p+1})$ can be extended to approximations represented by members of G, and not just of G_1.

The $C(\xi)$ and $D(\xi)$ conditions can also be represented using subgroups.

Definition 388D *A member* α *of* G_1 *is in* $C(\xi)$ *if, for any tree* t *such that* $r(t) \leq \xi$, $\alpha(t) = \gamma(t)^{-1}\alpha(\tau)^{r(t)}$ *and also*

$$\alpha([t\,t_1 t_2 \cdots t_m]) = \frac{1}{\gamma(t)}\alpha([\tau^{r(t)} t_1 t_2 \cdots t_m]), \tag{388c}$$

for any $t_1 t_2 \cdots t_m \in T$.

Theorem 388E *The set* $C(\xi)$ *is a normal subgroup of* G_1.

A proof of this result, and of Theorem 388G below, is given in Butcher (1972).

The $D(\xi)$ condition is also represented by a subset of G_1, which is also known to generate a normal subgroup.

Definition 388F *A member* α *of* G_1 *is a member of* $D(\xi)$ *if*

$$\alpha(tu) + \alpha(ut) = \alpha(t)\alpha(u), \tag{388d}$$

whenever $t, u \in T$ *and* $r(t) \leq \xi$.

Theorem 388G *The set* $D(\xi)$ *is a normal subgroup of* G_1.

The importance of these semi-groups is that E is a member of each of them and methods can be constructed which also lie in them. We first prove the following result:

Theorem 388H *For any real θ and positive integer ξ, $E^{(\theta)} \in C(\xi)$ and $E^{(\theta)} \in D(\xi)$.*

Proof. To show that $E^{(\theta)} \in C(\xi)$, we note that $E^{(\theta)}(t) = \gamma(t)^{-1}\theta^{r(t)}$ and that if $E^{(\theta)}$ is substituted for α in (388c), then both sides are equal to

$$\frac{\theta^{r(t)+r(t_1)+\cdots+r(t_m)+1}}{(r(t)+r(t_1)+\cdots+r(t_m)+1)\gamma(t)\gamma(t_1)\cdots\gamma(t_m)}.$$

To prove that $E^{(\theta)} \in D(\xi)$, substitute E into (388d). We find

$$\frac{r(t)}{(r(t)+r(u))\gamma(t)\gamma(u)} + \frac{r(u)}{(r(t)+r(u))\gamma(t)\gamma(u)} = \frac{1}{\gamma(t)} \cdot \frac{1}{\gamma(u)}. \qquad \square$$

389   An algebraic interpretation of effective order

The concept of conjugacy in group theory provides an algebraic interpretation of effective order. Two members of a group, x and z, are conjugate if there exists a member y of the group such that $yxy^{-1} = z$. We consider the group $G_1/(1+H_p)$ whose members are cosets of G_1 corresponding to sets of Runge–Kutta methods, which give identical numerical results in a single step to within $O(h^{p+1})$. In particular, $E(1+H_p)$ is the coset corresponding to methods which reproduce the exact solution to within $O(h^{p+1})$. This means that a method, with corresponding group element α, is of order p if

$$\alpha \in E(1 + H_p).$$

If a second method with corresponding group element β exists so that the conjugacy relation

$$\beta\alpha\beta^{-1} \in E(1 + H_p) \tag{389a}$$

holds, then the method corresponding to α has effective order p and the method corresponding to β has the role of perturbing method.

We use this interpretation to find conditions for effective orders up to 5. To simplify the calculation, we use a minor result:

Lemma 389A *A Runge–Kutta method with corresponding group element α has effective order p if and only if (389a) holds, where β is such that $\beta(\tau) = 0$.*

Proof. Suppose that (389a) holds with β replaced by $\widehat{\beta}$. Let $\beta = E^{(-\widehat{\beta}(\tau))}\widehat{\beta}$, so that $\beta(\tau) = 0$. We then find

$$\begin{aligned}
\beta\alpha\beta^{-1} &= E^{-\widehat{\beta}(\tau)}\widehat{\beta}\alpha\left(E^{-\widehat{\beta}(\tau)}\widehat{\beta}\right)^{-1} \\
&= E^{-\widehat{\beta}(\tau)}\widehat{\beta}\alpha\widehat{\beta}^{-1}E^{\widehat{\beta}(\tau)} \\
&\in E^{-\widehat{\beta}(\tau)}EE^{\widehat{\beta}(\tau)}(1+H_p) \\
&= E(1+H_p). \qquad \square
\end{aligned}$$

Once we have found effective order conditions on α and found a corresponding choice of β for α satisfying these conditions, we can use Lemma 389A in reverse to construct a family of possible perturbing methods.

To obtain the conditions we need on α we have constructed Table 389(I) based on Table 386(II). In this table, the trees up to order 5 are numbered, just as in the earlier table, and $\beta\alpha\beta^{-1} \in E(1+H_p)$ is replaced by $\beta\alpha \in E\beta(1+H_p)$, for convenience. In the order conditions formed from Table 389(I), we regard β_2, β_3, ... as free parameters. Simplifications are achieved by substituting values of α_1, α_2, ..., as they are found, into later equations that make use of them. The order conditions are

$$\alpha_1 = 1,$$
$$\alpha_2 = \tfrac{1}{2},$$
$$\alpha_3 = 2\beta_2 + \tfrac{1}{3},$$
$$\alpha_4 = \tfrac{1}{6},$$
$$\alpha_5 = 3\beta_2 + 3\beta_3 + \tfrac{1}{4},$$
$$\alpha_6 = \beta_2 + \beta_3 + \beta_4 + \tfrac{1}{8},$$
$$\alpha_7 = \beta_2 - \beta_3 + 2\beta_4 + \tfrac{1}{12},$$
$$\alpha_8 = \tfrac{1}{24},$$
$$\alpha_9 = 4\beta_2 + 6\beta_3 + 4\beta_5 + \tfrac{1}{5},$$
$$\alpha_{10} = \tfrac{5}{3}\beta_2 + \tfrac{5}{2}\beta_3 + \beta_4 + \beta_5 + 2\beta_6 + \tfrac{1}{10},$$
$$\alpha_{11} = \tfrac{4}{3}\beta_2 + \tfrac{1}{2}\beta_3 + 2\beta_4 + 2\beta_6 + \beta_7 + \tfrac{1}{15},$$
$$\alpha_{12} = \tfrac{1}{3}\beta_2 - 2\beta_2^2 + \tfrac{1}{2}\beta_3 + \tfrac{1}{2}\beta_4 + \beta_6 + \beta_8 + \tfrac{1}{30},$$
$$\alpha_{13} = \tfrac{2}{3}\beta_2 - \beta_2^2 + \beta_3 + \beta_4 + 2\beta_6 + \tfrac{1}{20},$$
$$\alpha_{14} = \beta_2 + 3\beta_4 - \beta_5 + 3\beta_7 + \tfrac{1}{20},$$
$$\alpha_{15} = \tfrac{1}{3}\beta_2 + \tfrac{3}{2}\beta_4 - \beta_6 + \beta_7 + \beta_8 + \tfrac{1}{40},$$
$$\alpha_{16} = \tfrac{1}{3}\beta_2 - \tfrac{1}{2}\beta_3 + \beta_4 - \beta_7 + 2\beta_8 + \tfrac{1}{60},$$
$$\alpha_{17} = \tfrac{1}{120}.$$

For explicit Runge–Kutta methods with fourth (effective) order, four stages are still necessary, but there is much more freedom than for methods with the same classical order. For fifth effective order there is a real saving in that only five stages are necessary. For the fourth order case, we need to choose the coefficients of the method so that

$$\alpha_1 = 1,$$
$$\alpha_2 = \tfrac{1}{2},$$
$$\alpha_4 = \tfrac{1}{6},$$
$$\alpha_8 = \tfrac{1}{24},$$

Table 389(I)   Effective order conditions

i	$r(t_i)$	$(\beta\alpha)(t_i)$	$(E\beta)(t_i)$
1	1	α_1	1
2	2	$\alpha_2 + \beta_2$	$\beta_2 + \frac{1}{2}$
3	3	$\alpha_3 + \beta_3$	$\beta_3 + 2\beta_2 + \frac{1}{3}$
4	3	$\alpha_4 + \beta_2\alpha_1 + \beta_4$	$\beta_4 + \beta_2 + \frac{1}{6}$
5	4	$\alpha_5 + \beta_5$	$\beta_5 + 3\beta_3 + 3\beta_2 + \frac{1}{4}$
6	4	$\alpha_6 + \beta_2\alpha_2 + \beta_6$	$\beta_6 + \beta_4 + \beta_3 + \frac{3}{2}\beta_2 + \frac{1}{8}$
7	4	$\alpha_7 + \beta_3\alpha_1 + \beta_7$	$\beta_7 + 2\beta_4 + \beta_2 + \frac{1}{12}$
8	4	$\alpha_8 + \beta_2\alpha_2 + \beta_4\alpha_1 + \beta_8$	$\beta_8 + \beta_4 + \frac{1}{2}\beta_2 + \frac{1}{24}$
9	5	$\alpha_9 + \beta_9$	$\beta_9 + 4\beta_5 + 6\beta_3 + 4\beta_2 + \frac{1}{5}$
10	5	$\alpha_{10} + \beta_2\alpha_3 + \beta_{10}$	$\beta_{10}+2\beta_6+\beta_5+\beta_4+\frac{5}{2}\beta_3+2\beta_2+\frac{1}{10}$
11	5	$\alpha_{11} + \beta_3\alpha_2 + \beta_{11}$	$\beta_{11}+\beta_7+2\beta_6+2\beta_4+\beta_3+\frac{4}{3}\beta_2+\frac{1}{15}$
12	5	$\alpha_{12} + \beta_2\alpha_3 + \beta_4\alpha_2 + \beta_{12}$	$\beta_{12}+\beta_8+\beta_6+\beta_4+\frac{1}{2}\beta_3+\frac{2}{3}\beta_2+\frac{1}{30}$
13	5	$\alpha_{13} + 2\beta_2\alpha_4 + \beta_2^2\alpha_1 + \beta_{13}$	$\beta_{13} + 2\beta_6 + \beta_4 + \beta_3 + \beta_2 + \frac{1}{20}$
14	5	$\alpha_{14} + \beta_5\alpha_1 + \beta_{14}$	$\beta_{14} + 3\beta_7 + 3\beta_4 + \beta_2 + \frac{1}{20}$
15	5	$\alpha_{15} + \beta_2\alpha_4 + +\beta_6\alpha_1 + \beta_{15}$	$\beta_{15} + \beta_8 + \beta_7 + \frac{3}{2}\beta_4 + \frac{1}{2}\beta_2 + \frac{1}{40}$
16	5	$\alpha_{16} + \beta_3\alpha_2 + \beta_7\alpha_1 + \beta_{16}$	$\beta_{16} + 2\beta_8 + \beta_4 + \frac{1}{3}\beta_2 + \frac{1}{60}$
17	5	$\alpha_{17}+\beta_2\alpha_4+\beta_4\alpha_2+\beta_8\alpha_1+\beta_{17}$	$\beta_{17} + \beta_8 + \frac{1}{2}\beta_4 + \frac{1}{6}\beta_2 + \frac{1}{120}$

and so that the equation formed by eliminating the various β values from the equations for α_3, α_5, α_6 an α_7 is satisfied. This final effective order condition is

$$\alpha_3 - \alpha_5 + 2\alpha_6 - \alpha_7 = \tfrac{1}{4},$$

and the five condition equations written in terms of the coefficients in a four-stage method are

$$b_1 + b_2 + b_3 + b_4 = 1,$$
$$b_2c_2 + b_3c_3 + b_4c_4 = \tfrac{1}{2},$$
$$b_3a_{32}c_2 + b_4a_{42}c_2 + b_4a_{43}c_3 = \tfrac{1}{6},$$
$$b_4a_{43}a_{32}c_2 = \tfrac{1}{24},$$

$$b_2c_2^2(1-c_2) + b_3c_3^2(1-c_3) + b_4c_4^2(1-c_4)$$
$$+ b_3a_{32}c_2(2c_3 - c_2) + b_4a_{42}c_2(2c_4 - c_2) + b_4a_{43}c_3(2c_4 - c_3) = \tfrac{1}{4}.$$

Table 389(II)   Group elements associated with a special effective order 4 method

t	$E(t)$	$\alpha(t)$	$\beta(t)$	$(\beta^{-1}E)(t)$	$(\beta^{-1}E\beta^{(r)})(t)$
(tree)	1	1	0	1	1
(tree)	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$
(tree)	$\frac{1}{3}$	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{1}{3}$
(tree)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{72}$	$\frac{11}{72}$	$\frac{11+r^3}{72}$
(tree)	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{108}$	$\frac{13}{54}$	$\frac{26+r^4}{108}$
(tree)	$\frac{1}{8}$	$\frac{5}{36}$	$\frac{1}{216}$	$\frac{13}{108}$	$\frac{26+3r^3+r^4}{216}$
(tree)	$\frac{1}{12}$	$\frac{1}{9}$	$-\frac{1}{216}$	$\frac{19}{216}$	$\frac{19+6r^3-r^4}{216}$
(tree)	$\frac{1}{24}$	$\frac{1}{24}$	0	$\frac{1}{36}$	$\frac{2+r^3}{72}$

We do not attempt to find a general solution to these equations, but instead explore a mild deviation from full classical order. In fact, we assume that the perturbing method has $\beta_2 = \beta_3 = 0$, so that we now have the conditions

$$b_1 + b_2 + b_3 + b_4 = 1,$$
$$b_2c_2 + b_3c_3 + b_4c_4 = \tfrac{1}{2},$$
$$b_2c_2^2 + b_3c_3^2 + b_4c_4^2 = \tfrac{1}{3},$$
$$b_3a_{32}c_2 + b_4a_{42}c_2 + b_4a_{43}c_3 = \tfrac{1}{6},$$
$$b_2c_2^3 + b_3c_3^3 + b_4c_4^3 = \tfrac{1}{4},$$
$$b_3a_{32}c_2(2c_3 - c_2) + b_4a_{42}c_2(2c_4 - c_2) + b_4a_{43}c_3(2c_4 - c_3) = \tfrac{1}{4},$$
$$b_4a_{43}a_{32}c_2 = \tfrac{1}{24}.$$

Methods satisfying these more general conditions do not need to have $c_4 = 1$ and we can find, for example, the tableau

$$
\begin{array}{c|cccc}
0 \\
\frac{1}{3} & \frac{1}{3} \\
\frac{2}{3} & \frac{1}{6} & \frac{1}{2} \\
\frac{5}{6} & \frac{5}{24} & 0 & \frac{5}{8} \\
\hline
 & \frac{1}{10} & \frac{1}{2} & 0 & \frac{2}{5}
\end{array}
\tag{389b}
$$

A suitable starting method, which does not advance the solution forward but introduces the correct perturbation so that (389b) faithfully reproduces this perturbation to within order 4, is given by the tableau

$$
\begin{array}{c|cccc}
0 \\
1 & 1 \\
\frac{2}{3} & \frac{2}{3} & 0 \\
\frac{1}{3} & 0 & -\frac{1}{3} & \frac{2}{3} \\
\hline
& -\frac{1}{24} & \frac{1}{24} & -\frac{1}{8} & \frac{1}{8}
\end{array} \,.
\tag{389c}
$$

The freedom that lay at our disposal in selecting this starting procedure was used to guarantee a certain simplicity in the choice of finishing procedure. This was in fact decided on first, and has a tableau identical with (389b) except for the b^T vector. The reason for this choice is that no extra work is required to obtain an output value because the stages in the final step will already have been completed. The tableau for this final step is

$$
\begin{array}{c|cccc}
0 \\
\frac{1}{3} & \frac{1}{3} \\
\frac{2}{3} & \frac{1}{6} & \frac{1}{2} \\
\frac{5}{6} & \frac{5}{24} & 0 & \frac{5}{8} \\
\hline
& \frac{3}{20} & \frac{1}{3} & \frac{1}{4} & \frac{4}{15}
\end{array} \,.
\tag{389d}
$$

This example method has not been optimized in any way, and is therefore not proposed for a practical computation. On the other hand, it shows that the search for efficient methods need not be restricted to the class of Runge–Kutta methods satisfying classical order conditions. It might be argued that methods with only effective order cannot be used in practice because stepsize change is not possible without carrying out a finishing step followed by a new start with the modified stepsize. However, if, after carrying out a step with the method introduced here, a stepsize change from h to rh is required, then this can be done by simply adding one additional stage and choosing the vector b^T which depends on r. The tableau for this h-adjusting step is

$$
\begin{array}{c|ccccc}
0 \\
\frac{1}{3} & \frac{1}{3} \\
\frac{2}{3} & \frac{1}{6} & \frac{1}{2} \\
\frac{5}{6} & \frac{5}{24} & 0 & \frac{5}{8} \\
\frac{1}{2} & \frac{13}{40} & \frac{1}{6} & \frac{1}{24} & -\frac{1}{30} \\
\hline
& \frac{3+r^3-2r^4}{20} & \frac{2-3r^3+4r^4}{6} & \frac{1-3r^3+2r^4}{4} & \frac{4+3r^3-r^4}{15} & r^3 - r^4
\end{array} \,.
\tag{389e}
$$

Rather than carry out detailed derivations of the various tableaux we have introduced, we present in Table 389(II) the values of the group elements in $G_1/(1 + H_4)$ that arise in the computations. These group elements are β, corresponding to the starting method (389c), α for the main method (389b),

$\beta^{-1}E$ corresponding to the finishing method (389d) and, finally, $\beta^{-1}E\beta^{(r)}$ for the stepsize-adjusting method (389e). For convenience in checking the computations, E is also provided.

Exercises 38

38.1 Find the B-series for the Euler method

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}.$$

38.2 Find the B-series for the implicit Euler method

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}.$$

38.3 Show that the two Runge–Kutta methods

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ \hline & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{array} \quad \text{and} \quad \begin{array}{c|ccc} 0 & -1 & 0 & 1 \\ 1 & \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & 2 & 0 & -2 \\ \hline & -\frac{3}{2} & \frac{1}{2} & 1 \end{array}$$

are P-equivalent. Find a method with only two stages equivalent to each of them.

38.4 Let m_1 and m_2 denote the Runge–Kutta methods

$$m_1 = \begin{array}{c|cc} \frac{1}{2} - \frac{1}{6}\sqrt{3} & \frac{1}{4} & \frac{1}{4} - \frac{1}{6}\sqrt{3} \\ \frac{1}{2} + \frac{1}{6}\sqrt{3} & \frac{1}{4} + \frac{1}{6}\sqrt{3} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array},$$

$$m_2 = \begin{array}{c|cc} -\frac{1}{2} - \frac{1}{6}\sqrt{3} & -\frac{1}{4} & -\frac{1}{4} - \frac{1}{6}\sqrt{3} \\ -\frac{1}{2} + \frac{1}{6}\sqrt{3} & -\frac{1}{4} + \frac{1}{6}\sqrt{3} & -\frac{1}{4} \\ \hline & -\frac{1}{2} & -\frac{1}{2} \end{array}.$$

Show that $[m_2] = [m_1]^{-1}$.

38.5 Show that $D \in X$ is the homomorphic partner of $[m]$, where

$$m = \begin{array}{c|c} 0 & 0 \\ \hline 0 & 1 \end{array}.$$

39 Implementation Issues

390 *Introduction*

In this section we consider several issues arising in the design and construction of practical algorithms for the solution of initial value problems based on Runge–Kutta methods.

An automatic code needs to be able to choose an initial stepsize and then adjust the stepsize from step to step as the integration progresses. Along with the need to choose appropriate stepsizes to obtain an acceptable accuracy in a given step, there is a corresponding need to reject some steps, because they will evidently contribute too large an error to the overall inaccuracy of the final result. The user of the software needs to have some way of indicating a preference between cheap, but low accuracy, results on the one hand and expensive, but accurate, results on the other. This is usually done by supplying a 'tolerance' as a parameter. We show that this tolerance can be interpreted as a Lagrange multiplier T. If E is a measure of the total error to plan for, and W is a measure of the work that is to be allocated to achieve this accuracy, then we might try as best we can to minimize $E + TW$. This will mean that a high value of T will correspond to an emphasis on reducing computing costs, and a low value of T will correspond to an emphasis on accuracy. It is possible to achieve something like an optimal value of this weighted objective function by requiring the local truncation error to be maintained as constant from step to step. However, there are other views as to how the allocation of resources should be appropriately allocated, and we discuss these in Subsection 393.

If the local truncation error committed in a step is to be the main determining criterion for the choice of stepsize, then we need a means of estimating the local error. This will lead to a control system for the stepsize, and we need to look at the dynamics of this system to ensure that good behaviour is achieved.

It is very difficult to find suitable criteria for adjusting order amongst a range of alternative Runge–Kutta methods. Generally, software designers are happy to construct fixed order codes. However, it is possible to obtain useful variable order algorithms if the stage order is sufficiently high. This applies especially to implicit methods, intended for stiff problems, and we devote at least some attention to this question.

For stiff problems, the solution of the algebraic equations inherent to the implementation of implicit methods is a major issue. The efficiency of a stiff solver will often depend on the management of the linear algebra, associated with a Newton type of solution, more than on any other aspect of the calculation.

391 *Optimal sequences*

Consider an integration over an interval $[a, b]$. We can interpret a as the point x_0 at which initial information $y(x_0) = y_0$ is given and b as a final point, which

we have generally written as $\overline{x}$ where we are attempting to approximate $y(\overline{x})$. As steps of a Runge–Kutta method are carried out we need to choose h for a new step starting at a point $x \in [a, b]$, assuming previous steps have taken the solution forward to this point. From information gleaned from details of the computation, it will be possible to obtain some sort of guide as to what the truncation error is likely to do in a step from x to $x+h$ and, assuming that the method has order p, the norm of this truncation error will be approximately like $C(x)h^{p+1}$, where C is some positively valued function. Write the choice of h for this step as $H(x)$. Assuming that all stepsizes are sufficiently small, we can write the overall error approximately as an integral

$$E(H) = \int_a^b C(x)H(x)^p dx.$$

The total work carried out will be taken to be the simply the number of steps. For classical Runge–Kutta methods the cost of carrying out each step will be approximately the same from step to step. However, the number of steps is approximately equal to the integral

$$W(H) = \int_a^b H(x)^{-1} dx.$$

To obtain an optimal rule for defining values of $H(x)$, as x varies, we have to ensure that it is not possible, by altering H, to obtain, at the same time, lower values of both $E(H)$ and $W(H)$. This means that the optimal choice is the same as would be obtained by minimizing $E(H)$, for a specified upper bound on $W(H)$, or, dually, minimizing $W(H)$, subject to an upper bound on $E(H)$. Thus we need to optimize the value of $E(H) + TW(H)$ for some positive value of the Lagrange multiplier T.

From calculus of variation arguments, the optimal is achieved by setting to zero the expression $(d/dH)(E(H) + TW(H))$. Assuming that $W(H)$ has the constant value p, chosen for convenience, this means that

$$pC(x)H(x)^{p-1} = pTH(x)^{-2},$$

for all x. Hence, $C(x)H(x)^{p+1}$ should be kept equal to the constant value T. In other words, optimality is achieved by keeping the magnitude of the local truncation error close to constant from step to step. In practice, the truncation error associated with a step about to be carried out is not known. However, an estimation of the error in the last completed step is usually available, using techniques such as those described in Section 33, and this can be taken as a usable guide. On the other hand, if a previous attempt to carry out this step has been rejected, because the truncation error was regarded as excessive, then this gives information about the correct value of h to use in a second attempt.

For robustness, a stepsize controller has to respond as smoothly as possible to (real or apparent) abrupt changes in behaviour. This means that the stepsize should not decrease or increase from one step to the next by an excessive ratio. Also, if the user-specified tolerance, given as a bound on the norm of the local truncation error estimate, is ever exceeded, recomputation and loss of performance will result. Hence, to guard against this as much as possible, a 'safety factor' is usually introduced into the computation. If h is the estimated stepsize to give a predicted truncation error equal to the tolerance, then some smaller value, such as $0.9h$, is typically used instead. Combining all these ideas, we can give a formula for arriving at a factor r, to give a new stepsize rh, following a step for which the error estimate is est. The tolerance is written as tol, and it is assumed that this previous step has been accepted. The ratio r is given by

$$r = \max\left(0.5, \min\left(2.0, 0.9\left(\frac{\text{tol}}{\text{est}}\right)^{1/(p+1)}\right)\right). \tag{391a}$$

The three constants, given here with values 0.5, 2.0 and 0.9, are all somewhat arbitrary and have to be regarded as design parameters.

392   *Acceptance and rejection of steps*

It is customary to test the error estimate in a step against T and to accept the step only when the estimated error is smaller. To reduce the danger of rejecting too many steps, the safety factor in (391a) is inserted. Thus there would have to be a very large increase in the rate of error production for a step to be rejected. We now consider a different way of looking at the question of acceptance and rejection of steps. This is based on removing the safety factor but allowing for the possible acceptance of a step as long as the ratio of the error to the tolerance is not too great. We need to decide what 'too great' should mean.

The criterion will be based on attempting to minimize the rate of error production plus T times the rate of doing work. Because we are considering the rejection of a completed step with size h, we need to add the work already carried out to the computational costs in some way. Suppose that the error estimated for the step is $r^{-(p+1)}T$, and that we are proposing to change the stepsize to rh. This will mean that, until some other change is made, the rate of growth of error $+ T \times$ work will be $T(1 + p)/rh$. By the time the original interval of size h has been traversed, the total expenditure will be $T(1+p)/rh$. Add the contribution from the work in the rejected step and the total expenditure will be $T((p + 1)/r + p)$.

If, instead, the step had been accepted, the expenditure (linear combination of error and work) would be $T(r^{-(p+1)} + p)$. Comparing the two results, we

Table 392(I)   Minimal value of stepsize ratio and maximal value of error$/T$ for
step acceptance

p	$(p+1)^{-1/p}$	$(p+1)^{(p+1)/p}$
1	0.500	4.00
2	0.577	5.20
3	0.630	6.35
4	0.669	7.48
5	0.700	8.59
6	0.723	9.68
7	0.743	10.77
8	0.760	11.84
9	0.774	12.92
10	0.787	13.98

conclude that the step should be accepted if $r^{-(p+1)} \leq (p+1)/r$, that is, when

$$r \geq (p+1)^{-1/p},$$

and rejected otherwise. Looked at another way, the step should be accepted
if the error estimated in a step, divided by the tolerance, does not exceed
$(p+1)^{(p+1)/p}$. Values of $(p+1)^{-1/p}$ and $(p+1)^{(p+1)/p}$ are given in Table
392(I).

393   *Error per step versus error per unit step*

The criterion we have described for stepsize selection is based on the principle
of 'error per step'. That is, a code designed on this basis attempts to
maintain the error committed in each step as close to constant as possible. An
alternative point of view is to use 'error per unit step', in which error *divided
by stepsize* is maintained approximately constant. This idea is attractive from
many points of view. In particular, it keeps the rate of error production under
control and is very natural to use. In an application, the user has to choose a
tolerance which indicates how rapidly he or she is happy to accept errors to
grow as the solution approximation evolves with time.

Furthermore, there is a reasonable expectation that, if a problem is
attempted with a range of tolerances, the total truncation error will vary
in more or less the same ratio as the tolerances. This state of affairs is known
as 'proportionality', and is widely regarded as being desirable. On the other
hand, if the error per step criterion is used we should hope only for the global
errors to vary in proportion to $\text{tol}^{p/(p+1)}$. The present author does not regard

this as being in any way inferior to simple proportionality. The fact that error per step is close to producing optimal stepsize sequences, in the sense we have described, seems to be a reason for considering, and even preferring, this choice in practical codes.

From the user point of view, the interpretation of the tolerance as a Lagrange multiplier is not such a difficult idea, especially if tol is viewed not so much as 'error per step' as 'rate of error production per unit of work'. This interpretation also carries over for algorithms for which p is still constant, but the work might vary, for some reason, from one step to the next.

394   Control-theoretic considerations

Controlling the stepsize, using a ratio of h in one step to h in the previous step, based on (391a), can often lead to undesirable behaviour. This can come about because of over-corrections. An error estimate in one step may be accidentally low and this can lead to a greater increase in stepsize than is justified by the estimate found in the following step. The consequent rejection of this second step, and its re-evaluation with a reduced stepsize, can be the start of a series of similarly disruptive and wasteful increases and decreases.

In an attempt to understand this phenomenon and to guard against its damaging effects, an analysis of stepsize management using the principles of control theory was instituted by Gustafsson, Lundh and Söderlind (1988). The basic idea that has come out of these analyses is that PI control should be used in preference to I control. Although these concepts are related to continuous control models, they have a discrete interpretation. Under the discrete analogue, I control corresponds to basing each new stepsize on the most recently available error estimate, whereas PI control would make use of the estimates found in the two most recently completed steps.

If we were to base a new stepsize on a simplified alternative to (391a), using the ratio $r = (\text{est}/\text{tol})^{1/(p+1)}$, this would correspond to what is known in control theory as 'dead-beat' control. On the other hand, using the ratio $r = (\text{tol}/\text{est})^{\alpha/(p+1)}$, where $0 < \alpha < 1$, would correspond to a damped version of this control system. This controller would not respond as rapidly to varying accuracy requirements, but would be less likely to change too quickly for future behaviour to deal with. Going further, and adopting PI control, would give a stepsize ratio equal to

$$r_n = \left(\frac{\text{tol}}{\text{est}_{n-1}} \right)^{\alpha/(p+1)} \left(\frac{\text{tol}}{\text{est}_{n-2}} \right)^{\beta/(p+1)}. \tag{394a}$$

In this equation, r_n is the stepsize ratio for determining the stepsize h_n to be used in step n. That is, if h_{n-1} is the stepsize in step $n-1$, then $h_n = r_n h_{n-1}$. The quantities est_{n-1} and est_{n-2}, denote the error estimates found in steps $n-1$ and $n-2$, respectively.

For convenience, we work additively, rather than multiplicatively, by dealing with $\log(h_n)$ and $\log(r_n)$ rather than with h_n and r_n themselves. Let ξ_{n-1} denote the logarithm of the stepsize that would be adopted in step n, if dead-beat control were to be used. That is,

$$\xi_{n-1} = \log(h_{n-1}) + \frac{1}{p+1}(\log(\text{tol}) - \log(\text{est}_{n-1})).$$

Now let η_n denote the logarithm of the stepsize actually adopted in step n. Thus we can write dead-beat control as

$$\eta_n = \xi_{n-1}$$

and the modification with damping factor α as

$$\eta_n = (1 - \alpha)\eta_{n-1} + \alpha\xi_{n-1}.$$

For the PI controller (394a), we have

$$\eta_n = (1 - \alpha)\eta_{n-1} - \beta\eta_{n-2} + \alpha\xi_{n-1} + \beta\xi_{n-2}. \tag{394b}$$

Appropriate choices for the parameters α and β have been discussed by the original authors. Crucial considerations are the stable behaviour of the homogeneous part of the difference equation (394b) and the ability of the control system to respond sympathetically, but not too sensitively, to changing circumstances. For example, $\alpha = 0.7$ and $\beta = -0.4$, as proposed by Gustafsson (1991), works well. Recently, further work has been done on control-theoretic approaches to stepsize control by Söderlind (2002).

395 Solving the implicit equations

For stiff problems, the methods of choice are implicit. We discuss some aspects of the technical problem of evaluating the stages of an implicit Runge–Kutta method. For a one-stage method, the evaluation technique is also similar for backward difference methods and for Runge–Kutta and general linear methods that have a lower triangular coefficient matrix.

For these simple methods, the algebraic question takes the form

$$Y - h\gamma f(X, Y) = U, \tag{395a}$$

where X and U are known. Let $J(X, Y)$ denote the Jacobian matrix with elements given by

$$J(X, Y)_{ij} = \frac{\partial f_i}{\partial y_j}(X, Y), \qquad i, j = 1, 2, \ldots, N.$$

A full Newton scheme would start with the use of a predictor to obtain a first approximation to Y. Denote this by $Y^{[0]}$ and update it with a sequence of approximations $Y^{[i]}$, $i = 1, 2, \ldots$, given by

$$Y^{[i]} = Y^{[i-1]} - \Delta,$$

where

$$(I - h\gamma J(X, Y^{[i-1]}))\Delta = Y^{[i-1]} - h\gamma f(X, Y^{[i-1]}) - U. \qquad (395b)$$

Although the full scheme has the advantage of quadratic convergence, it is usually not adopted in practice. The reason is the excessive cost of evaluating the Jacobian J and of carrying out the LU factorization of the matrix $I - h\gamma J$. The Newton scheme can be modified in various ways to reduce this cost. First, the re-evaluation of J after each iteration can be dispensed with. Instead the scheme (395b) can be replaced by

$$(I - h\gamma J(X, Y^{[0]}))\Delta = Y^{[i-1]} - h\gamma f(X, Y^{[i-1]}) - U,$$

and for many problems this is almost as effective as the full Newton method. Even if more iterations are required, the additional cost is often less than the saving in J evaluations and LU factorizations.

Secondly, in the case of diagonally implicit methods, it is usually possible to evaluate J only once per step, for example at the start of the first stage. Assuming the Jacobian is sufficiently slowly varying, this can be almost as effective as evaluating the Jacobian once for each stage.

The third, and most extreme, of the Jacobian update schemes is the use of the same approximation over not just one step but over many steps. A typical algorithm signals the need to re-evaluate J only when the rate of convergence is sufficiently slow as to justify this expenditure of resources to achieve an overall improvement. When J is maintained at a constant value over many steps, we have to ask the further question about when $I - h\gamma J$ should be refactorized. Assuming that γ is unchanged, any change in h will affect the convergence by using a factorization of this matrix which is based not only on an incorrect value of J, but on what may be a vastly different value of h.

It may be possible to delay the refactorization process by introducing a 'relaxation factor' into the iteration scheme. That is, when Δ has been computed in a generalized form of (395b), the update takes the form

$$Y^{[i]} = Y^{[i-1]} - \theta\Delta,$$

where θ is a suitably chosen scalar factor. To analyse how this works, suppose for simplicity that J is constant but that h has changed from $\overline{h}$ at the time the factorization took place to $r\overline{h}$ at the time a generalized Newton step is being carried out. As a further simplification, assume that $f(x, y) = Jy + V$ and that we are exploring the behaviour in a direction along along an eigenvector corresponding to an eigenvalue λ. Write $z = \overline{h}\gamma\lambda$. Under these assumptions the iteration scheme effectively seeks a solution to an equation of the form

$$\eta - rz\eta = a,$$

with solution $\eta = \eta^* = a/(1 - r)$, using an iteration scheme which replaces $\eta^* + \epsilon$ by $\eta^* + \phi(z)\epsilon$, where

$$\phi(z) = 1 - \theta \frac{1 - rz}{1 - z}.$$

Convergence will depend on the magnitude of $\phi(z)$ for all z that are likely to arise. Values of z near zero correspond to non-stiff components of the problem, and values of z with large magnitude in the left half-plane correspond to stiff components. Hence, it seems desirable to choose θ to minimize $|\phi(z)|$ for z in the left half-plane. The value that achieves this is

$$\theta = \frac{2}{1 + r}.$$

For fully implicit Runge–Kutta methods, the problem of evaluating the stages becomes much more complicated and potentially more costly. For a method with coefficient matrix A, we need to consider all stages at the same time. Let Y denote the sN-dimensional vector made up from Y_1, Y_2, ..., Y_s. Furthermore the approximation sequence will be written as $Y^{[j]}$, $j = 0, 1, \ldots$, each also made up from s subvectors, and Δ will denote a vector in $\mathbb{R}^{sN}$ made up from the subtrahends in each of the s components in iteration i. Thus

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix}, \quad Y^{[i]} = \begin{bmatrix} Y_1^{[i]} \\ Y_2^{[i]} \\ \vdots \\ Y_s^{[i]} \end{bmatrix}, \quad \Delta = \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_s \end{bmatrix} = \begin{bmatrix} Y_1^{[i-1]} - Y_1^{[i]} \\ Y_2^{[i-1]} - Y_2^{[i]} \\ \vdots \\ Y_s^{[i-1]} - Y_s^{[i]} \end{bmatrix}.$$

In place of (395a), the algebraic equations to solve in a step take the form

$$Y - hA \otimes f(X, Y) = U \in \mathbb{R}^{sN}. \tag{395c}$$

Note that $f(X, Y)$ denotes a vector in $\mathbb{R}^{sN}$ made up from subvectors of the form $f(X_j, Y_j)$, $j = 1, 2, \ldots, s$. The iteration scheme consists of solving the equations

$$\Delta_j - h \sum_{k=1}^{s} a_{jk} J(X_k, Y_k^{[i]}) \Delta_k = Y_j - h \sum_{k=1}^{s} a_{jk} f(X_k, Y_k^{[i]}) - U_i,$$

and then carrying out the update $Y_j^{[i]} = Y_j^{[i-1]} - \Delta_j$, $j = 1, 2, \ldots, s$. If it is assumed that Jacobians are evaluated only once per step, or even less frequently, then we can write (395c) in the simplified form

$$(I_s \otimes I_N - hA \otimes J)\Delta = Y^{[i-1]} - hA \otimes F^{[i-1]} - U, \tag{395d}$$

where $F^{[i-1]}$ is the vector with kth subvector equal to $f\left(X_k, Y_k^{[i-1]}\right)$. Here J is a single approximation to the $n \times n$ Jacobian matrix. One of the advantages of using a single J approximation is the fact that it is possible to operate, for example, with similarity transformations, on the coefficient matrix A and J independently.

If no such transformation is carried out, the computational costs can become very severe. The LU factorization of the matrix on the left-hand side of (395d) requires a number of operations proportional to $s^3 N^3$, compared with just N^3 if $s = 1$. However, if $A = T^{-1}\widehat{A}T$, where $\widehat{A}$ has a structure close to diagonal, then the cost reduces to something like sN^3.

Exercises 39

39.1 An implicit Runge–Kutta method is to be implemented for the solution of non-stiff problems using functional iteration to solve the nonlinear equations. How should the stepsize be selected?

39.2 A Runge–Kutta method of order p is used over an interval of length X. Suppose that for a subinterval of length $(1 - \theta)X$ the error in a step of length h is Ch^{p+1}, and for the remaining distance θX the error is αCh^5. Assume that a large number N of steps are performed, of which $(1-\phi)N$ are in the first subinterval and ϕN are in the second subinterval. Determine the value of ϕ which will minimize the total error committed in the integration.

39.3 Compare the result found in Exercise 39.2 with the result that would be obtained from an 'error per unit step' argument.

Chapter 4

Linear Multistep Methods

40 Preliminaries

400 Fundamentals

This chapter, devoted entirely to the analysis of linear multistep methods, follows on from the introduction to these methods presented in Section 24. We use the notation and ideas introduced there, but attempt to fill in missing details. In particular, we show in the present section how the concepts of consistency, stability and convergence are interrelated and give more of a theoretical justification for the concept of 'order'. This analysis depends heavily on the use of difference equations, especially on the conditions for the solution of a linear difference equation to be bounded. For a difference equation,

$$y_n = \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + \cdots + \alpha_k y_{n-k}, \qquad (400a)$$

we recall that all solutions are bounded if and only if the polynomial

$$z^k - \alpha_1 z^{k-1} - \alpha_2 z^{k-2} - \cdots - \alpha_k$$

has all its zeros in the closed unit disc and all multiple zeros in the interior of this disc.

The direct applicability of this result to a linear multistep method $[\alpha, \beta]$, in which the approximate solution at x_n is computed by

$$
\begin{aligned}
y_n = {}& \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + \cdots + \alpha_k y_{n-k} \\
& + \beta_0 h f(x_n, y_n) + \beta_1 h f(x_{n-1}, y_{n-1}) + \cdots + \beta_k h f(x_{n-k}, y_{n-k}), \quad (400b)
\end{aligned}
$$

is clear. We wish to be able to solve a wide variety of initial value problems in a reliable manner, and amongst the problems for which we need good answers is certainly the simple problem for which $f(x, y) = 0$. In this case the solution approximations are related by (400a), and stable behaviour for this problem becomes essential. It is a remarkable fact that convergence hinges on this stability result alone, as well as on consistency requirements.

As in Section 24 we write the method as $[\alpha, \beta]$, where

$$\alpha(z) = 1 - \alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_k z^k,$$

$$\beta(z) = \beta_0 + \beta_1 z + \beta_2 z^2 + \cdots + \beta_k z^k,$$

or in the more traditional formulation as (ρ, σ), where

$$\rho(z) = z^k - \alpha_1 z^{k-1} - \alpha_2 z^{k-2} - \cdots - \alpha_k,$$

$$\sigma(z) = \beta_0 z^k + \beta_1 z^{k-1} + \beta_2 z^{k-2} + \cdots + \beta_k.$$

401   Starting methods

As we pointed out in Subsection 246, linear multistep methods require starting methods even to carry out a single step. We consider, in general terms, some of the procedures used to obtain starting values; we then discuss any unifying characteristics they might have.

One obvious approach to starting a k-step method is to carry out $k - 1$ steps with a Runge–Kutta method, preferably of the same order as the linear multistep method itself. An interesting variation of this standard procedure is to use specially constructed Runge–Kutta methods which make it possible to move forward several steps at a time (Gear, 1980).

A second approach, which fits naturally into the style of linear multistep methods, is to solve a system of equations representing the integrals of $y'(x)$ from x_0 to each of $x_1, x_2, \ldots, x_{k-1}$ written, in each case, as a quadrature formula with abscissae at these same points. We illustrate this in the case of the third order Adams–Bashforth method

$$y_n = y_{n-1} + \frac{h}{12}\big(23f(x_{n-1}, y_{n-1}) - 16f(x_{n-2}, y_{n-2}) + 5f(x_{n-3}, y_{n-3})\big),$$

for which appropriate quadrature formulae, adapted to a differential equation, are

$$y_1 = y_0 + \frac{h}{12}\big(5f(x_0, y_0) + 8f(x_1, y_1) - f(x_2, y_2)\big), \qquad \text{(401a)}$$

$$y_2 = y_0 + \frac{h}{3}\big(f(x_0, y_0) + 4f(x_1, y_1) + f(x_2, y_2)\big). \qquad \text{(401b)}$$

These equations are solved by functional iteration to yield approximations $y_1 \approx y(x_1)$ and $y_2 \approx y(x_2)$.

In modern variable order codes, it is usual to start with order 1 or order 2, and to adapt to higher orders when this becomes possible and when it becomes advantageous from an efficiency point of view. This means that order k may be reached after many steps with varying stepsize.

The common feature of these approaches to starting a linear multistep method is that each is, in reality, a Runge–Kutta method possessing multiple outputs, to furnish approximations at a number of equally spaced points. For example, the iteration scheme given by (401a) and (401b) can be represented by the Runge–Kutta scheme

$$
\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
1 & \frac{5}{12} & \frac{2}{3} & -\frac{1}{12} \\
2 & \frac{1}{3} & \frac{4}{3} & \frac{1}{3} \\
\hline
 & \frac{5}{12} & \frac{2}{3} & -\frac{1}{12} \\
 & \frac{1}{3} & \frac{4}{3} & \frac{1}{3}
\end{array}
$$

in which the two output approximations are for y_1 and y_2, respectively. This scheme, like any starting procedure of Runge–Kutta type, has a property we assume for starting schemes used for the definition of convergence. This is that the quantities computed as approximations to y_i, $i = 1, 2, \ldots, k-1$, all converge to $y(x_0)$ as $h \to 0$.

402   Convergence

We consider the approximation of $y(\overline{x})$ by a linear multistep method, with $h = (\overline{x} - x_0)/m$, using initial values

$$
\begin{aligned}
y_0 &= \phi_0\big(y(x_0), h\big), \\
y_1 &= \phi_1\big(y(x_0), h\big), \\
&\vdots \qquad \vdots \\
y_{k-1} &= \phi_{k-1}\big(y(x_0), h\big).
\end{aligned}
$$

After the initial values have been evaluated, the values of y_n, for $n = k, k+1, \ldots, m$, are found in turn, using the linear k-step method $[\alpha, \beta]$. It is assumed that for $i = 1, 2, \ldots, k-1$,

$$
\big\|\phi_i\big(y(x_0), h\big) - y(x_0)\big\| \to 0, \qquad \text{as } h \to 0.
$$

Definition 402A *Consider a linear multistep method used with a starting method as described in the previous discussion. Let Y_m denote the approximation to $y(\overline{x})$ found using m steps with $h = (\overline{x} - x_0)/m$. The function f is assumed to be continuous and to satisfy a Lipschitz condition in its second variable. The linear multistep method is said to be 'convergent' if, for any such initial value problem,*

$$
\|Y_m - y(\overline{x})\| \to 0, \qquad \text{as } m \to \infty.
$$

403   Stability

For a general initial value problem, the computed solution satisfies

$$y_n = \sum_{i=1}^{k} \alpha_i y_{n-i} + h \sum_{i=0}^{k} \beta_i f(x_{n-i}, y_{n-i}).$$

However, for the one-dimensional problem for which $f(x, y) = 0$, we have the simpler difference equation

$$y_n = \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + \cdots + \alpha_k y_{n-k}. \tag{403a}$$

Definition 403A *A linear multistep method* $[\alpha, \beta]$ *is 'stable' if the difference equation* (403a) *has only bounded solutions.*

Because stability concepts of one sort or another abound in the theory of initial value problems, 'stability' is often referred to as 'zero-stability' – for example, in Lambert (1991)) – or as 'stability in the sense of Dahlquist'.

404   Consistency

Just as the initial value problem $y'(x) = 0$, with initial condition $y(x_0) = 0$, motivated the concept of stability, so the same problem, with initial value $y(x_0) = 1$, can be used to introduce preconsistency. We want to ensure that this problem can be solved exactly, starting from the exact initial value. Suppose the numerical solution is known to have the correct value at $x = x_{n-k}, x_{n-k+1}, \ldots, x_{n-1}$ so that $y_i = y(x_i) = 1$, for $i = n-k, n-k+1, \ldots, n-1$. Under these assumptions, the result computed at step n will be

$$y_n = \alpha_1 + \alpha_2 + \cdots + \alpha_k,$$

and this will equal the correct value $y_n = 1$ if and only if

$$1 = \alpha_1 + \alpha_2 + \cdots + \alpha_k. \tag{404a}$$

Definition 404A *A linear multistep method satisfying* (404a) *is said to be* '*preconsistent*'.

Now consider the differential equation

$$y'(x) = 1, \qquad y(x_0) = 0,$$

with exact solution at the step values

$$y_i = hi.$$

If this solution has been found for $i = n - k, n - k + 1, \ldots, n - 1$, then it is also correct for $i = n$ if and only if

$$nh = \alpha_1(n-1)h + \alpha_2(n-2)h + \cdots + \alpha_k(n-k)h + h(\beta_0 + \beta_1 + \cdots + \beta_k).$$

Assuming the method is preconsistent, the factor h can be cancelled and then n times (404a) can be subtracted. We then find

$$\alpha_1 + 2\alpha_2 + \cdots + k\alpha_k = \beta_0 + \beta_1 + \cdots + \beta_k. \tag{404b}$$

This leads to the following definition:

Definition 404B *A linear multistep method satisfying* (404a) *and* (404b) *is said to be 'consistent'.*

Another way of looking at the consistency conditions is to suppose that $y_i = y(x_i) + O(h^2)$ and that $f(x_i, y_i) = y'(x_i) + O(h)$, for $i = n-k, n-k+1, \ldots, n-1$, and to consider the computation of y_n using the equation

$$
\begin{aligned}
y_n - h\beta_0 & f(x_n, y_n) \\
&= \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + \cdots + \alpha_k y_{n-k} \\
&\quad + h(\beta_1 f(x_{n-1}, y_{n-1}) + \beta_2 f(x_{n-2}, y_{n-2}) + \cdots + \beta_k f(x_{n-k}, y_{n-k})) \\
&= \alpha_1 y(x_{n-1}) + \alpha_2 y(x_{n-2}) + \cdots + \alpha_k y(x_{n-k}) \\
&\quad + h(\beta_1 y'(x_{n-1}) + \beta_2 y'(x_{n-2}) + \cdots + \beta_k y'(x_{n-k})).
\end{aligned}
$$

Expand the right-hand side by Taylor's theorem about x_n, and we find

$$
\begin{aligned}
\big(\alpha_1 + \alpha_2 + &\cdots + \alpha_k\big)y(x_n) \\
&+ \big(\beta_1 + \cdots + \beta_k - \alpha_1 - 2\alpha_2 - \cdots - k\alpha_k\big)hy'(x_n) + O(h^2).
\end{aligned}
$$

This will give the correct answer of

$$y(x_n) - h\beta_0 y'(x_n),$$

to within $O(h^2)$, if and only if

$$\alpha_1 + \alpha_2 + \cdots + \alpha_k = 1$$

and

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

Hence, we can view the two requirements of consistency as criteria that the computed solution is capable of maintaining accuracy to within $O(h^2)$ over one step, and therefore over several steps.

