Chapter 5
B-series and Runge–Kutta methods

5.1 Introduction

The aim of this chapter is to continue a selected study of Runge–Kutta methods. While the emphasis will necessarily be on the application of the B-series approach to order questions, we will also attempt to carry out a traditional analysis following in the footsteps of the pioneers of these methods. This will be based on the scalar test equation $y'(x) = f(x,y)$.

In modern computing there is little interest in numerical methods which are applicable only to scalar problem and it comes as a cautionary tale that the derivation of these methods does not automatically yield a method that works more widely. This will be illustrated by deriving a family of methods for which order 5 is achieved for scalar problems, whereas only the order 4 conditions hold for a vector problem.

For the derivation of practical methods, we need to use the key result:

Theorem 5.1A (Reprise of Theorem 3.6C (p. 127)) For an initial value problem of arbitrary dimension, a Runge–Kutta method $(A, b^{\intercal}, c)$ has order p if and only if

$$\Phi(\mathrm{t}) = \frac{1}{\mathrm{t!}}, \tag{5.1 a}$$

for all trees such that $|\mathrm{t}| \leq p$.

Chapter outline

The theory of order for scalar problems is presented in Section 5.2. The stability of Runge–Kutta methods is surveyed in Section 5.3, followed by the derivation of explicit methods in Section 5.4. Order barriers will be introduced in Section 5.5 through the simplest case (that order $p = s$ is impossible for explicit methods with $p > 4$). This is followed in Section 5.6 by a consideration of implicit methods. The generalization to effective (or conjugate) order is surveyed in Section 5.7.

J. C. Butcher, *B-Series*, Springer Series in Computational Mathematics 55,
https://doi.org/10.1007/978-3-030-70956-3_5

5.2 Order analysis for scalar problems

In contrast to the B-series approach to order conditions, it is also instructive to explore order conditions in the same way as the early pioneers. Hence, we will review the work in [82] (Runge, 1895), [56] (Heun, 1900), [66] (Kutta, 1901), who took as their starting point the scalar initial value problem

$$y'(x) = f(x,y), \qquad y(x_0) = y_0. \tag{5.2 a}$$

In this derivation of the order conditions, $\partial_x f := \partial f/\partial x$, $\partial_y f := \partial f/\partial y$, with similar notations for higher partial derivatives.

We start with (5.2 a) and find the second derivative of y by the chain rule

$$y'' = \partial_x f + (\partial_y f) f.$$

Similarly, we find the third derivative

$$\begin{aligned} y^{(3)} &= (\partial_x^2 f + (\partial_x \partial_y f) f) + \partial_y f(\partial_x f + (\partial_y f) f) + (\partial_x \partial_y f) f + (\partial_y^2 f) f^2 \\ &= \partial_x^2 f + 2(\partial_x \partial_y f) f + (\partial_y^2 f) f^2 + (\partial_x f\, \partial_y f) f + (\partial_y f)^2 f \end{aligned}$$

and carry on to find fourth and higher derivatives. By evaluating the $y^{(n)}$ at $x = x_0$, we find the Taylor expansions to use in (5.2 a). A more complicated calculation leads us to the detailed series (5.2 d) in the case of any particular Runge–Kutta method and hence to the determination of its order. Details of this line of enquiry will be followed below.

The greatest achievement in this line of work was given in [59] (Hut'a, 1956), where sixth order methods involving 8 stages were derived. In all derivations of new methods up to the publication of this tour de force, a tacit assumption is made. This is that a method derived to have a specific order for a general scalar problem will have this same order for a coupled system of scalar problems; that is, it will have this order for a problem with $N > 1$. This unproven assumption is untrue and it becomes necessary to carry out the order analysis in a multi-dimensional setting.

Systematic derivation of Taylor series

The evaluation of $y^{(n)}$, $n = 1, 2, \ldots, 5$, will now be carried out in a systematic manner. Let

$$D_{mn} = \sum_{i=0}^{m} \binom{m}{i} (\partial_x^{m-i} \partial_y^{n+i} f) f^i. \tag{5.2 b}$$

We will also write $\boldsymbol{D}_{mn}$ to denote D_{mn} evaluated at (x_0, y_0).

Lemma 5.2A

$$\frac{\mathrm{d}}{\mathrm{d}x}D_{mn} = D_{m+1,n} + mD_{m-1,n+1}D_{10}. \tag{5.2c}$$

Proof.

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}x}\sum_{i=0}^{m}\binom{m}{i}(\partial_x^{m-i}\partial_y^{n+i}f)f^i \\
&= \left(\sum_{i=0}^{m}\binom{m}{i}(\partial_x^{m-i+1}\partial_y^{n+i}f)f^i + \sum_{i=0}^{m}\binom{m}{i}(\partial_x^{m-i}\partial_y^{n+i+1}f)f^{i+1}\right) \\
&\qquad + \sum_{i=0}^{m}\binom{m}{i} i(\partial_x f)(\partial_x^{m-i}\partial_y^{n+i}\partial_y f)f^{i-1} \\
&= \sum_{i=0}^{m+1}\left(\binom{m}{i} + \binom{m}{i-1}\right)\partial_x^{m-i+1}(\partial_y^{n+i}f)f^i \\
&\qquad + \sum_{i=0}^{m}\left(i\frac{m!}{i!(m-i)!}\right)(\partial_x f)(\partial_x^{m-i}\partial_y^{n+i}\partial_y f)f^{i-1} \\
&= \sum_{i=0}^{m+1}\binom{m+1}{i}(\partial_x^{m-i+1}\partial_y^{n+i}f)f^i + m\sum_{i=0}^{m-1}\binom{m-1}{i}(\partial_x f)(\partial_x^{m-i-1}\partial_y^{n+1+i}\partial_y f)f^i \\
&= D_{m+1,n} + mD_{m-1,n+1}D_{10}.
\end{aligned}$$

□

Using Lemma 5.2A, we find in turn,

$$\begin{aligned}
y' &= D_{00}, \\
y'' &= D_{10}, \\
y''' &= D_{20} + D_{01}D_{10}, \\
y^{(4)} &= D_{30} + 2D_{11}D_{10} + D_{11}D_{