405   *Necessity of conditions for convergence*

We formally prove that stability and consistency are necessary for convergence. Note that the proofs are based on the same simple problems that were introduced in Subsections 403 and 404.

Theorem 405A *A convergent linear multistep method is stable.*

Proof. If the method were not stable, there would exist an unbounded sequence η satisfying the difference equation

$$\eta_n = \alpha_1 \eta_{n-1} + \alpha_2 \eta_{n-2} + \cdots + \alpha_k \eta_{n-k}.$$

Define the sequence ζ by

$$\zeta_n = \max_{i=0}^{n} |\eta_i|,$$

so that ζ converges monotonically to ∞. Consider the solution of the initial value problem

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

with $\overline{x} = 1$. Assuming that n steps are to be performed, we use a stepsize $h = 1/n$ and initial values $y_i = \eta_i/\zeta_n$, for $i = 0, 1, \ldots, k-1$. The condition that $y_i \to 0$ for $0 \le i \le k-1$ is satisfied because $\zeta_n \to \infty$. The approximation computed for $y(\overline{x})$ is equal to η_n/ζ_n. Because the ζ sequence is unbounded, there will be an infinite number of values of n for which $|\zeta_n|$ is greater than the greatest magnitude amongst previous members of this sequence. For such values of n, $|\eta_n/\zeta_n| = 1$, and therefore the sequence $n \mapsto \eta_n/\zeta_n$ cannot converge to 0. □

Theorem 405B *A convergent linear multistep method is preconsistent.*

Proof. By Theorem 405A, we can assume that the method is stable. Let η be defined as the solution to the difference equation

$$\eta_n = \alpha_1 \eta_{n-1} + \alpha_2 \eta_{n-2} + \cdots + \alpha_k \eta_{n-k},$$

with initial values $\eta_0 = \eta_1 = \cdots = \eta_{k-1} = 1$. The computed solution of the problem

$$y'(x) = 0, \qquad y(0) = 1, \qquad \overline{x} = 1,$$

using n steps, is equal to $y_n = \eta_n$. Since this converges to 1 as $n \to \infty$, it follows that, for any $\epsilon > 0$, there exists an n sufficiently large so that $|y_i - 1| \le \epsilon$

for $i = n - k, n - k + 1, \ldots, n$. Hence,

$$\left|1 - \alpha_1 - \alpha_2 - \cdots - \alpha_k\right| \leq \left|\eta_n - \sum_{i=1}^{k} \alpha_i \eta_{n-i}\right| + \left(1 + \sum_{i=1}^{k} |\alpha_i|\right)\epsilon$$

$$= \left(1 + \sum_{i=1}^{k} |\alpha_i|\right)\epsilon.$$

Because this can be arbitrarily small, it follows that

$$1 - \alpha_1 - \alpha_2 - \cdots - \alpha_k = 0. \qquad \square$$

Theorem 405C *A convergent linear multistep is consistent.*

Proof. We note first that

$$\alpha_1 + 2\alpha_2 + \cdots + k\alpha_k \neq 0,$$

since, if the expression were zero, the method would not be stable. Define the sequence η by

$$\eta_i = \frac{\beta_0 + \beta_1 + \cdots + \beta_k}{\alpha_1 + 2\alpha_2 + \cdots + k\alpha_k} i, \qquad i = 0, 1, 2, \ldots.$$

Consider the numerical solution of the initial value problem

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

with the output computed at $\overline{x} = 1$, and with n steps computed with stepsize $h = 1/n$. Choose starting approximations as

$$y_i = \frac{1}{n}\eta_i, \tag{405a}$$

for $i = 0, 1, 2, \ldots, k - 1$, so that these values converge to zero as $n \to \infty$. We verify that the computed solution for *all* values of $i = 0, 1, 2, \ldots, n$ is given also by (405a), and it follows that the approximation at $x = 1$ is

$$\frac{\beta_0 + \beta_1 + \cdots + \beta_k}{\alpha_1 + 2\alpha_2 + \cdots + k\alpha_k},$$

independent of n. Because convergence implies that the limit of this is 1, it follows that

$$\beta_0 + \beta_1 + \cdots + \beta_k = \alpha_1 + 2\alpha_2 + \cdots + k\alpha_k. \qquad \square$$

406 *Sufficiency of conditions for convergence*

Given that a linear multistep is stable and consistent, we prove that it is convergent. We assume that the differential equation under consideration has the autonomous form

$$y'(x) = f(y(x)) \tag{406a}$$

and that f satisfies a Lipschitz condition with constant L. These assumptions can be weakened in various ways with no change to the final result, but with considerable complication to the details. If the Lipschitz condition holds only locally, then it becomes necessary to restrict the stepsize so that it is possible to guarantee that all approximations which enter into the discussion are sufficiently close to the exact trajectory for the condition to apply. If the problem is not autonomous, so that $f(y)$ is replaced by $f(x, y)$, then it is possible to allow f to be Lipschitz continuous in the y variable, but merely continuous in x.

However, we now press ahead with consideration of the possible convergence of the solution to (406a), together with the initial information given at x_0 and the requirement that the approximate solution is to be evaluated at $\overline{x}$. We always assume that $\overline{x} > x_0$, to avoid the inconvenience of having to allow for negative stepsizes.

For the rest of this subsection, it will be assumed, without further comment, that the differential equation we are attempting to solve is (406a) and that the solution is to be approximated on the interval $[x_0, \overline{x}]$ with initial value information given at x_0. The stepsize h will always be positive, and the Lipschitz condition holds with constant L. We refer to the problem as 'the standard initial value problem'. One further notation we use throughout is to write M for a bound on $\left\| f\big(y(x)\big) \right\|$ for $x \in [x_0, \overline{x}]$. Such a bound clearly exists because

$$\|f(y(x)) - f(y(x_0))\| \leq L\|y(x) - y(x_0)\|,$$

and the latter quantity is bounded.

As a first step towards understanding the relationship between an approximation to $y(\overline{x})$ and the exact value of this quantity, we consider a quantity which measures the error generated in a single step.

Definition 406A *Let $[\alpha, \beta]$ be a consistent linear multistep method. The 'local truncation error' associated with a differentiable function y at a point x with stepsize h is the value of*

$$\mathcal{L}(y, x, h) = y(x) - \sum_{i=1}^{k} \alpha_i y(x - ih) - h \sum_{i=0}^{k} \beta_i y'(x - ih).$$

We estimate the value of $\mathcal{L}(y, x, h)$ when y is the exact solution to (406a), and where not only x but also each $x - hi$, for $i = 1, 2, \ldots, k$, lies in the interval $[x_0, \overline{x}]$.

Lemma 406B *If y is the exact solution to the standard initial value problem and $x \in [x_0 + kh, \overline{x}]$, then*

$$\|\mathcal{L}(y, x, h)\| \leq \sum_{i=1}^{k} \left(\frac{1}{2} i^2 |\alpha_i| + i |i\alpha_i - \beta_i| \right) LMh^2.$$

Proof. We first estimate $y(x) - y(x - ih) - ihy'(x)$ using the identity

$$y(x) - y(x - ih) - hiy'(x) = h \int_{-i}^{0} \left(f(y(x + h\xi)) - f(y(x)) \right) d\xi,$$

so that

$$\|y(x) - y(x - ih) - ihy'(x)\| \leq hL \int_{-i}^{0} \|y(x + h\xi) - y(x)\| \, d\xi,$$

and noting, that for $\xi \leq 0$,

$$\|y(x + h\xi) - y(x)\| \leq h \int_{\xi}^{0} \|f(x + h\overline{\xi})\| d\overline{\xi} \leq h|\xi|M, \qquad (406b)$$

so that

$$\|y(x) - y(x - ih) - ihy'(x)\| \leq \tfrac{1}{2} i^2 h^2 LM.$$

From (406b), we see also that

$$\|f(y(x)) - f(y(x - ih))\| \leq ihLM.$$

Because of the consistency of the method, we have $\sum_{i=1}^{k} \alpha_i = 1$ and $\sum_{i=1}^{k} (i\alpha_i - \beta_i) = \beta_0$. We now write $\mathcal{L}(y, x, h)$ in the form

$$\mathcal{L}(y, x, h) = \sum_{i=1}^{k} \alpha_i (y(x) - y(x - ih) - ihy'(x))$$

$$+ h \sum_{i=1}^{k} (i\alpha_i - \beta_i)(y'(x) - y'(x - ih));$$

this is bounded by

$$\frac{1}{2} \sum_{i=1}^{k} i^2 |\alpha_i| LMh^2 + \sum_{i=1}^{k} i |i\alpha_i - \beta_i| LMh^2$$

and the result follows. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square$

Theorem 406C *Let ϵ_n denote the vector*

$$\epsilon_n = y(x_n) - y_n.$$

Then for h_0 sufficiently small so that $h_0|\beta_0|L < 1$ and $h < h_0$, there exist constants C and D such that

$$\left\| \epsilon_n - \sum_{i=1}^{k} \alpha_i \epsilon_{n-i} \right\| \leq Ch \max_{i=1}^{k} \|\epsilon_{n-i}\| + Dh^2. \qquad (406c)$$

Proof. The value of $\epsilon_n - \sum_{i=1}^{k} \alpha_i \epsilon_{n-i} - h \sum_{i=0}^{k} \beta_i(f(y(x_{n-i})) - f(y_{n-i}))$ is the difference of two terms, of which the first can be bounded by a constant times h^2, by Theorem 406B, and the second is zero. This means that

$$\epsilon_n - \sum_{i=1}^{k} \alpha_i \epsilon_{n-i} = T_1 + T_2 + T_3, \qquad (406d)$$

where

$$\|T_1\| = h|\beta_0| \|f(y(x_n)) - f(y_n)\| \qquad \leq hL|\beta_0| \cdot \|\epsilon_n\|, \qquad (406e)$$

$$\|T_2\| = h \left\| \sum_{i=1}^{k} \beta_i(f(y(x_{n-i})) - f(y_{n-i})) \right\| \leq hL \sum_{i=1}^{k} |\beta_i| \max_{i=1}^{k} \|\epsilon_{n-i}\|, \quad (406f)$$

and $\|T_3\|$ can be bounded in terms of a constant times h^2. We now use (406d) twice. First, assuming $h \leq h_0$, obtain a bound on $(1 - hL|\beta_0|)\|\epsilon_n\|$ in terms of $\max_{i=1}^{k} \|\epsilon_{n-i}\|$ and terms that are bounded by a constant times h^2. Hence, obtain a bound on $\|\epsilon_n\|$. Then, by inserting this preliminary result in the bound on T_1, we obtain the result of the theorem. $\qquad \square$

Theorem 406D *A stable consistent linear multistep method is convergent.*

Proof. Write (406c) in the form

$$\epsilon_n = \sum_{i=1}^{k} \alpha_i \epsilon_{n-i} + \psi_n,$$

where, according to Theorem 406C,

$$\|\psi_n\| \leq Ch \max_{i=1}^{k} \|\epsilon_{n-i}\| + Dh^2,$$

for h sufficiently small. Define θ_1, θ_2, ... as in Subsection 141, and note that, because the method is convergent, the θ sequence is bounded. From Theorem 141A, we have

$$\epsilon_n = \sum_{i=0}^{k-1} \theta_{n-i}\widetilde{\epsilon}_i + \sum_{i=k}^{n} \theta_{n-i}\psi_i,$$

where $\widetilde{\epsilon}_i$, for $i = 0, 1, \ldots, k-1$, are linear combinations of the errors in y_i and tend to zero as $h \to 0$. Hence we have

$$\|\epsilon_n\| \leq \Theta \sum_{i=0}^{k-1} \|\widetilde{\epsilon}_i\| + \Theta Chk \sum_{i=k}^{n-1} \|\epsilon_i\| + \Theta D(n-k)h^2, \qquad (406\text{g})$$

where $\Theta = \sup_{i=1}^{\infty} |\theta_i|$ and the factor k is introduced in the second summation in (406g) because the same maximum value of $\|\epsilon_{n-i}\|$ may arise in up to k adjacent terms. We rewrite (406g) in the form

$$\|\epsilon_n\| \leq \phi(h) + \Theta Chk \sum_{i=1}^{n-1} \|\epsilon_i\| + \Theta Dnh^2, \quad \|\epsilon_0\| \leq \phi(h),$$

where $\phi(h)$ takes positive values and will converge to zero as $h \to 0$. It now follows that $\|\epsilon_n\| \leq u_n$, where the sequence u is defined by

$$u_n = \Theta Chk \sum_{i=1}^{n-1} u_i + \Theta Dnh^2 + \phi(h), \quad u_0 = \phi(h). \qquad (406\text{h})$$

By subtracting (406h) with n replaced by $n-1$, we find that

$$u_n + \frac{Dh}{Ck} = (1 + \Theta Chk)\left(u_{n-1} + \frac{Dh}{Ck}\right),$$

which leads to the bound

$$\|\epsilon_n\| \leq u_n = (1 + \Theta Chk)^n \phi(h) + ((1 + \Theta Chk)^n - 1)\frac{Dh}{Ck}$$

$$\leq \exp(\Theta Cknh)\phi(h) + (\exp(\Theta Cknh) - 1)\frac{Dh}{Ck}.$$

To complete the proof, substitute $n = m$ where $mh = \overline{x} - x_0$, so that the error in the approximation at $x = \overline{x}$ using m steps with stepsize h is bounded by

$$\exp(\Theta Ck(\overline{x} - x_0))\phi(h) + \exp(\Theta Ck(\overline{x} - x_0))\frac{Dh}{Ck} \to 0. \qquad \square$$

Exercises 40

40.1 Find a four-stage Runge–Kutta method with $c_2 = \frac{1}{3}$, $c_3 = \frac{2}{3}$, $c_4 = 1$, which satisfies the order conditions

$$\sum_{i=1}^{4} b_i = \xi,$$

$$\sum_{i=1}^{4} b_i c_i = \frac{1}{2}\xi^2,$$

$$\sum_{i=1}^{4} b_i c_i^2 = \frac{1}{3}\xi^3,$$

$$\sum_{i,j=1}^{4} b_i a_{ij} c_j = \frac{1}{6}\xi^3,$$

where ξ is a real parameter and the elements of A are independent of ξ. Show how this method can be used as a starter for the predictor–corrector pair consisting of the third order Adams–Bashforth and Adams–Moulton methods.

40.2 For each of the following polynomial pairs, written as $[\alpha(z), \beta(z)]$, determine if the corresponding numerical method is consistent and stable:

1. $[1 - z, 2z - z^2]$,
2. $[1 - z^2, 2z - z^2]$,
3. $[1 + z - 3z^2 + z^3, 3z - z^2]$,
4. $[1 + z - z^2 - z^3, 3 + z]$.

40.3 Translate the conditions for stability, preconsistency and consistency from the $[\alpha, \beta]$ representation to the (ρ, σ) representation.

40.4 For a linear multistep method $[\alpha, \beta]$, define polynomials a and b by

$$a(z) = (1 + z)^k - \alpha_1(1 + z)^{k-1}(1 - z) - \alpha_2(1 + z)^{k-2}(1 - z)^2 - \cdots$$
$$- (1 - z)^k \alpha_k,$$
$$b(z) = \beta_0(1 + z)^k + \beta_1(1 + z)^{k-1}(1 - z) + \beta_2(1 + z)^{k-2}(1 - z)^2 + \cdots$$
$$+ (1 - z)^k \beta_k.$$

Find the conditions for stability, preconsistency and stability in terms of the polynomials a and b.

41   The Order of Linear Multistep Methods

410   Criteria for order

Given a linear multistep method $[\alpha, \beta]$, we seek conditions on the coefficients in the polynomials α and β that will guarantee that, locally, errors are $O(h^{p+1})$. By this we mean that if starting values satisfy $y_i = y(x_i) + O(h^{p+1})$, for $i = 0, 1, \ldots, k-1$, then this will imply that a similar estimate persists for $i \geq k$. We emphasize that this is a local property in the sense that it cannot be used in a limiting case in which integration is carried to a fixed point $\overline{x} > x_0$, because the number of steps required to achieve this is approximately $(\overline{x} - x_0)/h$, and this is unbounded as $h \to 0$. To verify that $y_n = y(x_n) + O(h^{p+1})$, assuming the same is true for the previous k step values, it will be enough to estimate the value of

$$y(x_n) - \sum_{i=1}^{k} \alpha_i y(x_{n-i}) - \sum_{i=0}^{k} \beta_i h y'(x_{n-i}) \qquad (410a)$$

and to show that, under appropriate smoothness assumptions, it is $O(h^{p+1})$. The smoothness assumptions will be that the problem under consideration has a solution with continuous derivatives up to order $p + 1$. This will enable us to expand (410a) in a Taylor series

$$C_0 y(x_n) + C_1 h y'(x_n) + C_2 h^2 y''(x_n) + \cdots$$
$$+ C_p h^p y^{(p)}(x_n) + C_{p+1} h^{p+1} y^{(p+1)}(x_n) + \cdots \quad (410b)$$

and order p will mean that $C_0 = C_1 = \cdots = C_p$. The value of C_{p+1} is closely related to the error constant and is non-zero unless the order is actually higher than p.

Theorem 410A *The constants $C_0, C_1, C_2, \ldots$ in (410b) are given by*

$$\alpha(\exp(-z)) - z\beta(\exp(-z)) = C_0 + C_1 z + C_2 z^2 + \cdots . \qquad (410c)$$

Proof. The coefficient of $y(x_n)$ in the Taylor expansion of (410a) is equal to $1 - \sum_{i=1}^{k} \alpha_i$, and this equals the constant term in the Taylor expansion of $\alpha(\exp(-z)) - z\beta(\exp(-z))$. Now suppose that $j = 1, 2, \ldots$ and calculate the coefficient of $y^{(j)}(x_n)$ in the Taylor expansion of (410a). This equals

$$-\sum_{i=1}^{k} \alpha_i \frac{(-i)^j}{j!} - \sum_{i=0}^{k} \beta_i \frac{(-i)^{j-1}}{(j-1)!},$$

where the coefficient of β_0 is -1 if $j = 1$ and zero for $j > 1$. This is identical to the coefficient of z^j in the Taylor expansion of $\alpha(\exp(-z)) - z\beta(\exp(-z))$. $\square$

Altering the expression in (410c) slightly, we can state without proof a criterion for order:

Theorem 410B *A linear multistep method $[\alpha, \beta]$ has order p (or higher) if and only if*

$$\alpha(\exp(z)) + z\beta(\exp(z)) = O(z^{p+1}).$$

Because we have departed from the traditional (ρ, σ) formulation for linear multistep methods, we restate this result in that standard notation:

Theorem 410C *A linear multistep method (ρ, σ) has order p if and only if*

$$\rho(\exp(z)) - z\sigma(\exp(z)) = O(z^{p+1}).$$

Return now to Theorem 410B and replace $\exp(z)$ by $(1+z)^{-1}$. It is found that

$$\alpha((1+z)^{-1}) - \log(1+z)\beta((1+z)^{-1}) = O(z^{p+1}), \tag{410d}$$

where $\log(1+z)$ is defined only in $\{z \in \mathbb{C} : |z| < 1\}$ by its power series

$$\log(1+z) = z - \tfrac{1}{2}z^2 + \tfrac{1}{3}z^3 - \cdots.$$

Because both $\alpha(1+z)$ and $\log(1+z)$ vanish when $z = 0$, it is possible to rearrange (410d) in the form given in the following result, which we present without further proof.

Theorem 410D *A linear multistep formula $[\alpha, \beta]$ has order p if and only if*

$$\frac{z}{\log(1+z)} \frac{\alpha(1+z)}{z} + \beta(1+z) = O(z^p).$$

411   Derivation of methods

Given the relationship between the coefficients in the α and β polynomials under the condition that they have a specified order, the choice of actual methods remains. The first approach we consider is when α is given and β is then chosen to achieve the required order. In Subsection 412 we consider the derivation of backward difference methods, in which β is first specified and α is then derived.

To find the coefficients in Adams–Moulton methods, use $\alpha(z) = 1 - z$, so

that from Theorem 410D we find

$$
\begin{aligned}
\beta(1+z) &= \frac{1}{1 - \frac{1}{2}z + \frac{1}{3}z^2 - \frac{1}{4}z^3 + \frac{1}{5}z^4 - \frac{1}{6}z^5 + \cdots} \\
&= 1 + \frac{1}{2}z - \frac{1}{12}z^2 + \frac{1}{24}z^3 - \frac{19}{720}z^4 + \frac{3}{160}z^5 \\
&\quad - \frac{863}{60480}z^6 + \frac{275}{24192}z^7 - \frac{33953}{3628800}z^8 \\
&\quad + \frac{8183}{1036800}z^9 - \frac{3250433}{479001600}z^{10} + \cdots .
\end{aligned}
\tag{411a}
$$

It is clear that order $k+1$ can be obtained using a k-step method because the expansion can be truncated at the term in z^k, leading to an $O(z^{k+1})$ error and degree k polynomial $\beta(1+z)$. For example, for $k = 1$ we have

$$
\beta(1+z) = 1 + \tfrac{1}{2}z,
$$

implying that

$$
\beta(z) = 1 + \tfrac{1}{2}(z - 1) = \tfrac{1}{2} + \tfrac{1}{2}z,
$$

giving the coefficients $\beta_0 = \beta_1 = \frac{1}{2}$. If $k = 2$ we have

$$
\beta(1+z) = 1 + \tfrac{1}{2}z - \tfrac{1}{12}z^2
$$

and

$$
\beta(z) = 1 + \tfrac{1}{2}(z - 1) - \tfrac{1}{12}(z - 1)^2 = \tfrac{5}{12} + \tfrac{2}{3}z + -\tfrac{1}{12}z^2,
$$

giving $\beta_0 = \frac{5}{12}$, $\beta_1 = \frac{2}{3}$, $\beta_2 = -\frac{1}{12}$. In general, we can find the coefficients by rewriting (411a) in the form

$$
\begin{aligned}
\beta(z) &= 1 + \tfrac{1}{2}(z - 1) - \tfrac{1}{12}(z - 1)^2 + \tfrac{1}{24}(z - 1)^3 - \tfrac{19}{720}(z - 1)^4 \\
&\quad + \tfrac{3}{160}(z - 1)^5 - \tfrac{863}{60480}(z - 1)^6 + \tfrac{275}{24192}(z - 1)^7 - \tfrac{33953}{3628800}(z - 1)^8 \\
&\quad + \tfrac{8183}{1036800}(z - 1)^9 - \tfrac{3250433}{479001600}(z - 1)^{10} + \cdots ,
\end{aligned}
$$

and truncating at the term in $(z - 1)^k$ to obtain the coefficients in the k-step order $k + 1$ method.

For Adams–Bashforth methods, in which β_0 necessarily vanishes, we write $\beta(z) = z\widehat{\beta}(z)$, where $\widehat{\beta}$ has degree $k - 1$ for a k-step method. In this case Theorem 410D can be written in the form

$$
\frac{z}{(1+z)\log(1+z)} \frac{\alpha(1+z)}{z} + \widehat{\beta}(1+z) = O(z^p),
$$

and we aim for order $p = k$. It is found that

$$\widehat{\beta}(1+z) = \frac{1}{(1+z)\left(1 - \frac{1}{2}z + \frac{1}{3}z^2 - \frac{1}{4}z^3 + \cdots\right)}$$

$$= 1 - \frac{1}{2}z + \frac{5}{12}z^2 - \frac{3}{8}z^3 + \frac{251}{720}z^4 - \frac{95}{288}z^5$$

$$+ \frac{19087}{60480}z^6 - \frac{5257}{17280}z^7 + \frac{1070017}{3628800}z^8$$ (411b)

$$- \frac{25713}{89600}z^9 + \frac{26842253}{95800320}z^{10} - \cdots,$$

so that the coefficients $\beta_1, \beta_2, \ldots, \beta_k$ can be found by selecting the coefficients of $z^0, z^1, \ldots, z^{k-1}$ in the truncation to the term in $(z-1)^{k-1}$ in the expansion

$$\widehat{\beta}(z) = 1 - \frac{1}{2}(z-1) + \frac{5}{12}(z-1)^2 - \frac{3}{8}(z-1)^3 + \frac{251}{720}(z-1)^4$$

$$- \frac{95}{288}(z-1)^5 + \frac{19087}{60480}(z-1)^6 - \frac{5257}{17280}(z-1)^7 + \frac{1070017}{3628800}(z-1)^8$$

$$- \frac{25713}{89600}(z-1)^9 + \frac{26842253}{95800320}(z-1)^{10} - \cdots.$$

For example, when $k = 2$ we have $\widehat{\beta}(z) = 1 - \frac{1}{2}(z-1) = \frac{3}{2} - \frac{1}{2}z$ leading to $\beta_1 = \frac{3}{2}$ and $\beta_2 = -\frac{1}{2}$ for the Adams–Bashforth method with order $p = 2$. When $k = 3$ we have $\widehat{\beta}(z) = 1 - \frac{1}{2}(z-1) + \frac{5}{12}(z-1)^2 = \frac{23}{12} - \frac{4}{3}z + \frac{5}{12}z^2$ so that, for the Adams–Bashforth method with order $p = 3$, we have $\beta_1 = \frac{23}{12}$, $\beta_2 = -\frac{4}{3}$, $\beta_3 = \frac{5}{12}$.

Values of the Adams–Bashforth and Adams–Moulton coefficients have previously been given in Tables 244(I) and 244(II), respectively.

412   Backward difference methods

These methods are also known as 'backward difference formulae' or BDF methods. Sometimes the notation BDFk is used for the order k member of this family. Instead of choosing a specific α polynomial, we consider the choice $\beta = \beta_0$, where β_0 is to be chosen for consistency. From Theorem 410D we have

$$\alpha(1+z) = -\beta_0 \log(1+z) + O(z^{p+1}).$$

Expand $\beta_0 \log(1+z)$ to terms in z^k, for order $p = k$, and then substitute $z-1$ in place of z. It is found that

$$\alpha(z) = \beta_0\left(-(z-1) + \frac{1}{2}(z-1)^2 - \frac{1}{3}(z-1)^3 + \cdots\right),$$

and β_0 is chosen so that $\alpha(0) = 1$. For $k = p = 1$, we have $\alpha(z) = \beta_0(1-z)$, so that $\beta_0 = 1$ and $\alpha_1 = 1$. For $k = p = 2$,

$$\alpha(z) = \beta_0\left((1-z) + \frac{1}{2}(1-z)^2\right) = \beta_0\left(\frac{3}{2} - 2z + \frac{1}{2}z^2\right),$$

Table 412(I)   Coefficients of the backward difference methods up to order 7

k	α_1	α_2	α_3	α_4	α_5	α_6	α_7	β_0	C
1	1							1	$\frac{1}{2}$
2	$\frac{4}{3}$	$-\frac{1}{3}$						$\frac{2}{3}$	$\frac{2}{9}$
3	$\frac{18}{11}$	$-\frac{9}{11}$	$\frac{2}{11}$					$\frac{6}{11}$	$\frac{3}{22}$
4	$\frac{48}{25}$	$-\frac{36}{25}$	$\frac{16}{25}$	$-\frac{3}{25}$				$\frac{12}{25}$	$\frac{12}{125}$
5	$\frac{300}{137}$	$-\frac{300}{137}$	$\frac{200}{137}$	$-\frac{75}{137}$	$\frac{12}{137}$			$\frac{60}{137}$	$\frac{10}{137}$
6	$\frac{120}{49}$	$-\frac{150}{49}$	$\frac{400}{147}$	$-\frac{75}{49}$	$\frac{24}{49}$	$-\frac{10}{147}$		$\frac{20}{49}$	$\frac{20}{343}$
7	$\frac{980}{363}$	$-\frac{490}{121}$	$\frac{4900}{1089}$	$-\frac{1225}{363}$	$\frac{196}{121}$	$-\frac{490}{1089}$	$\frac{20}{363}$	$\frac{140}{363}$	$\frac{35}{726}$

giving $\beta_0 = \frac{2}{3}$ and

$$\alpha_1 = \frac{4}{3}, \quad \alpha_2 = -\frac{1}{3}.$$

The coefficients for these methods are given up to $p = k = 7$ in Table 412(I), where the error constant C is found to be $\beta_0/(p+1)$.

   Note that the method with $p = k = 7$ is of no practical value, in terms of the criteria for convergence, because it is not stable. This remark also applies to methods with $k > 7$.

Exercises 41

41.1 Given α_2, find α_1, β_1 and β_2 such that the linear multistep method $(1 - \alpha_1 z - \alpha_2 z^2, \beta_1 z + \beta_2 z^2)$ has order 2. What are the bounds on α_2 for which the method is convergent?

41.2 Show that all backward difference methods with $k \leq 6$ are stable.

41.3 Show that the order 7 backward difference method is not stable.

41.4 Find a stable seventh order linear multistep method of the form $(1 - \alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_8 z^8, \beta_0)$.

42   Errors and Error Growth

420   Introduction

The result computed in a step is generally not exact, even if we ignore any errors introduced in previous steps. However, once a significant departure from the exact solution has occurred, we are in effect solving a different problem. Hence, a proper analysis of error takes account of errors generated locally, and

also the accumulated effect of errors generated in previous steps. We present a simplified discussion of this phenomenon in this subsection, and discuss the limitations of this discussion in Subsection 421.

Suppose a sequence of approximations

$$y_1 \approx y(x_1),$$
$$y_2 \approx y(x_2),$$
$$\vdots \qquad \vdots$$
$$y_{n-1} \approx y(x_{n-1}),$$

has been computed, and we are now computing step n. If, for the moment, we ignore errors in previous steps, the value of y_n can be evaluated using a Taylor expansion where, for implicit methods, we need to take account of the fact that $f(y_n)$ is also being calculated. We have

$$y(x_n) - y_n - h\beta_0(f(y(x_n)) - f(y_n))$$
$$= y(x_n) - \sum_{i=1}^{k} \alpha_i y(x_{n-i}) - h \sum_{i=0}^{k} \beta_i y'(x_{n-1}),$$

which is equal to

$$C_{p+1} h^{p+1} y^{(p+1)}(x_n) + O(h^{p+2}).$$

In this informal discussion, we not only ignore the term $O(h^{p+2})$ but also treat the value of $h^{p+1} y^{(p+1)}(x_{n-i})$ as constant. This is justified in a local sense. That is, if we confine ourselves to a finite sequence of steps preceding step n, then the variation in values of this quantity will also be $O(h^{p+2})$, and we ignore such quantities. Furthermore, if

$$y(x_n) - y_n - h\beta_0(f(y(x_n)) - f(y_n)) \approx C_{p+1} h^{p+1} y^{(p+1)}(x_n),$$

then the assumption that f satisfies a Lipschitz condition will imply that

$$y(x_n) - y_n \approx C_{p+1} h^{p+1} y^{(p+1)}(x_n)$$

and that

$$h(f(y(x_n)) - f(y_n)) = O(h^{p+2}).$$

With the contributions of terms of this type thrown into the $O(h^{p+2})$ category, and hence capable of being ignored from the calculation, we can write a difference equation for the error in step n, which will be written as $\epsilon_n = y(x_n) - y_n$, in the form

$$\epsilon_n - \sum_{i=1}^{k} \alpha_i \epsilon_{n-i} = K h^{p+1},$$

where K is a representative value of $C_{p+1}y^{(p+1)}$.

For a stable consistent method, the solution of this equation takes the form

$$\epsilon_n = -\alpha'(1)^{-1}h^{p+1}nK + \sum_{i=1}^{k}\eta_i\lambda_i^n, \qquad (420a)$$

where the coefficients η_i, $i = 1, 2, \ldots, k$, depend on initial values and λ_i, $i = 1, 2, \ldots, k$, are the solutions to the polynomial equation $\alpha(\lambda^{-1}) = 0$.

The factor $-\alpha'(1)^{-1}$ that occurs in (420a) can be written in a variety of forms, and we have

$$-\alpha'(1) = \rho'(1) = \beta(1) = \sigma(1) = \alpha_1 + 2\alpha_2 + \cdots + k\alpha_k.$$

The value of $-C\alpha'(1)^{-1}$ is known as the 'error constant' for the method and represents the factor by which $h^{p+1}y^{(p+1)}$ must be multiplied to give the contribution from each step to the accumulated error. Since the method is assumed to be stable, the terms of the form $\eta_i\lambda_i^n$ can be disregarded compared with the linearly growing term $-\alpha'(1)^{-1}h^{p+1}nK$. If the integration is carried out to a specific output value $\overline{x}$, and n steps are taken to achieve this result, then $hn = \overline{x} - x_0$. In this case we can make a further simplification and write the accumulated error as approximately

$$-(\overline{x} - x_0)\alpha'(1)^{-1}h^p C y^{(p+1)}(\overline{x}).$$

In the next subsection, these ideas will be discussed further.

421 Further remarks on error growth

In Subsection 420 we gave an informal argument that, over many steps, there is a contribution to the accumulated error from step n of approximately $-\alpha'(1)^{-1}C_{p+1}y^{(p+1)}(x_n)h^{p+1}$. Since we are interested in the effect of this contribution at some future point $\overline{x}$, we can consider the differential equation

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

with two possible initial values at the point $x = x_n$. These possible initial values are

$$y(x_n) \qquad \text{and} \qquad y(x_n) + \alpha'(1)^{-1}C_{p+1}y^{(p+1)}(x_n)h^{p+1},$$

and correspond respectively to the exact solution and to the solution perturbed by the error introduced in step n.

This suggests the possibility of analysing the development of numerical errors through the differential equation

$$z'(x) = \frac{\partial f(y(x))}{\partial y}z(x) + y^{(p+1)}(x), \qquad z(x_0) = 0. \qquad (421a)$$
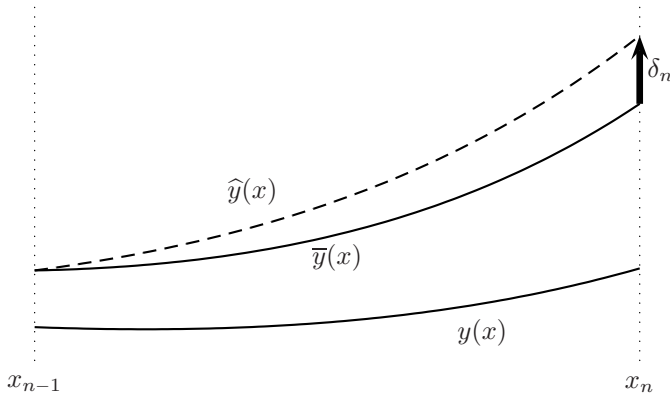
Figure 421(i)   Development of accumulated errors in a single step

Using this equation, we might hope to be able to approximate the error after n steps have been performed as

$$-\alpha'(1)^{-1}C_{p+1}h^p z(x_n),$$

because the linear term in (421a) expresses the rate of growth of the separation of an already perturbed approximation and the non-linear term, when scaled by $-\alpha'(1)^{-1}C_{p+1}h^p$, expresses the rate at which new errors are introduced as further steps are taken. The negative sign is consistent with the standard convention that errors are interpreted to mean the exact solution minus the approximation.

To turn this idea into a formal result it is possible to proceed in two steps. In the first step, asymptotic approximations are made. In the second, the errors in making these approximations are bounded and estimated so that they can all be bundled together in a single term which tends to zero more rapidly as $h \to 0$ than the asymptotic approximation to the error.

The second of these steps will not be examined in detail and the first step will be described in terms of the diagram given in Figure 421(i). In this figure, $y(x)$ is the exact solution and $\widehat{y}(x)$ is the function $y(x) + \alpha'(1)^{-1}C_{p+1}h^p z(x)$.

The function $\overline{y}(x)$ is the exact solution to the differential equation but with initial value at x_{n-1} set to $\widehat{y}(x_{n-1})$. In the single step from x_{n-1} to x_n, the perturbed approximation $\overline{y}$ drifts away from y at an approximate rate $\big(\partial f(y(x))/\partial y\big)\big(\overline{y}(x) - y(x)\big)$, to reach a value $\overline{y}(x_n)$. Add to this the contribution of local truncation error, corresponding to this step, denoted by $\delta_n = \alpha'(1)^{-1}C_{p+1}y^{(p+1)}(x_n)h^{p+1}$. With this local error added, the accumulated error moves to a value $\widehat{y}(x_n)$. However, following the smoothed-out curve $\widehat{y}(x)$ over the interval $[x_{n-1}, x_n]$ leads to the same point, to within $O(h^{p+2})$.

422   The underlying one-step method

Although linear multistep methods seem to be at the opposite end of the spectrum from Runge–Kutta methods, there is a very close link between them. Suppose the method $[\alpha, \beta]$ is preconsistent and stable, and consider the equation

$$
1 - \alpha_1 \eta^{-1} - \alpha_2 \eta^{-2} - \cdots - \alpha_k \eta^{-k}
$$
$$
- \beta_0 D - \beta_1 \eta^{-1} D - \beta_2 \eta^{-2} D - \cdots - \beta_k \eta^{-k} D = 0, \quad (422a)
$$

where $\eta \in G_1$. In Theorem 422A, we will show that (422a) has a unique solution.

Although η does not represent a Runge–Kutta method, it does represent a process for progressing a numerical approximation through a single time step. Suppose that the method is started using

$$
y_i = y(x_0) + \sum_{t \in T} \frac{\eta^i(t) h^{r(t)}}{\sigma(t)} F(t)(y(x_0)), \quad i = 0, 1, 2, \ldots, k - 1,
$$

corresponding to the group element η^i; then this value of y_i will persist for $i = k, k + 1, \ldots$. We will show this formally in Theorem 422C.

In the meantime, we remark that convergence of the formal series associated with η^i is not assured, even for $i = 1$, unless the function f and the value of h are restricted in some appropriate way. In this sense we can regard these 'B-series' as formal Taylor series.

What we really want is not η satisfying (422a) but the mapping Φ, say, which corresponds to it. If exponentiation of Φ is taken to denote compositions, or, for negative powers, compositions of the inverse mapping, then we want to be able to define Φ by

$$
\mathrm{id} - \alpha_1 \Phi^{-1} - \alpha_2 \Phi^{-2} - \cdots - \alpha_k \Phi^{-k}
$$
$$
- h\beta_0 f - h\beta_1(f \circ \Phi^{-1}) - h\beta_2(f \circ \Phi^{-2}) - \cdots - h\beta_k(f \circ \Phi^{-k}) = 0. \quad (422b)
$$

Because the corresponding member of G_1 can be evaluated up to any required order of tree, it is regarded as satisfactory to concentrate on this representation.

Theorem 422A *For any preconsistent, stable linear multistep method $[\alpha, \beta]$, there exists a member of the group G_1 satisfying* (422a).

Proof. By preconsistency, $\sum_{i=1}^{k} \alpha_i = 1$. Hence, (422a) is satisfied in the case of $t = \emptyset$, in the sense that if both sides are evaluated for the empty tree, then they each evaluate to zero. Now consider a tree t with $r(t) > 0$ and assume

that

$$1(u) - \alpha_1 \eta^{-1}(u) - \alpha_2 \eta^{-2}(u) - \cdots - \alpha_k \eta^{-k}(u)$$
$$- \beta_0 D(u) - \beta_1 \eta^{-1} D(u) - \beta_2 \eta^{-2} D(u) - \cdots - \beta_k \eta^{-k} D(u) = 0,$$

is satisfied for every tree u satisfying $r(u) < r(t)$. We will prove that there exists a value of $\eta(t)$ such that this equation is also satisfied if u is replaced by t. The coefficient of $\eta(t)$ in $\eta^{-i}(t)$ is equal to $i(-1)^{r(t)}$ and there are no other terms in $\eta^{-i}(t)$ with orders greater than $r(t) - 1$. Furthermore, all terms on the right-hand side contain only terms with orders less than $r(t)$. Hence, to satisfy (422a), with both sides evaluated at t, it is only necessary to solve the equation

$$(-1)^{r(t)-1} \sum_{i=1}^{k} i \alpha_i \eta(t) = C,$$

where C depends only on lower order trees. The proof by induction on $r(t)$ is now complete, because the coefficient of $\eta(t)$ is non-zero, by the stability of the method. $\qquad \square$

Definition 422B *Corresponding to a linear multistep method* $[\alpha, \beta]$, *the member of G_1 represents the 'underlying one-step method'.*

As we have already remarked, the mapping Φ in (422b), if it exists in more than a notional sense, is really the object of interest and this really is the underlying one-step method.

Theorem 422C *Let $[\alpha, \beta]$ denote a preconsistent, stable linear multistep method and let η denote a solution of* (422a). *Suppose that y_i is represented by η^i for $i = 0, 1, 2, \ldots, k-1$; then y_i is represented by η^i for $i = k, k+1, \ldots$.*

Proof. The proof is by induction, and it will only be necessary to show that y_k is represented by η^k, since this is a typical case. Multiply (422a) on the left by η^k and we find that

$$\eta^k - \alpha_1 \eta^{k-1} - \alpha_2 \eta^{k-2} - \cdots - \alpha_k$$
$$- \beta_0 \eta^k D - \beta_1 \eta^{k-1} D - \beta_2 \eta^{k-2} D - \cdots - \beta_k D = 0,$$

so that y_k is represented by η^k. $\qquad \square$

The concept of an underlying one-step method was introduced by Kirchgraber (1986). Although the underlying method cannot be represented as a Runge–Kutta method, it can be represented as a B-series or, what is equivalent, in the manner that has been introduced here. Of more recent developments, the extension to general linear methods (Stoffer, 1993) is of particular interest. This generalization will be considered in Subsection 535.

423   Weakly stable methods

The stability requirement for linear multistep methods specifies that all zeros of the polynomial ρ should lie in the closed unit disc with only simple zeros on the boundary. There is always a zero at 1, because of consistency, and there may or may not be other zeros on the boundary. We show in Subsection 441 that for a k-step method, with k even, the maximum possible order is $k + 2$. For methods with this maximal order, it turns out that *all* zeros of ρ lie on the unit circle and we are forced to take these methods seriously. We will write methods in the $[\alpha, \beta]$ terminology. A classic example is

$$\alpha(z) = 1 - z^2, \tag{423a}$$
$$\beta(z) = 2z \tag{423b}$$

and this is known as the 'leapfrog method'. Methods based on Newton–Cotes formulae were promoted by Milne (1953), and these all fall into this family.

The presence of additional zeros (that is, in addition to the single zero required by consistency) on the unit circle leads to the phenomenon known as 'weak stability'.

A characteristic property of weakly stable methods is their difficulty in dealing with the long term integration of dissipative problems. For example, if an approximation to the solution of $y' = -y$ is attempted using (423a), the difference equation for the computed results is

$$y_n + 2hy_{n-1} - y_{n-2} = 0. \tag{423c}$$

The general solution to (423c) is

$$y_n = A\lambda^n + B\mu^n, \tag{423d}$$

where

$$\lambda = -h + \sqrt{1 + h^2} \approx 1 - h + \tfrac{1}{2}h^2 \approx \exp(-h),$$
$$\mu = -h - \sqrt{1 + h^2} \approx -1 - h - \tfrac{1}{2}h^2 \approx -\exp(h),$$

where A and B depend on initial values. Substitute the approximate values of λ and μ into (423d) and we find

$$y_n \approx A\exp(-nh) + B(-1)^n \exp(nh).$$

For high values of n, the second term, which represents a parasitic solution, eventually dominates the solution and produces a very poor approximation. This is in contrast to what happens for the differential equation $y' = y$, for which the solution to the corresponding difference equation takes the form $y_n \approx A\exp(nh) + B(-1)^n \exp(-nh)$. In this case, the first term again corresponds to the true solution, but the second term will always be less significant.

424 *Variable stepsize*

If a sequence of approximations has already been computed using a specific stepsize and, for some reason, a decision is made to alter the stepsize, then a number of options arise as to how this might be done. For example, if a doubling of the stepsize is called for, then the necessary data might already be available without further computation. Halving the stepsize is not so convenient because new approximations to $y(x)$ and $y'(x)$ are required at points intermediate to the information that has already been computed. However, both these are special cases and it is usually required to change the stepsize by a ratio that is perhaps greater than 0.5 and less than 2.0. We consider a very simple model example in which new values are simply found by interpolation and the integration resumed using the modified data. Another approach which we will also consider is where a generalized version of the numerical method is defined specific to whatever sequence of stepsizes actually arises.

We now examine some basic stability questions arising from the interpolation option applied to an Adams method. At the end of step n, besides an approximation to $y(x_n)$, approximations are available for $hy'(x_n)$, $hy'(x_n - h)$, ..., $hy'(x_n - (p-1)h)$. We need to replace these derivative approximations by approximations to $rhy'(x_n), rhy'(x_n - rh), \ldots, rhy'(x_n - (p-1)rh)$, and these can be evaluated by the interpolation formula

$$
\begin{bmatrix}
rhy'(x_n) \\
rhy'(x_n - rh) \\
\vdots \\
rhy'(x_n - (p-1)rh)
\end{bmatrix}
\approx VD(r)V^{-1}
\begin{bmatrix}
hy'(x_n) \\
hy'(x_n - h) \\
\vdots \\
hy'(x_n - (p-1)h)
\end{bmatrix},
$$

where V is the Vandermonde matrix

$$
V =
\begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
1 & 1 & 1 & \cdots & 1 \\
1 & 2 & 2^2 & \cdots & 2^{p-1} \\
\vdots & \vdots & \vdots & & \vdots \\
1 & p-1 & (p-1)^2 & \cdots & (p-1)^{p-1}
\end{bmatrix}
$$

and

$$
D(r) = \operatorname{diag}(r, r^2, r^3, \ldots, r^p).
$$

The additional errors introduced into the computation by this change of stepsize technique can be significant. However, we are concerned here by the effect on stability. With constant stepsize, the stability of the difference equation system related to the derivative approximations is determined by

the influence matrix

$$
J = \begin{bmatrix}
0 & 0 & 0 & \cdots & 0 & 0 \\
1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0
\end{bmatrix}
$$

and because J is nilpotent, the dependence of quantities computed in a particular step eventually becomes insignificant. However, whenever the stepsize is altered by a factor r, the influence matrix becomes

$$
VD(r)V^{-1}J, \tag{424a}
$$

and this is, in general, not nilpotent. If, for example, the interpolation approach with stepsize ratio r is repeated over many steps, then (424a) might not be power-bounded and unstable behaviour will result. In the case $p = 3$, (424a) becomes

$$
\begin{bmatrix}
0 & 0 & 0 \\
2r^2 - r^3 & -\frac{1}{2}r^2 + \frac{1}{2}r^3 & 0 \\
4r^2 - 4r^3 & -r^2 + 2r^3 & 0
\end{bmatrix}, \tag{424b}
$$

and this is not power-bounded unless $r \leq 1.69562076955986$, a zero of the polynomial $r^3 - r^2 - 2$.

As an example of the alternative technique, in which the numerical method is modified to allow for irregular mesh spacing, consider the BDF3 method. Suppose that approximate solution values are known at x_{n-1}, $x_n - h(1 + r_2^{-1})$ and $x_n - h(1 + r_2^{-1} + (r_2 r_1)^{-1})$, where r_2 and r_1 are the most recent stepsize ratios. We now wish to compute $y(x_n)$ using a formula of the form

$$
y(x_n) \approx h\beta y'(x_n) + \alpha_1(r_1, r_2)y(x_n - h) + \alpha_2(r_1, r_2)y(x_n - h(1 + r_2^{-1}))
$$
$$
+ \alpha_3(r_1, r_2)y(x_n - h(1 + r_2^{-1} + (r_2 r_1)^{-1})).
$$

Using a result equivalent to Hermite interpolation, we find that, to maintain third order accuracy,

$$
\alpha_1 = \frac{(r_2 + 1)^2(r_1 r_2 + r_1 + 1)^2}{(3r_2^2 r_1 + 4r_1 r_2 + 2r_2 + r_1 + 1)(r_1 + 1)},
$$

$$
\alpha_2 = -\frac{r_2^2(r_1 r_2 + r_1 + 1)^2}{3r_2^2 r_1 + 4r_1 r_2 + 2r_2 + r_1 + 1},
$$

$$
\alpha_3 = \frac{r_2^2 r_1^3(r_2 + 1)^2}{(3r_2^2 r_1 + 4r_1 r_2 + 2r_2 + r_1 + 1)(r_1 + 1)}.
$$

Stability of this variable stepsize version of the BDF3 method will hinge on the boundedness of products of matrices of the form

$$M = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

where the values of r_1 and r_2 for successive members of the product sequence are appropriately linked together.

An extreme case will be where r_1 and r_2 are equal and as large as possible, subject to M having bounded powers. It is easy to verify that this greatest rate of continual increase in stepsize corresponds to

$$r_1 = r_2 = r^* = \frac{1 + \sqrt{5}}{2}.$$

It is interesting that an arbitrary sequence of stepsize change ratios, in the interval $(0, r^*]$, still guarantees stable behaviour.

Exercises 42

42.1 Let $C(\theta)$ denote the error constant for the third order linear multistep method $(1-(1-\theta)z-\theta z^2, \frac{5-\theta}{12} + \frac{2+2\theta}{3} + \frac{5\theta-1}{12}z^2)$. Show that $C = \frac{1-\theta}{24(1+\theta)}$.

42.2 Show that weakly stable behaviour is experienced with the linear multistep method $(1 - z^3, \frac{3}{8}(1 + z)^3)$.

42.3 Show that the norm of the product of an arbitrary sequence of matrices of the form (424b) is bounded as long as each r lies in the interval $[0, r^*]$, where $r^* \approx 1.69562076955986$.

43   Stability Characteristics

430   Introduction

In contrast to Runge–Kutta methods, in which stability regions are determined by a single stability function, the stability properties of linear multistep methods are inextricably bound up with difference equations. We consider the example of the second order Adams–Bashforth method

$$y_n = y_{n-1} + \frac{3}{2}hf(x_{n-1}, y_{n-1}) - \frac{1}{2}hf(x_{n-2}, y_{n-2}). \tag{430a}$$

For the differential equation $y' = qy$, this becomes

$$y_n = y_{n-1} + \frac{3}{2}hqy_{n-1} - \frac{1}{2}hqy_{n-2},$$
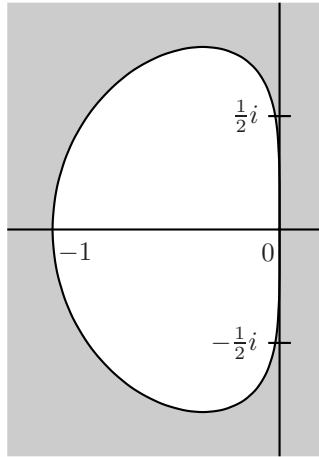
Figure 430(i)   Stability region for the second order Adams–Bashforth method

so that stable behaviour occurs if $hq = z$, where z is such that the equation

$$y_n = \left(1 + \frac{3}{2}z\right)y_{n-1} - \frac{1}{2}zy_{n-2}$$

has only bounded solutions. This occurs when the polynomial equation

$$w^2 - \left(1 + \frac{3}{2}z\right)w + \frac{1}{2}z = 0$$

has each of its two solutions in the closed unit disc and in the interior if they happen to coincide. The stability region for this method turns out to be the unshaded part of the complex plane shown in Figure 430(i), including the boundary.

Just as for Runge–Kutta methods, a consistent explicit linear multistep method has a bounded stability region and therefore cannot be A-stable. We therefore explore implicit methods as a source of appropriate algorithms for the solution of stiff problems. It will be found that A-stability is a very restrictive property in that it is incompatible with an order greater than 2. Also in this section, we consider a non-linear stability property, known as G-stability, which is a multistep counterpart of algebraic stability introduced in Chapter 3.

431   *Stability regions*

For a linear multistep method $[\alpha, \beta]$, the difference equation associated with the linear test problem, $y' = qy$, is

$$(1 - z\beta_0)y_n - (\alpha_1 + z\beta_1)y_{n-1} - (\alpha_2 + z\beta_1)y_{n-2} - \cdots$$
$$- (\alpha_k + z\beta_k)y_{n-k} = 0, \quad \text{(431a)}$$

and the stability region is the set of points hq in the complex plane for which (431a) has only bounded solutions as $n \to \infty$. To simplify the discussion, we will consider the *interior* of the stability region so that, for z in this set, all solutions to (431a) converge to zero as $n \to \infty$. We will refer to this interior set as the open stability region. Write the difference equation in the form

$$\alpha(E^{-1}) - z\beta(E^{-1}) = 0,$$

and we see that the open stability region can be defined in terms of the relation

$$\alpha(w^{-1}) - z\beta(w^{-1}) = 0. \quad \text{(431b)}$$

That is, z is in the open stability region if there does not exist w outside the open unit disc such that the pair (z, w) satisfies (431b). Stated another way, this means that if w outside the open unit disc this implies that z satisfying (431b) is *not* in the open stability region.

As a starting point in determining the stability region, it is convenient to evaluate the points on the boundary of the unit circle and to note that the mapping

$$w \mapsto \frac{\alpha(w^{-1})}{\beta(w^{-1})} \quad \text{(431c)}$$

traces out a set of points which includes the boundary of the stability region. In particular cases it is easy to determine the exact boundary. Since $w \mapsto w^{-1}$ maps the unit circle to itself, while changing the sense of rotation, it is equivalent to replace (431c) by

$$w \mapsto \frac{\alpha(w)}{\beta(w)}. \quad \text{(431d)}$$

This procedure is known as the 'boundary locus method' for determining stability regions, and we give some examples of its use in the next subsection.

A second procedure for determining stability regions is based on the idea of the 'type of a polynomial'. That is, if P is a polynomial of degree n then the type is a triple (n_1, n_2, n_3), where n_1, n_2 and n_3 are non-negative integers with sum exactly n. The interpretation is that n_1 is the number of zeros of P

in the open unit disc, n_2 is the number of zeros on the unit circle and n_3 is the number of zeros outside the closed unit disc. If we are willing to concentrate on the open stability region of a specific method, we can simplify the discussion to the question of determining whether or not the type of P is $(n, 0, 0)$. We will refer to such a polynomial as being 'strongly stable'. Polynomials can be tested for this property recursively, using the following result:

Theorem 431A *A polynomial P_n, given by*

$$P_n(w) = a_0 w^n + a_1 w^{n-1} + \cdots + a_{n-1} w + a_n,$$

where $a_0 \neq 0$ and $n \geq 2$, is strongly stable if and only if

$$|a_0|^2 > |a_n|^2 \tag{431e}$$

and P_{n-1} is strongly stable, where

$$\begin{aligned} &P_{n-1}(w) \\ &= (\overline{a}_0 a_0 - a_n \overline{a}_n) w^{n-1} + (\overline{a}_0 a_1 - a_n \overline{a}_{n-1}) w^{n-2} + \cdots + (\overline{a}_0 a_{n-1} - a_n \overline{a}_1). \end{aligned}$$

Proof. First note that (431e) is necessary for strong stability because if it were not true, the product of the zeros could not have a magnitude less than 1. Hence, we assume that this is the case and it remains to prove that P_n is strongly stable if and only if the same property holds for P_{n-1}. It is easy to verify that

$$w P_{n-1}(w) = \overline{a}_0 P_n(w) - a_n w^n \overline{P_n(\overline{w}^{-1})}.$$

By Rouché's theorem, $w P_{n-1}(w)$ has n zeros in the open unit disc if and only if the same property is true for $P_n(w)$, and the result follows. $\quad\square$

The result of this theorem is often referred to as the Schur criterion. In the case of $n = 2$, it leads to the two conditions

$$|a_0|^2 - |a_2|^2 > 0, \tag{431f}$$

$$(|a_0|^2 - |a_2|^2)^2 - |\overline{a}_0 a_1 - a_2 \overline{a}_1|^2 > 0. \tag{431g}$$

To apply the Schur criterion to the determination of the stability region for a k-step method, we need to ask for which z the polynomial given by

$$P(w) = w^k (\alpha(w^{-1}) - z\beta(w^{-1}))$$

is strongly stable. We present some examples of the use of this test in Subsection 433.

Algorithm 432α   Boundary locus method for low order Adams–Bashforth
methods

```
% Second order
% ------------
w = exp(i*linspace(0,2*pi));
z = 2*w.*(w-1)./(3*w-1);
plot(z)


% Third order
% -----------
w=exp(i*linspace(0,2*pi));
z=12*(1-w)./(23*w-16*w.^2+5*w.^3);
plot(z)


% Fourth order
% ------------
w=exp(i*linspace(0,2*pi));
z=24*(1-w)./(55*w-59*w.^2+37*w.^3-9*w.^4);
plot(z)
```

432   Examples of the boundary locus method

The first example is for the second order Adams–Bashforth method (430a)
for which (431c) takes the form

$$w \mapsto \frac{1 - w^{-1}}{\frac{3}{2}w^{-1} - \frac{1}{2}w^{-2}}.$$

For $w = \exp(i\theta)$ and $\theta \in [0, 2\pi]$, for points on the unit circle, we have z values
on the (possibly extended) boundary of the stability region given by

$$z = \frac{\exp(2i\theta) - \exp(i\theta)}{\frac{3}{2}\exp(i\theta) - \frac{1}{2}}.$$

The MATLAB code given in Algorithm 432α shows how this is done, and the
boundary traced out is exactly as in Figure 430(i).

No confusion is possible as to which part of the complex plane divided by
the boundary locus is the inside and which is the outside because, using an
argument based on the Cauchy–Riemann equations, we note that the inside
is always to the left of the path traced out as w increases from 0 to 2π. If
we had used (431d) in place of (431c) then, of course, the path would have
been traced in the opposite direction and the inside of the stability region
would have been on the right. Note that in Algorithm 432α the third and
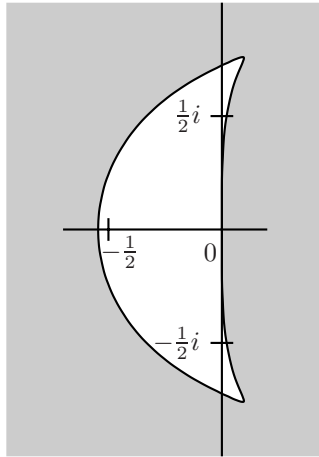
Figure 432(i) Stability region for the third order Adams–Bashforth method
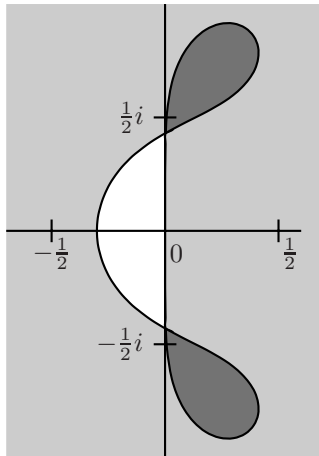
Figure 432(ii) Stability region for the fourth order Adams–Bashforth method

fourth order cases are traced in the reverse direction. The stability region of the third Adams–Bashforth method, as computed by this algorithm, is given as the unshaded region of Figure 432(i).

In the case of the fourth order method in this family, the root locus method traces out more than the boundary of the stability region, as we see in Figure 432(ii). Because crossing the locus corresponds to the shift of one of the growth factors from stable to unstable, the more heavily shaded region is doubly unstable in that it contains two unstable terms.
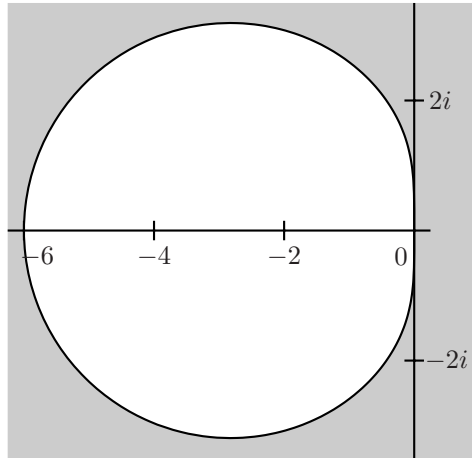
Figure 432(iii)   Stability region for the third order Adams–Moulton method
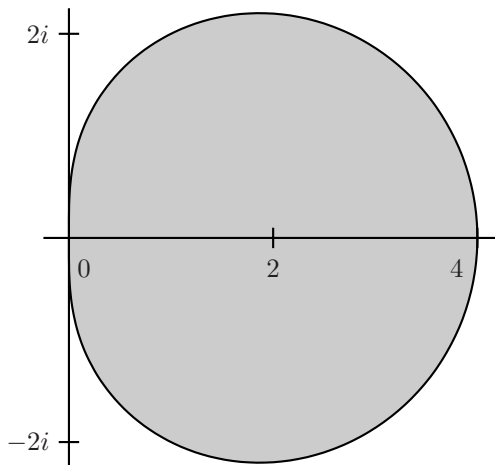
Figure 432(iv)   Stability region for the second order backward difference method

We present three final examples. The Adams–Moulton method of order 3 is given in Figure 432(iii); we see that even though this method is implicit it has a bounded stability region.

Now look at the stability regions of the backward difference methods of orders 2 and 3. The first of these, shown in Figure 432(iv), indicates that the second order method is A-stable and the second, Figure 432(v), shows that the third order method is *not* A-stable.
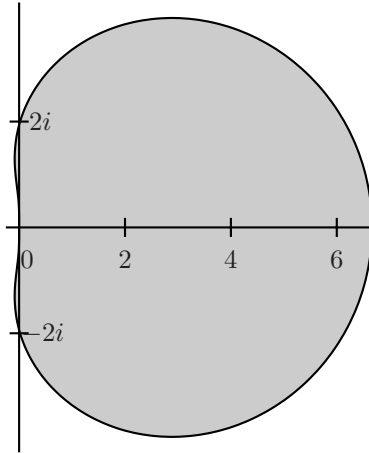
Figure 432(v) Stability region for the third order backward difference method

433 An example of the Schur criterion

We first recompute the stability region of the second order Adams–Bashforth method. We need to find for what values of the complex number z the polynomial $a_0 w^2 + a_1 w + a_2$ has its zeros in the open unit disc, where

$$a_0 = 1, \qquad a_1 = -1 - \frac{3}{2}z, \qquad a_2 = \frac{z}{2}.$$

The condition $|a_0|^2 - |a_2|^2 > 0$ is equivalent to

$$|z| < 2, \tag{433a}$$

while the second condition $(|a_0|^2 - |a_2|^2)^2 - |\overline{a}_0 a_1 - a_2 \overline{a}_1|^2 > 0$ simplifies to

$$\operatorname{Re}(z)(3|z|^2 - 4) < |z|^4. \tag{433b}$$

It is easy to verify that (433b) implies (433a). Thus, by plotting the points for which (433b) holds, we recover Figure 430(i).

434 Stability of predictor–corrector methods

We consider examples of PEC and PECE methods. For the PEC method based on second order Adams–Bashforth as predictor and Adams–Moulton as corrector, we have the following equations for the predicted and corrected values:

$$y_n^* = y_{n-1} + \frac{3}{2}h f_{n-1}^* - \frac{1}{2}h f_{n-2}^*, \tag{434a}$$

$$y_n = y_{n-1} + \frac{1}{2}h f_n^* + \frac{1}{2}h f_{n-1}^*. \tag{434b}$$
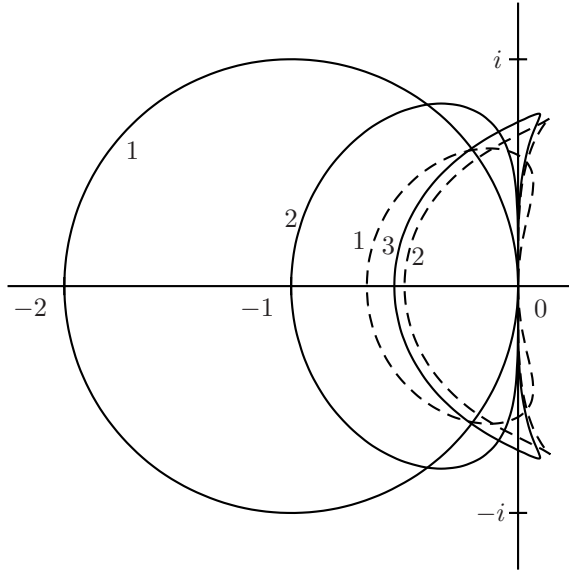
Figure 434(i)   Stability regions for Adams–Moulton methods (solid lines) and
PEC methods (dashed lines)

Superficially, this system describes two sequences, the y and the y^* which develop together. However, it is only the y^* sequence that has derivative values associated with it. Hence, the y sequence can conveniently be eliminated from consideration. Replace n by $n + 1$ in (434a), and we find

$$y^*_{n+1} = y_n + \frac{3}{2}hf^*_n - \frac{1}{2}hf^*_{n-1}. \tag{434c}$$

Add (434b) to this equation and subtract (434a), and we find

$$y^*_{n+1} = y^*_n + 2hf^*_n - \frac{3}{2}hf^*_{n-1} + \frac{1}{2}hf^*_{n-2}.$$

Apart from the actual values of the coefficients, this resembles an Adams–Bashforth method, and its stability region can be found in a similar way. If β^* and β are the respective generating polynomials for an order p Adams–Bashforth method and the corresponding Adams–Moulton method, then the general form of the generating polynomial for y^* in a PEC method is equal to $\widehat{\beta}$, where

$$\widehat{\beta}(z) = \beta^*(z) + \beta_0 z(1 - z)^p.$$

The value of β_0 could be replaced by any value we wish without sacrificing the order p. In fact, it could be replaced by the value of $(-1)^p \beta^*_{p+1}$ so that the method would actually be of order $p+1$. It would in this case be precisely
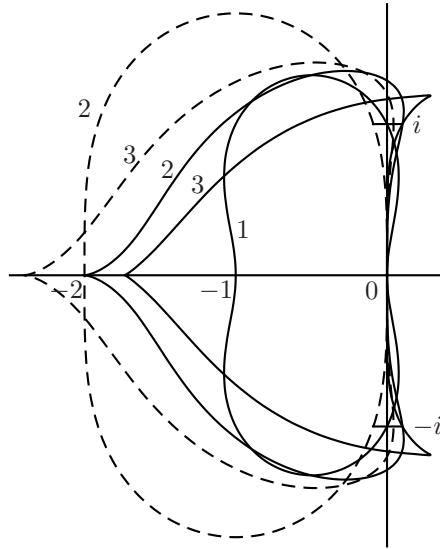
Figure 434(ii)   Stability regions for PECE methods with $q = p$ (solid lines) and
$q = p - 1$ methods (dashed lines). In each case p is attached to the curves

the order $p + 1$ Adams–Bashforth method. Another advantage of pushing the
order up one unit rather than accepting the standard PEC result, is that the
stability region seems to be less desirable for PEC methods. This is illustrated
in Figure 434(i), where the boundaries of some of these regions are shown.

PECE methods are more interesting because two derivatives are computed
in each step. Thus they are in reality two-stage general linear methods. From
the stability point of view, they can be analysed by eliminating y_n^* so that the
method

$$y_n^* = y_{n-1} + h \sum_{i=1}^{k} \beta_i^* f_{n-i},$$

$$y_n = y_{n-1} + h \beta_0 f_n^* + h \sum_{i=1}^{k} \beta_i f_{n-i}$$

yields the difference equation

$$y_n = (1 + (\beta_0 + \beta_1)z + \beta_0 \beta_1^* z^2) y_{n-1} + \sum_{i=2}^{k} (\beta_i z + \beta_0 \beta_i^* z^2) y_{n-i}.$$

Note that the step k may be higher for the predictor than for the corrector
but we assume that, if this is the case, sufficient zero values are added to the
sequence of β_i values to make the two k values effectively equal. In practice

there are two options. Either both the predictor and corrector have the same order p, in which case $k = p$ for the predictor and $k = p - 1$ for the corrector; or $k = p - 1$ for both predictor and corrector; in this case the predictor has order only $p - 1$. The boundaries of the stability regions are shown in Figure 434(ii) for each of these cases.

The relatively more generous stability regions for the PECE methods, when compared with PEC methods, for $p > 1$ are regarded as constituting a significant advantage in carrying out a final evaluation in implementations of predictor–corrector methods. Similar comparisons apparently favour PECECE over PECE methods.

Exercises 43

43.1 Use the Schur criterion to show that all zeros of the polynomial $7z^3 - 11z^2 + 5z + 1$ lie in the unit disc.

43.2 Use the Schur criterion to show that not all zeros of the polynomial $7z^3 - 11z^2 + 6z + 1$ lie in the unit disc.

43.3 Determine whether or not all zeros of the polynomial $7z^3 - 11z^2 + (5 + i)z + 1$ lie in the unit disc.

43.4 Find the stability regions for the PEC and PECE methods based on the fourth order Adams–Bashforth and Adams–Moulton methods.

44   Order and Stability Barriers

440   Survey of barrier results

It is a simple matter to construct a linear k-step method with order $2k$. This can be done, for example, by finding coefficients A_j, B_j, $j = 0, 1, \ldots, k$, such that

$$\frac{1}{z^2(z+1)^2(z+2)^2 \cdots (z+k)^2} = \sum_{j=0}^{k} \frac{A_j}{z+j} + \sum_{j=0}^{k} \frac{B_j}{(z+j)^2}$$

and then defining

$$\alpha_j = -\frac{A_j}{A_0}, \quad j = 1, 2, \ldots, k, \qquad \beta_j = \frac{B_j}{A_0}, \quad j = 0, 1, \ldots, k.$$

To justify this remark, consider the contour integral

$$\frac{1}{2\pi i} \oint_C \frac{\phi(z)dz}{\prod_{j=0}^{k}(z+j)^2} = \sum_{j=0}^{k} \frac{1}{2\pi i} \oint_C \phi(z) \left(\frac{A_j}{z+j} + \frac{B_j}{(z+j)^2} \right) dz,$$

where the contour C consists of a counter-clockwise circle of radius $R > k$ and centre at the origin and ϕ is a polynomial of degree not exceeding $2k$. By

taking R large the value of the integral can be estimated by $O(R^{-1})$; because it is constant, it must be zero. On the other hand, the terms in the partial fraction representation of the integral are

$$\sum_{j=0}^{k} \left(A_j \phi(-j) + B_j \phi'(-j)\right).$$

For example, if $k = 3$, we have

$$\frac{1}{z^2(z+1)^2(z+2)^2} = -\frac{11}{108}\frac{1}{z} - \frac{1}{4}\frac{1}{z+1} + \frac{1}{4}\frac{1}{z+2} + \frac{11}{108}\frac{1}{z+3}$$
$$+ \frac{1}{36}\frac{1}{z^2} + \frac{1}{4}\frac{1}{(z+1)^2} + \frac{1}{4}\frac{1}{(z+2)^2} + \frac{1}{36}\frac{1}{(z+3)^2},$$

leading to the values

$$\alpha_1 = -\frac{27}{11}, \quad \alpha_2 = \frac{27}{11}, \quad \alpha_3 = 1,$$

so that the method is unstable.

This is an example of a result found by Dahlquist (1956), that order p is impossible for a convergent method unless $p \leq k+1$ if k is odd, and $p \leq k+2$ if k is even.

With the recognition of the importance of stiffness came the property of A-stability (Dahlquist, 1963). It has been shown, also by Dahlquist, for A-stable linear multistep methods that p cannot exceed 2. This result is known as the second Dahlquist barrier, in contrast to the result about the order of a convergent k-step method, which is usually referred to as the first Dahlquist barrier.

441   Maximum order for a convergent k-step method

As a starting point for the proof we present of the Dahlquist first barrier, use Theorem 410B. Modify this by substituting z in (410d) with the function

$$\frac{2z}{1-z}$$

and then multiplying throughout by $(1+z)^k$. We then have

$$(1+z)^k \alpha\left(\frac{1-z}{1+z}\right) - \log\left(\frac{1+z}{1-z}\right)(1+z)^k \beta\left(\frac{1-z}{1+z}\right) = O(z^{p+1}),$$

or, what is equivalent,

$$\frac{(1+z)^k \alpha\left(\frac{1-z}{1+z}\right)}{z} \frac{z}{\log\left(\frac{1+z}{1-z}\right)} - (1+z)^k \beta\left(\frac{1-z}{1+z}\right) = O(z^p). \qquad (441a)$$

For the rest of this subsection, including assumptions within lemmas and theorems, we write

$$a(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_k z^k = (1+z)^k \alpha\left(\tfrac{1-z}{1+z}\right),$$

$$b(z) = b_0 + b_1 z + b_2 z^2 + \cdots + b_k z^k = (1+z)^k \beta\left(\tfrac{1-z}{1+z}\right).$$

By consistency, $a_0 = 0$ so that (441a) can be written in the form

$$(a_1 + a_2 z + \cdots + a_k z^{k-1})(c_0 + c_2 z^2 + c_4 z^4 + \cdots)$$
$$- (b_0 + b_1 z + b_2 z^2 + \cdots + b_k z^k) = O(z^p),$$

where

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

The way we use this result, when we consider the possibility that $p > k$, is to note that this implies that the coefficients of z^{k+1}, ..., z^{p-1} in

$$(a_1 + a_2 z + \cdots + a_k z^{k-1})(c_0 + c_2 z^2 + c_4 z^4 + \cdots) \qquad (441b)$$

are zero.

We will go about this is by establishing some results on the signs of the coefficients a_1, a_2, ..., a_k, c_2, c_4, ...

Lemma 441A *If the method under consideration is stable then $a_1 > 0$ and $a_i \geq 0$, for $i = 2, 3, \ldots, k$.*

Proof. Write the polynomial a in the form

$$a(z) = (1+z)^k - \alpha_1(1+z)^{k-1}(1-z) - \alpha_2(1+z)^{k-2}(1-z)^2 - \cdots - \alpha_k(1-z)^k.$$

We calculate the value of a_1, the coefficient of z, to be

$$k - (k-2)\alpha_1 - (k-4)\alpha_2 - \cdots - (-k)\alpha_k = k\alpha(1) - 2\alpha'(1) = -2\alpha'(1),$$

because $\alpha(1) = 0$. The polynomial ρ, which we recall is defined by

$$\rho(z) = z^k - \alpha_1 z^{k-1} - \alpha_2 z^{k-2} - \cdots - \alpha_k,$$

has no real zeros greater than 1, and hence, because $\rho(1) = 0$ and because $\lim_{z->\infty} \rho(z) = \infty$, it is necessary that $\rho'(1) > 0$. Calculate this to be

$$\rho'(1) = k - (k-1)\alpha_1 - (k-2)\alpha_2 - \cdots - \alpha_{k-1} = a_1.$$

This completes the proof that $a_1 > 0$.

Write ζ for a possible zero of a so that, because of the relationship between this polynomial and α, it follows that

$$\frac{1-\zeta}{1+\zeta}$$

is a zero of α, unless it happens that $\zeta = -1$, in which case there is a drop in the degree of α. In either case, we must have $\mathrm{Re}(\zeta) \leq 0$. Because all zeros of a are real, or occur in conjugate pairs, the polynomial a can be decomposed into factors of the form $z - \xi$ or of the form $z^2 - 2\xi z + (\xi^2 + \eta^2)$, where the real number ξ cannot be positive. This means that all factors have only terms with coefficients of the same sign, and accordingly this also holds for a itself. These coefficients must in fact be non-negative because $a_1 > 0$. $\qquad\square$

Lemma 441B *The coefficients c_2, c_4, ... are all negative.*

Proof. Using the series for $\log\big((1+z)/(1-z)\big)/z$, we see that c_0, c_2, c_4, ... satisfy

$$\Big(2 + \frac{2}{3}z^2 + \frac{2}{5}z^4 + \cdots\Big)(c_0 + c_2 z^2 + c_4 z^4 + \cdots) = 1. \qquad (441c)$$

It follows that $c_0 = \frac{1}{2}$, $c_2 = -\frac{1}{6}$. We prove $c_{2n} < 0$ by induction for $n = 2$, $n = 3$, ... If $c_{2i} < 0$ for $i = 1, 2, \ldots, n-1$ then we multiply (441c) by $2n + 1 - (2n-1)z^2$. We find

$$\sum_{i=0}^{\infty} d_{2i} z^{2i} \cdot \sum_{i=0}^{\infty} c_{2i} z^{2i} = 2n + 1 - (2n-1)z^2, \qquad (441d)$$

where, for $i = 1, 2, \ldots, n$,

$$d_{2i} = \frac{2(2n+1)}{2i+1} - \frac{2(2n-1)}{2i-1} = -\frac{8(n-i)}{(2i+1)(2i-1)},$$

so that $d_{2i} < 0$, for $i = 1, 2, \ldots, n-1$, and $d_{2n} = 0$. Equate the coefficients of z^{2n} in (441d) and we find that

$$c_{2n} = -\frac{c_2 d_{2n-2} + c_4 d_{2n-4} + \cdots + c_{2n-2} d_2}{d_0} < 0. \qquad\square$$

We are now in a position to prove the Dahlquist barrier result.

Theorem 441C *Let $[\alpha, \beta]$ denote a stable linear multistep method with order p. Then*

$$p \leq \begin{cases} k+1, & k \text{ odd}, \\ k+2, & k \text{ even}. \end{cases}$$

Proof. Consider first the case k odd and evaluate the coefficient of z^{k+1} in (441b). This equals

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

and, because no term is positive, the total can be zero only if each term is zero. However, this would mean that $a_1 = 0$, which is inconsistent with stability.

In the case k even, we evaluate the coefficient of z^{k+2} in (441b). This is

$$a_{k-1} c_4 + a_{k-3} c_6 + \cdots + a_1 c_{k+2}.$$

Again, every term is non-positive and because the total is zero, it again follows that $a_1 = 0$ which contradicts the assumption of stability. $\qquad\square$

There is some interest in the methods with maximal order $2k + 2$, for k even. For these methods, α has all its zeros on the unit circle. This evidently gives the methods a symmetry that suggests it might be advantageous to use them for problems whose behaviour is dominated by linear terms with purely imaginary eigenvalues. Against this possible advantage is the observation that the stability regions necessarily have empty interiors.

442 Order stars for linear multistep methods

In their historic paper, Wanner, Hairer and Nørsett (1978) introduced order stars on Riemann surfaces. Suppose that $\Phi(w, z)$ is a polynomial function of two complex variables, $w \in W$ and $z \in Z$. We assume that $Z = W = \mathbb{C}$. The subset R_Φ of $W \times Z$ defined by the relation $\Phi(w, z) = 0$ is a Riemann surface. Suppose that Φ has degree r in w and s in z. We may interpret R as a mapping from the Z plane which takes $z \in Z$ to the set of zeros of the equation $\Phi(w, z) = 0$ or as a mapping which takes $w \in W$ to the set of zeros of this same equation, but with z now the unknown. The main interpretation will be that $\Phi(w, z)$ is the characteristic polynomial $\det(wI - M(z))$ of the stability matrix of a multivalue method. If this method has order p then $\Phi(\exp(z), z) = O(z^{p+1})$. For ease of notation, we carry over concepts such as A-stability from multivalue methods, such as linear multistep methods, to the functions Φ used to characterize their stability.

Definition 442A *The function Φ is A-stable if R_Φ has no intersection with the product set*

$$\{w \in \mathbb{C} : |w| > 1\} \times \{z \in \mathbb{C} : \text{Re}(z) \le 0\}.$$

This definition is equivalent to the requirement that for any z in the left half complex plane, all eigenvalues of the stability matrix are in the closed unit disc. Just as in the case of Runge–Kutta methods, for which the Riemann surface has only a single sheet, scaling the eigenvalues by $\exp(-z)$ does not affect the behaviour on the imaginary axis or introduce or remove any poles.
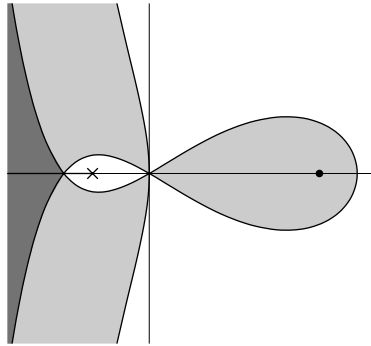
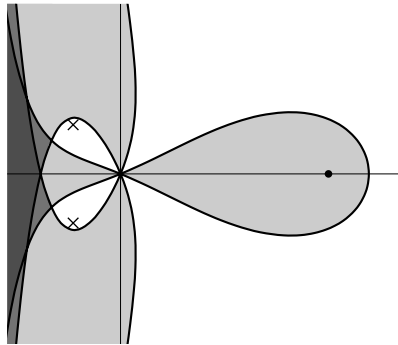**Figure 442(i)** Order star for the second order BDF method

Figure 442(ii) Order star for the third order BDF method

Hence we can consider a modified Riemann surface based on the function $\Phi(w \exp(z), z)$. Just as for the Runge–Kutta case, one of the sheets, known as the 'principal sheet', behaves like $w = 1 + O(z^{p+1})$ and order stars appear.

We illustrate this by considering the case of the second order backward difference method, for which

$$\Phi(w \exp(z), z) = \left(1 - \frac{2}{3}z\right) \exp(2z)w^2 - \frac{4}{3}\exp(z)w + \frac{1}{3},$$

and the third order backward difference method, for which

$$\Phi(w \exp(z), z) = \left(1 - \frac{6}{11}z\right)\exp(3z)w^3 - \frac{18}{11}\exp(2z)w^2 + \frac{9}{11}\exp(z)w - \frac{2}{11}.$$

For the second order case, shown in Figure 442(i), a pole at $z = \frac{3}{2}$ is marked, together with a branch point at $z = -\frac{1}{2}$. Note that for $z \in (\infty, -\frac{1}{2})$, the two roots of the equation $\Phi(w \exp(z), z) = 0$, for all z in this real interval, have equal magnitudes. In this figure, light shading grey indicates that a region

has exactly one of the sheets with magnitude greater than 1. A darker grey is used to indicate that *both* sheets have magnitudes greater than 1.

This method is A-stable, as we already know. This can be seen from the order star by noting that the only pole is in the right half-plane, and that the fingers do not intersect the imaginary axis. On the other hand, the third order method (Figure 442(ii)) is not A-stable because, in this case, the intersection of the imaginary axis with one the fingers is now not empty. Note that for the third order case, there is a single pole at $z = \frac{11}{6}$ and that three shades of grey are used to distinguish regions where one, two or three sheets have magnitudes greater than 1.

Although A-stable Runge–Kutta methods can have arbitrarily high orders, the order of A-stable linear multistep methods is restricted to 2. This was first proved using order stars (Wanner, Hairer and Nørsett, 1978), but we will use the closely related approach of order arrows (Butcher, 2002). These will be introduced in the Riemann surface case in the next subsection.

443   Order arrows for linear multistep methods

Given a relationship between complex numbers z and w defined by an equation of the form

$$\Phi(w \exp(z), z) = 0,$$

we can define order arrows as the set of points for which w is real and positive. In particular, the order arrows that emanate from zero correspond to w with increasing real parts (the up arrows) and, on these arrows, $w \in (1, \infty)$, or decreasing real parts (the down arrows) and for which $w \in [0, 1)$.

Order arrows on Riemann surfaces are illustrated for the BDF2 method (Figure 443(i)) and for the BDF3 method (Figure 443(ii)). Just as for Runge–Kutta methods, the up arrows either terminate at the pole $z = \beta_0^{-1}$ or at $-\infty$, and down arrows terminate at the zero $z = -\alpha_k \beta_k^{-1}$ or at $+\infty$. In interpreting these remarks, we need to allow for the possibility that the path traced out by an up or down arrow meets another arrow at a branch point of the Riemann surface. However, this special case is easily included in the general rule with a possible freedom to choose between two continuations of the incoming arrow.

The 'principal sheet' of the Riemann surface will refer to a neighbourhood of $(0, 1)$ for which the relationship between z and w is injective; that is, it behaves as though w is a function of z. As long as $\Phi(w, 0)$ has only a single zero with value $w = 1$, this idea makes sense. On the principal sheet, $w \exp(z) = \exp(z) + O(z^{p+1})$, and the behaviour at zero is similar to what happens for one-step methods. These simple ideas are enough to prove the Dahlquist second order bound:

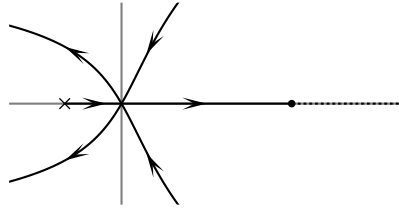**Theorem 443A** *An A-stable linear multistep method cannot have order greater than 2.*

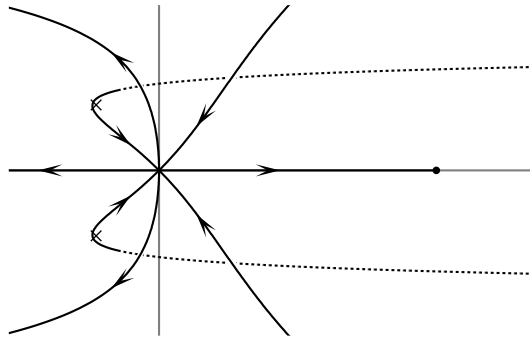**Figure 443(i)**   Order arrows for order 2 BDF method

Figure 443(ii)   Order arrows for order 3 BDF method

Proof. If the order were greater than 2, there would be more than three up arrows emanating from the origin. At least three of these up arrows would come out in the positive direction (or possibly would be tangential to the imaginary axis). Since there is only one pole, at least two of these arrows would cross the imaginary axis (or be tangential to it). Hence, the stability region does not include all of the imaginary axis and the method is not A-stable. $\qquad\square$

We can make this result more precise by obtaining a bound on the error constant for second order A-stable methods. The result yields an optimal role for the second order Adams–Moulton method, for which the error constant is $-\frac{1}{12}$, because

$$\exp(z) - \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z} = -\frac{1}{12}z^3 + O(z^4).$$

It is not possible to obtain a positive error constant amongst A-stable second order methods, and it is not possible to obtain an error constant smaller in magnitude than for the one-step Adams–Moulton method. To prove the result we use, in place of $\exp(z)$, the special stability function $(1 + \frac{1}{2}z)/(1 - \frac{1}{2}z)$ in forming a relative stability function.

Theorem 443B *Let C denote the error constant for an A-stable second order linear multistep method. Then*

$$C \le -\frac{1}{12},$$

with equality only in the case of the second order Adams–Moulton method.

Proof. Consider the relation

$$\Phi\left(w\frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}, z\right) = 0.$$

On the principal sheet, $w = 1 - (C + \frac{1}{12})z^3 + O(z^4)$. It is not possible that $C + \frac{1}{12} = 0$, because there would then be at least four up arrows emanating from 0 and, as in the proof of Theorem 443A, this is impossible because there is at most one pole in the right half-plane. On the other hand, if $C + \frac{1}{12} > 0$, there would be at least two up arrows emanating from zero in the positive direction and these must cross the imaginary axis.   □

Exercises 44

44.1 Show that, for a stable linear multistep method with order $k + 2$, all zeros of α are on the unit circle.

44.2 Show that the BDF3 method is not A-stable, by selecting a complex number x with negative real part for which the corresponding difference equation is not stable.

45   One-Leg Methods and G-stability

450   The one-leg counterpart to a linear multistep method

In Dahlquist (1976) one-leg methods were introduced. Given a linear multistep method defined by the generating polynomial pair $[\alpha, \beta]$, an alternative method can be found by replacing the weighted sum of derivative values

$$h\beta_0 f(x_n, y_n) + h\beta_1 f(x_{n-1}, y_{n-1}) + \cdots + h\beta_k f(x_{n-k}, y_{n-k}),$$

by the single term

$$h\left(\sum_{i=0}^{k}\beta_i\right) f\left(x_n - \theta h, \left(\sum_{i=0}^{k}\beta_i\right)^{-1}\sum_{i=0}^{k}\beta_i y_{n-i}\right),$$

where θ is a weighted combination of the step numbers

$$\theta = \frac{\sum_{i=0}^{k} i\beta_i}{\sum_{i=0}^{k}\beta_i}.$$

For convenience, we write

$$\widehat{\beta}_i = \frac{\beta_i}{\sum_{i=0}^{k} \beta_i}, \qquad i = 0, 1, 2, \ldots, k.$$

It is obvious that the linear stability of a one-leg method is the same as for the corresponding linear multistep method. However, it is possible to investigate the stability of numerical solutions of non-linear dissipative equations in a relatively simple way if the computation is carried out using one-leg methods. By contrast, the corresponding analysis for linear multistep methods becomes hopelessly complicated because of the occurrence of the same derivative terms in several steps in sequence.

Even though these stability results are derived for one-leg methods, they can be regarded as having a relevance to linear multistep method, because of a transformation that links them.

In later papers by Dahlquist and others (Dahlquist, 1983; Wantanabe and Sheikh, 1984; Hundsdorfer and Steininger, 1991), the feasibility of using one-leg methods directly, as a practical numerical algorithm, came into serious consideration. In this brief introduction to these methods, we also discuss an interpretation in terms of effective order, and review the main results on G-stability.

451   The concept of G-stability

We recall the non-linear stability property introduced in Subsection 357. The corresponding property for one-leg methods was introduced in Dahlquist (1976) and given the name G-stability. For convenience, we consider applications only to autonomous problems

$$y'(x) = f(y(x)), \tag{451a}$$

and we assume that the dissipativity property holds in the sense that solution values lie in an N-dimensional inner-product space, and that

$$\langle f(u) - f(v), u - v \rangle \le 0, \tag{451b}$$

for all $u, v \in \mathbb{R}^N$.

For Runge–Kutta methods, in the study of the non-linear stability property applicable to those methods, in Subsection 357, it was possible to use the norm $\|u\| = \sqrt{\langle u, u \rangle}$ to measure the drift between two approximately equal numerical approximations that takes place in step n. However, for linear k-step methods, each of the k subvectors making up the current state vector of each approximate solution has to be taken into account. Hence, we need to construct a suitable norm on the vector space $\mathbb{R}^{kN}$.

For $U \in \mathbb{R}^{kN}$, write U_i, $i = 1, 2, \ldots, k$, for subvectors in $\mathbb{R}^N$. That is,

$$U = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_k \end{bmatrix}.$$

If $U, V \in \mathbb{R}^{kN}$ then, given a positive definite symmetric $k \times k$ matrix G, we can define an inner product $\langle \cdot \rangle_G$ by

$$\langle U, V \rangle_G = \sum_{i,j=1}^{k} g_{ij} \langle U_i, V_j \rangle,$$

with corresponding norm

$$\|U\|_G = \sqrt{\sum_{i,j=1}^{k} g_{ij} \langle U_i, U_j \rangle}.$$

The aim of G-stability is to discover, for a given one-leg method, if G exists so that, for a problem satisfying (451b),

$$\left\|Y^{(n)} - Z^{(n)}\right\|_G^2 - \left\|Y^{(n-1)} - Z^{(n-1)}\right\|_G^2 \tag{451c}$$

cannot be positive, where

$$Y^{(n)} = \begin{bmatrix} y_n \\ y_{n-1} \\ y_{n-2} \\ \vdots \\ y_{n-k+1} \end{bmatrix}, \qquad Z^{(n)} = \begin{bmatrix} z_n \\ z_{n-1} \\ z_{n-2} \\ \vdots \\ z_{n-k+1} \end{bmatrix},$$

and the y and z sequences are numerical approximations corresponding to two different solutions to (451a).

The only inequality at our disposal that could be used to ensure that (451c) is not positive is the dissipativity requirement applied to the only evaluations of f that take place in the step. That is, we can use the fact that

$$\left\langle f\left(\sum_{i=0}^{k} \widehat{\beta}_i y_{n-i}\right) - f\left(\sum_{i=0}^{k} \widehat{\beta}_i z_{n-i}\right), \sum_{i=0}^{k} \widehat{\beta}_i (y_{n-i} - z_{n-i}) \right\rangle \leq 0. \tag{451d}$$

Because

$$y_n - \sum_{i=1}^{k} \alpha_i y_{n-i} = \left(\sum_{i=0}^{k} \beta_i\right)^{-1} f\left(\sum_{i=0}^{k} \widehat{\beta}_i y_{n-i}\right),$$

with a similar formula for the z sequence, it follows that

$$\left\langle y_n - z_n - \sum_{i=1}^{k} \alpha_i(y_{n-i} - z_{n-i}), \sum_{i=0}^{k} \beta_i(y_{n-i} - z_{n-i}) \right\rangle \leq 0,$$

and this will imply that (451c) has the correct sign if G can be selected so that the $(k+1) \times (k+1)$ matrix M is positive semi-definite, where

$$M = \alpha\beta^{\mathsf{T}} + \beta\alpha^{\mathsf{T}} - \begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & G \end{bmatrix}, \tag{451e}$$

where, in this context, α and β are the vectors

$$\alpha = \begin{bmatrix} 1 \\ -\alpha_1 \\ -\alpha_2 \\ \vdots \\ -\alpha_k \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}.$$

Let m_{ij}, $i, j = 0, 1, 2, \ldots, k$, denote the elements of M. For any vector $U \in \mathbb{R}^{(k+1)N}$, the fact that M is positive semi-definite implies that

$$\sum_{i,j=0}^{k} m_{ij} \langle U_i, U_j \rangle \geq 0.$$

Choose the vector

$$U = \begin{bmatrix} y_n - z_n \\ y_{n-1} - z_{n-1} \\ y_{n-2} - z_{n-2} \\ \vdots \\ y_{n-k+1} - z_{n-k+1} \\ y_{n-k} - z_{n-k} \end{bmatrix},$$

and we have the identity

$$\sum_{i,j=0}^{k} m_{ij} \langle y_{n-i} - z_{n-i}, y_{n-j} - z_{n-j} \rangle$$
$$= 2 \left\langle y_n - z_n - \sum_{i=1}^{k} \alpha_i(y_{n-i} - z_{n-i}), \sum_{i=0}^{k} \beta_i(y_{n-i} - z_{n-i}) \right\rangle$$
$$+ \| Y^{(n-1)} - Z^{(n-1)} \|_G^2 - \| Y^{(n)} - Z^{(n)} \|_G^2.$$

If the left-hand side is non-negative, and the first term on the right is non-positive, it follows that

$$\|Y^{(n)} - Z^{(n)}\|_G \leq \|Y^{(n-1)} - Z^{(n-1)}\|_G.$$

The positive semi-definiteness of M was recognized by Dahlquist (1976) as just the right condition to identify methods that behave stably for the type of non-linear problem we are considering. Accordingly we state the following definition:

Definition 451A *A one-leg method* $[\alpha, \beta]$ *is 'G-stable' if* M *given by* (451e) *is positive semi-definite.*

We present the example of the BDF2 method with

$$[\alpha(z), \beta(z)] = \left(1 - \frac{4}{3}z + \frac{1}{3}z^2, \frac{2}{3}\right).$$

Write

$$G = \begin{bmatrix} g_{11} & g_{12} \\ g_{12} & g_{22} \end{bmatrix}$$

and we find

$$M = \begin{bmatrix} \frac{4}{3} - g_{11} & -\frac{8}{9} - g_{12} & \frac{2}{9} \\ -\frac{8}{9} - g_{12} & g_{11} - g_{22} & g_{12} \\ \frac{2}{9} & g_{12} & g_{22} \end{bmatrix},$$

which is positive semi-definite if and only if G is the positive definite matrix

$$G = \begin{bmatrix} \frac{10}{9} & -\frac{4}{9} \\ -\frac{4}{9} & \frac{2}{9} \end{bmatrix}.$$

452   Transformations relating one-leg and linear multistep methods

Denote the point at which the derivative is calculated in step n of a one-leg method by $\widehat{y}_n$. Also denote the corresponding x argument as $\widehat{x}_n$. Hence, we have

$$\widehat{x}_n = x_n - \frac{\sum_{i=0}^{k} i\beta_i}{\sum_{i=0}^{k} \beta_i} h, \tag{452a}$$

$$\widehat{y}_n = \left(\sum_{i=0}^{k} \beta_i\right)^{-1} \sum_{i=0}^{k} \beta_i y_{n-i}, \tag{452b}$$

$$y_n = \sum_{i=1}^{k} \alpha_{n-i} y_{n-i} + \left(\sum_{i=0}^{k} \beta_i\right) f(\widehat{x}_n, \widehat{y}_n). \tag{452c}$$

Form a linear combination of $\widehat{y}_{n-i}$, $i = 0, 1, \ldots, k$, given by (452b), based on the coefficients in the α polynomial, and note that the operators $\alpha(E^{-1})$ and $\beta(E^{-1})$ are commutative. We have

$$\widehat{y}_n - \sum_{i=1}^{k} \alpha_i \widehat{y}_{n-i} = h \sum_{i=1}^{k} \beta_i f(\widehat{x}_n, \widehat{y}_n). \qquad (452\text{d})$$

The relationship between the y and $\widehat{y}$ sequences given by (452b) and (452d) was suggested by Dahlquist (1976) as an indication that stability questions for a linear multistep method can be replaced by similar questions for the corresponding one-leg method.

453   Effective order interpretation

The concept of effective order, introduced in Subsections 365 and 389, gives an alternative interpretation of the relationship between the computed approximation and the exact solution.

Define the function $\gamma(z)$ by

$$\gamma(z) = \left(\sum_{i=0}^{k} \widehat{\beta}_i \exp(-iz) \right)^{-1} = \gamma_0 + \gamma_1 z + \gamma_2 z^2 + \cdots,$$

where $\gamma_0 = 1$, and the starting approximation by

$$S(y)(x) = \sum_{i=0}^{p} \gamma_i h^i y^{(i)}(x),$$

assuming the linear multistep method $[\alpha, \beta]$ has order p. Write $\widehat{y}(x) = S(y)(x)$. We then have

$$\widehat{y}(x_n) - \sum_{i=1}^{k} \alpha_i \widehat{y}(x_{n-i}) = h \sum_{i=1}^{k} \beta_i f\left(x_n, \sum_{i=0}^{k} \widehat{\beta}_i \widehat{y}(x_{n-i}) \right) + O(h^{p+1}).$$

454   Concluding remarks on G-stability

It might be suspected that amongst A-stable linear multistep methods, G-stable methods stand out as being in some way superior. Such considerations turn out to be vacuous because a linear multistep method is A-stable if and only if it is G-stable. That G-stable methods are A-stable is shown simply as in Theorem 454A below. However, the converse result is much deeper. This was proved in Dahlquist (1978). Now the easy result:

Theorem 454A *A G-stable linear multistep method is A-stable.*

Proof. We use the criterion that if $|w| < 1$, then $z = \alpha(w)/\beta(w)$ is in the right half-plane. Form the inner product W^*MW, where M is the matrix given by (451e) and

$$W = \begin{bmatrix} 1 \\ w \\ w^2 \\ \vdots \\ w^k \end{bmatrix}.$$

We find that

$$\alpha(\overline{w})\beta(w) + \alpha(w)\beta(\overline{w}) = W^*MW + (1 - |w|^2) \sum_{j,l=1}^{k} g_{jl}\overline{w}^{j-1}w^{l-1} > 0,$$

so that $\mathrm{Re}\big(\alpha(w)/\beta(w)\big) > 0$.   □

<div align="center">

Exercises 45

</div>

45.1 Show that the method defined by $\alpha(z) = 1 - \frac{3}{2}z + \frac{1}{2}z^2$, $\beta(z) = \frac{3}{4} - \frac{1}{4}z$, is G-stable, by finding the corresponding matrix G.

45.2 Show that if $q_1 + iq_2$ is in the left half-plane, then the differential equation

$$y'(x) = qy(x)$$

can be written as a system

$$\begin{bmatrix} y_1'(x) \\ y_2'(x) \end{bmatrix} = \begin{bmatrix} q_1 & -q_2 \\ q_2 & q_1 \end{bmatrix} \begin{bmatrix} y_1(x) \\ y_2(x) \end{bmatrix},$$

where $y(x) = y_1(x) + iy_2(x)$. Furthermore, show that this system satisfies (451b), using the usual inner product.

46   Implementation Issues

460   Survey of implementation considerations

In addition to the basic algorithm giving the value of y_n in terms of y_{n-1}, y_{n-2}, ..., y_{n-k}, effective use of linear multistep methods requires further tools. We have already discussed, albeit briefly, the starting process for a method with fixed order and fixed stepsize. However, linear multistep methods are seldom

used in such a manner. It is usually efficient to adapt both the stepsize and the order to suit local behaviour of the computed solution, and this leads to the need for representations of the methods that will make adaptivity possible. Given that a variable order implementation is going to be used, it is easier to start at order 1 and build the order upwards as the solution develops. Reducing order is relatively easy and also needs to be built in as an option within a variable order formulation.

It is natural to make a comparison between implementation techniques for Runge–Kutta methods and for linear multistep methods. Unlike for explicit Runge–Kutta methods, interpolation and error estimation are regarded as straightforward for linear multistep methods. Not only is it possible to obtain an asymptotically correct estimate of the local truncation error, but it is a simple extension of the approximation technique to obtain a usable approximation for the local error that might have been expected if the next higher order had instead been used.

461   Representation of data

After a number of steps, with constant size h, have been carried out using an order p method, for example by a PECE combination of Adams–Bashforth and Adams–Moulton methods, approximations are available to $y(x_n)$, $hy'(x_n)$, $hy'(x_{n-1})$, ..., $hy'(x_{n-p+1})$. If the stepsize is to be altered by a factor r to a new value rh, then there seem to be two distinct approaches to proceeding further.

The first approach is to use a modified form of the Adams formulae which enables $y(x_n + rh)$ to be written in terms of $y(x_n)$, $hy'(x_n)$, $hy'(x_{n-1})$, ..., $hy'(x_{n-p+1})$. Of course this only works for a single step. For the step after that, the data on which to base the approximation would be $y(x_n + hr)$, $hy'(x_n + hr)$, $hy'(x_n)$, ..., $hy'(x_{n-p+2})$ and the results computed would be approximations to $y(x_n + hr + hr\hat{r})$, where $\hat{r}$ is the stepsize ratio for this new step. Rather than explore the form of the modified Adams formula in this rather ad hoc manner, write the exact quantities that the incoming data is supposed to approximate as the sequence consisting of

$$y(x_n - h\theta_1), \quad hy'(x_n - h\theta_1), \quad hy'(x_n - h\theta_2), \quad \ldots, \quad hy'(x_n - h\theta_k).$$

The Adams–Bashforth method would then generalize to an approximation of the form

$$y(x_n) \approx y(x_n - h\theta_1) + \sum_{i=1}^{k} \beta_i^* hy'(x_n - h\theta_i), \tag{461a}$$

and the Adams–Moulton to an approximation of the form

$$y(x_n) \approx \beta_0 hy'(x_n) + y(x_n - h\theta_1) + \sum_{i=1}^{k} \beta_i hy'(x_n - h\theta_i). \tag{461b}$$

To obtain order $p = k$ for (461a), the coefficients β_i^*, $i = 1, 2, \ldots, k$, have to be chosen so that

$$1 = \exp(-\theta_1 z) + z \sum_{i=1}^{k} \beta_i^* \exp(-\theta_i z) + O(z^{p+1}),$$

and to obtain order $p = k+1$ for (461b), β_i, $i = 1, 2, \ldots, k$, are chosen so that

$$1 = \exp(-\theta_1 z) + z\beta_0 + z \sum_{i=1}^{k} \beta_i \exp(-\theta_i z) + O(z^{p+1}).$$

To use this approach in practice, the coefficients β_1^*, β_2^*, ... and β_0, β_1, ... have to be evaluated afresh every step, before any differential equation solutions are approximated. For many problems this is justified, and many codes use some sort of approach based on this technique.

The second main approach to stepsize adjustment was proposed by Nordsieck (1962) and further developed by Gear (1967, 1971). For a Nordsieck method of order p, the data imported into step n consists of approximations to

$$y(x_{n-1}), \quad hy'(x_{n-1}), \quad \frac{1}{2!}h^2 y''(x_{n-1}), \quad \frac{1}{p!}h^p y^{(p)}(x_{n-1}),$$

and the quantities exported from this step are approximations to

$$y(x_n), \quad hy'(x_n), \quad \frac{1}{2!}h^2 y''(x_n), \quad \frac{1}{p!}h^p y^{(p)}(x_n). \tag{461c}$$

Note that the factors $(i!)^{-1}$ are inserted for convenience. When a stepsize change from h to rh is required, the simple adjustment of scaling the quantities in (461c) by powers of the scale factor r is used. This means that they become approximations to

$$y(x_n), \quad rhy'(x_n), \quad \frac{1}{2!}(rh)^2 y''(x_n), \quad \frac{1}{p!}(rh)^p y^{(p)}(x_n).$$

Denote the vector of Nordsieck approximations imported into step n by

$$\eta_0^{[n-1]} \approx y(x_{n-1}),$$

$$\eta_1^{[n-1]} \approx hy'(x_{n-1}),$$

$$\eta_2^{[n-1]} \approx \frac{1}{2!}h^2 y''(x_{n-1}),$$

$$\vdots \qquad \vdots$$

$$\eta_p^{[n-1]} \approx \frac{1}{p!}h^p y^{(p)}(x_{n-1}),$$

Table 461(I)   Coefficients, $\gamma_0, \gamma_1, \ldots, \gamma_p$, for Nordsieck methods

	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$	$p = 7$	$p = 8$
γ_0	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$\frac{95}{288}$	$\frac{19087}{60480}$	$\frac{5257}{17280}$
γ_1	1	1	1	1	1	1	1
γ_2	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{11}{12}$	$\frac{25}{24}$	$\frac{137}{120}$	$\frac{49}{40}$	$\frac{363}{280}$
γ_3		$\frac{1}{6}$	$\frac{1}{3}$	$\frac{35}{72}$	$\frac{5}{8}$	$\frac{203}{270}$	$\frac{469}{540}$
γ_4			$\frac{1}{24}$	$\frac{5}{48}$	$\frac{17}{96}$	$\frac{49}{192}$	$\frac{967}{2880}$
γ_5				$\frac{1}{120}$	$\frac{1}{40}$	$\frac{7}{144}$	$\frac{7}{90}$
γ_6					$\frac{1}{720}$	$\frac{7}{1440}$	$\frac{23}{2160}$
γ_7						$\frac{1}{5040}$	$\frac{1}{1260}$
γ_8							$\frac{1}{40320}$

so that the result computed by the Adams–Bashforth predictor will be

$$y_n^* = \eta_0^{[n-1]} + \eta_1^{[n-1]} + \cdots + \eta_p^{[n-1]}.$$

If an approximation is also required for the scaled derivative at x_n, this can be found from the formula, also based on a Taylor expansion,

$$hy'(x_n) \approx \eta_1^{[n-1]} + 2\eta_2^{[n-1]} + \cdots + p\eta_p^{[n-1]}. \tag{461d}$$

To find the Nordsieck equivalent to the Adams–Moulton corrector formula, it is necessary to add β_0 multiplied by the difference between the corrected value of the scaled derivative and the extrapolated value computed by (461d). That is, the corrected value of $\eta_0^{[n]}$ becomes

$$\eta_0^{[n]} = \beta_0 \Delta_n + \eta_0^{[n-1]} + \eta_1^{[n-1]} + \cdots + \eta_p^{[n-1]},$$

where

$$\Delta_n = hf(x_n, y_n^*) - \sum_{i=1}^{s} i\eta_i^{[n-1]}.$$

In this formulation we have assumed a PECE mode but, if further iterations are carried out, the only essential change will be that the second argument of $hf(x_n, y_n^*)$ will be modified.

For constant stepsize, the method should be equivalent to the Adams predictor–corrector pair and this means that all the output values will be modified in one way or another from the result that would have been formed by simple extrapolation from the incoming Nordsieck components. Thus we can write the result computed in a step as

$$
\begin{bmatrix}
\eta_0^{[n]} \\
\eta_1^{[n]} \\
\eta_2^{[n]} \\
\vdots \\
\eta_{p-1}^{[n]} \\
\eta_p^{[n]}
\end{bmatrix}
=
\begin{bmatrix}
\gamma_0 \\
\gamma_1 \\
\gamma_2 \\
\vdots \\
\gamma_{p-1} \\
\gamma_p
\end{bmatrix}
\Delta_n +
\begin{bmatrix}
1 & 1 & 1 & \cdots & 1 & 1 \\
0 & 1 & 2 & \cdots & p-1 & p \\
0 & 0 & 1 & \cdots & \binom{p-1}{2} & \binom{p}{2} \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & p \\
0 & 0 & 0 & \cdots & 0 & 1
\end{bmatrix}
\begin{bmatrix}
\eta_0^{[n-1]} \\
\eta_1^{[n-1]} \\
\eta_2^{[n-1]} \\
\vdots \\
\eta_{p-1}^{[n-1]} \\
\eta_p^{[n-1]}
\end{bmatrix}.
\tag{461e}
$$

The quantities γ_i, $i = 0, 1, 2, \ldots, p$, have values determined by the equivalence with the standard fixed stepsize method and we know at least that

$$
\gamma_0 = \beta_0, \qquad \gamma_1 = 1.
$$

The value selected for γ_1 ensures that $\eta_1^{[n]}$ is precisely the result evaluated from $\eta_0^{[n]}$ using the differential equation. We can arrive at the correct values of γ_2, $\ldots$, γ_p, by the requirement that the matrix

$$
\begin{bmatrix}
1 & 3 & \cdots & \binom{p-1}{2} & \binom{p}{2} \\
0 & 1 & \cdots & \binom{p-1}{3} & \binom{p}{3} \\
\vdots & \vdots & & \vdots & \vdots \\
0 & 0 & \cdots & 1 & p \\
0 & 0 & \cdots & 0 & 1
\end{bmatrix}
-
\begin{bmatrix}
\gamma_2 \\
\gamma_3 \\
\vdots \\
\gamma_{p-1} \\
\gamma_p
\end{bmatrix}
\begin{bmatrix}
2 & 3 & \cdots & p-1 & p
\end{bmatrix}
$$

has zero spectral radius.

Values of the coefficients γ_i, $i = 0, 1, \ldots, p$, are given in Table 461(I) for $p = 2, 3, \ldots, 8$.

Adjustment of stepsize is carried out by multiplying the vector of output approximations formed in (461e) at the completion of step n, by the diagonal matrix $D(r)$ before the results are accepted as input to step $n + 1$, where

$$
D(r) = \operatorname{diag}(1, r, r^2, \ldots, r^p).
$$

It was discovered experimentally by Gear that numerical instabilities can result from using this formulation. This can be seen in the example $p = 3$, where we find the values $\gamma_2 = \frac{3}{4}$, $\gamma_3 = \frac{1}{6}$ Stability is determined by products of matrices of the form

$$
\begin{bmatrix}
-\frac{1}{2}r^2 & \frac{3}{4}r^2 \\
-\frac{1}{3}r^3 & \frac{1}{2}r^3
\end{bmatrix},
$$

and for $r \geq 1.69562$, this matrix is no longer power-bounded.

Gear's pragmatic solution was to prohibit changes for several further steps after a stepsize change had occurred. An alternative to this remedy will be considered in the next subsection.

462   Variable stepsize for Nordsieck methods

The motivation we have presented for the choice of γ_1, γ_2, ... in the formulation of Nordsieck methods was to require a certain matrix to have zero spectral radius. Denote the vector γ and the matrix V by

$$
\gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{bmatrix}, \qquad
V = \begin{bmatrix}
1 & 2 & 3 & \cdots & p \\
0 & 1 & 3 & \cdots & \frac{1}{2}p(p-1) \\
0 & 0 & 1 & \cdots & \frac{1}{6}p(p-1)(p-2) \\
\vdots & \vdots & \vdots & & \vdots \\
0 & 0 & 0 & \cdots & 1
\end{bmatrix},
$$

and denote by e_1^T the basis row vector $e_1^\mathsf{T} = [\,1 \quad 0 \quad \cdots \quad 0\,]$. The characteristic property of γ is that the matrix

$$(I - \gamma e_1^\mathsf{T})V \tag{462a}$$

has zero spectral radius. When variable stepsize is introduced, the matrix in (462a) is multiplied by $D(r) = \mathrm{diag}(r, r^2, r^3, \ldots, r^p)$ and, as we have seen, if γ is chosen on the basis of constant h, there is a deterioration in stable behaviour. We consider the alternative of choosing γ as a function of r so that

$$\rho(D(r)(I - \gamma e_1^\mathsf{T})V) = 0.$$

The value of γ_1 still retains the value 1 but, in the only example we consider, $p = 3$, it is found that

$$\gamma_2 = \frac{1 + 2r}{2(1 + r)}, \qquad \gamma_3 = \frac{r}{3(1 + r)},$$

and we have

$$
D(r)(I - \gamma e_1^\mathsf{T})V = \begin{bmatrix}
0 & 0 & 0 \\
0 & -\frac{r^3}{1+r} & \frac{3r^2}{2(1+r)} \\
0 & -\frac{2r^4}{3(1+r)} & \frac{r^3}{2(1+r)}
\end{bmatrix}. \tag{462b}
$$

It is obvious that this matrix is power-bounded for all positive values of r. However, if a sequence of n steps is carried out with stepsize changes r_1, r_2, ..., r_n then the product of matrices of the form given by (462b) for these values of r to be analysed to determine stability. The spectral radius of such a product is found to be

$$\frac{|r_1 - r_n|r_1^2}{1 + r_1} \cdot \frac{|r_2 - r_1|r_2^2}{1 + r_2} \cdot \frac{|r_3 - r_2|r_3^2}{1 + r_3} \cdots \frac{|r_n - r_{n-1}|r_n^2}{1 + r_n},$$

and this will be bounded by 1 as long as $r_i \in [0, r^\star]$, where $r^\star$ has the property that

$$\frac{r_1 r_2 |r_2 - r_1|}{\sqrt{(1 + r_1)(1 + r_2)}} \leq 1, \quad \text{whenever } r_1, r_2 \in [0, r^\star].$$

It is found after some calculations that stability, in the sense of this discussion, is achieved if $r^\star \approx 2.15954543$.

463   Local error estimation

The standard estimator for local truncation error is based on the Milne device. That is, the difference between the predicted and corrected values provides an approximation to some constant multiplied by $h^{p+1} y^{(p+1)}(x_n)$, and the local truncation error can be estimated by multiplying this by a suitable scale factor.

This procedure has to be interpreted in a different way if, as in some modern codes, the predictor and corrector are accurate to different orders. We no longer have an asymptotically correct approximation to the local truncation error but to the error in the predictor, assuming this has the lower order. Nevertheless, stepsize control based on this approach often gives reliable and useful performance.

To allow for a possible increase in order, estimation is also needed for the scaled derivative one order higher than the standard error estimator. It is very difficult to do this reliably, because any approximation will be based on a linear combination of $hy'(x)$ for different x arguments. These quantities in turn will be of the form $hf(x, y(x) + Ch^{p+1} + O(h^{p+2}))$, and the terms of the form $Ch^{p+1} + O(h^{p+2})$ will distort the result obtained. However, it is possible to estimate the scaled order $p+2$ derivative reliably, at least if the stepsize has been constant over recent steps, by forming the difference of approximations to the order $p+1$ derivative over two successive steps. If the stepsize has varied moderately, the approximation this approximation will still be reasonable. In any case, if the criterion for increasing order turns out to be too optimistic for any specific problem, then after the first step with the new order a rejection is likely to occur, and the order will either be reduced again or else the stepsize will be lowered while still maintaining the higher order.

Exercises 46

46.1 Show how to write $y(x_n + rh)$ in terms of $y(x_n)$, $hy'(x_n)$ and $hy'(x_n - h)$, to within $O(h^3)$. Show this approximation might be used to generalize the order 2 Adams–Bashforth method to variable stepsize.

46.2 How should the formulation of Subsection 461 be modified to represent Adams–Bashforth methods?

Chapter 5

General Linear Methods

50 Representing Methods in General Linear Form

500 Multivalue–multistage methods

The systematic computation of an approximation to the solution of an initial value problem usually involves just two operations: evaluation of the function f defining the differential equation and the forming of linear combinations of previously computed vectors. In the case of implicit methods, further complications arise, but these can also be brought into the same general linear formulation.

We consider methods in which a collection of vectors forms the input at the beginning of a step, and a similar collection is passed on as output from the current step and as input into the following step. Thus the method is a multivalue method, and we write r for the number of quantities processed in this way. In the computations that take place in forming the output quantities, there are assumed to be s approximations to the solution at points near the current time step for which the function f needs to be evaluated. As for Runge–Kutta methods, these are known as stages and we have an s-stage or, in general, multistage method.

The intricate set of connections between these quantities make up what is known as a general linear method. Following Burrage and Butcher (1980), we represent the method by four matrices which we will generally denote by A, U, B and V. These can be written together as a partitioned $(s+r) \times (s+r)$ matrix

$$\begin{bmatrix} A & U \\ B & V \end{bmatrix}.$$

The input vectors available at step n will be denoted by $y_1^{[n-1]}$, $y_2^{[n-1]}$, ..., $y_r^{[n-1]}$. During the computations which constitute the step, stage values Y_1, Y_2, ..., Y_s, are computed and derivative values $F_i = f(Y_i)$, $i = 1, 2, \ldots, s$, are computed in terms of these. Finally, the output values are computed and, because these will constitute the input at step $n+1$, they will be denoted by

$y_i^{[n]}$, $i = 1, 2, \ldots, r$. The relationships between these quantities are defined in terms of the elements of A, U, B and V by the equations

$$Y_i = \sum_{j=1}^{s} a_{ij} h F_j + \sum_{j=1}^{r} u_{ij} y_j^{[n-1]}, \quad i = 1, 2, \ldots, s, \tag{500a}$$

$$y_i^{[n]} = \sum_{j=1}^{s} b_{ij} h F_j + \sum_{j=1}^{r} v_{ij} y_j^{[n-1]}, \quad i = 1, 2, \ldots, r. \tag{500b}$$

It will be convenient to use a more concise notation, and we start by defining vectors $Y, F \in \mathbb{R}^{sN}$ and $y^{[n-1]}, y^{[n]} \in \mathbb{R}^{rN}$ as follows:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix}, \quad F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_s \end{bmatrix}, \quad y^{[n-1]} = \begin{bmatrix} y_1^{[n-1]} \\ y_2^{[n-1]} \\ \vdots \\ y_r^{[n-1]} \end{bmatrix}, \quad y^{[n]} = \begin{bmatrix} y_1^{[n]} \\ y_2^{[n]} \\ \vdots \\ y_r^{[n]} \end{bmatrix}.$$

Using these supervectors, it is possible to write (500a) and (500b) in the form

$$\begin{bmatrix} Y \\ y^{[n]} \end{bmatrix} = \begin{bmatrix} A \otimes I_N & U \otimes I_N \\ B \otimes I_N & V \otimes I_N \end{bmatrix} \begin{bmatrix} hF \\ y^{[n-1]} \end{bmatrix}. \tag{500c}$$

In this formulation, I_N denotes the $N \times N$ unit matrix and the Kronecker product is given by

$$A \otimes I_N = \begin{bmatrix} a_{11} I_N & a_{12} I_N & \cdots & a_{1s} I_N \\ a_{21} I_N & a_{22} I_N & \cdots & a_{2s} I_N \\ \vdots & \vdots & & \vdots \\ a_{s1} I_N & a_{s2} I_N & \cdots & a_{ss} I_N \end{bmatrix}.$$

When there is no possibility of confusion, we simplify the notation by replacing

$$\begin{bmatrix} A \otimes I_N & U \otimes I_N \\ B \otimes I_N & V \otimes I_N \end{bmatrix} \quad \text{by} \quad \begin{bmatrix} A & U \\ B & V \end{bmatrix}.$$

In Subsections 502–505, we illustrate these ideas by showing how some known methods, as well as some new methods, can be formulated in this manner. First, however, we will discuss the possibility of transforming a given method into one using a different arrangement of the data passed from step to step.

501   Transformations of methods

Let T denote a non-singular $r \times r$ matrix. Given a general linear method characterized by the matrices (A, U, B, V), we consider the construction of a second method for which the input quantities, and the corresponding output quantities, are replaced by linear combinations of the subvectors in $y^{[n-1]}$ (or in $y^{[n]}$, respectively). In each case the rows of T supply the coefficients in the linear combinations. These ideas are well known in the case of Adams methods, where it is common practice to represent the data passed between steps in a variety of configurations. For example, the data imported into step n may consist of approximations to $y(x_{n-1})$ and further approximations to $hy'(x_{n-i})$, for $i = 1, 2, \ldots, k$. Alternatively it might, as in Bashforth and Adams (1883), be expressed in terms of $y(x_{n-1})$ and of approximations to a sequence of backward differences of the derivative approximations. It is also possible, as proposed in Nordsieck (1962), to replace the approximations to the derivatives at equally spaced points in the past by linear combinations which will approximate scaled first and higher derivatives at x_{n-1}.

Let $z_i^{[n-1]}$, $i = 1, 2, \ldots, r$, denote a component of the transformed input data where

$$z_i^{[n-1]} = \sum_{j=1}^{r} t_{ij} y_j^{[n-1]}, \qquad z_i^{[n]} = \sum_{j=1}^{r} t_{ij} y_j^{[n]}.$$

This transformation can be written more compactly as

$$z^{[n-1]} = T y^{[n-1]}, \qquad z^{[n]} = T y^{[n]}.$$

Hence the method which uses the y data and the coefficients (A, U, B, V), could be rewritten to produce formulae for the stages in the form

$$Y = hAF + Uy^{[n-1]} = hAF + UT^{-1} z^{[n-1]}. \tag{501a}$$

The formula for $y^{[n]} = hBF + Vy^{[n-1]}$, when transformed to give the value of $z^{[n]}$, becomes

$$z^{[n]} = T(hBF + Vy^{[n-1]}) = h(TB)F + (TVT^{-1})z^{[n-1]}. \tag{501b}$$

Combine (501a) and (501b) into the single formula to give

$$\begin{bmatrix} Y \\ z^{[n]} \end{bmatrix} = \begin{bmatrix} A & UT^{-1} \\ TB & TVT^{-1} \end{bmatrix} \begin{bmatrix} hF \\ z^{[n-1]} \end{bmatrix}.$$

Thus, the method with coefficient matrices $(A, UT^{-1}, TB, TVT^{-1})$ is related to the original method (A, U, B, V) by an equivalence relationship with a natural computational significance. The significance is that a sequence of approximations, using one of these formulations, can be transformed into the sequence that would have been generated using the alternative formulation.

It is important to ensure that any definitions concerning the properties of a generic general linear method transform in an appropriate manner, when the coefficient matrices are transformed.

Even though there may be many interpretations of the same general linear method, there may well be specific representations which have advantages of one sort or another. Some examples of this will be encountered later in this section.

502   Runge–Kutta methods as general linear methods

Since Runge–Kutta methods have a single input, it is usually convenient to represent them, as general linear methods, with $r = 1$. Assuming the input vector is an approximation to $y(x_{n-1})$, it is only necessary to write $U = \mathbf{1}$, $V = 1$, write B as the single row b^T of the Runge–Kutta tableau and, finally, identify A with the $s \times s$ matrix of the same name also in this tableau.

A very conventional and well-known example is the classical fourth order method

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

which, in general linear formulation, is represented by the partitioned matrix

$$
\left[
\begin{array}{cccc|c}
0 & 0 & 0 & 0 & 1 \\
\frac{1}{2} & 0 & 0 & 0 & 1 \\
0 & \frac{1}{2} & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 1 \\
\hline
\frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & 1
\end{array}
\right].
$$

A more interesting example is the Lobatto IIIA method

$$
\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\
1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\
\hline
& \frac{1}{6} & \frac{2}{3} & \frac{1}{6}
\end{array}
$$

for which the straightforward representation, with $s = 3$ and $r = 1$, is misleading. The reason is that the method has the 'FSAL property' in the sense that the final stage evaluated in a step is identical with the *first* stage of the following step. It therefore becomes possible, and even appropriate, to

use a representation with $s = r = 2$ which expresses, quite explicitly, that the FSAL property holds. This representation would be

$$
\left[
\begin{array}{cc|cc}
\frac{1}{3} & -\frac{1}{12} & 1 & \frac{5}{12} \\
\frac{2}{3} & \frac{1}{6} & 1 & \frac{1}{6} \\
\hline
\frac{2}{3} & \frac{1}{6} & 1 & \frac{1}{6} \\
0 & 1 & 0 & 0
\end{array}
\right] , \tag{502a}
$$

and the input quantities are supposed to be approximations to

$$
y_1^{[n-1]} \approx y(x_{n-1}), \quad y_2^{[n-1]} \approx hy'(x_{n-1}).
$$

Finally, we consider a Runge–Kutta method introduced in Subsection 322, with tableau

$$
\begin{array}{c|cccc}
0 & & & & \\
-\frac{1}{2} & -\frac{1}{2} & & & \\
\frac{1}{2} & \frac{3}{4} & -\frac{1}{4} & & \\
1 & -2 & 1 & 2 & \\
\hline
& \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6}
\end{array} . \tag{502b}
$$

As we pointed out when the method was introduced, it can be implemented as a two-value method by replacing the computation of the second stage derivative by a quantity already computed in the previous step. The method is now not equivalent to any Runge–Kutta method but, as a general linear method, it has coefficient matrix

$$
\left[
\begin{array}{ccc|cc}
0 & 0 & 0 & 1 & 0 \\
\frac{3}{4} & 0 & 0 & 1 & -\frac{1}{4} \\
-2 & 2 & 0 & 1 & 1 \\
\hline
\frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 1 & 0 \\
0 & 1 & 0 & 0 & 0
\end{array}
\right] . \tag{502c}
$$

503   Linear multistep methods as general linear methods

For a linear k-step method $[\alpha, \beta]$ of the special form $\alpha(z) = 1 - z$, the natural way of writing this as a general linear method is to choose $r = k + 1$, $s = 1$ and the input approximations as

$$
y^{[n-1]} \approx
\left[
\begin{array}{c}
y(x_{n-1}) \\
hy'(x_{n-1}) \\
hy'(x_{n-2}) \\
\cdots \\
hy'(x_{n-k})
\end{array}
\right] .
$$

The matrix representing the method now becomes

$$
\left[
\begin{array}{c|ccccccc}
\beta_0 & 1 & \beta_1 & \beta_2 & \beta_3 & \cdots & \beta_{k-1} & \beta_k \\
\hline
\beta_0 & 1 & \beta_1 & \beta_2 & \beta_3 & \cdots & \beta_{k-1} & \beta_k \\
1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0
\end{array}
\right].
$$

Because $y_1^{[n-1]}$ and $y_{k+1}^{[n-1]}$ occur in the combination $y_1^{[n-1]} + \beta_k y_{k+1}^{[n-1]}$ in each of the two places where these quantities are used, we might try to simplify the method by transforming using the matrix

$$
T =
\left[
\begin{array}{cccccc}
1 & 0 & 0 & \cdots & 0 & \beta_k \\
0 & 1 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0 \\
0 & 0 & 0 & \cdots & 0 & 1
\end{array}
\right].
$$

The transformed coefficient matrices become

$$
\left[
\begin{array}{cc}
A & UT^{-1} \\
TB & TVT^{-1}
\end{array}
\right]
=
\left[
\begin{array}{c|ccccccc}
\beta_0 & 1 & \beta_1 & \beta_2 & \beta_3 & \cdots & \beta_{k-1} & 0 \\
\hline
\beta_0 & 1 & \beta_1 & \beta_2 & \beta_3 & \cdots & \beta_{k-1}+\beta_k & 0 \\
1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0
\end{array}
\right],
$$

and we see that it is possible to reduce r from $k+1$ to k, because the $(k+1)$th input vector is never used in the calculation.

The well-known technique of implementing an implicit linear multistep method by combining it with a related explicit method to form a predictor–corrector pair fits easily into a general linear formulation. Consider, for example, the PECE method based on the third order Adams–Bashforth and Adams–Moulton predictor–corrector pair. Denote the predicted

approximation by y_n^* and the corrected value by y_n. We then have

$$y_n^* = y_{n-1} + \frac{23}{12}hf(x_{n-1}, y_{n-1}) - \frac{4}{3}hf(x_{n-2}, y_{n-2}) + \frac{5}{12}hf(x_{n-3}, y_{n-3}),$$

$$y_n = y_{n-1} + \frac{5}{12}hf(x_n, y_n^*) + \frac{2}{3}hf(x_{n-1}, y_{n-1}) - \frac{1}{12}hf(x_{n-2}, y_{n-2}).$$

As a two-stage general linear method, we write $Y_1 = y_n^*$ and $Y_2 = y_n$. The $r = 4$ input approximations are the values of y_{n-1}, $hf(x_{n-1}, y_{n-1})$, $hf(x_{n-2}, y_{n-2})$ and $hf(x_{n-3}, y_{n-3})$. The $(s+r) \times (s+r)$ coefficient matrix is now

$$\left[\begin{array}{cc|cccc} 0 & 0 & 1 & \frac{23}{12} & -\frac{4}{3} & \frac{5}{12} \\ \frac{5}{12} & 0 & 1 & \frac{2}{3} & -\frac{1}{12} & 0 \\ \hline \frac{5}{12} & 0 & 1 & \frac{2}{3} & -\frac{1}{12} & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array} \right]. \tag{503a}$$

The one-leg methods, introduced by Dahlquist (1976) as counterparts of linear multistep methods, have their own natural representations as general linear methods. For the method characterized by the polynomial pair $[\alpha(z), \beta(z)]$, the corresponding one-leg method computes a single stage value Y, with stage derivative F, using the formula

$$y_n = \sum_{i=1}^{k} \alpha_i y_{n-i} + \left(\sum_{i=0}^{k} \beta_i \right) hF, \tag{503b}$$

where

$$Y = \frac{\sum_{i=0}^{k} \beta_i y_{n-i}}{\sum_{i=0}^{k} \beta_i}. \tag{503c}$$

This does not fit into the standard representation for general linear methods but it achieves this format when Y and y_n are separated out from the two expressions (503b) and (503c). We find

$$Y = \beta_0 hF + \left(\sum_{i=0}^{k} \beta_i \right)^{-1} \sum_{i=1}^{k} (\beta_0 \alpha_i + \beta_i) y_{n-i},$$

$$y_n = \left(\sum_{i=0}^{k} \beta_i \right) hF + \sum_{i=1}^{k} \alpha_i y_{n-i}.$$

As a general linear method, it has the form

$$
\left[
\begin{array}{c|cccccc}
\beta_0 & \gamma_1 & \gamma_2 & \gamma_3 & \cdots & \gamma_{k-1} & \gamma_k \\
\hline
\sum_{i=0}^{k} \beta_i & \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_{k-1} & \alpha_k \\
0 & 1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & 0 & \cdots & 1 & 0
\end{array}
\right],
$$

where

$$
\gamma_i = \left(\sum_{j=0}^{k} \beta_j \right)^{-1} \left(\beta_0 \alpha_i + \beta_i \right), \qquad i = 1, 2, \ldots, k.
$$

504   Some known unconventional methods

Amongst the methods that do not fit under the conventional Runge–Kutta or linear multistep headings, we consider the cyclic composite methods of Donelson and Hansen (1971), the pseudo Runge–Kutta methods of Byrne and Lambert (1966) and the hybrid methods of Gragg and Stetter (1964), Butcher (1965) and Gear (1965). We illustrate, by examples, how methods of these types can be cast in general linear form.

To overcome the limitations of linear multistep methods imposed by the conflicting demands of order and stability, Donelson and Hansen proposed a procedure in which two or more linear multistep methods are used in rotation over successive steps. Write the constituent methods as $(\alpha^{(1)}, \beta^{(1)})$, $(\alpha^{(2)}, \beta^{(2)})$, ..., $(\alpha^{(m)}, \beta^{(m)})$, so that the formula for computing y_n will be

$$
y_n = \sum_{i=1}^{k} \alpha_i^{(j)} y_{n-i} + \sum_{i=0}^{k} \beta_i^{(j)} h f(x_{n-i}, y_{n-i}),
$$

where $j \in \{1, 2, \ldots, m\}$ is chosen so that $n - j$ is a multiple of m.

The step value – that is the maximum of the degrees of $\alpha^{(j)}$ and $\beta^{(j)}$ – may vary amongst the m constituent methods, but they can be assumed to have a common value k equal to the maximum over all the basic methods. We illustrate these ideas in the case $k = 3$, $m = 2$. As a consequence of the Dahlquist barrier, order $p = 5$ with $k = 3$ is inconsistent with stability and therefore convergence. Consider the following two linear multistep methods:

$$
[\alpha^{(1)}(z), \beta^{(1)}(z)] = [1 + \tfrac{8}{11}z - \tfrac{19}{11}z^2, \tfrac{10}{33} + \tfrac{19}{11}z + \tfrac{8}{11}z^2 - \tfrac{1}{33}z^3],
$$
$$
[\alpha^{(2)}(z), \beta^{(2)}(z)] = [1 - \tfrac{449}{240}z - \tfrac{19}{30}z^2 + \tfrac{361}{240}z^3, \tfrac{251}{720} + \tfrac{19}{30}z - \tfrac{449}{240}z^2 - \tfrac{35}{72}z^3].
$$

Each of these has order 5 and is, of course, unstable. To combine them, used alternately, into a single step of a general linear method, it is convenient to regard h as the stepsize for the complete cycle of two steps. We denote the incoming approximations as $y_{n-3/2}$, y_{n-1}, hf_{n-2}, $hf_{n-3/2}$ and hf_{n-1}. The first half-step, relating $y_{n-1/2}$ and $hf_{n-1/2}$ to the input quantities, gives

$$y_{n-\frac{1}{2}} = \tfrac{5}{33}hf_{n-\frac{1}{2}} + \tfrac{19}{11}y_{n-\frac{3}{2}} - \tfrac{8}{11}y_{n-1} - \tfrac{1}{66}hf_{n-2} + \tfrac{4}{11}hf_{n-\frac{3}{2}} + \tfrac{19}{22}hf_{n-1}.$$

Substitute this into the corresponding formula for y_n and we find

$$y_n = \tfrac{4753}{7920}hf_{n-\frac{1}{2}} + \tfrac{251}{1440}hf_n + \tfrac{19}{11}y_{n-\frac{3}{2}} - \tfrac{8}{11}y_{n-1}$$
$$- \tfrac{449}{15840}hf_{n-2} + \tfrac{3463}{7920}hf_{n-\frac{3}{2}} + \tfrac{449}{660}hf_{n-1}.$$

Translating these formulae into the (A, U, B, V) formulation gives

$$\begin{bmatrix} A & U \\ B & V \end{bmatrix} = \left[\begin{array}{cc|ccccc} \frac{5}{33} & 0 & \frac{19}{11} & -\frac{8}{11} & -\frac{1}{66} & \frac{4}{11} & \frac{19}{22} \\ \frac{4753}{7920} & \frac{251}{1440} & \frac{19}{11} & -\frac{8}{11} & -\frac{449}{15840} & \frac{3463}{7920} & \frac{449}{660} \\ \hline \frac{5}{33} & 0 & \frac{19}{11} & -\frac{8}{11} & -\frac{1}{66} & \frac{4}{11} & \frac{19}{22} \\ \frac{4753}{7920} & \frac{251}{1440} & \frac{19}{11} & -\frac{8}{11} & -\frac{449}{15840} & \frac{3463}{7920} & \frac{449}{660} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

This formulation can be simplified, in the sense that r can be reduced, and we have, for example, the following alternative coefficient matrices:

$$\begin{bmatrix} A & U \\ B & V \end{bmatrix} = \left[\begin{array}{cc|ccccc} \frac{5}{33} & 0 & 1 & -\frac{1}{66} & \frac{4}{11} & \frac{19}{22} \\ \frac{4753}{7920} & \frac{251}{1440} & 1 & -\frac{449}{15840} & \frac{3463}{7920} & \frac{449}{660} \\ -\frac{173}{990} & -\frac{251}{1980} & 1 & -\frac{1}{180} & \frac{307}{990} & \frac{329}{330} \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right].$$

Because of the natural way in which we have written this particular composite cyclic pair in general linear form, and then rewritten it, using equally simple operations, into a less recognizable form, an obvious question arises. The question is whether it might have been more appropriate to use the general linear formulation from the start, and then explore the existence of suitable methods that have no connection with linear multistep methods.

We now turn to pseudo Runge–Kutta methods. Consider the method given by (261a). Even though four input values are used in step n (y_{n-1}, $hF_1^{[n-1]}$, $hF_2^{[n-1]}$ and $hF_3^{[n-1]}$), this can be effectively reduced to two because, in addition to y_{n-1}, only the combination $\frac{1}{12}hF_1^{[n-1]} - \frac{1}{3}hF_2^{[n-1]} - \frac{1}{4}hF_3^{[n-1]}$ is actually used. This means that a quantity of this form, but with $n-1$ replaced by n, has to be computed in step n for use in the following step. The $(3+2) \times (3+2)$ matrix representing this method is

$$
\left[
\begin{array}{ccc|cc}
0 & 0 & 0 & 1 & 0 \\
\frac{1}{2} & 0 & 0 & 1 & 0 \\
-\frac{1}{3} & \frac{4}{3} & 0 & 1 & 0 \\
\hline
\frac{11}{12} & \frac{1}{3} & \frac{1}{4} & 1 & 1 \\
\frac{1}{12} & -\frac{1}{3} & -\frac{1}{4} & 0 & 0
\end{array}
\right].
$$

For a seventh order method taken from Butcher (1965), the solution at the end of the step is approximated using 'predictors' at $x_n - \frac{1}{2}h$ and at x_n, in preparation for a final 'corrector' value, also at x_n. The input quantities correspond to solution approximations $y_1^{[n-1]} \approx y(x_{n-1})$, $y_2^{[n-1]} \approx y(x_{n-2})$ and $y_3^{[n-1]} \approx y(x_{n-3})$, and the corresponding scaled derivative approximations $y_4^{[n-1]} \approx hy'(x_{n-1})$, $y_5^{[n-1]} \approx hy'(x_{n-2})$ and $y_6^{[n-1]} \approx hy'(x_{n-3})$. The general linear representation is

$$
\left[
\begin{array}{ccc|cccccc}
0 & 0 & 0 & -\frac{225}{128} & \frac{200}{128} & \frac{153}{128} & \frac{225}{128} & \frac{300}{128} & \frac{45}{128} \\
\frac{384}{155} & 0 & 0 & \frac{540}{128} & -\frac{297}{31} & -\frac{212}{31} & -\frac{1395}{155} & -\frac{2130}{155} & -\frac{309}{155} \\
\frac{2304}{3085} & \frac{465}{3085} & 0 & \frac{783}{617} & -\frac{135}{617} & -\frac{31}{617} & -\frac{135}{3085} & -\frac{495}{3085} & -\frac{39}{3085} \\
\hline
\frac{2304}{3085} & \frac{465}{3085} & 0 & \frac{783}{617} & -\frac{135}{617} & -\frac{31}{617} & -\frac{135}{3085} & -\frac{495}{3085} & -\frac{39}{3085} \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{array}
\right].
$$

505   Some recently discovered general linear methods

The methods already introduced in this section were inspired as modifications of Runge–Kutta or linear multistep methods. We now consider two example methods motivated not by either of the classical forms, but by the general linear structure in its own right.

The first of these is known as an 'Almost Runge–Kutta' method. That is, although it uses three input and output approximations, it behaves like a Runge–Kutta method from many points of view. The input vectors can be thought of as approximations to $y(x_{n-1})$, $hy'(x_{n-1})$ and $h^2 y''(x_{n-1})$ and the output vectors are intended to be approximations to these same quantities, but evaluated at x_n rather than at x_{n-1}:

$$
\left[
\begin{array}{cccc|ccc}
0 & 0 & 0 & 0 & 1 & 1 & \frac{1}{2} \\
\frac{1}{16} & 0 & 0 & 0 & 1 & \frac{7}{16} & \frac{1}{16} \\
-\frac{1}{4} & 2 & 0 & 0 & 1 & -\frac{3}{4} & -\frac{1}{4} \\
0 & \frac{2}{3} & \frac{1}{6} & 0 & 1 & \frac{1}{6} & 0 \\
\hline
0 & \frac{2}{3} & \frac{1}{6} & 0 & 1 & \frac{1}{6} & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-\frac{1}{3} & 0 & -\frac{2}{3} & 2 & 0 & -1 & 0
\end{array}
\right]. \tag{505a}
$$

The particular example given here has order 4, in contrast to the third order method introduced in Section 27 to illustrate implementation principles. Further details concerning Almost Runge–Kutta methodsare presented in Subsection 543.

The second example is given by the coefficient matrix

$$
\left[
\begin{array}{ccc|ccc}
0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 \\
\frac{1}{4} & 1 & 0 & 0 & 0 & 1 \\
\hline
\frac{5}{4} & \frac{1}{3} & \frac{1}{6} & -\frac{2}{3} & \frac{4}{3} & \frac{1}{3} \\
\frac{35}{24} & -\frac{1}{3} & \frac{1}{8} & -\frac{2}{3} & \frac{4}{3} & \frac{1}{3} \\
\frac{17}{12} & 0 & \frac{1}{12} & -\frac{2}{3} & \frac{4}{3} & \frac{1}{3}
\end{array}
\right]. \tag{505b}
$$

In the case of (505b), the input values are given respectively as approximations to
$$y(x_{n-1}),$$
$$y(x_{n-1} + \tfrac{1}{2}h) + hy'(x_{n-1})$$
and to
$$y(x_{n-1}) - \tfrac{1}{4}hy'(x_{n-1}) + \tfrac{1}{24}h^3 y'''(x_{n-1}),$$
and the output consists of the same three quantities, to within $O(h^4)$, with x_{n-1} advanced one step to x_n. Thus the method has order 3. This is an example of a 'type 1 DIMSIM method', to be introduced in Subsection 541.

Both (505a) and (505b) possess the property of RK stability, which guarantees that the method behaves, at least in terms of linear stability, like a Runge–Kutta method. While their multivalue structure is a disadvantage compared with Runge–Kutta methods, they have some desirable properties. For (505a) the stage order is 2, and for (505b) the stage order is 3.

Exercises 50

50.1 Write the general linear method given by (503a) in transformed form using the matrix

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{3}{4} & -1 & \frac{1}{4} \\ 0 & \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} \end{bmatrix}.$$

Note that this converts the method into Nordsieck form.

50.2 Write the general linear method given by (502a) in transformed form using the matrix

$$T = \begin{bmatrix} 1 & \frac{1}{6} \\ 0 & 1 \end{bmatrix}.$$

50.3 Write the implicit Runge–Kutta method

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

as a general linear method with $r = 2$, $s = 1$, by taking advantage of the FSAL property.

50.4 Show that it is possible, by using a suitable transformation, to reduce the general linear method derived in Exercise 50.3 to an equivalent method with $r = s = 1$. Show that this new method is equivalent to the implicit mid-point rule Runge–Kutta method.

50.5 Write the PEC predictor–corrector method based on the order 2 Adams–Bashforth method and the order 2 Adams–Moulton method in general linear form.

50.6 The following two methods were once popular, but are now regarded as flawed because they are 'weakly stable':

$$y_n = y_{n-2} + 2hf(x_{n-1}, y_{n-1}),$$
$$y_n = y_{n-3} + \tfrac{3}{2}h(f(x_{n-1}, y_{n-1}) + f(x_{n-2}, y_{n-2})).$$

This means that, although the methods are stable, the polynomial α for each of them has more than one zero on the unit circle. Show how to write them as a cyclic composite pair, using general linear formulation, and that they no longer have such a disadvantage.

50.7 Consider the Runge–Kutta method

$$
\begin{array}{c|cccc}
0 & & & & \\
-1 & -1 & & & \\
\frac{1}{2} & \frac{5}{8} & -\frac{1}{8} & & \\
1 & -\frac{3}{2} & \frac{1}{2} & 2 & \\
\hline
& \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6}
\end{array}.
$$

Modify this method in the same way as was proposed for (502b), and write the resulting two-value method in general linear form.

51 Consistency, Stability and Convergence

510 *Definitions of consistency and stability*

Since a general linear method operates on a vector of approximations to some quantities computed in the preceding step, we need to decide something about the nature of this information. For most numerical methods, it is obvious what form this takes, but for a method as general as the ones we are considering here there are many possibilities. At least we assume that the ith subvector in $y^{[n-1]}$ represents $u_i y(x_{n-1}) + v_i h y'(x_{n-1}) + O(h^2)$. The vectors u and v are characteristic of any particular method, subject to the freedom we have to alter v by a scalar multiple of u; because we can reinterpret the method by changing x_n by some fixed multiple of h. The choice of u must be such that the stage values are each equal to $y(x_n) + O(h)$. This means that $Uu = \mathbf{1}$. We always require the output result to be $u_i y(x_n) + v_i h y'(x_n) + O(h^2)$ and this means that $Vu = u$ and that $Vv + B\mathbf{1} = u + v$. If we are given nothing about a method except the four defining matrices, then V must have an eigenvalue equal to 1 and u must be a corresponding eigenvector. It then has to be checked that the space of such eigenvectors contains a member such that $Uu = \mathbf{1}$ and such that $B\mathbf{1} - u$ is in the range of $V - I$.

If a method has these properties then it is capable of solving $y' = 1$, with $y(0) = a$ exactly, in the sense that if $y_i^{[0]} = u_i a + v_i h$, then for all $n = 1, 2, \ldots, y_i^{[n]} = u_i(a + nh) + v_i h$. This suggests the following definitions:

Definition 510A *A general linear method (A, U, B, V) is 'preconsistent' if there exists a vector u such that*

$$Vu = u, \tag{510a}$$

$$Uu = \mathbf{1}. \tag{510b}$$

The vector u is the 'preconsistency vector'.

Definition 510B *A general linear method (A, U, B, V) is 'consistent' if it is preconsistent with preconsistency vector u and there exists a vector v such that*

$$B\mathbf{1} + Vv = u + v. \tag{510c}$$

Just as for linear multistep methods, we need a concept of stability. In the general linear case this is defined in terms of the power-boundedness of V and, as we shall see, is related to the solvability of the problem $y' = 0$.

Definition 510C *A general linear method* (A, U, B, V) *is 'stable' if there exists a constant* C *such that, for all* $n = 1, 2, \ldots$, $\|V^n\| \leq C$.

511   Covariance of methods

Assume the interpretation of a method is agreed to, at least in terms of the choice of the preconsistency vector. We want to ensure that numerical approximations are transformed appropriately by a shift of origin. Consider the two initial value problems

$$y'(x) = f(y(x)), \qquad y(x_0) = y_0, \tag{511a}$$
$$z'(x) = f(z(x) - \eta), \qquad z(x_0) = y_0 + \eta, \tag{511b}$$

where $\eta \in \mathbb{R}^N$ is arbitrary. If (511a) has a solution, then (511b) also has a solution, which is identical to the former solution except that each point on the trajectory is translated by η. If the solution is required at some $\overline{x} > x_0$, then the solution to (511a) at this point can be found by subtracting η from the solution of (511b).

When each of these problems is solved by a numerical method, it is natural to expect that the numerical approximations should undergo the same covariance rule as for the exact solution. This means that in a single step of a method (A, U, B, V), interpreted as having a preconsistency vector u, we want to be able to shift component i of $y^{[0]}$ by $u_i \eta$, for all $i = 1, 2, \ldots, r$, and be assured that component i of $y^{[1]}$ is also shifted by the same amount. At the same time the internal approximations (the stage values) should be shifted by η. Of course no shift will take place to the stage derivatives.

The idea of covariance is illustrated in Figure 511(i). For an initial value problem (f, y_0) as given by (511a), the operation ν represents the computation of a numerical approximation to the solution on an interval $[x_0, \overline{x}]$, or at a single value of x. Furthermore, σ represents a shift of coordinates by a specific vector η, as in the transformation to the problem (511b). Covariance is just the statement that the diagram in Figure 511(i) commutes, that is, that $\sigma \circ \nu = \nu \circ \sigma$. The diagonal arrow representing these equal composed functions corresponds to the operation of solving the problem and then shifting coordinates, or else shifting first and then solving.

The covariance of the output values is equivalent to (510a) and the covariance of the stage values is equivalent to (510b). We have no interest in methods that are not covariant even though it is possible to construct artificial methods which do not have this property but can still yield satisfactory numerical results.
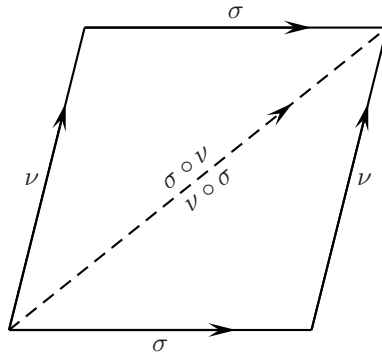
Figure 511(i)   A commutative diagram for covariance

512   *Definition of convergence*

Just as for linear multistep methods, the necessity of using a starting procedure complicates the idea of convergence. We deal with this complication by assuming nothing more from the starting procedure than the fact that, for sufficiently small h, it produces an approximation arbitrarily close to

$$
\begin{bmatrix}
u_1 y(x_0) \\
u_2 y(x_0) \\
\vdots \\
u_r y(x_0)
\end{bmatrix},
$$

where u is *some* non-zero vector in $\mathbb{R}^r$. Here $y(x_0)$ is the given initial data and it will be our aim to obtain a good approximation at some $\overline{x} > x_0$. This approximation should converge to

$$
\begin{bmatrix}
u_1 y(\overline{x}) \\
u_2 y(\overline{x}) \\
\vdots \\
u_r y(\overline{x})
\end{bmatrix}, \tag{512a}
$$

for any problem satisfying a Lipschitz condition. For notational convenience, (512a) will usually be abbreviated as $uy(\overline{x})$.

Formally, we write $\phi(h)$ for the starting approximation associated with the method and with a given initial value problem.

Definition 512A *A general linear method (A, U, B, V), is 'convergent' if for any initial value problem*

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

subject to the Lipschitz condition $\|f(y) - f(z)\| \le L\|y - z\|$, there exist a non-zero vector $u \in \mathbb{R}^r$, and a starting procedure $\phi : (0, \infty) \to \mathbb{R}^r$, such that for all $i = 1, 2, \ldots, r$, $\lim_{h \to 0} \phi_i(h) = u_i y(x_0)$, and such that for any $\overline{x} > x_0$, the sequence of vectors $y^{[n]}$, computed using n steps with stepsize $h = (\overline{x} - x_0)/n$ and using $y^{[0]} = \phi(h)$ in each case, converges to $u y(\overline{x})$.

The necessity of stability and consistency, as essential properties of convergent methods, are proved in the next two subsections, and this is followed by the converse result that all stable and consistent methods are convergent.

513 The necessity of stability

Stability has the effect of guaranteeing that errors introduced in any step of a computation do not have disastrous effects on later steps. The necessity of this property is expressed in the following result:

Theorem 513A *A general linear method (A, U, B, V) is convergent only if it is stable.*

Proof. Suppose, on the contrary, that $\{\|V^n\| : n = 1, 2, 3, \ldots\}$ is unbounded. This implies that there exists a sequence of vectors $w_1, w_2, w_3, \ldots$ such that $\|w_n\| = 1$, for all $n = 1, 2, 3, \ldots$, and such that the sequence $\{\|V^n w_n\| : n = 1, 2, 3, \ldots\}$ is unbounded. Consider the solution of the initial value problem

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

using (A, U, B, V), where n steps are taken with stepsize $h = 1/n$, so that the solution is approximated at $\overline{x} = 1$. Irrespective of the choice of the vector u in Definition 512A, the convergence of the method implies that the sequence of approximations converges to zero. For the approximation carried out with n steps, use as the starting approximation

$$\phi\left(\frac{1}{n}\right) = \frac{1}{\max_{i=1}^{n} \|V^i w_i\|} w_n.$$

This converges to zero, because $\|\phi(1/n)\| = \left(\max_{i=1}^{n} \|V^i w_i\|\right)^{-1}$. The result, computed after n steps, will then be

$$V^n \phi\left(\frac{1}{n}\right) = \frac{1}{\max_{i=1}^{n} \|V^i w_i\|} V^n w_n,$$

with norm

$$\left\| V^n \phi\left(\frac{1}{n}\right) \right\| = \frac{\|V^n w_n\|}{\max_{i=1}^n \|V^i w_i\|}. \tag{513a}$$

Because the sequence $n \mapsto \|V^n w_n\|$ is unbounded, an infinite set of n values will have the property that the maximum value of $\|V^i w_i\|$, for $i \le n$, will occur with $i = n$. This means that (513a) has value 1 arbitrarily often, and hence is not convergent to zero as $n \to \infty$. $\qquad\square$

514   The necessity of consistency

By selecting a specific differential equation, as in Subsection 513, we can prove that for covariant methods, consistency is necessary.

Theorem 514A *Let (A, U, B, V) denote a convergent method which is, moreover, covariant with preconsistency vector u. Then there exists a vector $v \in \mathbb{R}^r$, such that (510c) holds.*

Proof. Consider the initial value problem

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

with constant starting values $\phi(h) = 0$ and $\overline{x} = 1$. The sequence of approximations, when n steps are to be taken with $h = 1/n$, is given by

$$y^{[i]} = \frac{1}{n} B\mathbf{1} + V y^{[i-1]}, \qquad i = 1, 2, \ldots, n.$$

This means that the error vector, after the n steps have been completed, is given by

$$\begin{aligned}
y^{[n]} - u &= \frac{1}{n}\left(I + V + V^2 + \cdots + V^{n-1}\right) B\mathbf{1} - u \\
&= \frac{1}{n}\left(I + V + V^2 + \cdots + V^{n-1}\right)(B\mathbf{1} - u).
\end{aligned}$$

Because V has bounded powers, it can be written in the form

$$V = S^{-1} \begin{bmatrix} I & 0 \\ 0 & W \end{bmatrix} S,$$

where I is $\widetilde{r} \times \widetilde{r}$ for $\widetilde{r} \le r$ and W is power-bounded and is such that $1 \notin \sigma(W)$. This means that

$$y^{[n]} - u = S^{-1} \begin{bmatrix} I & 0 \\ 0 & \frac{1}{n}(I - W)^{-1}(I - W^n) \end{bmatrix} S(B\mathbf{1} - u),$$

whose limit as $n \to \infty$ is

$$S^{-1} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} S(B\mathbf{1} - u).$$

If $y^{[n]} - u$ is to converge to 0 as $n \to \infty$, then $S(B\mathbf{1} - u)$ has only zero in its first $\widetilde{r}$ components. Write this vector in the form

$$S(B\mathbf{1} - u) = \begin{bmatrix} 0 \\ (I - W)\widetilde{v} \end{bmatrix}$$

$$= \left(I - \begin{bmatrix} I & 0 \\ 0 & W \end{bmatrix} \right) Sv$$

$$= S(I - V)v,$$

where

$$v = S^{-1} \begin{bmatrix} 0 \\ \widetilde{v} \end{bmatrix}.$$

Thus $B\mathbf{1} + Vv = u + v$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

515   *Stability and consistency imply convergence*

We show that stable and consistent methods are convergent. This is done in three steps. The first is to analyse the internal and the external local truncation error; the second is to obtain a difference inequality relating the total error at the end of a step with the total error at the end of the previous step. Finally, we find a bound on the global error and show that it converges to zero.

In the truncation error estimation, we need to decide what the input and output approximations and the internal stages are intended to approximate. The choice we make here is determined by a wish for simplicity: we do not need good error bounds, only bounds sufficiently strong to enable us to establish convergence. Our assumption will be that $y_i^{[n]}$ approximates $u_i y(x_n) + h v_i y'(x_n)$, and that the internal stage Y_i approximates $y(x_{n-1} + hc_i)$, where c_i is determined by what happens to the time variable.

We need to make some assumptions about the problem whose solution is being approximated. What we shall suppose is that there exists a closed set S in $\mathbb{R}^N$ such that all values of $y(x)$ that will ever arise in a trajectory lie in the interior of S. Furthermore, we suppose that for any $y \in S$, $\|y\| \leq M$ and $\|f(y)\| \leq LM$. Also, we suppose that for $y, z \in S$, $\|f(y) - f(z)\| \leq L\|y - z\|$. Since we are concerned with the limit as $h \to 0$, we restrict the value of h to an interval $(0, h_0]$, for some $h_0 > 0$.

With this in mind, we find bounds as follows:

Lemma 515A *Assume that $h \leq h_0$, chosen so that $h_0 L \|A\|_\infty < 1$. Define ϵ as the vector in $\mathbb{R}^s$ satisfying*

$$\sum_{j=1}^{s} (\delta_{ij} - h_0 L |a_{ij}|) \epsilon_j = \tfrac{1}{2} c_i^2 + \sum_{j=1}^{s} |a_{ij} c_j|.$$

Let $\widehat{y}_i^{[n-1]} = u_i y(x_{n-1}) + v_i h y'(x_{n-1})$, $\widehat{y}_i^{[n]} = u_i y(x_n) + v_i h y'(x_n)$, for $i = 1, 2, \ldots, r$, and $\widehat{Y}_i = y(x_{n-1} + h c_i)$, for $i = 1, 2, \ldots, s$, where $c = A\mathbf{1} + Uv$. Also let $\widetilde{Y}_i$ denote the value of Y_i that would be computed exactly using $\widehat{y}^{[n-1]}$ as input vector $y^{[n-1]}$. Assume the function f satisfies a Lipschitz condition with constant L and that the exact solution to the initial value problem satisfies $\|y(x)\| \leq M$, $\|y'(x)\| \leq LM$. Then

$$\left\| \widehat{Y}_i - h \sum_{j=1}^{s} a_{ij} f(\widehat{Y}_j) - \sum_{j=1}^{r} U_{ij} \widehat{y}_j^{[n-1]} \right\|$$
$$\leq h^2 L^2 M \left(\tfrac{1}{2} c_i^2 + \sum_{j=1}^{s} |a_{ij} c_j| \right), \tag{515a}$$

$$\left\| \widehat{y}_i^{[n]} - h \sum_{j=1}^{s} b_{ij} f(\widehat{Y}_j) - \sum_{j=1}^{r} V_{ij} \widehat{y}_j^{[n-1]} \right\|$$
$$\leq h^2 L^2 M \left(\tfrac{1}{2} |u_i| + |v_i| + \sum_{j=1}^{s} |b_{ij} c_j| \right), \tag{515b}$$

$$\left\| \widehat{y}_i^{[n]} - h \sum_{j=1}^{s} b_{ij} f(\widetilde{Y}_j) - \sum_{j=1}^{r} V_{ij} \widehat{y}_j^{[n-1]} \right\|$$
$$\leq h^2 L^2 M \left(\tfrac{1}{2} |u_i| + |v_i| + \sum_{j=1}^{s} |b_{ij} c_j| + h_0 L \sum_{j=1}^{s} |b_{ij}| \epsilon_j \right). \tag{515c}$$

Proof. We first note that

$$\|y(x_{n-1} + h c_i) - y(x_{n-1})\| = h \left\| \int_0^{c_i} y'(x_{n-1} + h\xi) d\xi \right\|$$
$$\leq h \int_0^{c_i} \left\| y'(x_{n-1} + h\xi) \right\| d\xi$$
$$\leq |c_i| h L M.$$

We now have

$$\widehat{Y}_i - h \sum_{j=1}^{s} a_{ij} f(\widehat{Y}_j) - \sum_{j=1}^{r} U_{ij} \widehat{y}_j^{[n-1]} = T_1 + T_2 + T_3 + T_4,$$

where

$$T_1 = \widehat{Y}_i - y(x_{n-1}) - h \int_0^{c_i} f(y(x_{n-1} + h\xi))d\xi,$$

$$T_2 = y(x_{n-1}) + c_i h y'(x_{n-1}) - \sum_{j=1}^r U_{ij} \widehat{y}_j^{[n-1]} - \sum_{j=1}^s a_{ij} h y'(x_{n-1}),$$

$$T_3 = h \int_0^{c_i} \Big(f(y(x_{n-1} + h\xi)) - y'(x_{n-1}) \Big) d\xi,$$

$$T_4 = -h \sum_{j=1}^s a_{ij} \Big(f(y(x_{n-1} + hc_j)) - y'(x_{n-1}) \Big).$$

Simplify and estimate these terms, and we find

$$T_1 = y(x_{n-1} + hc_i) - y(x_{n-1}) - h \int_0^{c_i} y'(x_{n-1} + h\xi)d\xi = 0,$$

$$T_2 = y(x_{n-1}) + c_i h y'(x_{n-1})$$

$$- \sum_{j=1}^r U_{ij} \Big(u_j y(x_{n-1}) + h v_j y'(x_{n-1}) \Big) - \sum_{j=1}^s a_{ij} h y'(x_{n-1})$$

$$= 0, \quad \text{because } Uu = \mathbf{1} \text{ and } Uv + A\mathbf{1} = c,$$

$$\|T_3\| = h \left\| \int_0^{c_i} \Big(f(y(x_{n-1} + h\xi)) - f(y(x_{n-1})) \Big) d\xi \right\|$$

$$\leq h \int_0^{c_i} \left\| f(y(x_{n-1} + h\xi)) - f(y(x_{n-1})) \right\| d\xi$$

$$\leq hL \int_0^{c_i} \left\| y(x_{n-1} + h\xi) - y(x_{n-1}) \right\| d\xi$$

$$\leq h^2 L^2 M \int_0^{c_i} \xi d\xi$$

$$= \tfrac{1}{2} h^2 L^2 M c_i^2,$$

$$\|T_4\| = h \left\| \sum_{j=1}^s a_{ij} \Big(f(y(x_{n-1} + hc_j)) - f(y(x_{n-1})) \Big) \right\|$$

$$\leq h \sum_{j=1}^s |a_{ij}| \cdot \| f(y(x_{n-1} + hc_j)) - f(y(x_{n-1})) \|$$

$$\leq hL \sum_{j=1}^s |a_{ij}| \cdot \| y(x_{n-1} + hc_j) - y(x_{n-1}) \|$$

$$\leq h^2 L^2 M \sum_{j=1}^s |a_{ij} c_j|,$$

so that, combining these estimates, we arrive at (515a).

To verify (515b), we write

$$\widehat{y}_i^{[n]} - h \sum_{j=1}^{s} b_{ij} f(\widehat{Y}_j) - \sum_{j=1}^{r} V_{ij} \widehat{y}_j^{[n-1]} = T_1 + T_2 + T_3 + T_4,$$

where

$$T_1 = u_i \Big(y(x_{n-1} + h) - y(x_{n-1}) - h \int_0^1 y'(x_{n-1} + h\xi)d\xi \Big),$$

$$T_2 = v_i h y'(x_{n-1} + h) + \Big(u_i - \sum_{j=1}^{s} b_{ij} - \sum_{j=1}^{r} V_{ij} v_j \Big) h y'(x_{n-1}),$$

$$T_3 = h u_i \int_0^1 \big(y'(x_{n-1} + h\xi) - y'(x_{n-1}) \big) d\xi,$$

$$T_4 = -h \sum_{j=1}^{s} b_{ij} \big(y'(x_{n-1} + h c_j) - y'(x_{n-1}) \big).$$

We check that $T_1 = 0$ and that, because $\sum_{j=1}^{s} b_{ij} + \sum_{j=1}^{r} V_{ij} v_j = u_i + v_i$, T_2 simplifies to $h v_i (y'(x_{n-1} + h) - y'(x_{n-1}))$ so that $\|T_2\| \le h^2 L^2 M |v_i|$. Similarly, $\|T_3\| \le \frac{1}{2} h^2 L^2 M |u_i|$ and $\|T_4\| \le h^2 L^2 M \sum_{j=1}^{s} |b_{ij} c_j|$. To prove (515c) we first need to estimate the elements of $\widetilde{Y} - \widehat{Y}$ by deducing from (515a) that

$$\Big\| \big(\widetilde{Y}_i - \widehat{Y}_i \big) - h \sum_{j=1}^{s} a_{ij} \big(f(\widetilde{Y}_j) - f(\widehat{Y}_j) \big) \Big\| \le \Big(\tfrac{1}{2} c_i^2 + \sum_{j=1}^{s} |a_{ij} c_j| \Big) h^2 L^2 M,$$

and hence that

$$\| \widetilde{Y}_j - \widehat{Y}_j \| \le h^2 L^2 M \epsilon_j.$$

Thus,

$$\Big\| h \sum_{j=1}^{s} b_{ij} \big(f(\widetilde{Y}_j) - f(\widehat{Y}_j) \big) \Big\| \le h^2 L^3 M h_0 \sum_{j=1}^{s} |b_{ij}| \epsilon_j.$$

Add this estimate of $\Big\| h \sum_{j=1}^{s} b_{ij} \big(f(\widetilde{Y}_j) - f(\widehat{Y}_j) \big) \Big\|$ to (515b) to obtain (515c). □

The next step in the investigation is to find a bound on the local truncation error.

Lemma 515B *Under the conditions of Lemma 515A, the exact solution and the computed solution in a step are related by*

$$\widehat{y}_i^{[n]} - y_i^{[n]} = \sum_{j=1}^{r} V_{ij} \Big(\widehat{y}_j^{[n-1]} - y_j^{[n-1]} \Big) + K_i^{[n]}, \qquad i = 1, 2, \ldots, r,$$

where

$$\|K^{[n]}\| \le h\alpha \max_{i=1}^{r} \left\|\widehat{y}_i^{[n-1]} - y_i^{[n-1]}\right\| + \beta h^2,$$

and α and β are given by

$$\alpha = L \max_{i=1}^{s} |\overline{\epsilon}_i|,$$

where $\overline{\epsilon}$ is given by

$$\sum_{j=1}^{s} (\delta_{ij} - h_0 L |a_{ij}|)\overline{\epsilon}_j = \sum_{j=1}^{s} |U_{ij}|, \qquad i = 1, 2, \ldots, s,$$

and

$$\beta = L^2 M \max_{i=1}^{s} \left(\frac{1}{2}|u_i| + |v_i| + \sum_{j=1}^{s} |b_{ij}c_j| + h_0 L \sum_{j=1}^{s} |b_{ij}|\epsilon_j \right),$$

where ϵ is as in Lemma 515A.

Proof. From (515c), and the relation

$$y_i^{[n]} - h\sum_{j=1}^{s} b_{ij} f(Y_j) - \sum_{j=1}^{r} V_{ij} y_j^{[n-1]} = 0,$$

we have

$$\left\| \widehat{y}_i^{[n]} - y_i^{[n]} - \sum_{j=1}^{r} V_{ij} \left(\widehat{y}_j^{[n-1]} - y_j^{[n-1]} \right) \right\|$$

$$\le h\sum_{j=1}^{s} |b_{ij}| \left\| f(\widetilde{Y}_j) - f(Y_j) \right\|$$

$$+ h^2 L^2 M \left(\frac{1}{2}|u_i| + |v_i| + \sum_{j=1}^{s} |b_{ij}c_j| + h_0 L \sum_{j=1}^{s} |b_{ij}|\epsilon_j \right)$$

$$\le hL\sum_{j=1}^{s} |b_{ij}| \left\| Y_j - \widetilde{Y}_j \right\|$$

$$+ h^2 L^2 M \left(\frac{1}{2}|u_i| + |v_i| + \sum_{j=1}^{s} |b_{ij}c_j| + h_0 L \sum_{j=1}^{s} |b_{ij}|\epsilon_j \right). \tag{515d}$$

Bound $\eta_j = \|\widetilde{Y}_j - Y_j\|$ using the estimate

$$\left\| \widetilde{Y}_j - Y_j - \sum_{k=1}^{r} U_{jk} \left(\widehat{y}_k^{[n-1]} - y_k^{[n-1]} \right) \right\| \le hL\sum_{k=1}^{s} |a_{jk}| \cdot \|\widetilde{Y}_k - Y_k\|,$$

which leads to

$$\sum_{k=1}^{s}(\delta_{jk} - h_0 L|a_{jk}|)\eta_k \leq \sum_{k=1}^{r}|U_{jk}|\max_{k=1}^{r}\left\|\widehat{y}_k^{[n-1]} - y_k^{[n-1]}\right\|$$

and to

$$\|\widetilde{Y}_j - Y_j\| \leq h\bar{\epsilon}_j \max_{k=1}^{s}\|\widetilde{Y}_k - Y_k\|.$$

Substitute this bound into (515d) and we obtain the required result.   □

To complete the argument that stability and consistency imply convergence, we estimate the global error in the computation of $y(\overline{x})$ by carrying out n steps from an initial value $y(x_0)$ using a stepsize equal to $h = (\overline{x} - x_0)/n$.

Lemma 515C *Using notations already introduced in this subsection, together with*

$$E^{[i]} = \begin{bmatrix} \widehat{y}_1^{[i]} - y_1^{[i]} \\ \widehat{y}_2^{[i]} - y_2^{[i]} \\ \vdots \\ \widehat{y}_r^{[i]} - y_r^{[i]} \end{bmatrix}, \qquad i = 0, 1, 2, \ldots, n,$$

for the accumulated error in step i, we have the estimate

$$\|E^{[n]}\| \leq \begin{cases} \exp(\alpha C(\overline{x} - x_0))\|E^{[0]}\| + \frac{\beta h}{\alpha}(\exp(\alpha C(\overline{x} - x_0)) - 1), & \alpha > 0, \\ \exp(\alpha C(\overline{x} - x_0))\|E^{[0]}\| + \beta C(\overline{x} - x_0)h, & \alpha = 0, \end{cases}$$

where $C = \sup_{i=0,1,\ldots}\|V^i\|_{\infty}$ and the norm of $E^{[n]}$ is defined as the maximum of the norms of its r subvectors.

Proof. The result of Lemma 515B can be written in the form

$$E^{[i]} = (V \otimes I)E^{[i-1]} + K^{[i]},$$

from which it follows that

$$E^{[i]} = (V^i \otimes I)E^{[0]} + \sum_{j=1}^{i}(V^{j-1} \otimes I)K^{[i+1-j]},$$

and hence that

$$\|E^{[i]}\| \leq C\|E^{[0]}\| + \sum_{j=0}^{i-1}C\|K^{[i-j]}\|.$$

Insert the known bounds on the terms on the right-hand side, and we find

$$\|E^{[i]}\| \leq \alpha hC\sum_{j=0}^{i-1}\|E^{[j]}\| + Ci\beta h^2 + C\|E^{[0]}\|.$$

This means that $\|E^{[i]}\|$ is bounded by η_i defined by

$$\eta_i = \alpha h C \sum_{j=0}^{i-1} \eta_j + C i \beta h^2 + \eta_0, \qquad \eta_0 = C \|E^{[0]}\|.$$

To simplify this equation, find the difference of the formulae for η_i and η_{i-1} to give the difference equation

$$\eta_i - \eta_{i-1} = \alpha h C \eta_{i-1} + C \beta h^2$$

with solution

$$\eta_i = (1 + h\alpha C)^i \eta_0 + \frac{\beta h}{\alpha}((1 + h\alpha C)^i - 1),$$

or, if $\alpha = 0$,

$$\eta_i = \eta_0 + i C \beta h^2.$$

Substitute $i = n$ and we complete the proof.   □

We summarize the implications of these results:

Theorem 515D *A stable and consistent general linear method is convergent.*

Exercises 51

51.1 Show that the general linear method

$$\left[\begin{array}{c|cc} 0 & 1 & a \\ \hline b & 1 & 0 \\ c & 0 & 0 \end{array}\right]$$

is preconsistent with $u = [1, 0]^{\mathsf{T}}$. For what values of a, b and c is the method consistent?

51.2 Show that a linear multistep method, interpreted as a general linear method, is convergent if and only if the corresponding one-leg method is convergent.

51.3 For what values of a, b, c, d, e, f and g is the method

$$\left[\begin{array}{c|cc} 0 & a & b \\ \hline c & d & e \\ f & 0 & g \end{array}\right]$$

capable of producing convergent approximations?

52 The Stability of General Linear Methods

520 Introduction

The linear stability analysis of general linear methods, as for the special cases of Runge–Kutta and linear multistep methods, is based on the differential equation

$$y'(x) = qy(x). \tag{520a}$$

The idea will be to consider the influence of a single step of the method on an incoming vector $y^{[n-1]}$. We obtain a relation of the form

$$y^{[n]} = M(z)y^{[n-1]}, \tag{520b}$$

where $z = hq$ and $M(z)$ is an $r \times r$ matrix-valued function of the complex variable z.

Definition 520A *For a general linear method (A, U, B, V), the 'stability matrix' $M(z)$ is defined by*

$$M(z) = V + zB(I - zA)^{-1}U.$$

As we have anticipated, we have the following result:

Theorem 520B *Let $M(z)$ denote the stability matrix for a general linear method. Then, for a linear differential equation (520a), (520b) holds with $z = hq$.*

Proof. For the special problem defined by $f(y) = qy$, the vector of stage derivatives F is related to the vector of stage values Y by $F = qY$. Hence, (500c) reduces to the form

$$
\begin{bmatrix} Y \\ y^{[n]} \end{bmatrix} = \begin{bmatrix} A & U \\ B & V \end{bmatrix} \begin{bmatrix} zY \\ y^{[n-1]} \end{bmatrix}.
$$

It follows that $Y = (I - zA)^{-1}Uy^{[n-1]}$, and that

$$y^{[n]} = zBY + Vy^{[n-1]} = M(z)y^{[n-1]}. \qquad \square$$

If the method is stable, in the sense of Section 51, then $M(0) = V$ will be power-bounded. The idea now is to extend this to values of z in the complex plane where $M(z)$ has bounded powers.

Just as for Runge–Kutta and linear multistep methods, associated with each method is a stability region. This, in turn, is related to the characteristic polynomial of $M(z)$.

Definition 520C *Let (A, U, B, V) denote a general linear method and $M(z)$ the corresponding stability matrix. The 'stability matrix' for the method is the polynomial $\Phi(w, z)$ given by*

$$\Phi(w, z) = \det(wI - M(z)),$$

and the 'stability region' is the subset of the complex plane such that if z is in this subset, then

$$\sup_{n=1}^{\infty} \|M(z)^n\| < \infty.$$

We refer to the 'instability region' as the complement of the stability region.

Note that in applications of these definitions, $\Phi(w, z)$ may be a rational function. Quite often, the essential properties will be contained in just the numerator of this expression. We equally refer to the numerator of this rational function as the stability function.

We state the following obvious result without proof.

Theorem 520D *The instability region for (A, U, B, V) is a subset of the set of points z, such that $\Phi(w, z) = 0$, where $|w| \geq 1$. The instability region is a superset of the points defined by $\Phi(w, z) = 0$, where $|w| > 1$.*

The unanswered question in this result is: 'Which points on the boundary of the stability region are actually members of it?' This is not always a crucial question, and we quite often interpret the stability region as the 'strict stability region', consisting of those z for which

$$\lim_{n \to \infty} \|M(z)^n\| = 0.$$

This will correspond to the set of z values such that $|w| < 1$, for any w satisfying $\Phi(w, z) = 0$.

In particular, we can define A-stability.

Definition 520E *A general linear method is 'A-stable' if $M(z)$ is power-bounded for every z in the left half complex plane.*

Just as for Runge–Kutta and linear multistep methods, A-stability is the ideal property for a method to possess for it to be applicable to stiff problems. Corresponding to the further requirement for Runge–Kutta methods that $R(\infty) = 0$, we have the generalization of L-stability to general linear methods.

Definition 520F *A general linear method is L-stable if it is A-stable and $\rho(M(\infty)) = 0$.*

521   Methods with maximal stability order

Although a full discussion of the order of general linear methods will be postponed until Section 53, we look here at the relationship between stability and methods with a property closely related to order.

Definition 521A *A method with stability function $\Phi(w, z)$ has 'stability order' $\widetilde{p}$ if*

$$\Phi(\exp(z), z) = O(z^{\widetilde{p}+1}).$$

Suppose the stability function is given by

$$\Phi(w, z) = \sum_{j=0}^{k} w^{k-j} \sum_{l=0}^{\nu_j} \alpha_{jl} z^j,$$

where k is the w-degree of Φ and ν_j is the z-degree of the coefficient of w^{k-j}. We can regard the sequence of integers

$$\nu = [\nu_0, \nu_1, \ldots, \nu_k],$$

as representing the complexity of the stability function Φ. To include all sensible cases without serious redundancies, we always assume that $\nu_j \geq -1$ for $j = 0, 1, 2, \ldots, k$ with strict inequality in the cases $j = 0$ and $j = k$.

It is interesting to ask the question: 'For a given sequence ν, what is the highest possible stability order?'. The question can be looked at in two parts. First, there is the question of determining for what $\widetilde{p}$ it is possible to find a function Φ with a given complexity and with stability order $\widetilde{p}$. Secondly, there is the question of finding a general linear method corresponding to a given Φ, with order p as close as possible to $\widetilde{p}$. The first half of the question can be firmly answered and is interesting since it gives rise to speculations about possible generalizations of the Ehle results on rational approximations to the exponential function. The definitive result that we have referred to is as follows:

Theorem 521B *For given ν, the maximum possible stability order is given by*

$$\widetilde{p} = \sum_{j=0}^{k} (\nu_j + 1) - 2. \tag{521a}$$

Proof. If order higher than $\widetilde{p}$ given by (521a) is possible, then

$$\sum_{j=0}^{k} \exp((k-j)z) \sum_{l=0}^{\nu_j} \alpha_{jl} z^l = C_{\widetilde{p}+2} z^{\widetilde{p}+2} + C_{\widetilde{p}+3} z^{\widetilde{p}+3} + \cdots,$$

where the right-hand side is convergent for any z. Differentiate $\nu_k + 1$ times and multiply the result by $\exp(-z)$. We now have a stability function with complexity $[\nu_0, \nu_1, \ldots, \nu_{k-1}]$, where the w-degree can be reduced even further if $\nu_{k-1} = -1$. Furthermore, the new approximation also has a stability order contrary to the bound we are trying to prove. Thus, by an induction argument

we reduce to the case $k = 0$, and it remains to prove that there does not exist a non-zero polynomial P of degree ν_0 such that

$$P(z) = O(z^{\nu_0+1}).$$

To show that an approximation with stability order $\widetilde{p}$ given by (521a) exists, it is possible to reverse the non-existence argument and to construct the required stability function recursively, but we use a different approach.

Consider the rational function

$$\phi(t) = \prod_{j=0}^{k}(t + j)^{-\nu_j-1}, \qquad (521b)$$

with partial fraction expansion which can be written in the form

$$\phi(t) = \sum_{j=0}^{k}\sum_{l=0}^{\nu_j} \frac{l!\alpha_{jl}}{(j + t)^{l+1}}.$$

Calculate the integral

$$\frac{1}{2\pi i}\oint_C \phi(t)\exp_{\widetilde{p}}(tz)dt, \qquad (521c)$$

where

$$\exp_{\widetilde{p}}(z) = \sum_{j=0}^{\widetilde{p}} \frac{z^j}{j!}$$

is the polynomial of degree $\widetilde{p}$ approximating the exponential function to within $O(z^{\widetilde{p}+1})$ and C is a circular counter-clockwise contour, centred at 0 and with radius $R > k$. Using the partial fraction form of ϕ, (521c) is found to be

$$\sum_{j=0}^{k}\sum_{l=0}^{\nu_j} \alpha_{jl}z^l \exp_{\widetilde{p}-l}(-zj), \qquad (521d)$$

but using (521b), the integral can be bounded in terms of R^{-1} for large R, and is therefore zero. Use the fact that $z^l \exp_{\widetilde{p}-l}(-zj) = z^l \exp(-zj) + O(z^{\widetilde{p}+1})$ and the result follows. □

Because of the maximal order properties of these approximations, they will be known as 'generalized Padé approximations'. Some examples are given in Table 521(I). In each case, $\Phi(w, z)$ is scaled so that the coefficient of $w^k z^0$ is 1. Some of these functions correspond to A-stable methods, and this is indicated in the table. The entry for $\nu = [1, 0, 1]$ is reducible, in the sense that $\Phi(w, z)$ factorizes into the approximation for $[1, 1]$ multiplied by $w - 1$; the order 3 suggested for this method is, of course, an illusion.

Table 521(I) Some generalized Padé approximations

ν	$\widetilde{p}$	$\Phi(w, z)$	Remarks
$[1, 0, 0]$	2	$(1 - \frac{2}{3}z)w^2 - \frac{4}{3}w + \frac{1}{3}$	A-stable
$[1, 0, 1]$	3	$(1 - \frac{1}{2}z)w^2 - 2w + 1 + \frac{1}{2}z$	A-stable
$[1, 1, 0]$	3	$(1 - \frac{2}{5}z)w^2 - (\frac{4}{5} + \frac{4}{5}z)w - \frac{1}{5}$	
$[2, 0, 0]$	3	$(1 - \frac{6}{7}z + \frac{2}{7}z^2)w^2 - \frac{8}{7}w + \frac{1}{7}$	A-stable
$[2, 0, 1]$	4	$(1 - \frac{8}{11}z + \frac{2}{11}z^2)w^2 - \frac{16}{11}w + \frac{5}{11} + \frac{2}{11}z$	A-stable
$[2, 1, 0]$	4	$(1 - \frac{10}{17}z + \frac{2}{17}z^2)w^2 - (\frac{16}{17} + \frac{8}{17}z)w - \frac{1}{17}$	A-stable
$[2, 0, 2]$	5	$(1 - \frac{5}{8}z + \frac{1}{8}z^2)w^2 - 2w + 1 + \frac{5}{8}z + \frac{1}{8}z^2$	see text
$[2, 1, 2]$	6	$(1 - \frac{7}{15}z + \frac{1}{15}z^2)w^2 - \frac{16}{15}zw - 1 - \frac{7}{15}z - \frac{1}{15}z^2$	
$[3, 0, 0]$	4	$(1 - \frac{14}{15}z + \frac{2}{5}z^2 - \frac{4}{45}z^3)w^2 - \frac{16}{15}w + \frac{1}{15}$	A-stable
$[4, 0, 0]$	5	$(1 - \frac{30}{31}z + \frac{14}{31}z^2 - \frac{4}{31}z^3 + \frac{2}{93}z^4)w^2 - \frac{32}{31}w + \frac{1}{31}$	

The approximation based on $\nu = [2, 0, 2]$ is especially interesting. According to the result formerly known as the Daniel–Moore conjecture (Daniel and Moore, 1970), it cannot correspond to an A-stable method and also have order $p = 5$, because it does not satisfy the necessary condition $p \leq 2s$. However, the solutions to the equation $\Phi(w, z) = 0$ for $z = iy$ satisfy

$$|w|^2 = \left| \frac{8 \pm iy\sqrt{9 + y^2}}{8 - y^2 - 5iy} \right|^2 = 1.$$

By the maximum modulus principle, the bound $|w| \leq 1$ holds in the left half-plane and the only point in the *closed* left half-plane where the two w roots have equal values on the unit circle is when $z = 0$. For Obreshkov methods we have to regard this as representing instability in the sense of Dahlquist. On the other hand, general linear methods with this stability function exist with $V = I$ and therefore convergent methods are definitely possible. A possible method satisfying this requirement is

$$\left[\begin{array}{cc|cc} \frac{5}{16} & \frac{107}{48} & 1 & 0 \\ -\frac{21}{1712} & \frac{5}{16} & 0 & 1 \\ \hline \frac{775}{856} & -\frac{99}{8} & 1 & 0 \\ -\frac{459}{91592} & \frac{295}{856} & 0 & 1 \end{array} \right].$$

Although $\Phi(\exp(z), z) = O(z^6)$, the order is only 4 because the solution to $\Phi(w, z) = 0$ which is 'principal' in the sense that it is a good approximation to $\exp(z)$, is

$$w = \frac{1 + \frac{3}{8}z\sqrt{1 - \frac{1}{9}z^2}}{1 - \frac{5}{8}z + \frac{1}{8}z^2} = \exp(z) - \frac{1}{270}z^5 + O(z^6).$$

In Butcher and Chipman (1992), the search for possible ν corresponding to A-stable methods was focused on the cases $2\nu_0 - \widetilde{p} \in \{0, 1, 2\}$. For $k = 1$ (the one-step case), this is necessary and sufficient for A-stability. It seems to be the case that, even for $k > 1$, those methods for which $2\nu_0 - \widetilde{p} > 2$ cannot be A-stable. This proposition has become known as the 'Butcher–Chipman conjecture'. A partial proof was given in Butcher (2002), restricted to the cases $2\nu_0 - \widetilde{p} = 3, 4, 7, 8, 11, 12, \ldots$, and a complete proof is given in Butcher (2008). An outline of the argument will be given in Subsection 522.

522   Outline proof of the Butcher–Chipman conjecture

The essential elements of the proof are just as in the proof of Theorem 355G. That is, the result hinges on the fact that if $2\nu_0 - \widetilde{p} > 2$, then an up arrow from zero must be tangential to the imaginary axis, or protrude into the left half-plane, and terminate at a pole. This will mean that this pole will be in the left half-plane or else the arrow will have to cross the imaginary axis to reach this pole.

The missing detail, which we will now focus on, is the fact that each pole is at the termination of an up arrow from zero. We cannot prove this in a simple way based on non-crossing of up and down arrows, because the relation $\Phi(w\exp(z), z) = 0$ now defines a Riemann surface, rather that w as a function of z. The way we will proceed is (i) to modify the order arrow diagram slightly to avoid the need to deal in a special way with special points which arise in the diagram and (ii) to look at changes in the structure of the diagram as the approximation is changed smoothly from one approximation to another.

The modification to arrow diagrams is illustrated in the case of the $[4, 2]$ Padé approximation. Consider Figure 522(i), where two versions of the arrow system are presented. On the left is the standard diagram and on the right is its modified form. The modifications are of two types. First, all arrows are moved an infinitesimal distance to the right to avoid an ambiguity caused by 'stagnation points', such as at $z = 0$. The ambiguity is that an up arrow arriving at a stagnation point is equally related to arrows leaving this point on the left and on the right. Under the modification, this arrow can be regarded as being continued as an up arrow to the right. For example, in the approximation shown in Figure 522(i), arrows arrive in directions $0, 2\pi/7, 4\pi/7, \ldots, 12\pi/7$. In the diagram on the right these are continued unambiguously as outgoing up arrows in the directions $0\pi/7, 3\pi/7, 5\pi/7, \ldots, 13\pi/7$, respectively. The second
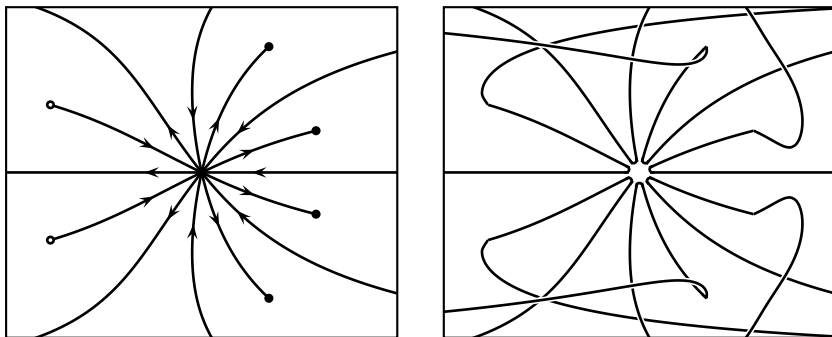
Figure 522(i)   Unmodified (left) and modified (right) order arrows for the
approximation [4, 2]

modification is to replace poles and zeros as termination points for up and down
arrows respectively, by additional sheets in the Riemann surface. The way this
done, in the case of poles, is to introduce the approximation defined by

$$(1 - t)\Phi(w, z) + t\widetilde{\Phi}(w, z),$$

where $\widetilde{\Phi}$ is defined from $\begin{bmatrix} 0 & \nu_0 & \nu_1 & \cdots & \nu_k \end{bmatrix}$, normalized so that $\widetilde{P}_1(0) = 1$.
If we take the limit as $t \to 0$, the Riemann surface limit does not exist but
the projection of the new sheet onto the z plane does exist. This new plane
has the same projection as the order arrow system for

$$\pm \exp(z) + P_1(z),$$

where the sign is chosen to agree with the coefficient of w^{r+1} in $\widetilde{\Phi}(w, z)$. A
similar construction is used for a new bottom sheet defined from the zeros of
P_k. This means that the artificial bottom sheet is found as the limit as $t \to 0$
of the arrow system for $w\Phi(w, z) \pm t$. There is no reason why this should not
be replaced by $w^n\Phi(w, z) \pm t$ where n is any positive integer and we would
obtain similar behaviour.

   Given an order p approximation $[\nu_0, \nu_1, \ldots, \nu_k]$, denoted by Φ, we can
construct, for any $t \in [0, 1]$, the approximation $\Phi_t = t\Phi + (1-t)\Phi_0$, where Φ_0 is
the $[\nu_0, \nu_1, \ldots, \nu_r - 1]$ approximation of order $p-1$. Because of the uniqueness
of generalized Padé approximations, Φ_t will have order only $p - 1$ if $t < 1$.
The parameter t now takes the role of homotopy variable and we will consider
the structure of the arrow system as t moves from 0 to 1. We illustrate in
Figure 522(ii) what happens in a series of diagrams in the case $p = 4, \nu_0 = 2$,
for $t = 0$, $t = 0+$ (a small positive value), $t = 1-$ (a value less than but
close to 1) and $t = 1$. Note that these are stylized diagrams and apply to
a generic situation. That is, they could apply to any of the approximations,
$[2, 2], [2, 1, 0], [2, 0, 1], [2, 0, 0]$ etc. Furthermore, the diagrams are distorted to

| $t = 0$ | $t = 0+$ | $t = 1-$ | $t = 1$ |
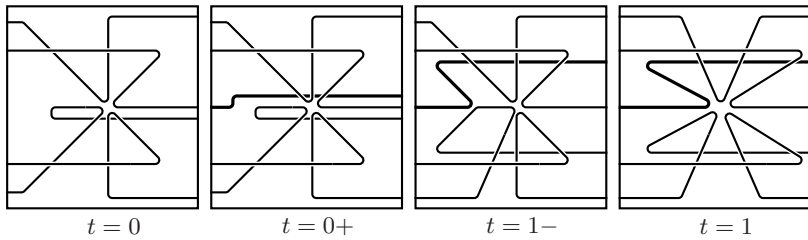
Figure 522(ii)   Homotopy from an order 3 to an order 4 approximation

(a) up arrow vertical   (b) pole on left   (c) pole on right
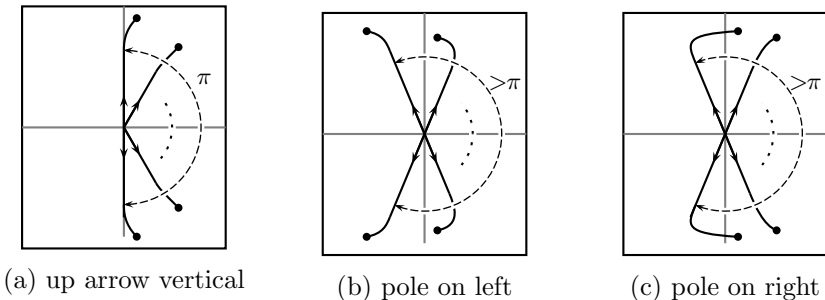
Figure 522(iii)   Illustrating the impossibility of A-stable methods with
$$2\nu_0 - p > 2$$

avoid overlapping lines. For $t > 0$, a new arrow is introduced; this is shown as a prominent line. As t approaches 1, it moves into position as an additional up arrow to 0 and an additional up arrow away from 0.

In such a homotopic sequence as this, it is not possible that an up arrow associated with a pole is detached from 0 because either this would mean a loss of order or else the new arrow would have to pass through 0 to compensate for this. However, at the instant when this happens, the order would have been raised to p, which is impossible because of the uniqueness of the $[\nu_0, \nu_1, \ldots, \nu_k]$ approximation.

To complete this outline proof, we recall the identical final step in the proof of Theorem 355G which is illustrated in Figure 522(iii). If $2\nu_0 > p+2$, then the up arrows which terminate at poles subtend an angle $(\nu_0 - 1)2\pi/(p+1) \geq \pi$. If this angle is π, as in (a) in this figure, then there will be an up arrow leaving 0 in a direction tangential to the imaginary axis. Thus there will be points on the imaginary axis where $|w| > 1$. In the case of (b), an up arrow terminates at a pole in the left half-plane, again making A-stability impossible. Finally, in (c), where an up arrow leaves 0 and passes into the left half-plane, but returns to the right half-plane to terminate at a pole, it must have crossed the imaginary axis. Hence, as in (a), there are points on the imaginary axis where $|w| > 1$ and A-stability is not possible.

523 Non-linear stability

We will consider an example of an A-stable linear multistep method based on the function

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

As a linear multistep method this is

$$\left[\begin{array}{c|cccc} 1 & \tfrac{1}{2} & \tfrac{1}{2} & -\tfrac{1}{4} & \tfrac{3}{4} \\ \hline 1 & \tfrac{1}{2} & \tfrac{1}{2} & -\tfrac{1}{4} & \tfrac{3}{4} \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right],$$

where the input to step n consists of the vectors $y_{n-1}, y_{n-2}, hf(y_{n-1}), hf(y_{n-2})$, respectively.

To understand the behaviour of this type of method with a dissipative problem, Dahlquist (1976) analysed the corresponding one-leg method. However, with the general linear formulation, the analysis can be carried out directly. We first carry out a transformation of the input and output variables to the form

$$\left[\begin{array}{cc} A & UT^{-1} \\ TB & TVT^{-1} \end{array}\right],$$

where

$$T = \left[\begin{array}{cccc} \tfrac{2}{3} & \tfrac{1}{3} & \tfrac{1}{3} & \tfrac{1}{2} \\ \tfrac{1}{3} & -\tfrac{1}{3} & \tfrac{7}{6} & -\tfrac{1}{2} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right].$$

The resulting method is found to be

$$\left[\begin{array}{c|cccc} 1 & 1 & -\tfrac{1}{2} & 0 & 0 \\ \hline 1 & 1 & 0 & 0 & 0 \\ \tfrac{3}{2} & 1 & -\tfrac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right].$$

Because the first two output values in the transformed formulation do not depend in any way on the final two input values, these values, and the final two output values, can be deleted from the formulation. Thus, we have the reduced method

$$\left[\begin{array}{c|cc} 1 & 1 & -\tfrac{1}{2} \\ \hline 1 & 1 & 0 \\ \tfrac{3}{2} & 0 & -\tfrac{1}{2} \end{array}\right]. \tag{523a}$$

From the coefficients in the first two rows of T, we identify the inputs in (523a) with specific combinations of the input values in the original formulation:

$$y_1^{[n-1]} = \tfrac{2}{3}y_{n-1} + \tfrac{1}{3}y_{n-2} + \tfrac{1}{3}hf(y_{n-1}) + \tfrac{1}{2}hf(y_{n-2}),$$
$$y_2^{[n-1]} = \tfrac{1}{3}y_{n-1} - \tfrac{1}{3}y_{n-2} + \tfrac{7}{6}hf(y_{n-1}) - \tfrac{1}{2}hf(y_{n-2}).$$

Stable behaviour of this method with a dissipative problem hinges on the verifiable identity

$$\|y_1^{n]}\|^2 + \tfrac{1}{3}\|y_2^{[n]}\|^2 = \|y_1^{n-1]}\|^2 + \tfrac{1}{3}\|y_2^{[n-1]}\|^2$$
$$+ 2\langle hf(Y), Y\rangle - \tfrac{1}{4}\|y_2^{[n-1]} - hf(Y)\|^2.$$

This means that if $2\langle hf(Y), Y\rangle \leq 0$, then $\|y^{[n]}\|_G \leq \|y^{[n-1]}\|_G$, where $G = \mathrm{diag}(1, \tfrac{1}{3})$.

Given an arbitrary general linear method, we ask when a similar analysis can be performed. It is natural to restrict ourselves to methods without unnecessary inputs, outputs or stages; such irreducible methods are discussed in Butcher (1987a).

As a first step we consider how to generalize the use of the G norm. Let G denote an $r \times r$ positive semi-definite matrix. For $u, v \in \mathbb{R}^{rN}$ made up from subvectors $u_1, u_2, \ldots, u_r \in \mathbb{R}^N$, $v_1, v_2, \ldots, v_r \in \mathbb{R}^N$, respectively, define $\langle \cdot, \cdot \rangle_G$ and the corresponding semi-norm $\| \cdot \|_G$ as

$$\langle u, v \rangle_G = \sum_{i,j=1}^{r} g_{ij}\langle u_i, v_j\rangle,$$

$$\|u\|_G^2 = \langle u, u \rangle_G.$$

We will also need to consider vectors $U \oplus u \in \mathbb{R}^{(s+r)N}$, made up from subvectors $U_1, U_2, \ldots, U_s, u_1, u_2, \ldots, u_r \in \mathbb{R}^N$. Given a positive semi-definite $(s+t) \times (s+r)$ matrix M, we will define $\|U \oplus u\|_M$ in a similar way. Given a diagonal $s \times s$ matrix D, with diagonal elements $d_i \geq 0$, we will also write $\langle U, V \rangle_D$ as $\sum_{i=1}^{s} d_i\langle U_i, V_i\rangle$. Using this terminology we have the following result:

Theorem 523A *Let Y denote the vector of stage values, F the vector of stage derivatives and $y^{[n-1]}$ and $y^{[n]}$ the input and output respectively from a single step of a general linear method (A, U, B, V). Assume that M is a positive semi-definite $(s+r) \times (s+r)$ matrix, where*

$$M = \begin{bmatrix} DA + A^\mathsf{T}D - B^\mathsf{T}GB & DU - B^\mathsf{T}GV \\ U^\mathsf{T}D - V^\mathsf{T}GB & G - V^\mathsf{T}GV \end{bmatrix}, \qquad (523b)$$

with G a positive semi-definite $r \times r$ matrix and D a positive semi-definite diagonal $s \times s$ matrix. Then

$$\|y^{[n]}\|_G^2 = \|y^{[n-1]}\|_G^2 + 2\langle hF, Y\rangle_D - \|hF \oplus y^{[n-1]}\|_M^2.$$

Proof. The result is equivalent to the identity

$$M = \begin{bmatrix} 0 & 0 \\ 0 & G \end{bmatrix} - \begin{bmatrix} B^\mathsf{T} \\ V^\mathsf{T} \end{bmatrix} G \begin{bmatrix} B & V \end{bmatrix} + \begin{bmatrix} D \\ 0 \end{bmatrix} \begin{bmatrix} A & U \end{bmatrix} + \begin{bmatrix} A^\mathsf{T} \\ U^\mathsf{T} \end{bmatrix} \begin{bmatrix} D & 0 \end{bmatrix}. \qquad \square$$

We are now in a position to extend the algebraic stability concept to the general linear case.

Theorem 523B *If M given by (523b) is positive semi-definite, then*

$$\|y^{[n]}\|_G^2 \le \|y^{[n-1]}\|_G^2.$$

524   Reducible linear multistep methods and G-stability

We consider the possibility of analysing the possible non-linear stability of linear multistep methods without using one-leg methods. First note that a linear k-step method, written as a general linear method with $r = 2k$ inputs, is reducible to a method with only k inputs. For the standard k-step method written in the form (400b), we interpret $hf(x_{n-i}, y_{n-i})$, $i = 1, 2, \ldots, k$, as having already been evaluated from the corresponding y_{n-i}. Define the input vector $y^{[n-1]}$ by

$$y_i^{[n-1]} = \sum_{j=i}^k \left(\alpha_j y_{n-j+i-1} + \beta_j h f(x_{n-j+i}, y_{n-j+i-1}) \right), \quad i = 1, 2, \ldots, k,$$

so that the single stage $Y = y_n$ satisfies

$$Y = h\beta_0 f(x_n, Y) + y_1^{[n-1]}$$

and the output vector can be found from

$$y_i^{[n]} = \alpha_i y_1^{[n-1]} + y_{i+1}^{[n]} + (\beta_0 \alpha_i + \beta_i) h f(x_n, Y),$$

where the term $y_{i+1}^{[n]}$ is omitted when $i = k$. The reduced method has the defining matrices

$$\begin{bmatrix} A & U \\ B & V \end{bmatrix} = \left[\begin{array}{c|ccccccc} \beta_0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ \hline \beta_0\alpha_1 + \beta_1 & \alpha_1 & 1 & 0 & \cdots & 0 & 0 \\ \beta_0\alpha_2 + \beta_2 & \alpha_2 & 0 & 1 & \cdots & 0 & 0 \\ \beta_0\alpha_3 + \beta_3 & \alpha_3 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \beta_0\alpha_{k-1} + \beta_{k-1} & \alpha_{k-1} & 0 & 0 & \cdots & 0 & 1 \\ \beta_0\alpha_k + \beta_k & \alpha_k & 0 & 0 & \cdots & 0 & 0 \end{array} \right], \qquad (524a)$$

and was shown in Butcher and Hill (2006) to be algebraically stable if it is A-stable.

525   G-symplectic methods

In the special case of Runge–Kutta methods, the matrix M, given by (357d), which arose in the study of non-linear stability, had an additional role. This was in Section 37 where M was used in the characterization of symplectic behaviour. This leads to the question: 'does M, given by (523b), have any significance in terms of symplectic behaviour'?.

For methods for which $M = 0$, although we cannot hope for quadratic invariants to be conserved, a 'G extension' of such an invariant may well be conserved. Although we will show this to be correct, it still has to be asked if there is any computational advantage in methods with this property. The author believes that these methods may have beneficial properties, but it is too early to be definite about this.

The definition, which we now present, will be expressed in terms of the submatrices making up M.

Definition 525A *A general linear method* (A, U, B, V) *is G-symplectic if there exists a positive semi-definite symmetric* $r \times r$ *matrix* G *and an* $s \times s$ *diagonal matrix* D *such that*

$$G = V^{\mathsf{T}}GV, \tag{525a}$$

$$DU = B^{\mathsf{T}}GV, \tag{525b}$$

$$DA + A^{\mathsf{T}}D = B^{\mathsf{T}}GB. \tag{525c}$$

The following example of a G-symplectic method was presented in Butcher (2006):

$$
\begin{bmatrix} A & U \\ B & V \end{bmatrix} = \left[\begin{array}{cc|cc} \frac{3+\sqrt{3}}{6} & 0 & 1 & -\frac{3+2\sqrt{3}}{3} \\ -\frac{\sqrt{3}}{3} & \frac{3+\sqrt{3}}{6} & 1 & \frac{3+2\sqrt{3}}{3} \\ \hline \frac{1}{2} & \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & -1 \end{array} \right]. \tag{525d}
$$

It can be verified that (525d) satisfies (525a)–(525c) with $G = \mathrm{diag}(1, 1 + \frac{2}{3}\sqrt{3})$ and $D = \mathrm{diag}(\frac{1}{2}, \frac{1}{2})$.

Although this method is just one of a large family of such methods which the author, in collaboration with Laura Hewitt and Adrian Hill of Bath University, is trying to learn more about, it is chosen for special attention here. An analysis in Theorem 534A shows that it has order 4 and stage order 2. Although it is based on the same stage abscissae as for the order 4 Gauss Runge–Kutta method, it has a convenient structure in that A is diagonally implicit.

For the harmonic oscillator, the Hamiltonian is supposed to be conserved, and this happens almost exactly for solutions computed by this method *for any number of steps*. Write the problem in the form $y' = iy$ so that for stepsize h, $y^{[n]} = M(ih)y^{[n-1]}$ where M is the stability matrix. Long term conservation
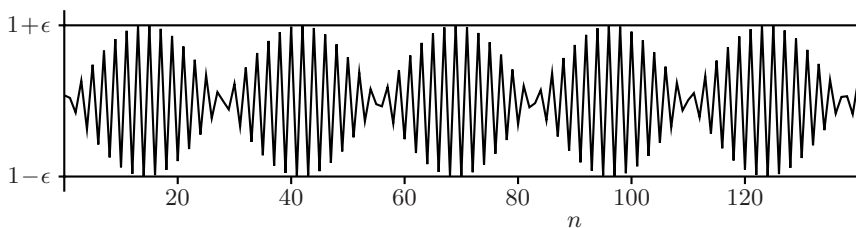
Figure 525(i)   Variation in $|y_1^{[n]}|$ for $n = 0, 1, \ldots, 140$, with $h = 0.1$;
note that $\epsilon = 0.000276$

requires that the characteristic polynomial of $M(ih)$ has both zeros on the unit circle. This characteristic polynomial is:

$$w^2 \left(1 - ih\frac{3+\sqrt{3}}{6}\right)^2 + w\left(\frac{2}{3}i\sqrt{3}\right)h - \left(1 + ih\frac{3+\sqrt{3}}{6}\right)^2.$$

Substitute

$$w = \frac{1 + ih\frac{3+\sqrt{3}}{6}}{1 - ih\frac{3+\sqrt{3}}{6}}iW,$$

and we see that

$$W^2 + h\frac{2\frac{\sqrt{3}}{3}}{1 + h^2(\frac{3+\sqrt{3}}{6})^2}W + 1.$$

The coefficient of W lies in $(-\sqrt{3} + 1, \sqrt{3} - 1)$ and the zeros of this equation are therefore on the unit circle for all real h. We can interpret this as saying that the two terms in

$$\left((p_1^{[n]})^2 + (q_1^{[n]})^2\right) + \left(1 + \tfrac{2}{3}\sqrt{3}\right)\left((p_2^{[n]})^2 + (q_2^{[n]})^2\right)$$

are not only conserved in total but are also approximately conserved individually, as long as there is no round-off error. The justification for this assertion is based on an analysis of the first component of $y_1^{[n]}$ as n varies. Write the eigenvalues of $M(ih)$ as $\lambda(h) = 1 + O(h)$ and $\mu(h) = -1 + O(h)$ and suppose the corresponding eigenvectors, in each case scaled with first component equal to 1, are $u(h)$ and $v(h)$ respectively. If the input $y^{[0]}$ is $au(h) + bv(h)$ then $y_1^{[n]} = a\lambda(h)^n + b\mu(h)^n$ with absolute value

$$|y_1^{[n]}| = \left(a^2 + b^2 + 2ab\mathrm{Re}\left((\overline{\lambda(h)}\mu(h))^n\right)\right)^{1/2}.$$

If $|b/a|$ is small, as it will be for small h if a suitable starting method is used, $|y_1^{[n]}|$ will never depart very far from its initial value. This is illustrated in Figure 525(i) in the case $h = 0.1$.

Exercises 52

52.1 Find the stability matrix and stability function for the general linear method

$$
\left[
\begin{array}{cc|cc}
\frac{1}{2} & 0 & 1 & -\frac{1}{2} \\[4pt]
\frac{4}{3} & \frac{1}{2} & 1 & -\frac{5}{6} \\[4pt]
\hline
\frac{19}{16} & \frac{9}{16} & 1 & -\frac{3}{4} \\[4pt]
\frac{1}{4} & \frac{3}{4} & 0 & 0
\end{array}
\right].
$$

Show that this method A-stable.

52.2 Find a general linear method with stability function equal to the $[2, 0, 0]$ generalized Padé approximation to exp.

52.3 Find the $[3, 0, 1]$ generalized Padé approximation to exp.

52.4 Show that the $[2, 0, 1]$ generalized Padé approximation to exp is A-stable.

53   The Order of General Linear Methods

530   Possible definitions of order

Traditional methods for the approximation of differential equations are designed with a clear-cut interpretation in mind. For example, linear multistep methods are constructed on the assumption that, at the beginning of each step, approximations are available to the solution and to the derivative at a sequence of step points; the calculation performed by the method is intended to obtain approximations to these same quantities but advanced one step ahead. In the case of Runge–Kutta methods, only the approximate solution value at the beginning of a step is needed, and at the end of the step this is advanced one time step further.

We are not committed to these interpretations for either linear multistep or Runge–Kutta methods. For example, in the case of Adams methods, the formulation can be recast so that the data available at the start and finish of a step is expressed in terms of backward difference approximations to the derivative values or in terms of other linear combinations which approximate Nordsieck vectors. For Runge–Kutta methods the natural interpretation, in which y_n is regarded as an approximation to $y(x_n)$, is not the only one possible. As we have seen in Subsection 389, the generalization to effective order is such an alternative interpretation.

For a general linear method, the r approximations, $y_i^{[n-1]}$, $i = 1, 2, \ldots, r$, are imported into step n and the r corresponding approximations, $y_i^{[n]}$, are exported at the end of the step. We do not specify anything about these quantities except to require that they are computable from an approximation to $y(x_n)$ and, conversely, the exact solution can be recovered, at least approximately, from $y_i^{[n-1]}$, $i = 1, 2, \ldots, r$.

This can be achieved by associating with each input quantity, $y_i^{[n-1]}$, a generalized Runge–Kutta method,

$$S_i = \frac{c^{(i)}}{b_0^{(i)}} \left|\begin{array}{c} A^{(i)} \\ \hline b^{(i)T} \end{array}\right. \quad . \tag{530a}$$

Write s_i as the number of stages in S_i. The aim will be to choose these input approximations in such a way that if $y_i^{[n-1]}$ is computed using S_i applied to $y(x_{n-1})$, for $i = 1, 2, \ldots, r$, then the output quantities computed by the method, $y_i^{[n]}$, are close approximations to S_i applied to $y(x_n)$, for $i = 1, 2, \ldots, r$.

We refer to the sequence of r generalized Runge–Kutta methods $S_1, S_2, \ldots, S_r$ as a 'starting method' for the general linear method under consideration and written as S. It is possible to interpret each of the output quantities computed by the method, on the assumption that S is used as a starting method, as itself a generalized Runge–Kutta method with a total of $s + s_1 + s_2 + \cdots + s_r$ stages. It is, in principle, a simple matter to calculate the Taylor expansion for the output quantities of these methods and it is also a simple matter to calculate the Taylor expansion of the result found by shifting the exact solution forward one step. We write SM for the vector of results formed by carrying out a step of M based on the results of computing initial approximations using S. Similarly, ES will denote the vector of approximations formed by advancing the trajectory forward a time step h and then applying each member of the vector of methods that constitutes S to the result of this.

A restriction is necessary on the starting methods that can be used in practice. This is that at least one of S_1, S_2, $\ldots$, S_r, has a non-zero value for the corresponding $b_0^{(i)}$. If $b_0^{(i)} = 0$, for all $i = 1, 2, \ldots, r$, then it would not be possible to construct preconsistent methods or to find a suitable finishing procedure, F say, such that SF becomes the identity method.

Accordingly, we focus on starting methods that are non-degenerate in the following sense.

Definition 530A *A starting method S defined by the generalized Runge–Kutta methods (530a), for $i = 1, 2, \ldots, r$, is 'degenerate' if $b_0^{(i)} = 0$, for $i = 1, 2, \ldots, r$, and 'non-degenerate' otherwise.*

Definition 530B *Consider a general linear method M and a non-degenerate starting method S. The method M has order p relative to S if the results found from SM and ES agree to within $O^{(p+1)}$.*

Definition 530C *A general linear method M has order p if there exists a non-degenerate starting method S such that M has order p relative to S.*
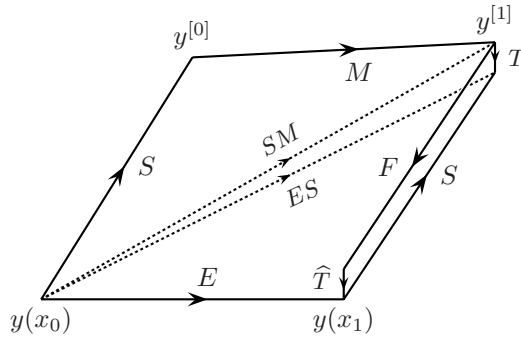
Figure 531(i)   Representation of local truncation error

In using Definition 530C, it is usually necessary to construct, or at least to identify the main features of, the starting method S which gives the definition a practical meaning. In some situations, where a particular interpretation of the method is decided in advance, Definition 530B is used directly. Even though the Taylor series expansions, needed to analyse order, are straightforward to derive, the details can become very complicated. Hence, in Subsection 532, we will build a framework for simplifying the analysis. In the meantime we consider the relationship between local and accumulated error.

531   Local and global truncation errors

Figure 531(i) shows the relationship between the action of a method M with order p, a non-degenerate starting method S, and the action of the exact solution E, related as in Definition 530C. We also include in the diagram the action of a finishing procedure F which exactly undoes the work of S, so that $SF = \text{id}$. In this figure, T represents the truncation error, as the correction that would have to be added to SM to obtain ES. Also shown is $\widehat{T}$, which is the error after carrying out the sequence of operations making up SMF, regarded as an approximation to E. However, in practice, the application of F to the computed result is deferred until a large number of steps have been carried out.

Figure 531(i) illustrates that the purpose of a general linear method is to approximate not the exact solution, but the result of applying S to every point on the solution trajectory. To take this idea further, consider Figure 531(ii), where the result of carrying the approximation over many steps is shown. In step k, the method M is applied to an approximation to $E^{k-1}S$ to yield an approximation to $E^k S$ without resorting to the use of the finishing method F. In fact the use of F is postponed until an output approximation is finally needed.
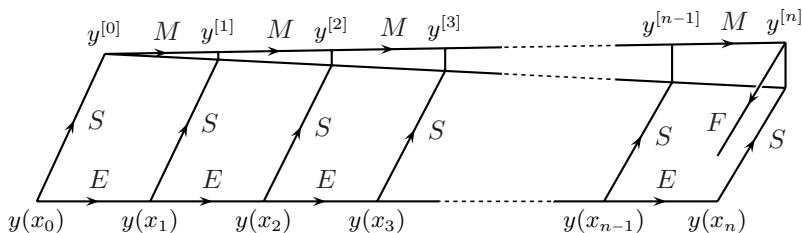
Figure 531(ii)   Representation of global truncation error

532   Algebraic analysis of order

Associated with each of the components of the vector of starting methods is a member of the algebra G introduced in Subsection 385. Denote ξ_i, $i = 1, 2, \ldots, r$, as the member corresponding to S_i. That is, ξ_i is defined by

$$\xi_i(\emptyset) = b_0^{(i)},$$
$$\xi_i(t) = \Phi^{(i)}(t), \quad t \in T,$$

where the elementary weight $\Phi^{(i)}(t)$ is defined from the tableau (530a). Associate $\eta_i \in G_1$ with stage $i = 1, 2, \ldots, s$, and define this recursively by

$$\eta_i = \sum_{j=1}^{s} a_{ij}\eta_j D + \sum_{j=1}^{r} U_{ij}\xi_j. \tag{532a}$$

Having computed η_i and $\eta_i D$, $i = 1, 2, \ldots, s$, we are now in a position to compute the members of G representing the output approximations. These are given by

$$\sum_{j=1}^{s} b_{ij}\eta_j D + \sum_{j=1}^{r} V_{ij}\xi_j, \qquad i = 1, 2, \ldots, r. \tag{532b}$$

If the method is of order p, this will correspond to $E\xi_i$, within H_p. Hence, we may write the algebraic counterpart to the fact that the method M is of order p, relative to the starting method S, as

$$E\xi_i = \sum_{j=1}^{s} b_{ij}\eta_j D + \sum_{j=1}^{r} V_{ij}\xi_j, \qquad \text{in } G/H_p, i = 1, 2, \ldots, r. \tag{532c}$$

Because (532b) represents a Taylor expansion, the expression

$$E\xi_i - \sum_{j=1}^{s} b_{ij}\eta_j D - \sum_{j=1}^{r} V_{ij}\xi_j, \qquad i = 1, 2, \ldots, r, \tag{532d}$$

represents the amount by which $y_i^{[n]}$ falls short of the value that would be found if there were no truncation error. Hence, (532d) is closely related to the local truncation error in approximation i.

Before attempting to examine this in more detail, we introduce a vector notation which makes it possible to simplify the way formulae such as (532a) and (532c) are expressed. The vector counterparts are

$$\eta = A\eta D + U\xi, \qquad (532\text{e})$$
$$E\xi = B\eta D + V\xi, \qquad (532\text{f})$$

where these formulae are to be interpreted in the space G/H_p. That is, the two sides of (532e) and of (532f) are to be equal when evaluated for all $t \in T^{\#}$ such that $r(t) \le p$.

Theorem 532A *Let $M = (A, U, B, V)$ denote a general linear method and let ξ denote the algebraic representation of a starting method S. Assume that (532e) and (532f) hold in G/H_p. Denote*

$$\epsilon = E\xi - B\eta D - V\xi, \qquad \text{in } G.$$

Then the Taylor expansion of $S(y(x_0 + h)) - M(S(y(x_0)))$ is

$$\sum_{r(t)>p} \frac{\epsilon(t)}{\sigma(t)} h^{r(t)} F(t)(y(x_0)). \qquad (532\text{g})$$

Proof. We consider a single step from initial data given at x_0 and consider the Taylor expansion of various expressions about x_0. The input approximation, computed by S, has Taylor series represented by ξ. Suppose the Taylor expansions for the stage values are represented by η so that the stage derivatives will be represented by ηD and these will be related by (532e). The Taylor expansion for the output approximations is represented by $B\eta D + V\xi$, and this will agree with the Taylor expansion of $S(y(x_0 + h))$ up to h^p terms if (532f) holds. The difference from the target value of $S(y(x_0 + h))$ is given by (532g). □

533   An example of the algebraic approach to order

We will consider the modification of a Runge–Kutta method given by (502c). Denote the method by M and a possible starting method by S. Of the two quantities passed between steps, the first is clearly intended to approximate the exact solution and we shall suppose that the starting method for this approximation is the identity method, denoted by 1. The second approximation is intended to be close to the scaled derivative at a nearby point

Table 533(I)   Calculations to verify order $p = 4$ for (502c)

i	0	1	2	3	4	5	6	7	8
t_i	$\emptyset$								
ξ_1	1	0	0	0	0	0	0	0	0
ξ_2	0	1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
η_1	1	0	0	0	0	0	0	0	0
$\eta_1 D$	0	1	0	0	0	0	0	0	0
η_2	1	$\frac{1}{2}$	$-\frac{1}{4}\theta_2$	$-\frac{1}{4}\theta_3$	$-\frac{1}{4}\theta_4$	$-\frac{1}{4}\theta_5$	$-\frac{1}{4}\theta_6$	$-\frac{1}{4}\theta_7$	$-\frac{1}{4}\theta_8$
$\eta_2 D$	0	1	$\frac{1}{2}$	$\frac{1}{4}$	$-\frac{1}{4}\theta_2$	$\frac{1}{8}$	$-\frac{1}{8}\theta_2$	$-\frac{1}{4}\theta_3$	$-\frac{1}{4}\theta_4$
η_3	1	1	$1+\theta_2$	$\frac{1}{2}+\theta_3$	$\theta_4-\frac{1}{2}\theta_2$	$\frac{1}{4}+\theta_5$	$\theta_6-\frac{1}{4}\theta_2$	$\theta_7-\frac{1}{2}\theta_3$	$\theta_8-\frac{1}{2}\theta_4$
$\eta_3 D$	0	1	1	1	$1+\theta_2$	1	$1+\theta_2$	$\frac{1}{2}+\theta_3$	$\theta_4-\frac{1}{2}\theta_2$
$\widehat{\xi}_1$	1	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{6}+\frac{1}{12}\theta_2$	$\frac{1}{12}$	$-\frac{1}{12}\theta_2$
$\widehat{\xi}_2$	0	1	$\frac{1}{2}$	$\frac{1}{4}$	$-\frac{1}{4}\theta_2$	$\frac{1}{8}$	$-\frac{1}{8}\theta_2$	$-\frac{1}{4}\theta_3$	$-\frac{1}{4}\theta_4$
$E\xi_1$	1	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{24}$
$E\xi_2$	0	1	$1+\theta_2$	$1+2\theta_2+\theta_3$	$\frac{1}{2}+\theta_2+\theta_4$	$1+3\theta_2+3\theta_3+\theta_5$	$\frac{1}{2}+\frac{3}{2}\theta_2+\theta_3+\theta_4+\theta_6$	$\frac{1}{3}+\theta_2+2\theta_4+\theta_7$	$\frac{1}{6}+\frac{1}{2}\theta_2+\theta_4+\theta_8$

and we will assume that this is represented by $\theta : T^{\#} \to \mathbb{R}$, where $\theta(\emptyset) = 0$, $\theta(\tau) = 1$. The values of $\theta(t)$ for other trees we will keep as parameters to be chosen. Are there possible values of these parameters for which M has order $p = 4$, relative to S?

We will start with $\xi_1 = 1$ and $\xi_2 = \theta$ and compute in turn η_1, $\eta_1 D$, η_2, $\eta_2 D$, η_3, $\eta_3 D$ and finally the representatives of the output approximations, which we will write here as $\widehat{\xi}_1$ and $\widehat{\xi}_2$. The order requirements are satisfied if and only if values of the free θ values can be chosen so that $\widehat{\xi}_1 = E\xi_1$ and $\widehat{\xi}_2 = E\xi_2$. Reading from the matrix of coefficients for the method, we see that

$$\eta_1 = \xi_1, \qquad\qquad\qquad \eta_2 = \xi_1 - \tfrac{1}{4}\xi_2 + \tfrac{3}{4}\eta_1 D,$$

$$\eta_3 = \xi_1 + \xi_2 - 2\eta_1 D + 2\eta_2 D,$$

$$\widehat{\xi}_1 = \xi_1 + \tfrac{1}{6}\eta_1 D + \tfrac{2}{3}\eta_2 D + \tfrac{1}{6}\eta_3 D, \qquad \widehat{\xi}_2 = \eta_2 D.$$

The details of these calculations are shown in Table 533(I). Comparing the entries in the $\widehat{\xi}_1$ and $E\xi_1$ rows in this table, we see that we get agreement if and only if $\theta_2 = -\frac{1}{2}$. Moving now to the $\widehat{\xi}_2$ and $E\xi_2$ rows, we find that these agree only with specific choices of θ_3, θ_4, ..., θ_8. Thus the method has order 4 relative to S for a unique choice of $\xi_2 = \theta$, which is found to be

$$[\theta_0 \ \theta_1 \ \theta_2 \ \theta_3 \ \theta_4 \ \theta_5 \ \theta_6 \ \theta_7 \ \theta_8] = [0 \ 1 \ -\tfrac{1}{2} \ \tfrac{1}{4} \ \tfrac{1}{8} \ -\tfrac{1}{8} \ -\tfrac{1}{16} \ -\tfrac{7}{48} \ -\tfrac{7}{96}].$$

It might seem from this analysis, that a rather complicated starting method is necessary to obtain fourth order behaviour for this method. However, the method can be started successfully in a rather simple manner. For S_1, no computation is required at all and we can consider defining S_2 using the generalized Runge–Kutta method

$$
\begin{array}{c|cc}
0 & & \\
-\frac{1}{2} & -\frac{1}{2} & \\
\hline
0 & 0 & 1
\end{array} .
$$

This starter, combined with a first step of the general linear method M, causes this first step of the method to revert to the Runge–Kutta method (502b), which was used to motivate the construction of the new method.

534   The order of a G-symplectic method

A second example, for the method (525d), introduced as an example of a G-symplectic method, is amenable to a similar analysis.

Theorem 534A *The following method has order 4 and stage order 2:*

$$
\begin{bmatrix} A & U \\ B & V \end{bmatrix} =
\left[
\begin{array}{cc|cc}
\frac{3+\sqrt{3}}{6} & 0 & 1 & -\frac{3+2\sqrt{3}}{3} \\
-\frac{\sqrt{3}}{3} & -\frac{3+\sqrt{3}}{6} & 1 & \frac{3+2\sqrt{3}}{3} \\
\hline
\frac{1}{2} & \frac{1}{2} & 1 & 0 \\
\frac{1}{2} & -\frac{1}{2} & 0 & -1
\end{array}
\right] . \tag{534a}
$$

Before verifying this result we need to specify the nature of the starting method S and the values of the stage abscissae, c_1 and c_2. From an initial point (x_0, y_0), the starting value is given by

$$ y_1^{[0]} = y_0, $$

$$ y_2^{[0]} = \frac{\sqrt{3}}{12} h^2 y''(x_0) - \frac{\sqrt{3}}{108} h^4 y^{(4)}(x_0) + \frac{9+5\sqrt{3}}{216} h^4 \frac{\partial f}{\partial y} y^{(3)}(x_0), $$

and the abscissa vector is $c = \left[\frac{1}{2} + \frac{1}{6}\sqrt{3} \quad \frac{1}{2} - \frac{1}{6}\sqrt{3} \right]^{\mathsf{T}}$.

Proof. Write ξ_1, ξ_2 as the representations of $y_1^{[0]}$, $y_2^{[0]}$ and η_1, η_2 to represent the stages. The stages have to be found recursively and only the converged values are given in Table 534(I), which shows the sequence of quantities occurring in the calculation. The values given for $\widehat{\xi}_i$ are identical to those for $E\xi_i$, $i = 1, 2$, verifying that the order is 4. Furthermore $\eta_i(t) = E^{(c_i)}(t)$, $i = 1, 2$, for $r(t) \leq 2$, showing stage order 2. □

Table 534(I)   Calculations to verify order $p = 4$ for (534a)

i	0	1	2	3	4	5	6	7	8
t_i	$\emptyset$	$\bullet$	$\mathbf{!}$	V	$\mathbf{!}$	V	V	Y	$\mathbf{!}$
ξ_1	1	0	0	0	0	0	0	0	0
ξ_2	0	0	$\frac{\sqrt3}{12}$	0	0	$-\frac{\sqrt3}{18}$	$-\frac{\sqrt3}{36}$	$\frac{3+\sqrt3}{36}$	$\frac{3+\sqrt3}{72}$
η_1	1	$\frac{3+\sqrt3}{6}$	$\frac{2+\sqrt3}{12}$	$\frac{9+5\sqrt3}{36}$	$\frac{9+5\sqrt3}{72}$	$\frac{11+6\sqrt3}{36}$	$\frac{11+6\sqrt3}{72}$	$\frac{2+\sqrt3}{36}$	$\frac{2+\sqrt3}{72}$
$\eta_1 D$	0	1	$\frac{3+\sqrt3}{6}$	$\frac{2+\sqrt3}{6}$	$\frac{2+\sqrt3}{12}$	$\frac{11+6\sqrt3}{36}$	$\frac{11+6\sqrt3}{72}$	$\frac{9+5\sqrt3}{36}$	$\frac{9+5\sqrt3}{72}$
η_2	1	$\frac{3-\sqrt3}{6}$	$\frac{2-\sqrt3}{12}$	$\frac{3+5\sqrt3}{36}$	$\frac{3+5\sqrt3}{72}$	$\frac{7+6\sqrt3}{36}$	$-\frac{7+6\sqrt3}{72}$	$-\frac{4+3\sqrt3}{36}$	$-\frac{4+3\sqrt3}{72}$
$\eta_2 D$	0	1	$\frac{3-\sqrt3}{6}$	$\frac{2-\sqrt3}{6}$	$\frac{2-\sqrt3}{12}$	$\frac{9-5\sqrt3}{36}$	$\frac{9-5\sqrt3}{72}$	$-\frac{3+5\sqrt3}{36}$	$-\frac{3+5\sqrt3}{72}$
$\widehat{\xi}_1$	1	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{24}$
$\widehat{\xi}_2$	0	0	$\frac{\sqrt3}{12}$	$\frac{\sqrt3}{6}$	$\frac{\sqrt3}{12}$	$\frac{7\sqrt3}{36}$	$\frac{7\sqrt3}{72}$	$\frac{3+4\sqrt3}{36}$	$\frac{3+4\sqrt3}{72}$

535   The underlying one-step method

In much the same way as a formal one-step method could be constructed as an underlying representation of a linear multistep method, as in Subsection 422, a one-step method can be constructed with the same underlying relationship to a general linear method. Consider a general linear method (A, U, B, V) and suppose that the preconsistency vector is u. We can ask if it is possible to find $\xi \in X^r$ and $\eta \in X_1^s$, such that (532e) and (532f) hold *exactly* but with E replaced by $\theta \in X_1$; that is, such that

$$\eta(t) = A(\eta D)(t) + U\xi(t), \tag{535a}$$

$$(\theta\xi)(t) = B(\eta D)(t) + V\xi(t), \tag{535b}$$

for *all* $t \in T^{\#}$. In this case we can interpret θ as representing an underlying one-step method. The notional method represented by θ is not unique, because another solution can be found equal to $\widehat{\theta} = \phi^{-1}\theta\phi$, where $\phi \in X_1$ is arbitrary. We see this by multiplying both sides of (535a) and (535b) by ϕ^{-1} to arrive at the relations

$$\widehat{\eta}(t) = A(\widehat{\eta}D)(t) + U\widehat{\xi}(t),$$

$$(\widehat{\theta\xi})(t) = B(\widehat{\eta}D)(t) + V\widehat{\xi}(t),$$

with $\widehat{\xi} = \phi^{-1}\xi$. We want to explore the existence and uniqueness of the underlying one-step method subject to an additional assumption that some

particular component of ξ has a specific value. As a step towards this aim, we remark that (535a) and (535b) transform in a natural way if the method itself is transformed in the sense of Subsection 501. That is, if the method (A, U, B, V) is transformed to $(A, UT^{-1}, TB, TVT^{-1})$, and (535a) and 535b) hold, then, in the transformed method, ξ transforms to $T\xi$ and θ transforms to $T\theta T^{-1}$. Thus

$$\eta(t) = A(\eta D)(t) + (UT^{-1})(T\xi)(t), \qquad (535c)$$

$$((T\theta T^{-1})(T\xi))(t) = TB(\eta D)(t) + V(T\xi)(t). \qquad (535d)$$

This observation means that we can focus on methods for which $u = e_1$, the first member of the natural basis for $\mathbb{R}^r$, in framing our promised uniqueness result.

Theorem 535A *Let (A, U, B, V) denote a consistent general linear method such that $u = e_1$ and such that*

$$U = [\,\mathbf{1} \quad \widetilde{U}\,], \qquad V = \begin{bmatrix} 1 & \widetilde{v}^{\mathsf{T}} \\ 0 & \widetilde{V} \end{bmatrix},$$

where $1 \notin \sigma(\widetilde{V})$. Then there exists a unique solution to (535a) and (535b) for which $\xi_1 = 1$.

Proof. By carrying out a further transformation if necessary, we may assume without loss of generality that $\widetilde{V}$ is lower triangular. The conditions satisfied by $\xi_i(t)$ $(i = 2, 3, \ldots, r)$, $\eta_i(t)$ $(i = 1, 2, \ldots, s)$ and $\theta(t)$ can now be written in the form

$$(1 - \widetilde{V}_{i,i})\xi_i(t) = \sum_{j=1}^{s} b_{ij}(\eta D)(t) + \sum_{j=2}^{i-1} \widetilde{V}_{i-1,j-1}\xi_j(t),$$

$$\eta_i(t) = \sum_{j=1}^{s} a_{ij}(\eta D)(t) + 1(t) + \sum_{j=2}^{r} \widetilde{U}_{i,j-1}\xi_j(t),$$

$$\theta(t) = \sum_{j=1}^{s} b_{1j}(\eta D)(t) + 1(t) + \sum_{j=2}^{r} \widetilde{v}_{j-1}\xi_j(t).$$

In each of these equations, the right-hand sides involve only trees with order lower than $r(t)$ or terms with order $r(t)$ which have already been evaluated. Hence, the result follows by induction on $r(t)$. □

The extension of the concept of underlying one-step method to general linear methods was introduced in Stoffer (1993).

Although the underlying one-step method is an abstract structure, it has practical consequences. For a method in which $\rho(\tilde{V}) < 1$, the performance of a large number of steps, using constant stepsize, forces the local errors to conform to Theorem 535A. When the stepsize needs to be altered, in accordance with the behaviour of the computed solution, it is desirable to commence the step following the change, with input approximations consistent with what the method would have expected if the new stepsize had been used for many preceding steps. Although this cannot be done precisely, it is possible for some of the most dominant terms in the error expansion to be adjusted in accordance with this requirement. With this adjustment in place, it becomes possible to make use of information from the input vectors, as well as information computed within the step, in the estimation of local truncation errors. It also becomes possible to obtain reliable information that can be used to assess the relative advantages of continuing the integration with an existing method or of moving onto a higher order method. These ideas have already been used to good effect in Butcher and Jackiewicz (2003) and further developments are the subject of ongoing investigations.

Exercises 53

53.1 A numerical method of the form

$$Y_1^{[n]} = y_{n-1} + h\widehat{a}_{11} f\left(x_{n-2} + hc_1, Y_1^{[n-1]}\right) + h\widehat{a}_{12} f\left(x_{n-2} + hc_2, Y_2^{[n-1]}\right)$$
$$+ ha_{11} f\left(x_{n-1} + hc_1, Y_1^{[n]}\right) + ha_{12} f\left(x_{n-1} + hc_2, Y_2^{[n]}\right),$$

$$Y_2^{[n]} = y_{n-1} + h\widehat{a}_{21} f\left(x_{n-2} + hc_1, Y_1^{[n-1]}\right) + h\widehat{a}_{22} f\left(x_{n-2} + hc_2, Y_2^{[n-1]}\right)$$
$$+ ha_{21} f\left(x_{n-1} + hc_1, Y_1^{[n]}\right) + ha_{22} f\left(x_{n-1} + hc_2, Y_2^{[n]}\right),$$

$$y_n = y_{n-1} + h\widehat{b}_1 f\left(x_{n-2} + hc_1, Y_1^{[n-1]}\right) + h\widehat{b}_2 f\left(x_{n-2} + hc_2, Y_2^{[n-1]}\right)$$
$$+ hb_1 f\left(x_{n-1} + hc_1, Y_1^{[n]}\right) + hb_2 f\left(x_{n-1} + hc_2, Y_2^{[n]}\right),$$

is sometimes known as a 'two-step Runge–Kutta method'. Find conditions for this method to have order 4.

53.2 Find an explicit fourth order method ($a_{11} = a_{12} = a_{22} = 0$) of the form given by Exercise 53.1.

53.3 Find an A-stable method of the form given by Exercise 53.1.

54   Methods with Runge–Kutta stability

540   *Design criteria for general linear methods*

We consider some of the structural elements in practical general linear
methods, which are not available together in any single method of either
linear multistep or Runge–Kutta type. High order is an important property,
but high stage order is also desirable. For single-value methods this is only
achievable when a high degree of implicitness is present, but this increases
implementation costs. To avoid these excessive costs, a diagonally implicit
structure is needed but this is incompatible with high stage order in the case
of one-value methods. Hence, we will search for good methods within the large
family of multistage, multivalue methods.

The additional complexity resulting from the use of diagonally implicit
general linear methods makes good stability difficult to analyse or even
achieve. Hence, some special assumptions need to be made. In Subsection 541
we present one attempt at obtaining a manageable structure using DIMSIM
methods. We then investigate further methods which have the Runge–Kutta
stability property so that the wealth of knowledge available for the stability
of Runge–Kutta methods becomes available. Most importantly we consider
methods with the Inherent Runge–Kutta stability property, introduced in
Subsection 551.

541   *The types of DIMSIM methods*

'Diagonally implicit multistage integration methods' (DIMSIMs) were
introduced in Butcher (1995a). A DIMSIM is loosely defined as a method
in which the four integers p (the order), q (the stage order), r (the number
of data vectors passed between steps) and s (the number of stages) are all
approximately equal. To be a DIMSIM, a method must also have a diagonally
implicit structure. This means that the $s \times s$ matrix A has the form

$$
A = \begin{bmatrix}
\lambda & 0 & 0 & \cdots & 0 \\
a_{21} & \lambda & 0 & \cdots & 0 \\
a_{31} & a_{32} & \lambda & \cdots & 0 \\
\vdots & \vdots & \vdots & & \vdots \\
a_{s1} & a_{s2} & a_{s3} & \cdots & \lambda
\end{bmatrix},
$$

where $\lambda \geq 0$. The rationale for this restriction on this coefficient matrix is that
the stages can be computed sequentially, or in parallel if the lower triangular
part of A is zero. This will lead to a considerable saving over a method in which
A has a general implicit structure. For Runge–Kutta methods, where $r = 1$,
this sort of method is referred to as explicit if $\lambda = 0$ or as diagonally implicit
(DIRK, or as singly diagonally implicit or SDIRK) if $\lambda > 0$; see Subsection 361.

Table 541(I)   Types of DIMSIM and related methods

Type	A					Application	Architecture
1	$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ a_{21} & 0 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ a_{s1} & a_{s2} & a_{s3} & \cdots & 0 \end{bmatrix}$					Non-stiff	Sequential
2	$\begin{bmatrix} \lambda & 0 & 0 & \cdots & 0 \\ a_{21} & \lambda & 0 & \cdots & 0 \\ a_{31} & a_{32} & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ a_{s1} & a_{s2} & a_{s3} & \cdots & \lambda \end{bmatrix}$					Stiff	Sequential
3	$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$					Non-stiff	Parallel
4	$\begin{bmatrix} \lambda & 0 & 0 & \cdots & 0 \\ 0 & \lambda & 0 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{bmatrix}$					Stiff	Parallel

While these Runge–Kutta methods suffer from the disadvantages associated with low stage order, there is no such difficulty associated with general linear methods.

For non-stiff problems, it is advantageous to choose $\lambda = 0$, whereas for stiff problems, it is necessary that $\lambda > 0$, if A-stability is to be achieved. Furthermore, as we have already remarked, parallel evaluation of the stages is only possible if A is a diagonal matrix; specifically, this would be the zero matrix in the non-stiff case. From these considerations, we introduce the 'types' of a DIMSIM method, and we retain this terminology for methods with a similar structure.

The four types, together with their main characteristics, are shown in Table 541(I). The aim in DIMSIM methods has been to find methods in which p, q, r and s are equal, or approximately equal, and at the same time to choose V as a simple matrix, for example a matrix with rank 1.

If $p = q$, it is a simple matter to write down conditions for this order and stage order. We have the following result:

Theorem 541A *A method*

$$\begin{bmatrix} A & U \\ B & V \end{bmatrix},$$

has order and stage order p if and only if there exists a function

$$\phi : \mathbb{C} \to \mathbb{C}^r,$$

analytic in a neighbourhood of 0, such that

$$\exp(cz) = zA\exp(cz) + U\phi(z) + O(z^{p+1}), \qquad (541a)$$
$$\exp(z)\phi(z) = zB\exp(cz) + V\phi(z) + O(z^{p+1}), \qquad (541b)$$

where $\exp(cz)$ denotes the vector in $\mathbb{C}^s$ for which component i is equal to $\exp(c_i z)$.

Proof. Assume that (541a) and (541b) are satisfied and that the components of $\phi(z)$ have Taylor series

$$\phi_i(z) = \sum_{j=0}^{p} \alpha_{ij} z^j + O(z^{p+1}).$$

Furthermore, suppose starting method i is chosen to give the output

$$\sum_{j=0}^{p} \alpha_{ij} h^j y^{(j)}(x_0) + O(h^{p+1}),$$

where y denotes the exact solution agreeing with a given initial value at x_0. Using this starting method, consider the value of

$$y(x_0 + hc_k) - h\sum_{i=1}^{s} a_{ki} y'(x_0 + hc_i) - \sum_{i=1}^{r} U_{ki} \sum_{j=0}^{p} \alpha_{ij} h^j y^{(j)}(x_0). \qquad (541c)$$

If this is $O(h^{p+1})$ then it will follow that $Y_k - y(x_0 + hc_k) = O(h^{p+1})$. Expand (541c) about x_0, and it is seen that the coefficient of $h^j y^{(j)}(x_0)$ is

$$\frac{1}{j!} c_k^j - \sum_{i=1}^{s} a_{ki} \frac{1}{(j-1)!} c_i^{j-1} - \sum_{i=1}^{r} U_{ki} \alpha_{ij}.$$

However, this is exactly the same as the coefficient of z^j in the Taylor expansion of the difference of the two sides of (541a). Given that the order

of the stages is p, and therefore that $hf(Y_i) = hy'(x_0 + hc_i) + O(h^{p+1})$, we can carry out a similar analysis of the condition for the kth output vector to equal

$$\sum_{j=0}^{p} \alpha_{kj} h^j y^{[j]}(x_0 + h) + O(h^{p+1}). \tag{541d}$$

Carry out a Taylor expansion about x_0 and we find that (541d) can be written as

$$\sum_{j=0}^{p}\sum_{i=j}^{p} \alpha_{kj} \frac{1}{(i-j)!} h^i y^{(i)}(x_0) + O(h^{p+1}). \tag{541e}$$

The coefficient of h^i in (541e) is identical to the coefficient of z^i in $\exp(z)\phi_k(z)$. Hence, combining this with the terms

$$\sum_{i=1}^{s} b_{ki} \frac{1}{(j-1)!} c_i^{j-1} + \sum_{i=1}^{r} V_{ki}\alpha_{ij},$$

we find (541b).

To prove necessity, use the definition of order given by (532e) and (532f) and evaluate the two sides of each of these equations for the sequence of trees $t_0 = \emptyset$, $t_1 = \tau$, $t_2 = [t_1]$, ..., $t_p = [t_{p-1}]$. Use the values of α_{ij} given by

$$\alpha_{ij} = \xi_i(t_j),$$

so that

$$(E\xi_i)(t_j) = \sum_{k=0}^{j} \frac{1}{k!} \xi_i(t_{j-k}),$$

which is the coefficient of z^j in $\exp(z)\sum_{k=0}^{p} \alpha_{ik} z^k$. We also note that

$$\eta_i(t_j) = \frac{1}{j!} c_i^j, \qquad (\eta_i D)(t_j) = \frac{1}{(j-1)!} c_i^{j-1},$$

which are, respectively, the z^j coefficients in $\exp(c_i z)$ and in $z\exp(c_i z)$. Write $\phi(z)$ as the vector-valued function with ith component equal to $\sum_{k=0}^{p} \alpha_{ik} z^k$, and we verify that coefficients of all powers of z up to z^p agree in the two sides of (541a) and (541b). $\square$

542   Runge–Kutta stability

For methods of types 1 and 2, a reasonable design criterion is that its stability region should be similar to that of a Runge–Kutta method. The reasons for this are that Runge–Kutta methods not only have convenient stability properties from the point of view of analysis but also that they have

stability properties that are usually superior to those of alternative methods. For example, A-stability is inconsistent with high order for linear multistep methods but is available for Runge–Kutta methods of any order.

The stability matrix for a general linear method has the form

$$M(z) = V + zB(I - zA)^{-1}U$$

and the characteristic polynomial is

$$\Phi(w, z) = \det(wI - M(z)). \tag{542a}$$

In general this is a complicated function, in which the coefficients of powers of w are rational functions of z. To obtain stability properties as close to those of a Runge–Kutta method as possible we will seek methods for which $\Phi(w, z)$ factorizes as in the following definition.

Definition 542A *A general linear method* (A, U, B, V) *has 'Runge–Kutta stability' if the characteristic polynomial given by* (542a) *has the form*

$$\Phi(w, z) = w^{r-1}(w - R(z)).$$

For a method with Runge–Kutta stability, the rational function $R(z)$ is known as the 'stability function' of the method.

We will usually abbreviate 'Runge–Kutta stability' by 'RK stability'. We present two examples of methods satisfying this condition with $p = q = r = s = 2$ and with $c = [\,0 \;\; 1\,]^\mathsf{T}$. The first is of type 1 and is assumed to have the form

$$\left[\begin{array}{cc} A & U \\ B & V \end{array} \right] = \left[\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ a_{21} & 0 & 0 & 1 \\ \hline b_{11} & b_{12} & 1 - V_{12} & V_{12} \\ b_{11} & b_{12} & 1 - V_{12} & V_{12} \end{array} \right].$$

The assumption that $U = I$ is not a serious restriction because, if U is non-singular, an equivalent method can be constructed with $U = I$ and B and V replaced by UB and UVU^{-1}, respectively. The form chosen for V makes it of rank 1 and preconsistent for the vector $c = [\,1 \;\; 1\,]^\mathsf{T}$.

By the stage order conditions, it is found that

$$\phi(z) = (I - zA) \exp(cz) = \left[\begin{array}{c} 1 \\ 1 + (1 - a_{21})z + \frac{1}{2}z^2 \end{array} \right].$$

To find B, we have

$$Bz \exp(cz) = (\exp(z)I - V)\phi(z) + O(z^3).$$

Write the coefficients of z and z^2 in separate columns and we deduce that

$$B\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 - V_{12} + a_{21}V_{12} & \frac{1}{2}(1 - V_{12}) \\ 2 - V_{12} - a_{21} + a_{21}V_{12} & 2 - a_{21} - \frac{1}{2}V_{12} \end{bmatrix},$$

so that

$$B = \begin{bmatrix} \frac{1}{2} - \frac{1}{2}V_{12} + a_{21}V_{12} & \frac{1}{2}(1 - V_{12}) \\ -\frac{1}{2}V_{12} + a_{21}V_{12} & 2 - a_{21} - \frac{1}{2}V_{12} \end{bmatrix}.$$

To achieve RK stability, impose the requirement that the stability function $V + zB(I - zA)^{-1}$ has zero determinant and it is found that $a_{21} = 2$ and $V_{12} = \frac{1}{2}$.

This gives the method

$$\begin{bmatrix} A & U \\ B & V \end{bmatrix} = \left[\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \\ \hline \frac{5}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{2} \\ \frac{3}{4} & -\frac{1}{4} & \frac{1}{2} & \frac{1}{2} \end{array}\right]. \tag{542b}$$

To derive a type 2 method with RK stability, carry out a similar calculation but with

$$A = \begin{bmatrix} \lambda & 0 \\ a_{21} & \lambda \end{bmatrix}.$$

In this case, the method is

$$\begin{bmatrix} A & U \\ B & V \end{bmatrix} = \left[\begin{array}{cc|cc} \lambda & 0 & 1 & 0 \\ \frac{2}{1+2\lambda} & \lambda & 0 & 1 \\ \frac{5-2\lambda+12\lambda^2+8\lambda^3}{4+8\lambda} & \frac{1}{4} - \lambda^2 & \frac{1}{2} + \lambda & \frac{1}{2} - \lambda \\ \frac{3-2\lambda+20\lambda^2+8\lambda^3}{4+8\lambda} & \frac{-1+10\lambda-12\lambda^2-8\lambda^3}{4+8\lambda} & \frac{1}{2} + \lambda & \frac{1}{2} - \lambda \end{array}\right],$$

or, with $\lambda = 1 - \frac{1}{2}\sqrt{2}$, for L-stability,

$$\begin{bmatrix} A & U \\ B & V \end{bmatrix} = \left[\begin{array}{cc|cc} 1 - \frac{\sqrt{2}}{2} & 0 & 1 & 0 \\ \frac{6+2\sqrt{2}}{7} & 1 - \frac{\sqrt{2}}{2} & 0 & 1 \\ \frac{73-34\sqrt{2}}{28} & \frac{4\sqrt{2}-5}{4} & \frac{3-\sqrt{2}}{2} & \frac{\sqrt{2}-1}{2} \\ \frac{87-48\sqrt{2}}{28} & \frac{34\sqrt{2}-45}{28} & \frac{3-\sqrt{2}}{2} & \frac{\sqrt{2}-1}{2} \end{array}\right]. \tag{542c}$$

Type 3 and type 4 methods do not exist with RK stability, and will not be explored in detail in this section. We do, however, give a single example of each. For the type 3 method we have

$$\begin{bmatrix} A & U \\ B & V \end{bmatrix} = \left[\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline -\frac{3}{8} & -\frac{3}{8} & -\frac{3}{4} & \frac{7}{4} \\ -\frac{7}{8} & \frac{9}{8} & -\frac{3}{4} & \frac{7}{4} \end{array}\right]. \tag{542d}$$

This method is designed for parallel computation in the sense that the two stages do not depend on each other, because $A = 0$, and hence they can be evaluated in parallel. Is there any advantage in the use of methods like this? Of course, the answer will depend on the specific coefficients in the method but, in the case of (542d), we might wish to compare it with the type 1 method given by (542b) whose error constant has magnitude $\frac{1}{6}$. In contrast, (542d) has error constant $\frac{19}{24}$ which is equivalent to $\frac{19}{96}$ when adjusted for the sequential cost of one f evaluation per step. Thus, in this case, the type 3 method is less efficient even under the assumption of perfect speed-up.

The type 4 method

$$
\begin{bmatrix} A & U \\ B & V \end{bmatrix} = \left[\begin{array}{cc|cc} \frac{3-\sqrt{3}}{2} & 0 & 1 & 0 \\ 0 & \frac{3-\sqrt{3}}{2} & 0 & 1 \\ \hline \frac{18-11\sqrt{3}}{4} & \frac{7\sqrt{3}-12}{4} & \frac{3-2\sqrt{3}}{2} & \frac{2\sqrt{3}-1}{2} \\ \frac{22-13\sqrt{3}}{4} & \frac{9\sqrt{3}-12}{4} & \frac{3-2\sqrt{3}}{2} & \frac{2\sqrt{3}-1}{2} \end{array} \right] \tag{542e}
$$

is found to be A-stable with the additional property that its stability matrix has zero spectral radius at infinity. Just as for the type 3 method we have introduced, while the advantages of this type of method are not clear, results found by Singh (1999) are encouraging.

For type 1 and 2 methods, increasing order presents great challenges in the solution of the order conditions combined with RK stability requirements. For an account of the techniques used to find particular methods of orders up to 8, see Butcher and Jackiewicz (1996, 1998).

543 Almost Runge–Kutta methods

The characteristic feature of explicit Runge–Kutta methods, that only minimal information computed in a step is passed on as input to the next step, is a great advantage of this type of method but it is also a perceived disadvantage. The advantage lies in excellent stability properties, while the disadvantage lies in the low stage order to which the second and later stages are restricted. Almost Runge–Kutta methods (ARK) are an attempt to retain the advantage but overcome some of the disadvantages.

Recall the method (505a). Evaluate its stability matrix and we find

$$
M(z) = V + zB(I - zA)^{-1}U
$$
$$
= \begin{bmatrix} 1 + \frac{5}{6}z + \frac{1}{3}z^2 + \frac{1}{48}z^3 & \frac{1}{6} + \frac{1}{6}z + \frac{7}{48}z^2 + \frac{1}{48}z^3 & \frac{1}{48}z^2 + \frac{1}{96}z^3 \\ z + \frac{5}{6}z^2 + \frac{1}{3}z^3 + \frac{1}{48}z^4 & \frac{1}{6}z + \frac{1}{6}z^2 + \frac{7}{48}z^3 + \frac{1}{48}z^4 & \frac{1}{48}z^3 + \frac{1}{96}z^4 \\ z + \frac{1}{2}z^2 + \frac{7}{12}z^3 + \frac{1}{24}z^4 & -1 + \frac{1}{2}z - \frac{1}{12}z^2 + \frac{5}{24}z^3 + \frac{1}{24}z^4 & \frac{1}{48}z^4 \end{bmatrix}.
$$

The eigenvalues of this matrix are

$$
\sigma(M(z)) = \left\{ 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4, 0, 0 \right\},
$$

Table 543(I)   Calculation of stages and stage derivatives for the method (505a)

α	$\alpha(\emptyset)$	$\alpha(\bullet)$	$\alpha(\mathbf{I})$	$\alpha(\mathbf{V})$	$\alpha(\mathbf{I\!I})$	$\alpha(\mathbf{V\!I})$	$\alpha(\mathbf{\Psi})$	$\alpha(\mathbf{Y})$	$\alpha(\mathbf{I\!I\!I})$
1	1	0	0	0	0	0	0	0	0
D	0	1	0	0	0	0	0	0	0
ξ_3	0	0	1	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
η_1	1	1	$\frac{1}{2}$	$\frac{\theta_3}{2}$	$\frac{\theta_4}{2}$	$\frac{\theta_5}{2}$	$\frac{\theta_6}{2}$	$\frac{\theta_7}{2}$	$\frac{\theta_8}{2}$
$\eta_1 D$	0	1	1	1	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{\theta_3}{2}$	$\frac{\theta_4}{2}$
η_2	1	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1+\theta_3}{16}$	$\frac{1+2\theta_4}{32}$	$\frac{1+\theta_5}{16}$	$\frac{1+2\theta_6}{32}$	$\frac{\theta_3+2\theta_7}{32}$	$\frac{\theta_4+2\theta_8}{32}$
$\eta_2 D$	0	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1+\theta_3}{16}$	$\frac{1+2\theta_4}{32}$
η_3	1	1	$\frac{1}{2}$	$\frac{1-\theta_3}{4}$	$\frac{1-2\theta_4}{8}$	$-\frac{\theta_5}{4}$	$-\frac{\theta_6}{4}$	$\frac{1-2\theta_7}{8}$	$\frac{1-4\theta_8}{16}$
$\eta_3 D$	0	1	1	1	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1-\theta_3}{4}$	$\frac{1-2\theta_4}{8}$
η_4	1	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{24}$
$\eta_4 D$	0	1	1	1	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$
$E\widehat{\xi_1}$	1	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{24}$
$E\widehat{\xi_2}$	0	1	1	1	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$
$E\widehat{\xi_3}$	0	0	1	1	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$
$\widehat{\xi_1}$	1	0	0	0	0	0	0	0	0
$\widehat{\xi_2}$	0	1	0	0	0	0	0	0	0
$\widehat{\xi_3}$	0	0	1	-1	$-\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$

so that it is RK stable. Other features of the method are that the minimal information passed between steps is enough to push the stage order up to 2, and that the third input and output vector need not be evaluated to great accuracy because of what will be called 'annihilation conditions'. These conditions ensure that errors like $O(h^3)$ in the input vector $y_3^{[n-1]}$ only affect the output results by $O(h^5)$.

Assume that the three input approximations are represented by $\xi_1 = 1$, $\xi_2 = D$ and ξ_3, where we assume only that

$$\xi_3(\emptyset) = \xi_3(\bullet) = 0 \qquad \text{and} \qquad \xi_3(\mathbf{I}) = 1.$$

Thus, $y_1^{[n-1]} = y(x_{n-1}), y_2^{[n-1]} = hy'(x_{n-1}), y_3^{[n-1]} = h^2 y''(x_{n-1}) + O(h^3)$. The output approximations are computed by first evaluating the representations of the stage values and stage derivatives. Since we are only working to order 5 accuracy in the output results, it will be sufficient to evaluate the stages only up to order 4. Denote the representations of the four stage values by η_i, $i = 1, 2, 3, 4$. Also, denote the values of $\xi_3(t)$ for trees of orders 3 and 4 by θ_i, $i = 3, 4, \ldots, 8$. Details of the calculation of stage values are shown in Table 543(I).

Table 543(II)   Output and input values for (505a) evaluated at fifth order trees

α	$\alpha(\text{tree})$	$\alpha(\text{tree})$	$\alpha(\text{tree})$	$\alpha(\text{tree})$	$\alpha(\text{tree})$	$\alpha(\text{tree})$	$\alpha(\text{tree})$	$\alpha(\text{tree})$	$\alpha(\text{tree})$
ξ_3	θ_9	θ_{10}	θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}
$\widehat{\xi_1}$	$\frac{1}{120}$	$\frac{1}{240}$	$-\frac{1+5\theta_3}{240}$	$-\frac{1+10\theta_4}{480}$	$\frac{1}{480}$	$-\frac{1}{120}$	$-\frac{1}{240}$	$\frac{1+5\theta_3}{240}$	$\frac{1+10\theta_4}{480}$
$\widehat{\xi_2}$	0	0	0	0	0	0	0	0	0
$\widehat{\xi_3}$	-1	$-\frac{1}{2}$	$-\frac{1}{3}$	$-\frac{1}{6}$	$-\frac{1}{4}$	$-\frac{1}{2}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{8}$

The output results are intended to represent approximations to $E\xi_1$, $E\xi_2$ and $E\xi_3$. Write the representation of $y_i^{[n]}$ by $E\widehat{\xi_i}$, for $i = 1, 2, 3$. We calculate $\widehat{\xi_i}$ up to order 5 trees so that we not only verify fourth order behaviour, but also obtain information on the principal terms in the local truncation error. As a first step in this analysis, we note that, to order 4, $E\widehat{\xi_1} = E$ and hence $\widehat{\xi_1} = 1$. Similarly $\widehat{\xi_2} = D$ to fourth order. Up to fourth order, we have calculated the value of $E\widehat{\xi_3} = -\frac{1}{3}\eta_1 D - \frac{2}{3}\eta_3 D + 2\eta_4 D - \xi_2$ and $\widehat{\xi_3}$ is also given in Table 543(I).

If the calculations are repeated using the specific values $[\theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8]$ $= [-1, -\frac{1}{2}, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{4}]$, then we have $\widehat{\xi_i} = \xi_i + H_4$ so that, relative to a starting method defined by ξ_i, $i = 1, 2, 3$, the method has order 4. However, a starting value defined for arbitrary values of $\theta_3, \theta_4, \ldots, \theta_8$ produces the specific choice given by the components of $\widehat{\xi_3}$ after a single step. To investigate this method more precisely, the values of $\widehat{\xi_1}$, $\widehat{\xi_2}$ and $\widehat{\xi_3}$ have been calculated also for fifth order trees and these are shown in Table 543(II).

A reading of this table suggests that the method not only exhibits fourth order behaviour but also has reliable behaviour in its principal error terms. This is in spite of the fact that the starting method provides incorrect contributions of third and higher order elementary differentials, because these inaccuracies have no long term effect. The components of the error terms in the first output component depend on θ_3 and θ_4 after a single step, but this effect disappears in later steps.

In Subsection 544 we consider order 3 ARK methods, and we then return in Subsection 545 to a more detailed study of order 4 methods. However, we first discuss some questions which apply to both orders.

Because we will require methods in these families to have stage order 2, the matrix U will need to be of the form

$$U = [\,\mathbf{1} \quad c - A\mathbf{1} \quad \tfrac{1}{2}c^2 - Ac\,] \tag{543a}$$

and we will assume this throughout. We also note that the stability matrix $M(z) = V + zB(I - zA)^{-1}U$ is always singular because $ze_1^\mathsf{T} - e_2^\mathsf{T}$ is an eigenvalue of this matrix. We see this by observing that $ze_p^\mathsf{T}(I - zA) = (-ze_1^\mathsf{T} + e_2^\mathsf{T})B$ and $(ze_1^\mathsf{T} - e_2^\mathsf{T})V = ze_p^\mathsf{T}U$.

544   Third order, three-stage ARK methods

Since $r = s = 3$, we will write the coefficient matrices as follows:

$$
\left[\begin{array}{cc} A & U \\ B & V \end{array}\right] =
\left[\begin{array}{ccc|ccc}
0 & 0 & 0 & 1 & u_{12} & u_{13} \\
a_{21} & 0 & 0 & 1 & u_{22} & u_{23} \\
b_1 & b_2 & 0 & 1 & b_0 & 0 \\
b_1 & b_2 & 0 & 1 & b_0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
\beta_1 & \beta_2 & \beta_3 & 0 & \beta_0 & 0
\end{array}\right].
$$

Denote the abscissa vector by $c = [c_1, c_2, 1]^\mathsf{T}$ and also write $b^\mathsf{T} = [b_1, b_2, 0]$ and $\beta^\mathsf{T} = [\beta_1, \beta_2, \beta_3]$.

Because we will require the method to have stage order 2, the matrix U will need to be of the form given by (543a). For the method to have order 3, and at the same time be RK stable, it is necessary that the trace of M is equal to the Taylor expansion of the non-zero eigenvalue. Thus,

$$
\mathrm{tr}(M) = \mathrm{tr}(V) + z\,\mathrm{tr}(BU) + z^2\,\mathrm{tr}(BAU) + z^3\,\mathrm{tr}(BA^2U) = 1 + z + \tfrac{1}{2}z^2 + \tfrac{1}{6}z^3,
$$

and it follows that

$$
\mathrm{tr}(BU) = 1, \quad \mathrm{tr}(BAU) = \tfrac{1}{2}, \quad \mathrm{tr}(BA^2U) = \tfrac{1}{6}, \tag{544a}
$$

where we note that $\mathrm{tr}(V) = 1$, because of the form of V.

To obtain order 3 accuracy for the first output value, it is only necessary that

$$
b_0 + b_1 + b_2 = 1, \quad b_1c_1 + b_2c_2 = \tfrac{1}{2}, \quad b_1c_1^2 + b_2c_2^2 = \tfrac{1}{3}, \tag{544b}
$$

and to obtain an order 2 approximation to the scaled second derivative for the third output value, we require that

$$
\beta_0 + \beta^\mathsf{T}\mathbf{1} = 0, \tag{544c}
$$
$$
\beta^\mathsf{T}c = 1. \tag{544d}
$$

Note that $b^\mathsf{T}Ac = \tfrac{1}{6}$ does not arise as an order condition, because the method has stage order 2. Expand the equations given in (544a), making use of (544b), and we find

$$
\beta^\mathsf{T}(\tfrac{1}{2}c^2 - Ac) = 0, \tag{544e}
$$
$$
\beta^\mathsf{T}A(\tfrac{1}{2}c^2 - Ac) = 0, \tag{544f}
$$
$$
b^\mathsf{T}Ac + \beta^\mathsf{T}A^2(\tfrac{1}{2}c^2 - Ac) = \tfrac{1}{6}. \tag{544g}
$$

Eliminating terms known to be zero, we see that (544g) simplifies to

$$b_2 a_{21} c_1 = \frac{1}{6(1 + \frac{1}{2}\beta_3 c_1)}. \tag{544h}$$

Consider the vector $v^{\mathsf{T}} = \beta_3 e_3^{\mathsf{T}} - \beta^{\mathsf{T}}(I + \beta_3 A)$ and note that $v^{\mathsf{T}} x_1 = v^{\mathsf{T}} x_2 = c^{\mathsf{T}} x_3 = 0$, where $x_1 = e_3$, $x_2 = \frac{1}{2}c^2 - Ac$ and $x_3 = A(\frac{1}{2}c^2 - Ac)$. It is not possible that x_1, x_2, x_3 are linearly dependent because this would imply $\beta_1 = \beta_2 = 0$, which is inconsistent with $\beta^{\mathsf{T}}\mathbf{1} = 0$ and $\beta^{\mathsf{T}}c = 1$. Hence, $v^{\mathsf{T}} = 0$ and we arrange this in the form

$$\beta^{\mathsf{T}} = \beta_3 e_3^{\mathsf{T}}(I + \beta_3 A)^{-1} = \beta_3 e_3^{\mathsf{T}} - \beta_3^2 b^{\mathsf{T}} + \beta_3^3 b^{\mathsf{T}}A. \tag{544i}$$

Multiply (544i) by c and use (544d), (544h) to obtain a relationship between β_3 and c_1:

$$c_1 = \frac{-2(1 - \beta_3 + \frac{1}{2}\beta_3^2 - \frac{1}{6}\beta_3^3)}{\beta_3(1 - \beta_3 + \frac{1}{2}\beta_3^2)}. \tag{544j}$$

The ingredients for constructing an ARK method with $p = r = s = 3$ are now all available and they are put together as follows:

1. Choose the value of β_3.
2. Evaluate c_1 from (544j).
3. Choose the value of c_2.
4. Evaluate b_0, b_1, b_2 to satisfy (544b).
5. Evaluate a_{21} to satisfy (544h).
6. Evaluate the remaining elements of β^{T} from (544i).
7. Evaluate the elements of U and V.

The following example method is found from $\beta_3 = 2$, leading to $c_1 = \frac{1}{3}$, together with the choice $c_2 = \frac{2}{3}$:

$$\left[\begin{array}{ccc|ccc}
0 & 0 & 0 & 1 & \frac{1}{3} & \frac{1}{18} \\
\frac{1}{2} & 0 & 0 & 1 & \frac{1}{6} & \frac{1}{18} \\
0 & \frac{3}{4} & 0 & 1 & \frac{1}{4} & 0 \\
\hline
0 & \frac{3}{4} & 0 & 1 & \frac{1}{4} & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
3 & -3 & 2 & 0 & -2 & 0
\end{array}\right].$$

Further examples of third order ARK methods, together with details on possible interpolation techniques, can be found in Rattenbury (2005).

545   Fourth order, four-stage ARK methods

We write specific coefficients of the method as shown in the tableau

$$
\left[
\begin{array}{cccc|ccc}
0 & 0 & 0 & 0 & 1 & u_{12} & u_{13} \\
a_{21} & 0 & 0 & 0 & 1 & u_{22} & u_{23} \\
a_{31} & a_{32} & 0 & 0 & 1 & u_{32} & u_{33} \\
b_1 & b_2 & b_3 & 0 & 1 & b_0 & 0 \\
\hline
b_1 & b_2 & b_3 & 0 & 1 & b_0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\beta_1 & \beta_2 & \beta_3 & \beta_4 & 0 & \beta_0 & 0
\end{array}
\right].
$$

As usual, c will denote the abscissa vector. We also write $b^{\mathsf{T}} = \begin{bmatrix} b_1 & b_2 & b_3 & 0 \end{bmatrix}$ and $\beta^{\mathsf{T}} = \begin{bmatrix} \beta_1, & \beta_2 & \beta_3 & \beta_4 \end{bmatrix}$.

As in the example method discussed in Subsection 543, the input approximations will be of the form $y(x_{n-1}) + O(h^5)$, $hy'(x_{n-1}) + O(h^5)$ and $h^2 y''(x_{n-1}) + O(h^3)$, The crucial assumptions we will make are that each of the stages is computed with order at least 2, and that the three output values are not affected by order 3 perturbations in the third input approximation. For stage order 2 it is necessary and sufficient that the matrix U should have the form

$$
U = \begin{bmatrix} \mathbf{1} & c - A\mathbf{1} & \tfrac{1}{2}c^2 - Ac \end{bmatrix}.
$$

Since $u_{42} = b_0$, this will mean that $b^{\mathsf{T}}\mathbf{1} + b_0 = 1$. The conditions for order 4 on the first output component yield the equations

$$b^{\mathsf{T}}c = \tfrac{1}{2}, \tag{545a}$$

$$b^{\mathsf{T}}c^2 = \tfrac{1}{3}, \tag{545b}$$

$$b^{\mathsf{T}}c^3 = \tfrac{1}{4}, \tag{545c}$$

$$b^{\mathsf{T}}Ac^2 = \tfrac{1}{12}, \tag{545d}$$

$$b^{\mathsf{T}}(\tfrac{1}{2}c^2 - Ac) = 0, \tag{545e}$$

where (545e) is included to ensure that an $O(h^3)$ error in the third input vector does not detract from the order 4 behaviour. Combining (545b) and (545e), we find

$$b^{\mathsf{T}}Ac = \tfrac{1}{6}. \tag{545f}$$

Either (545e) or the equivalent condition (545f), together with the related condition on β^{T} given in (545i) below, will be referred to as 'annihilation conditions'. The vector β^{T}, together with β_0, defines the third output approximation, which is required to give the result $h^2 y''(x_n) + O(h^3)$. Hence,

$$\beta^{\mathsf{T}}\mathbf{1} + \beta_0 = 0, \tag{545g}$$

$$\beta^{\mathsf{T}}c = 1. \tag{545h}$$

We now turn to the conditions for RK stability. If the stability matrix

$$M(z) = V + zBU + z^2 BAU + z^3 BA^2 U + z^4 BA^3 U$$

is to have only a single non-zero eigenvalue, this eigenvalue must be the trace of $M(z)$ and for order 4 must equal $1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4$. We therefore impose the conditions that the traces of BU, BAU, BA^2U and BA^3U have values 1, $\frac{1}{2}$, $\frac{1}{6}$, $\frac{1}{24}$, respectively. These can be written in the form

$$\beta^{\mathsf{T}}(\tfrac{1}{2}c^2 - Ac) = 0. \tag{545i}$$

$$\beta^{\mathsf{T}}A(\tfrac{1}{2}c^2 - Ac) = 0, \tag{545j}$$

$$\beta^{\mathsf{T}}A^2(\tfrac{1}{2}c^2 - Ac) = 0, \tag{545k}$$

$$b^{\mathsf{T}}A^2 c + \beta^{\mathsf{T}}A^3(\tfrac{1}{2}c^2 - Ac) = \frac{1}{24}. \tag{545l}$$

Because $A^4 = 0$, (545l) simplifies to

$$b^{\mathsf{T}}A^2 c = \frac{1}{24(1 + \frac{1}{2}\beta_4 c_1)}. \tag{545m}$$

We now show that β^{T} satisfies the equation

$$\beta_4 e_4^{\mathsf{T}} = \beta^{\mathsf{T}}(I + \beta_4 A). \tag{545n}$$

This follows by observing that $\beta_4 e_4^{\mathsf{T}} - \beta^{\mathsf{T}}(I + \beta_4 A)$ multiplied respectively by e_4, $\frac{1}{2}c^2 - Ac$, $A(\frac{1}{2}c^2 - Ac)$ and $A^2(\frac{1}{2}c^2 - Ac)$ are each zero if and only if each of (545j), (545k) and (545l) holds.

Multiply each side of (545n) by $(I + \beta_4 A)^{-1}c$ and use (545h) to show that

$$1 = \beta_4 - \frac{1}{2}\beta_4^2 + \frac{1}{6}\beta_4^3 - \frac{\beta_4^4}{24(1 + \frac{1}{2}\beta_4 c_1)},$$

from which it follows that

$$c_1 = \frac{-2(1 - \beta_4 + \frac{1}{2}\beta_4^2 - \frac{1}{6}\beta_4^3 + \frac{1}{24}\beta_4^4)}{\beta_4(1 - \beta_4 + \frac{1}{2}\beta_4^2 - \frac{1}{6}\beta_4^3)}. \tag{545o}$$

To construct a four-stage fourth order ARK method in detail, carry out the following steps:

1. Choose the value of β_4.
2. Evaluate c_1 from (545o).
3. Choose values of c_2 and c_3.
4. Evaluate b_0, b_1, b_2, b_3 to satisfy (545a), (545b), (545c), (545g).
5. Evaluate a_{21}, a_{31}, a_{32} to satisfy (545f), (545d), (545m).
6. Evaluate the remaining elements of β^{T} from (545n).
7. Evaluate the elements of U and V.

In contrast to the method given in (505a), the following method has the same $c = [\,1 \quad \frac{1}{2} \quad \frac{1}{2} \quad 1\,]^{\mathsf{T}}$ but different b^{T}:

$$
\left[
\begin{array}{cccc|cccc}
0 & 0 & 0 & 0 & 1 & 1 & \frac{1}{2} \\
\frac{1}{16} & 0 & 0 & 0 & 1 & \frac{7}{16} & \frac{1}{16} \\
-\frac{1}{16} & 1 & 0 & 0 & 1 & -\frac{7}{16} & -\frac{5}{16} \\
\frac{1}{6} & \frac{1}{3} & \frac{1}{3} & 0 & 1 & \frac{1}{6} & 0 \\
\hline
\frac{1}{6} & \frac{1}{3} & \frac{1}{3} & 0 & 1 & \frac{1}{6} & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-1 & \frac{4}{3} & -\frac{4}{3} & 2 & 0 & -1 & 0
\end{array}
\right].
$$

A further example with $c = [\,\frac{11}{24} \quad \frac{13}{24} \quad 1 \quad 1\,]^{\mathsf{T}}$ is given by the matrix

$$
\left[
\begin{array}{cccc|cccc}
0 & 0 & 0 & 0 & 1 & \frac{11}{24} & \frac{121}{1152} \\
\frac{104}{297} & 0 & 0 & 0 & 1 & \frac{455}{2376} & -\frac{143}{10368} \\
\frac{1820}{4653} & \frac{44}{47} & 0 & 0 & 1 & -\frac{1523}{4653} & -\frac{473}{2538} \\
\frac{48}{143} & \frac{48}{143} & \frac{47}{286} & 0 & 1 & \frac{47}{286} & 0 \\
\hline
\frac{48}{143} & \frac{48}{143} & \frac{47}{286} & 0 & 1 & \frac{47}{286} & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-\frac{354}{143} & \frac{162}{143} & -\frac{423}{286} & 3 & 0 & -\frac{51}{286} & 0
\end{array}
\right].
$$

These methods were introduced in Butcher (1997, 1998). Although it does not seem possible to find similar methods with $s = p$ stages where $p > 4$, we will see in the next subsection that something very similar can be achieved.

546 A fifth order, five-stage method

We will consider a special method constructed using a more general formulation of fourth order methods in which there is an additional fifth stage. There is enough freedom to ensure that the error constants are zero. This does not mean that, regarded as an ARK method, a method constructed this way has fifth order, because the trivial rescaling normally used to achieve variable stepsize does not preserve the correct behaviour up to h^5 terms. However, a slight modification to the way the method is implemented restores fifth order performance.

The derivation and the results of preliminary experiments are presented in Butcher and Moir (2003). A fuller description is given by Rattenbury (2005).

For constant stepsize, the tableau for the method is

$$
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{4} & \frac{1}{32} \\
\frac{2}{5} & 0 & 0 & 0 & 0 & 1 & \frac{1}{10} & \frac{1}{40} \\
\frac{27}{160} & \frac{75}{128} & 0 & 0 & 0 & 1 & -\frac{3}{640} & -\frac{69}{1280} \\
\frac{69}{35} & -\frac{51}{28} & \frac{8}{7} & 0 & 0 & 1 & -\frac{41}{140} & \frac{17}{280} \\
\frac{16}{45} & \frac{2}{15} & \frac{16}{45} & \frac{7}{90} & 0 & 1 & \frac{7}{90} & 0 \\
\hline
\frac{16}{45} & \frac{2}{15} & \frac{16}{45} & \frac{7}{90} & 0 & 1 & \frac{7}{90} & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-\frac{1352}{225} & \frac{34}{15} & -\frac{256}{75} & -\frac{196}{225} & \frac{24}{5} & 0 & \frac{242}{75} & 0
\end{bmatrix}. \qquad (546a)
$$

When the stepsize is changed at the end of step n from h to rh, an additional term has to be added to the scaled result. In this context $D(r)$ will denote the scaling matrix $D(r) = \mathrm{diag}(1, r, r^2)$ so that, for any of the lower order ARK methods, change of stepsize would be accompanied by the rescaling $y^{[n]} \to (D(r) \otimes I_N)y^{[n]}$. For (546a), this is corrected to

$$
y^{[n]} \to (D(r) \otimes I_N)y^{[n]} + r^2(1 - r)\delta,
$$

where

$$
\delta = \frac{496}{45}hF1 + \frac{224}{25}hF2 - \frac{4928}{225}hF3 - \frac{6482}{225}hF4 + 38hF5 - \frac{1636}{225}y_2^{[n-1]}.
$$

547   ARK methods for stiff problems

In Butcher and Rattenbury (2005), the ARK type of method was extended to the solution of stiff problems. Methods were presented with orders 3 and 4, subject to a number of criteria, and these were supported by preliminary numerical comparisons with standard methods. Because stiff ARK methods are still at an early stage of development, we will not attempt to give a full description, but will present a single third order method,

$$
\begin{bmatrix} A & U \\ B & V \end{bmatrix} =
\begin{bmatrix}
\frac{1}{3} & 0 & 0 & 1 & \frac{2}{3} & \frac{1}{6} \\
-\frac{1}{16} & \frac{1}{3} & 0 & 1 & \frac{11}{48} & \frac{1}{48} \\
-\frac{1}{6} & \frac{2}{3} & \frac{1}{3} & 1 & \frac{1}{6} & 0 \\
\hline
-\frac{1}{6} & \frac{2}{3} & \frac{1}{3} & 1 & \frac{1}{6} & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
\frac{1}{3} & -\frac{8}{3} & 2 & 0 & \frac{1}{3} & 0
\end{bmatrix}, \qquad (547a)
$$

together with a convenient starting method. This is not the most successful of the methods known so far, but it has simple coefficients and will serve for illustrative purposes.

To start the method, and simultaneously progress the method a single step forward, the starting method should be a three-output Runge–Kutta method. For input the value of $y(x_0)$, the method given in the following tableau gives suitable approximations to $y(x_1)$, $hy'(x_1)$ and $h^2 y''(x_1)$:

$$
\begin{bmatrix} A & U \\ B & V \end{bmatrix} =
\left[
\begin{array}{cccc|c}
\frac{1}{3} & 0 & 0 & 0 & 1 \\
\frac{1}{3} & \frac{1}{3} & 0 & 0 & 1 \\
-\frac{5}{3} & \frac{4}{3} & \frac{1}{3} & 0 & 1 \\
1 & -\frac{1}{4} & -\frac{1}{12} & \frac{1}{3} & 1 \\
\hline
1 & -\frac{1}{4} & -\frac{1}{12} & \frac{1}{3} & 1 \\
0 & 0 & 0 & 1 & 0 \\
-2 & -1 & \frac{2}{3} & \frac{7}{3} & 0
\end{array}
\right].
$$

For the method given by (547a), the stability function is

$$
R(z) = \frac{1 - \frac{1}{6}z^2 - \frac{1}{27}z^3}{(1 - \frac{1}{3}z)^3},
$$

and it can be verified to satisfy the conditions of A-stability.

Further details concerning stiff ARK methods, and of ARK methods in general, can be found in Rattenbury (2005).

Exercises 54

54.1 Find the stability matrix of the method given by (542b) and verify that it is RK-stable.

54.2 Does a transformation matrix exist such that the input to the transformed method approximates the two quantities $y(x_{n-1} + \theta h)$ and $hy'(x_{n-1} + \theta h)$, in each to within $O(h^3)$, for some θ?

54.3 Show that the method given by (542c) is L-stable.

54.4 Is the same true for the method in which $\sqrt{2}$ is replaced by $-\sqrt{2}$ throughout?

54.5 Which of the two methods (542c) and the method where the sign of $\sqrt{2}$ is reversed, is likely to be more accurate?

54.6 Find a third order ARK method with $\beta_3 = 2$ and $c_2 = 1$.

55   Methods with Inherent Runge–Kutta Stability

550   Doubly companion matrices

As a preliminary to a discussion of inherent RK stability, we recall the properties of the matrices introduced by Butcher and Chartier (1997). The original application was in the analysis of singly implicit methods with a specific effective order, but they also have a central role in the construction of the methods to be considered in Subsection 551. A review of doubly companion matrices is given in Butcher and Wright (2006).

Let $\alpha(z) = 1 + \alpha_1 z + \cdots + \alpha_n z^n$ and $\beta(z) = 1 + \beta_1 z + \cdots + \beta_n z^n$ denote given polynomials, and consider the $n \times n$ matrix

$$X = \begin{bmatrix} -\alpha_1 & -\alpha_2 & -\alpha_3 & \cdots & -\alpha_{n-1} & -\alpha_n - \beta_n \\ 1 & 0 & 0 & \cdots & 0 & -\beta_{n-1} \\ 0 & 1 & 0 & \cdots & 0 & -\beta_{n-2} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -\beta_1 \end{bmatrix}. \tag{550a}$$

Theorem 550A *The coefficients in the characteristic polynomial of X, $\det(wI - X) = w^n + \gamma_1 w^{n-1} + \gamma_2 w^{n-2} + \cdots + \gamma_n$, are given by*

$$1 + \gamma_1 z + \gamma_2 z^2 + \cdots + \gamma_n z^n = \det(I - zX) = \alpha(z)\beta(z) + O(z^{n+1}).$$

Proof. We assume that the eigenvalues of X are distinct and non-zero. There is no loss of generality in this assumption because, for given values of the α coefficients, the coefficients in the characteristic polynomial are continuous functions of the β coefficients; furthermore, choices of the β coefficients which lead to distinct non-zero eigenvalues form a dense set.

Let λ denote an eigenvalue of X, and let

$$v_k = \lambda^k + \beta_1 \lambda^{k-1} + \beta_2 \lambda^{k-2} + \cdots + \beta_k, \qquad k = 0, 1, 2, \ldots, n.$$

By comparing components numbered n, $n-1$, $\ldots$, 2 of Xv and λv, where

$$V = \begin{bmatrix} v_{n-1} & v_{n-2} & \cdots & 1 \end{bmatrix}^{\mathsf{T}}, \tag{550b}$$

we see that v is the eigenvector corresponding to λ. Now compare the first components of λv and Xv and it is found that

$$\lambda v_n + \alpha_1 v_{n-1} + \cdots + \alpha_n = 0$$

and contains all the terms with non-negative exponents in the product

$$v_n(1 + \alpha_1 \lambda^{-1} + \cdots + \alpha_n \lambda^{-n}).$$

Replace λ by z^{-1} and the result follows.   □

Write $\phi(z)$ for the vector (550b) with λ replaced by z. We now note that

$$z\phi(z) - X\phi(z) = \prod_{i=1}^{n}(z - \lambda_i)e_1, \qquad (550c)$$

because the expression vanishes identically except for the first component which is a monic polynomial of degree n which vanishes when z is an eigenvalue.

We are especially interested in choices of α and β such that X has a single n-fold eigenvalue, so that

$$\alpha(z)\beta(z) = (1 - \lambda z)^n + O(z^{n+1}) \qquad (550d)$$

and so that the right-hand side of (550c) becomes $(z - \lambda)^n e_1$. In this case it is possible to write down the similarity that transforms X to Jordan canonical form.

Theorem 550B *Let the doubly companion matrix X be chosen so that (550d) holds. Also let $\phi(z)$ denote the vector given by (550b) with λ replaced by z, and let S the matrix given by*

$$\Psi = \left[\frac{1}{(n-1)!}\phi^{(n-1)}(\lambda) \quad \frac{1}{(n-2)!}\phi^{(n-2)}(\lambda) \quad \cdots \quad \frac{1}{1!}\phi'(\lambda) \quad \phi(\lambda) \right].$$

Then

$$\Psi^{-1} X \Psi = \begin{bmatrix} \lambda & 0 & 0 & \cdots & 0 & 0 \\ 1 & \lambda & 0 & \cdots & 0 & 0 \\ 0 & 1 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \lambda \end{bmatrix}.$$

Proof. From the special case of (550c), we have

$$X\phi(z) = z\phi(z) - (z - \lambda)^n e_1. \qquad (550e)$$

Differentiate k times, divide by $k!$ and set $z = \lambda$, for $k = 1, 2, \ldots, n-1$. The result is

$$X \frac{1}{k!}\phi^{(k)}(\lambda) = \lambda I \frac{1}{k!}\phi^{(k)}(\lambda) + \frac{1}{(k-1)!}\phi^{(k-1)}(\lambda), \qquad k = 1, 2, \ldots, n-1.$$

Hence the vectors $\phi(\lambda), \frac{1}{1!}\phi'(\lambda), \frac{1}{2!}\phi''(\lambda), \ldots, \frac{1}{(n-1)!}\phi^{(n-1)}(\lambda)$ form a sequence of eigenvector and generalized eigenvectors, and the result follows. $\square$

The inverse of Ψ is easy to evaluate by interchanging the roles of rows and columns of X. We present the following result without further proof.

Corollary 550C *If*

$$\chi(\lambda) = [\,1 \quad \lambda + \alpha_1 \quad \lambda^2 + \alpha_1\lambda + \alpha_2 \quad \cdots \quad \lambda^{n-1} + \alpha_1\lambda^{n-2} + \cdots + \alpha_{n-1}\,],$$

then

$$\Psi^{-1} = [\,\chi(\lambda) \quad \tfrac{1}{1!}\chi'(\lambda) \quad \cdots \quad \tfrac{1}{(n-2)!}\chi^{(n-2)}(\lambda) \quad \tfrac{1}{(n-1)!}\chi^{(n-1)}(\lambda)\,]^{\mathsf{T}}.$$

551   *Inherent Runge–Kutta stability*

In this subsection we discuss a special type of general linear method based on several assumptions on the form of the method. The original formulation for stiff methods was given in Butcher (2001) and for non-stiff methods in Wright (2002). In Butcher and Wright (2003) it was shown how these ansätze are interrelated and this led to the current formulation in Butcher and Wright (2003a).

Besides making use of doubly companion matrices, we also use the special $r \times r$ matrix J and its transpose K, where

$$J = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

For the special type of inherently RK stable general linear method we consider, A has the diagonally implicit form

$$A = \begin{bmatrix} \lambda & 0 & 0 & \cdots & 0 \\ a_{21} & \lambda & 0 & \cdots & 0 \\ a_{31} & a_{32} & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ a_{s1} & a_{s2} & a_{s3} & \cdots & \lambda \end{bmatrix},$$

with $\lambda \geq 0$, and V has the form

$$V = \begin{bmatrix} 1 & v^{\mathsf{T}} \\ 0 & \dot{V} \end{bmatrix}, \tag{551a}$$

where $\rho(\dot{V}) = 0$, with ρ denoting spectral radius. We assume that $p = q$ and that $s = r = p + 1$. In some special cases, the last columns of U and V will vanish, thus making it possible for r to be reduced to $r = p$.

Definition 551A *A general linear method* (A, U, B, V) *is 'inherently Runge–Kutta stable' if* V *is of the form* (551a) *and the two matrices*

$$BA - XB \quad \text{and} \quad BU - XV + VX$$

are zero except for their first rows, where X *is some matrix.*

The significance of this definition is expressed in the following.

Theorem 551B *Let* (A, U, B, V) *denote an inherently RK stable general linear method. Then the stability matrix*

$$M(z) = V + zB(I - zA)^{-1}U$$

has only a single non-zero eigenvalue.

Proof. Calculate the matrix

$$(I - zX)M(z)(I - zX)^{-1},$$

which has the same eigenvalues as $M(z)$. We use the notation $\equiv$ to denote equality of two matrices, except for the first rows. Because $BA \equiv XB$ and $BU \equiv XV - VX$, it follows that

$$(I - zX)B \equiv B(I - zA),$$
$$(I - zX)V \equiv V(I - zX) - zBU,$$

so that

$$(I - zX)M(z) \equiv V(I - zX).$$

Hence $(I - zX)M(z)(I - zX)^{-1}$ is identical to V, except for the first row. Thus the eigenvalues of this matrix are its $(1, 1)$ element together with the p zero eigenvalues of $\dot{V}$. $\qquad\square$

Since we are adopting, as standard $r = p + 1$ and a stage order $q = p$, it is possible to insist that the vector-valued function of z, representing the input approximations, comprises a full basis for polynomials of degree p. Thus, we will introduce the function Z given by

$$Z = \begin{bmatrix} 1 \\ z \\ z^2 \\ \vdots \\ z^p \end{bmatrix}, \tag{551b}$$

which represents the input vector

$$y^{[n-1]} = \begin{bmatrix} y(x_{n-1}) \\ hy'(x_{n-1}) \\ h^2 y''(x_{n-1}) \\ \vdots \\ h^p y^{(p)}(x_{n-1}) \end{bmatrix}. \tag{551c}$$

This is identical, except for a simple rescaling by factorials, to the Nordsieck vector representation of input and output approximations, and it will be convenient to adopt this as standard.

Assuming that this standard choice is adopted, the order conditions are

$$\exp(cz) = zA \exp(cz) + UZ + O(z^{p+1}), \tag{551d}$$

$$\exp(z)Z = zB \exp(cz) + VZ + O(z^{p+1}). \tag{551e}$$

This result, and generalizations of it, make it possible to derive stiff methods of quite high orders. Furthermore, Wright (2003) has shown how it is possible to derive explicit methods suitable for non-stiff problems which satisfy the same requirements. Following some more details of the derivation of these methods, some example methods will be given.

552   *Conditions for zero spectral radius*

We will need to choose the parameters of IRKS methods so that the $p \times p$ matrix $\dot{V}$ has zero spectral radius. In Butcher (2001) it was convenient to force $\dot{V}$ to be strictly lower triangular, whereas in the formulation in Wright (2002) it was more appropriate to require $\dot{V}$ to be strictly upper triangular. To get away from these arbitrary choices, and at the same time to allow a wider range of possible methods, neither of these assumptions will be made and we explore more general options. To make the discussion non-specific to the application to IRKS methods, we assume we are dealing with $n \times n$ matrices related by a linear equation of the form

$$y = axb - c, \tag{552a}$$

and the aim will be to find lower triangular x such that y is strictly upper triangular. The constant matrices a, b and c will be assumed to be non-singular and LU factorizable. In this discussion only, define functions λ, μ and δ so that for a given matrix a,

  $\lambda(a)$ is unit lower triangular such that $\lambda(a)^{-1}a$ is upper triangular,

  $\mu(a)$ is the upper triangular matrix such that $a = \lambda(a)\mu(a)$,

  $\delta(a)$ is the lower triangular part of a.

Using these functions we can find the solution of (552a), when this solution exists. We have in turn

$$\delta(axb) = \delta(c),$$
$$\delta\big(\mu(a^{-1})^{-1}\lambda(a^{-1})^{-1}x\lambda(b)\mu(b)\big) = \delta(c),$$
$$\delta\big(\lambda(a^{-1})^{-1}x\lambda(b)\big) = \delta\big(\mu(a^{-1})\delta(c)\mu(b)^{-1}\big),$$

implying that

$$x = \delta\big(\lambda(a^{-1})\delta\big(\mu(a^{-1})\delta(c)\mu(b)^{-1}\big)\lambda(b)^{-1}\big). \tag{552b}$$

Thus, (552b) is the required solution of (552a).

This result can be generalized by including linear constraints in the formulation. Let d and e denote vectors in $\mathbb{R}^n$ and consider the problem

$$\delta(axb - c) = 0, \qquad xd = e.$$

Assume that d is scaled so that its first component is 1. The matrices a, b and c are now, respectively $n \times (n-1)$, $(n-1) \times n$ and $(n-1) \times (n-1)$. Partition these, and the vectors d and e, as

$$a = \begin{bmatrix} a_1 & a_2 \end{bmatrix}, \quad b = \begin{bmatrix} b_1^{\mathsf{T}} \\ b_2 \end{bmatrix}, \quad d = \begin{bmatrix} 1 \\ d_2 \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix},$$

where a_1 is a single column and b_1^{T} a single row.

The solution to this problem is

$$x = \begin{bmatrix} e_1 & 0 \\ e_2 - \widehat{x}d_2 & \widehat{x} \end{bmatrix},$$

where $\widehat{x}$ satisfies $\delta(\widehat{a}\widehat{x}\widehat{b} - c) = 0$, and

$$\widehat{a} = a_2, \quad \widehat{b} = b_2 - d_2 b_1^{\mathsf{T}}, \quad \widehat{c} = c - a e b_1^{\mathsf{T}}.$$

Finally we consider the addition of a second constraint so that the problem becomes

$$\delta(axb - c) = 0, \qquad xd = e, \qquad f^{\mathsf{T}}x = g^{\mathsf{T}},$$

where c is $(n-2) \times (n-2)$ and the dimensions of the various other matrix and vector partitions, including the specific values $d_1 = f_3 = 1$, are indicated in parentheses

$$a = \begin{matrix} \scriptstyle(1) & \scriptstyle(n-2) & \scriptstyle(1) \\ \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \end{matrix} \scriptstyle(n-2) \qquad b = \begin{bmatrix} b_1^{\mathsf{T}} & \scriptstyle(1) \\ b_2 & \scriptstyle(n-2) \\ b_3 & \scriptstyle(1) \end{bmatrix}^{\scriptstyle(n-2)} \qquad d = \begin{bmatrix} 1 & \scriptstyle(1) \\ d_2 & \scriptstyle(n-2) \\ d_3 & \scriptstyle(1) \end{bmatrix}^{\scriptstyle(1)}$$

$$e = \begin{bmatrix} e_1 & \scriptstyle(1) \\ e_2 & \scriptstyle(n-2) \\ e_3 & \scriptstyle(1) \end{bmatrix} \qquad f^{\mathsf{T}} = \begin{matrix} \scriptstyle(1) & \scriptstyle(n-2) & \scriptstyle(1) \\ \begin{bmatrix} f_1 & f_2^{\mathsf{T}} & 1 \end{bmatrix} \end{matrix} \scriptstyle(1) \qquad g^{\mathsf{T}} = \begin{matrix} \scriptstyle(1) & \scriptstyle(n-2) & \scriptstyle(1) \\ \begin{bmatrix} g_1 & g_2^{\mathsf{T}} & g_3 \end{bmatrix} \end{matrix} \scriptstyle(1)$$

For both linear constraints to be satisfied it is necessary that $f^\mathsf{T}e = f^\mathsf{T}Bd = g^\mathsf{T}d$. Assuming this consistency condition is satisfied, denote the common value of $f^\mathsf{T}e$ and $g^\mathsf{T}d$ by θ. The solution can now be written in the form

$$
x = \begin{bmatrix} e_1 & 0 & 0 \\ e_2 - \widehat{x}d_2 & \widehat{x} & 0 \\ e_3 + g_1 - \theta + f_2^\mathsf{T}\widehat{x}d_2 & g_2 - f_2^\mathsf{T}\widehat{x} & g_3 \end{bmatrix},
$$

where

$$
\delta(\widehat{a}\widehat{x}\widehat{b} - \widehat{c}) = 0,
$$

with

$$
\widehat{a} = a_2 - a_3 f_2^\mathsf{T}, \qquad \widehat{b} = b_2 - d_2 b_1^\mathsf{T}, \qquad \widehat{c} = c - aeb_1^\mathsf{T} - a_3 g^\mathsf{T}b + \theta a_3 b_1^\mathsf{T}.
$$

553   Derivation of methods with IRK stability

For the purpose of this discussion, we will always assume that the input approximations are represented by Z given by (551b), so that these approximations as input to step n are equal, to within $O(h^{p+1})$, to the quantities given by (551c).

Theorem 553A *If a general linear method with* $p = q = r - 1 = s - 1$ *has the property of IRK stability then the matrix* X *in Definition 551A is a* $(p + 1) \times (p + 1)$ *doubly companion matrix.*

Proof. Substitute (551d) into (551e) and compare (551d) with zX multiplied on the left. We find

$$
\exp(z)Z = z^2 BA \exp(cz) + zBUZ + VZ + O(z^{p+1}), \tag{553a}
$$
$$
z \exp(z)XZ = z^2 XB \exp(cz) + zXVZ + O(z^{p+1}). \tag{553b}
$$

Because $BA \equiv XB$ and $BU \equiv XV - VX$, the difference of (553a) and (553b) implies that

$$
zXZ \equiv Z + O(z^{p+1}).
$$

Because $zJZ \equiv Z + O(z^{p+1})$, it now follows that

$$
(X - J)Z \equiv O(z^p),
$$

which implies that $X - J$ is zero except for the first row and last column. $\square$

We will assume without loss of generality that $\beta_{p+1} = 0$.

By choosing the first row of X so that $\sigma(X) = \sigma(A)$, we can assume that the relation $BA = XB$ applies also to the first row. We can now rewrite the defining equations in Definition 551A as

$$BA = XB, \tag{553c}$$
$$BU = XV - VX + e_1\xi^{\mathsf{T}}, \tag{553d}$$

where $\xi^{\mathsf{T}} = [\xi_1 \quad \xi_2 \quad \cdots \quad \xi_{p+1}]$ is a specific vector. We will also write $\xi(z) = \xi_1 z + \xi_2 z^2 + \cdots + \xi_{p+1}z^{p+1}$. The transformed stability function in Theorem 551B can be recalculated as

$$(I - zX)M(z)(I - zX)^{-1} = V + ze_1\xi^{\mathsf{T}}(I - zX)^{-1},$$

with $(1,1)$ element equal to

$$1 + z\xi(I - zX)^{-1}e_1 = \frac{\det(I + z(e_1\xi - X))}{\det(I - zX)}$$
$$= \frac{(\alpha(z) + \xi(z))\beta(z)}{\alpha(z)\beta(z)} + O(z^{p+2}), \tag{553e}$$

where the formula for the numerator follows by observing that $X - e_1\xi$ is a doubly companion matrix, in which the α elements in the first row are replaced by the coefficients of $\alpha(z) + \xi(z)$.

The $(1,1)$ element of the transformed stability matrix will be referred to as the 'stability function' and denoted by $R(z)$. It has the same role for IRKS methods as the stability function of a Runge–Kutta method. For implicit methods, the stability function will be $R(z) = N(z)/(1 - \lambda z)^{p+1}$, where $N(z)$ is a polynomial of degree $p + 1$ given by

$$N(z) = \exp(z)(1 - \lambda z)^{p+1} - \epsilon_0 z^{p+1} + O(z^{p+2}).$$

The number ϵ_0 is the 'error constant' and is a design parameter for a particular method. It would normally be chosen so that the coefficient of z^{p+1} in $N(z)$ is zero. This would mean that if λ is chosen for A-stability, then this choice of ϵ_0 would give L-stability.

For non-stiff methods, $\lambda = 0$ and $N(z) = \exp(z) - \epsilon_0 z^{p+1} + O(z^{p+2})$. In this case, ϵ_0 would be chosen to balance requirements of accuracy against an acceptable stability region.

In either case, we see from (553e) that $N(z) = \alpha(z)(\beta(z) + \xi(z)) + O(z^{p+1})$, so that $\xi(z)$, and hence the coefficients $\xi_1, \xi_2, \ldots, \xi_{p+1}$ can be found.

Let C denote the $(p + 1) \times (p + 1)$ matrix with (i, j) element equal to $c_i^{j-1}/(j - 1)!$ and E the $(p + 1) \times (p + 1)$ matrix with (i, j) element equal to $1/(j - i)!$ (with the usual convention that this element vanishes if $i > j$). We can now write (551d) and (551e) as

$$U = C - ACK,$$
$$V = E - BCK.$$

Substitute into (553d) and make use of (553c) and we find

$$BC(I - KX) = XE - EX + e_1\xi^{\mathsf{T}}. \tag{553f}$$

Both $I - KX$ and $XE - EX + e_1\xi^{\mathsf{T}}$ vanish, except for their last columns, and (553f) simplifies to

$$BC
\begin{bmatrix}
\beta_p \\
\beta_{p-1} \\
\vdots \\
\beta_1 \\
1
\end{bmatrix}
=
\begin{bmatrix}
\frac{1}{1!} & \frac{1}{2!} & \cdots & \frac{1}{p!} & \frac{1}{(p+1)!} - \epsilon_0 \\
0 & \frac{1}{1!} & \cdots & \frac{1}{(p-1)!} & \frac{1}{(p)!} \\
\vdots & \vdots & & \vdots & \vdots \\
0 & 0 & \cdots & \frac{1}{1!} & \frac{1}{2!} \\
0 & 0 & \cdots & 0 & \frac{1}{1!}
\end{bmatrix}
\begin{bmatrix}
\beta_p \\
\beta_{p-1} \\
\vdots \\
\beta_1 \\
1
\end{bmatrix}.$$

Imposing conditions on the spectrum of V implies constraints on B. This principle is used to derive methods with a specific choice of the vector β and the abscissa vector c.

Rather than work in terms of B directly, we introduce the matrix $\widetilde{B} = \Psi^{-1}B$. Because

$$\widetilde{B}A = (J + \lambda I)\widetilde{B},$$

and because both A and $J + \lambda I$ are lower triangular, $\widetilde{B}$ is also lower triangular. In the derivation of a method, $\widetilde{B}$ will be found first and the method coefficient matrices found in terms of this as

$$A = \widetilde{B}^{-1}(J + \lambda I)\widetilde{B},$$
$$U = C - ACK,$$
$$B = \Psi\widetilde{B},$$
$$V = E - BCK.$$

To construct an IRKS method we need to carry out the following steps:

1. Choose the value of λ and ϵ_0 taking into account requirements of stability and accuracy.

2. Choose c_1, c_2, ..., c_{p+1}. These would usually be distributed more or less uniformly in $[0, 1]$.

3. Choose β_1, β_2, ..., β_p. This choice is to some extent arbitrary but can determine the magnitude of some of the elements in the coefficient matrices of the method.

4. Choose a non-singular $p \times p$ matrix P used to determine in what way $\dot{V}$ has zero spectral radius. If δ is defined as in Subsection 552, then we will impose the condition $\delta(P^{-1}\dot{V}P) = 0$. It would be normal to choose P as the product of a permutation matrix and a lower triangular matrix.

5. Solve the linear equations for the non-zero elements of $\widetilde{B}$ from a combination of the equations $\delta(P^{-1}\dot{\Psi}\widetilde{B}C\dot{K}P) = \delta(P^{-1}\dot{E}P)$ and

$$
\widetilde{B}C\begin{bmatrix} \beta_p \\ \beta_{p-1} \\ \vdots \\ \beta_1 \\ 1 \end{bmatrix} = \Psi^{-1}\begin{bmatrix} \frac{1}{1!} & \frac{1}{2!} & \cdots & \frac{1}{p!} & \frac{1}{(p+1)!} - \epsilon_0 \\ 0 & \frac{1}{1!} & \cdots & \frac{1}{(p-1)!} & \frac{1}{(p)!} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{1!} & \frac{1}{2!} \\ 0 & 0 & \cdots & 0 & \frac{1}{1!} \end{bmatrix}\begin{bmatrix} \beta_p \\ \beta_{p-1} \\ \vdots \\ \beta_1 \\ 1 \end{bmatrix}.
$$

554   Methods with property F

There is a practical advantage for methods in which

$$
e_1^\mathsf{T}B = e_{p+1}^\mathsf{T}A,
$$
$$
e_2^\mathsf{T}B = e_{p+1}^\mathsf{T}.
$$

A consequence of these assumptions is that $\beta_p = 0$.

For this subclass of IRKS methods, in addition to the existence of reliable approximations

$$
hF_i = hy'(x_{n-1} + hc_i) + O(h^{p+2}), \qquad i = 1, 2, \ldots, p+1, \tag{554a}
$$

where $y(x)$ is the trajectory such that $y(x_{n-1}) = y_1^{[n-1]}$, the value of $y_2^{[n-1]}$ provides an additional approximation

$$
hF_0 = hy'(x_{n-1}) + O(h^{p+2}),
$$

which can be used together with the $p+1$ scaled derivative approximations given by (554a).

This information makes it possible to estimate the values of

$$
h^{p+1}y^{(p+1)}(x_n) \quad \text{and} \quad h^{p+2}y^{(p+2)}(x_n),
$$

which are used for local error estimation purposes both for the method currently in use as well as for a possible method of one higher order. Thus we can find methods which provide rational criteria for stepsize selection as well as for order selection.

Using terminology established in Butcher (2006), we will refer to methods with this special property as possessing property F. They are an extension of FSAL Runge–Kutta methods.

The derivation of methods based on the ideas in Subsections 553 and 554 is joint work with William Wright and is presented in Wright (2002) and Butcher and Wright (2003, 2003a).

555   Some non-stiff methods

The following method, for which $c = [\frac{1}{3}, \frac{2}{3}, 1]^{\mathsf{T}}$, has order 2:

$$
\begin{bmatrix} A & U \\ B & V \end{bmatrix} =
\left[
\begin{array}{ccc|ccc}
0 & 0 & 0 & 1 & \frac{1}{3} & \frac{1}{18} \\
\frac{3}{10} & 0 & 0 & 1 & \frac{11}{30} & \frac{11}{90} \\
\frac{1}{5} & \frac{5}{12} & 0 & 1 & \frac{23}{60} & \frac{7}{45} \\
\hline
\frac{5}{3} & -\frac{29}{12} & \frac{4}{3} & 1 & \frac{5}{12} & \frac{2}{9} \\
-2 & 4 & -1 & 0 & 0 & 0 \\
3 & -9 & 6 & 0 & 0 & 0
\end{array}
\right].
\tag{555a}
$$

This method was constructed by choosing $\beta_1 = -\frac{1}{6}$, $\beta_2 = \frac{2}{9}$, $\epsilon_0 = 0$ and requiring $\dot{V}$ to be strictly upper triangular. It could be interpreted as having an enhanced order of 3, but of course the stage order is only 2.

The next method, with $c = [\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1]^{\mathsf{T}}$, has order 3:

$$
\left[
\begin{array}{cccc|cccc}
0 & 0 & 0 & 0 & 1 & \frac{1}{4} & \frac{1}{32} & \frac{1}{384} \\
\frac{224}{403} & 0 & 0 & 0 & 1 & -\frac{45}{806} & -\frac{45}{3224} & \frac{67}{19344} \\
\frac{1851}{2170} & \frac{93}{280} & 0 & 0 & 1 & -\frac{3777}{8680} & -\frac{681}{6944} & \frac{297}{138880} \\
\frac{305}{364} & \frac{5}{28} & \frac{5}{12} & 0 & 1 & -\frac{473}{1092} & -\frac{81}{728} & \frac{17}{17472} \\
\hline
\frac{305}{364} & \frac{5}{28} & \frac{5}{12} & 0 & 1 & -\frac{473}{1092} & -\frac{81}{728} & \frac{17}{17472} \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
-\frac{156}{7} & \frac{188}{7} & -20 & 8 & 0 & \frac{52}{7} & \frac{1}{7} & -\frac{1}{28} \\
-\frac{512}{7} & \frac{584}{7} & -\frac{160}{3} & 16 & 0 & \frac{568}{21} & \frac{4}{7} & -\frac{1}{7}
\end{array}
\right].
\tag{555b}
$$

For this method, possessing property F, $\beta_1 = \frac{1}{2}$, $\beta_2 = \frac{1}{16}$, $\epsilon_0 = 0$. The 3×3 matrix $\dot{V}$ is chosen so that $\delta(P^{-1}\dot{V}P) = 0$, where

$$
P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 4 & 1 & 0 \end{bmatrix}.
$$

556   Some stiff methods

The first example, with $\lambda = \frac{1}{4}$ and $c = [\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1]^\mathsf{T}$, has order 3:

$$
\begin{bmatrix} A & U \\ B & V \end{bmatrix}
$$

$$
= \left[\begin{array}{cccc|cccc}
\frac{1}{4} & 0 & 0 & 0 & 1 & 0 & -\frac{1}{32} & -\frac{1}{192} \\
\frac{11}{2124} & \frac{1}{4} & 0 & 0 & 1 & \frac{130}{531} & -\frac{11}{8496} & -\frac{719}{67968} \\
\frac{117761}{23364} & -\frac{189}{44} & \frac{1}{4} & 0 & 1 & -\frac{130}{531} & \frac{183437}{186912} & \frac{283675}{747648} \\
\frac{312449}{23364} & -\frac{4525}{396} & \frac{1}{36} & \frac{1}{4} & 1 & -\frac{650}{531} & \frac{121459}{46728} & \frac{130127}{124608} \\
\hline
-\frac{58405}{7788} & \frac{4297}{132} & -\frac{475}{12} & 15 & 1 & \frac{125}{236} & \frac{510}{649} & -\frac{733}{20768} \\
-\frac{64}{33} & \frac{746}{33} & -\frac{95}{3} & 12 & 0 & 0 & \frac{85}{44} & \frac{677}{1056} \\
-\frac{8}{3} & \frac{4}{3} & \frac{4}{3} & 0 & 0 & 0 & 0 & \frac{13}{24} \\
-32 & 112 & -128 & 48 & 0 & 0 & 0 & 0
\end{array}\right].
$$

$$(556a)$$

This method was constructed with $\beta_1 = -\frac{1}{4}$, $\beta_2 = \beta_3 = \frac{1}{4}$, $\epsilon_0 = \frac{1}{256}$ and $\delta(\dot{V}) = 0$. The choice of ϵ_0 was determined by requiring the stability function to be

$$
R(z) = \frac{1 - \frac{1}{8}z^2 - \frac{1}{48}z^3}{(1 - \frac{1}{4}z)^4},
$$

which makes the method L-stable.

The second example has order 4 and an abscissa vector $[\,1 \quad \frac{3}{4} \quad \frac{1}{4} \quad \frac{1}{2} \quad 1\,]$:

$$
A = \begin{bmatrix}
\frac{1}{4} & 0 & 0 & 0 & 0 \\
-\frac{513}{54272} & \frac{1}{4} & 0 & 0 & 0 \\
\frac{3706119}{69088256} & -\frac{488}{3819} & \frac{1}{4} & 0 & 0 \\
\frac{32161061}{197549232} & -\frac{111814}{232959} & \frac{134}{183} & \frac{1}{4} & 0 \\
-\frac{135425}{2948496} & -\frac{641}{10431} & \frac{73}{183} & \frac{1}{2} & \frac{1}{4}
\end{bmatrix},
$$

$$
U = \begin{bmatrix}
1 & \frac{3}{4} & \frac{1}{4} & \frac{1}{24} & 0 \\
1 & \frac{27649}{54272} & \frac{5601}{54272} & \frac{513}{108544} & -\frac{153}{54272} \\
1 & \frac{15366379}{207264768} & \frac{756057}{69088256} & \frac{1620299}{414529536} & -\frac{1615}{3636224} \\
1 & -\frac{32609017}{197549232} & \frac{929753}{65849744} & \frac{4008881}{197549232} & \frac{58327}{27726208} \\
1 & -\frac{367313}{8845488} & -\frac{22727}{2948496} & \frac{40979}{5896992} & \frac{323}{620736}
\end{bmatrix},
$$

$$B = \begin{bmatrix} -\frac{135425}{2948496} & -\frac{641}{10431} & \frac{73}{183} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 1 \\ \frac{2255}{1159} & -\frac{47125}{10431} & \frac{447}{61} & -\frac{11}{2} & \frac{7}{2} \\ \frac{25240}{3477} & -\frac{192776}{10431} & \frac{6728}{183} & -20 & 8 \\ \frac{9936}{1159} & -\frac{239632}{10431} & \frac{3120}{61} & -24 & 8 \end{bmatrix},$$

$$V = \begin{bmatrix} 1 & -\frac{367313}{8845488} & -\frac{22727}{2948496} & \frac{40979}{5896992} & \frac{323}{620736} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{28745}{10431} & -\frac{1937}{13908} & \frac{117}{18544} & \frac{65}{11712} \\ 0 & -\frac{141268}{10431} & -\frac{2050}{3477} & -\frac{187}{2318} & \frac{113}{1464} \\ 0 & -\frac{216416}{10431} & -\frac{452}{3477} & -\frac{491}{1159} & \frac{161}{732} \end{bmatrix}. \qquad (556b)$$

This property F method was constructed with $\beta_1 = \frac{3}{4}$, $\beta_2 = \frac{3}{16}$, $\beta_3 = \frac{1}{64}$, $\epsilon_0 = \frac{13}{15360}$ and $\delta(P^{-1}\dot{V}P) = 0$, where

$$P = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 8 & 1 & 0 & 0 \\ 16 & 4 & 1 & 0 \end{bmatrix}.$$

The method is L-stable with

$$R(z) = \frac{1 - \frac{1}{4}z - \frac{1}{8}z^2 + \frac{1}{96}z^3 + \frac{7}{768}z^4}{(1 - \frac{1}{4}z)^5}.$$

557   Scale and modify for stability

With the aim of designing algorithms based on IRKS methods in a variable order, variable stepsize setting, we consider what happens when h changes from step to step. If we use a simple scaling system, as in classical Nordsieck implementations, we encounter two difficulties. The first of these is that methods which are stable when h is fixed can become unstable when h is allowed to vary. The second is that attempts to estimate local truncation errors, for both the current method and for a method under consideration for succeeding steps, can become unreliable.

Consider, for example, the method (555b). If h is the stepsize in step n, which changes to rh in step $n + 1$, the output would be scaled from $y^{[n]}$ to $(D(r) \otimes I_N)y^{[n]}$, where $D(r) = \text{diag}(1, r, r^2, r^3)$. This means that the V matrix which determines stable behaviour for non-stiff problems, becomes effectively

$$
D(r)V = \begin{bmatrix}
1 & -\frac{473}{1092} & -\frac{81}{728} & \frac{17}{17472} \\
0 & 0 & 0 & 0 \\
0 & \frac{52}{7}r^2 & \frac{1}{7}r^2 & -\frac{1}{28}r^2 \\
0 & \frac{568}{21}r^3 & \frac{4}{7}r^3 & -\frac{1}{7}r^3
\end{bmatrix}.
$$

To guarantee stability we want all products of matrices of the form

$$
\widehat{V}(r) = \begin{bmatrix}
\frac{1}{7}r^2 & -\frac{1}{28}r^2 \\
\frac{4}{7}r^3 & -\frac{1}{7}r^3
\end{bmatrix}
\tag{557a}
$$

to be bounded. As a first requirement, we would need (557a) to be power-bounded. Because the determinant is zero, this means only that the trace $r^2(1-r)/7$ must lie in $[-1,1]$, so that $r \in [0,r^\star]$, where $r^\star \approx 2.310852163$ is a zero of $r^3 = r^2 + 7$. For a product $\widehat{V}(r_n)\widehat{V}(r_{n-1})\cdots\widehat{V}(r_1)$, the non-zero eigenvalue is $\prod_{i=1}^{n}\left((r^2 - r^3)/7\right)$ so that $r_1, r_2, \ldots, r_n \in [0, r^\star]$ is sufficient for variable stepsize stability.

While this is a very mild restriction on r values for this method, the corresponding restriction may be more severe for other methods. For example, for the scaled value of V given by (556b) the maximum permitted value of r is approximately 1.725419906.

Whatever restriction needs to be imposed on r for stability, we may wish to avoid even this restriction. We can do this using a modification to simple Nordsieck scaling. By Taylor expansion we find

$$
\begin{aligned}
&-\frac{40}{21}hy'(x_{n-1}+hc_1) - \frac{6}{7}hy'(x_{n-1}+hc_2) + \frac{40}{21}hy'(x_{n-1}+hc_3) \\
&- \frac{2}{3}hy'(x_{n-1}+hc_4) + \frac{32}{21}hy'(x_{n-1}) + \frac{1}{7}h^2y''(x_{n-1}) - \frac{1}{28}h^3y^{(3)}(x_{n-1}) \\
&\hspace{9cm} = O(h^4),
\end{aligned}
$$

so that it is possible to add a multiple of the vector

$$
d = \begin{bmatrix} -\frac{40}{21} & -\frac{6}{7} & \frac{40}{21} & -\frac{2}{3} \,\big|\, 0 & \frac{32}{21} & \frac{1}{7} & -\frac{1}{28} \end{bmatrix}
$$

to any row of the combined matrices $[B|V]$ without decreasing the order below 3. In the scale and modify procedure we can, after effectively scaling $[B|V]$ by $D(r)$, modify the result by adding $(1-r^2)d$ to the third row and $4(1-r^3)d$ to the fourth row. Expressed another way, write

$$
\delta = -\frac{40}{21}hF_1 - \frac{6}{7}hF_2 + \frac{40}{21}hF_3 - \frac{2}{3}hF_4 + \frac{32}{21}y_2^{[n-1]} + \frac{1}{7}y_3^{[n-1]} - \frac{1}{28}y_4^{[n-1]},
$$

so that the scale and modify process consists of replacing $y^{[n]}$ by

$$
D(r)y^{[n]} + \mathrm{diag}\left(0, 0, (1-r^2), 4(1-r^3)\right)\delta.
$$

558   Scale and modify for error estimation

Consider first the constant stepsize case and assume that, after many steps, there is an accumulated error in each of the input components to step n. If $y(x)$ is the particular trajectory defined by $y(x_{n-1}) = y_1^{[n-1]}$, then write the remaining input values as

$$y_i^{[n-1]} = h^{i-1} y^{(i-1)}(x_{n-1}) - \epsilon_{i-1} h^{p+1} y^{(p+1)}(x_{n-1}) + O(h^{p+2}),$$
$$i = 2, 3, \ldots, p+1. \quad (558a)$$

After a single step, the principal output will have acquired a truncation error so that its value becomes $y(x_n) - \epsilon_0 h^{p+1} y^{(p+1)}(x_n) + O(h^{p+2})$, where

$$\epsilon_0 = \frac{1}{(p+1)!} - \frac{1}{p!} \sum_{j=1}^{s} b_{1j} c_j^p + \sum_{j=2}^{r} v_{1j} \epsilon_{j-1}. \quad (558b)$$

Write ϵ as the vector with components ϵ_1, ϵ_2, ..., ϵ_p. The value of ϵ is determined by the fact that (558a) evolves after a single step to

$$y_i^{[n]} = h^{i-1} y^{(i-1)}(x_n) - \epsilon_{i-1} h^{p+1} y^{(p+1)}(x_n) + O(h^{p+2}),$$
$$i = 2, 3, \ldots, p+1. \quad (558c)$$

However,

$$y_i^{[n]} = h \sum_{j=1}^{s} b_{ij} y'(x_{n-1} + h c_j) + \sum_{j=2}^{r} v_{ij} y_j^{[n-1]} + O(h^{p+1}), \quad i = 2, 3, \ldots, p+1,$$
$$(558d)$$

so that substitution of (558a) and (558c) into (558d), followed by Taylor expansion about x_{n-1}, gives the result

$$\epsilon = \begin{bmatrix} \frac{1}{p!} \\ \frac{1}{(p-1)!} \\ \vdots \\ \frac{1}{1!} \end{bmatrix} - \frac{1}{p!} \dot{B} + \dot{V} \epsilon,$$

where $\dot{B}$ is the matrix B with its first row deleted. It was shown in Wright (2003) that

$$\epsilon_i = \beta_{p+1-i}, \quad i = 1, 2, \ldots, p.$$

Without a modification to the simple scaling process, the constancy of ϵ from step to step will be destroyed, and we consider how to correct for this. There are several reasons for wanting this correction. First, the reliability

of (558b), as providing an estimate of the local error in a step, depends on values of ϵ in the input to the current step. Secondly, asymptotically correct approximations to $h^{p+1}y^{(p+1)}(x_n)$ are needed for stepsize control purposes and, if these approximations are based on values of both hF and $y^{[n-1]}$, then these will also depend on ϵ in the input to the step. Finally, reliable estimates of $h^{p+2}y^{(p+2)}(x_n)$ are needed as a basis for dynamically deciding when an order increase is appropriate. It was shown in Butcher and Podhaisky (2006) that, at least for methods possessing property F, estimation of both $h^{p+1}y^{(p+1)}$ and $h^{p+2}y^{(p+2)}$ is possible, *as long as constant ϵ values are maintained.*

In Subsection 557 we considered the method (555b) from the point of view of variable stepsize stability. To further adjust to maintain the integrity of ϵ in a variable h regime, it is only necessary to add to the scaled and modified outputs $y_3^{[n]}$ and $y_4^{[n]}$, appropriate multiples of $-hF_1 + 3hF_2 - 3hF_3 + hF_4$.

Exercises 55

55.1 Show that the method given by (555a) has order 2, and that the stages are also accurate to this order.

55.2 Find the stability matrix of the method (555a), and show that it has two zero eigenvalues.

55.3 Show that the method given by (556a) has order 3, and that the stages are also accurate to this order.

55.4 Find the stability matrix of the method (556a), and show that it has two zero eigenvalues.

55.5 Show that (556a) is L-stable.

55.6 Show that the (i, j) element of Ψ^{-1} is equal to the coefficient of $w^{i-1}z^{j-1}$ in the power series expansion about $z = 0$ of $\alpha(z)/(1 - (\lambda + w)z)$.

References

Alexander R. (1977) Diagonally implicit Runge–Kutta methods for stiff ODEs. *SIAM J. Numer. Anal.,* 14, 1006–1021.

Axelsson O. (1969) A class of A-stable methods. *BIT,* 9, 185–199.

Axelsson O. (1972) A note on class of strongly A-stable methods. *BIT,* 12, 1–4.

Barton D., Willers I. M. and Zahar R. V. M. (1971) The automatic solution of systems of ordinary differential equations by the method of Taylor series. *Comput. J.,* 14, 243–248.

Bashforth F. and Adams J. C. (1883) *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 Employed in Constructing the Tables which Give the Theoretical Forms of Such Drops.* Cambridge University Press, Cambridge.

Brenan K. E., Campbell S. L. and Petzold L. R. (1989) *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations.* North-Holland, New York.

Brouder C. (2000) Runge–Kutta methods and renormalization. *Eur. Phys. J. C.,* 12, 521–534.

Burrage K. (1978) A special family of Runge–Kutta methods for solving stiff differential equations. *BIT,* 18, 22–41.

Burrage K. and Butcher J. C. (1980) Non-linear stability of a general class of differential equation methods. *BIT,* 20, 185–203.

Burrage K., Butcher J. C. and Chipman F. H. (1980) An implementation of singly-implicit Runge–Kutta methods. *BIT,* 20, 326–340.

Butcher J. C. (1963) Coefficients for the study of Runge–Kutta integration processes. *J. Austral. Math. Soc.,* 3, 185–201.

Butcher J. C. (1963a) On the integration processes of A. Huťa. *J. Austral. Math. Soc.,* 3, 202–206.

Butcher J. C. (1965) A modified multistep method for the numerical integration of ordinary differential equations. *J. Assoc. Comput. Mach.,* 12, 124–135.

Butcher J. C. (1965a) On the attainable order of Runge–Kutta methods. *Math. Comp.,* 19, 408–417.

Butcher J. C. (1966) On the convergence of numerical solutions to ordinary differential equations. *Math. Comp.,* 20, 1–10.

Butcher J. C. (1972) An algebraic theory of integration methods. *Math. Comp.,* 26, 79–106.

Butcher J. C. (1975) A stability property of implicit Runge–Kutta methods. *BIT,* 15, 358–361.

Butcher J. C. (1977) On A-stable implicit Runge–Kutta methods. *BIT,* 17, 375–378.

Butcher J. C. (1979) A transformed implicit Runge–Kutta method. *J. Assoc. Comput. Mach.,* 26, 731–738.

Butcher J. C. (1985) The nonexistence of ten-stage eighth order explicit Runge–Kutta methods. *BIT,* 25, 521–540.

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

Butcher J. C. (1987a) The equivalence of algebraic stability and AN-stability. *BIT,* 27, 510–533.

Butcher J. C. (1995) On fifth order Runge–Kutta methods. *BIT,* 35, 202–209.

Butcher J. C. (1995a) An introduction to DIMSIMs. *Comput. Appl. Math.,* 14, 59–72.

Butcher J. C. (1997) An introduction to 'Almost Runge–Kutta' methods. *Appl. Numer. Math.,* 24, 331–342.

Butcher J. C. (1998) ARK methods up to order five. *Numer. Algorithms,* 17, 193–221.

Butcher J. C. (2001) General linear methods for stiff differential equations. *BIT,* 41, 240–264.

Butcher J. C. (2002) The A-stability of methods with Padé and generalized Padé stability functions. *Numer. Algorithms,* 31, 47–58.

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

Butcher J. C. (2008) Order and stability of generalized Padé approximations. *Appl. Numer. Math.* (to appear).

Butcher J. C. and Cash J. R. (1990) Towards efficient Runge–Kutta methods for stiff systems. *SIAM J. Numer. Anal.,* 27, 753–761.

Butcher J. C. and Chartier P. (1997) A generalization of singly-implicit Runge–Kutta methods. *Appl. Numer. Math.,* 24, 343–350.

Butcher J. C. and Chipman F. H. (1992) Generalized Padé approximations to the exponential function. *BIT,* 32, 118–130.

Butcher J. C. and Hill A. T. (2006) Linear multistep methods as irreducible general linear methods. *BIT,* 46, 5–19.

Butcher J. C. and Jackiewicz Z. (1996) Construction of diagonally implicit general linear methods of type 1 and 2 for ordinary differential equations. *Appl. Numer. Math.,* 21, 385–415.

Butcher J. C. and Jackiewicz Z. (1998) Construction of high order diagonally implicit multistage integration methods for ordinary differential equations. *Appl. Numer. Math.,* 27, 1–12.

Butcher J. C. and Jackiewicz Z. (2003) A new approach to error estimation for general linear methods. *Numer. Math.,* 95, 487–502.

Butcher J. C. and Moir N. (2003) Experiments with a new fifth order method. *Numer. Algorithms,* 33, 137–151 .

Butcher J. C. and Podhaisky H. (2006) On error estimation in general linear methods for stiff ODEs. *Appl. Numer. Math.,* 56, 345–357.

Butcher J. C. and Rattenbury N. (2005) ARK methods for stiff problems. *Appl. Numer. Math.,* 53, 165–181 .

Butcher J. C. and Wright W. M. (2003) A transformation relating explicit and diagonally-implicit general linear methods. *Appl. Numer. Math.,* 44, 313–327.

Butcher J. C. and Wright W. M. (2003a) The construction of practical general linear methods. *BIT,* 43, 695–721.

Butcher J. C. and Wright W. M. (2006) Applications of doubly companion matrices. *Appl. Numer. Math.,* 56, 358–373.

Byrne G. D. and Lambert R. J. (1966) Pseudo-Runge–Kutta methods involving two points. *J. Assoc. Comput. Mach.,* 13, 114–123.

Cooper G. J. (1987) Stability of Runge-Kutta methods for trajectory problems. *IMA J. Numer. Anal.,* 7, 1–13.

Cooper G. J. and Verner J. H. (1972) Some explicit Runge–Kutta methods of high order. *SIAM J. Numer. Anal.,* 9, 389–405.

Curtis A. R. (1970) An eighth order Runge–Kutta process with eleven function evaluations per step. *Numer. Math.,* 16, 268–277.

Curtis A. R. (1975) High-order explicit Runge–Kutta formulae, their uses and limitations. *J. Inst. Math. Appl.,* 16, 35–55.

Curtiss C. F. and Hirschfelder J. O. (1952) Integration of stiff equations. *Proc. Nat. Acad. Sci. U.S.A.,* 38, 235–243.

Dahlquist G. (1956) Convergence and stability in the numerical integration of ordinary differential equations. *Math. Scand.,* 4, 33–53.

Dahlquist G. (1963) A special stability problem for linear multistep methods. *BIT,* 3, 27–43.

Dahlquist G. (1976) Error analysis for a class of methods for stiff non–linear initial value problems. In G. A. Watson (ed.) *Numerical Analysis*, Lecture Notes in Math. 506, Springer, Berlin, 60–72.

Dahlquist G. (1978) G-stability is equivalent to A-stability. *BIT,* 18, 384–401.

Dahlquist G. (1983) On one-leg multistep methods. *SIAM J. Numer. Anal.,* 20, 1130–1138.

Dahlquist G. and Jeltsch R. (1979) Generalized disks of contractivity for explicit and implicit Runge–Kutta methods, Technical Report TRITA NA–7906, Dept. of Numer. Anal. and Computing Sci., Roy. Inst. Tech.

Daniel J. W. and Moore R. E. (1970) *Computation and Theory in Ordinary Differential Equations.* W. H. Freeman, San Francisco.

Davis P. J. and Rabinowitz P. (1984) *Methods of Numerical Integration.* Academic Press, New York.

Donelson J. and Hansen E. (1971) Cyclic composite multistep predictor–corrector methods. *SIAM J. Numer. Anal.,* 8, 137–157.

Dormand J. R. and Prince P. J. (1980) A family of embedded Runge–Kutta formulae. *J. Comput Appl, Math.,* 6, 19–26.

Ehle B. L. (1969) On Padé approximations to the exponential function and *A*-stable methods for the numerical solution of initial value problems, Research Rep. CSRR 2010, Dept. of AACS, University of Waterloo, Ontario, Canada.

Ehle B. L. (1973) A-stable methods and Padé approximations to the exponential. *SIAM J. Math. Anal.,* 4, 671–680.

Ehle B. L. and Picel Z. (1975) Two parameter, arbitrary order, exponential approximations for stiff equations. *Math. Comp.,* 29, 501–511.

Euler L. (1913) De integratione aequationum differentialium per approximationem. In *Opera Omnia*, 1st series, Vol. 11, Institutiones Calculi Integralis, Teubner, Leipzig and Berlin, 424–434.

Fehlberg E. (1968) Classical fifth, sixth, seventh and eighth order Runge–Kutta formulas with stepsize control, NASA TR R-287.

Fehlberg E. (1969) Klassische Runge–Kutta-Formeln fünfter und siebenter Ordnung mit Schrittweiten-Kontrolle. *Computing,* 4, 93–106.

Gear C. W. (1965) Hybrid methods for initial value problems in ordinary differential equations. *SIAM J. Numer. Anal.,* 2, 69–86.

Gear C. W. (1967) The numerical integration of ordinary differential equations. *Math. Comp.,* 21, 146–156.

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

Gear C. W. (1971a) Algorithm 407, DIFSUB for solution of ordinary differential equations. *Comm. ACM,* 14, 185–190.

Gear C. W. (1980) Runge–Kutta starters for multistep methods. *ACM Trans. Math. Software,* 6, 263–279.

Gibbons A. (1960) A program for the automatic integration of differential equations using the method of Taylor series. *Comput. J.,* 3, 108–111.

Gill S. (1951) A process for the step-by-step integration of differential equations in an automatic computing machine. *Proc. Cambridge Philos. Soc.,* 47, 96–108.

Gragg W. B. and Stetter H. J. (1964) Generalized multistep predictor–corrector methods. *J. Assoc. Comput. Mach.,* 11, 188–209.

Gustafsson K. (1991) Control theoretic techniques for stepsize selection in explicit Runge–Kutta methods. *ACM Trans. Math. Software,* 17, 533–544.

Gustafsson K., Lundh M. and Söderlind G. (1988) A PI stepsize control for the numerical solution of ordinary differential equations. *BIT,* 28, 270–287.

Hairer E. (1978) A Runge–Kutta method of order 10. *J. Inst. Math. Appl.,* 21, 47–59.

Hairer E. and Leone P. (2000) Some properties of symplectic Runge–Kutta methods. *NZ J. Math.,* 29, 169–175.

Hairer E., Lubich C. and Roche M. (1989) *The Numerical Solution of Differential-Algebraic Systems by Runge–Kutta Methods,* Lecture Notes in Math. 1409. Springer, Berlin.

Hairer E., Lubich C. and Wanner G. (2006) *Geometric Numerical Integration: Structure-preserving Algorithms for Ordinary Differential Equations.* Springer, Berlin.

Hairer E., Nørsett S. P. and Wanner G. (1993) *Solving Ordinary Differential Equations I: Nonstiff Problems.* Springer, Berlin.

Hairer E. and Wanner G. (1974) On the Butcher group and general multi-value methods. *Computing,* 13, 1–15.

Hairer E. and Wanner G. (1981) Algebraically stable and implementable Runge–Kutta methods of high order. *SIAM J. Numer. Anal.,* 18, 1098–1108.

Hairer E. and Wanner G. (1982) Characterization of non-linearly stable implicit Runge–Kutta methods. In J. Hinze (ed.) *Numerical Integration of Differential Equations and Large Linear Systems*, Lecture Notes in Math. 968, Springer, Berlin, 207–219.

Hairer E. and Wanner G. (1996) *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems.* Springer, Berlin.

Henrici P. (1962) *Discrete Variable Methods in Ordinary Differential Equations.* John Wiley & Sons Inc, New York.

Henrici P. (1963) *Error Propagation for Difference Methods.* John Wiley & Sons Inc, New York.

Heun K. (1900) Neue Methoden zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen. *Z. Math. Phys.,* 45, 23–38.

Higham N. J. (1993) The accuracy of floating point summation. *SIAM J. Sci. Comput.,* 14, 783–799.

Hundsdorfer W. H. and Steininger B. I. (1991) Convergence of linear multistep and one-leg methods for stiff nonlinear initial value problems. *BIT,* 31, 124–143.

Huťa A. (1956) Une amélioration de la méthode de Runge–Kutta–Nyström pour la résolution numérique des équations différentielles du premier ordre. *Acta Fac. Nat. Univ. Comenian. Math.,* 1, 201–224.

Huťa A. (1957) Contribution à la formule de sixième ordre dans la méthode de Runge–Kutta–Nyström. *Acta Fac. Nat. Univ. Comenian. Math.,* 2, 21–24.

Iserles A., Munthe-Kaas H. Z., Nørsett S. P. and Zanna A. (2000) Lie-group methods. *Acta Numer.,* 9, 215–365.

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

Kahan W. (1965) Further remarks on reducing truncation errors. *Comm. ACM,* 8, 40.

Kirchgraber U. (1986) Multistep methods are essentially one-step methods. *Numer. Math.,* 48, 85–90.

Kutta W. (1901) Beitrag zur näherungsweisen Integration totaler Differential-gleichungen. *Z. Math. Phys.,* 46, 435–453.

Lambert J. D. (1991) *Numerical Methods for Ordinary Differential Systems: the Initial Value Problem.* John Wiley & Sons Ltd, Chichester.

Lasagni F. M. (1988) Canonical Runge–Kutta methods. *Z. Angew. Math. Phys.,* 39, 952–953.

López-Marcos M. A., Sanz-Serna J. M. and Skeel R. D. (1996) Cheap enhancement of symplectic integrators. In D. F. Griffiths and G. A. Watson (eds.) *Numerical Analysis*, Pitman Res. Notes Math. Ser., 344, Longman, Harlow, 107–122.

Lotka A. J. (1925) *Elements of Physical Biology.* Williams and Wilkins, Baltimore, Md.

Merson R. H. (1957) An operational method for the study of integration processes. In *Proc. Symp. Data Processing*, Weapons Research Establishment, Salisbury, S. Australia.

Milne W. E. (1926) Numerical integration of ordinary differential equations. *Amer. Math. Monthly,* 33, 455–460.

Milne W. E. (1953) *Numerical Solution of Differential Equations.* John Wiley & Sons Inc, New York.

Møller O. (1965) Quasi double-precision in floating point addition. *BIT,* 5, 37–50.

Møller O. (1965a) Note on quasi double-precision. *BIT,* 5, 251–255.

Moore R. E. (1964) The automatic analysis and control of error in digital computation based on the use of interval numbers. In L. B. Rall (ed.) *Error in Digital Computation*, vol. 1. John Wiley & Sons Inc, New York, 61–130.

Moulton F. R. (1926) *New Methods in Exterior Ballistics.* University of Chicago Press.

Nordsieck A. (1962) On numerical integration of ordinary differential equations. *Math. Comp.,* 16, 22–49.

Nørsett S. P. (1974) Semi-explicit Runge–Kutta methods, Report No. 6/74, Dept. of Math., Univ. of Trondheim.

Nyström E. J. (1925) Über die numerische Integration von Differentialgleichungen. *Acta Soc. Sci. Fennicae,* 50 (13), 55pp.

Obreshkov N. (1940) Neue Quadraturformeln. *Abh. der Preuß. Akad. der Wiss., Math.-naturwiss. Klasse,* 4, .

Prothero A. and Robinson A. (1974) On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations. *Math. Comp.,* 28, 145–162.

Rattenbury N. (2005) Almost Runge–Kutta methods for stiff and non-stiff problems, *PhD thesis, The University of Auckland.*

Richardson L. F. (1927) The deferred approach to the limit. *Philos. Trans. Roy. Soc. London Ser. A.*, 226, 299–361.

Robertson H. H. (1966) The solution of a set of reaction rate equations. In J. Walsh (ed.) *Numerical Analysis: An Introduction*, Academic Press, London, 178–182.

Romberg W. (1955) Vereinfachte numerische Integration. *Norske Vid. Selsk. Forh., Trondheim,* 28, 30–36.

Rosenbrock H. H. (1963) Some general implicit processes for the numerical solution of differential equations. *Comput. J.,* 5, 329–330.

Runge C. (1895) Über die numerische Auflösung von Differentialgleichungen. *Math. Ann.,* 46, 167–178.

Sanz-Serna J. M. (1988) Runge–Kutta schemes for Hamiltonian systems. *BIT,* 39, 877–883.

Sanz-Serna J. M. and Calvo M. P. (1994) *Numerical Hamiltonian Problems.* Chapman & Hall, London.

Scherer R. (1977) A note on Radau and Lobatto formulae for ODEs. *BIT,* 17, 235–238.

Scherer R. (1978) Spiegelung von Stabilitätsbereichen. In   R. Bulirsch, R. D. Grigorieff and J. Schröder (eds.)   *Numerical Treatment of Differential Equations*, Lecture Notes in Math. 631, Springer, Berlin, 147–152.

Singh A. D. (1999) Parallel diagonally implicit multistage integration methods for stiff ordinary differential equations, *PhD thesis, The University of Auckland.*

Söderlind G. (2002) Automatic control and adaptive time-stepping. *Numer. Algorithms,* 31, 281–310.

Stoffer D. (1993) General linear methods: connection to one step methods and invariant curves. *Numer. Math.,* 64, 395–408.

Suris Yu. B. (1988) Preservation of symplectic structure in the numerical solution of Hamiltonian systems (in Russian). *Akad. Nauk SSSR, Inst. Prikl. Mat., Moscow,,* 232, 148–160, 238–239.

Van der Pol B. (1926) On relaxation-oscillations. *Philos. Mag. Ser. 7,* 2, 978–992.

Verner J. H. (1978) Explicit Runge–Kutta methods with estimates of the local truncation error. *SIAM J. Numer. Anal.,* 15, 772–790.

Vitasek E. (1969) The numerical stability in solution of differential equations. In J.L. Morris (ed.)   *Conf. on Numerical Solution of Differential Equations*, Lecture Notes in Math. 109, Springer, Berlin, 87–111.

Volterra V. (1926) Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Memorie della R. Accad. Naz. dei Lincei, (Ser. VI),* 2, 31–131.

Wanner G., Hairer E. and Nørsett S. P. (1978) Order stars and stability theorems. *BIT,* 18, 475–489.

Wantanabe D. S. and Sheikh Q. M. (1984) One-leg formulas for stiff ordinary differential equations. *SIAM J. Sci. Statist. Comput.,* 2, 489–496.

Watts H. A. and Shampine L. F. (1972) *A*-stable block implicit one-step methods. *BIT,* 12, 252–266.

Wright K. (1970) Some relationships between implicit Runge–Kutta, collocation and Lanczos τ methods, and their stability properties. *BIT,* 10, 217–227.

Wright W. M. (2002) Explicit general linear methods with inherent Runge–Kutta stability. *Numer. Algorithms,* 31, 381–399.

Wright W. M. (2003) General linear methods with inherent Runge–Kutta stability, *PhD thesis, The University of Auckland.*

Index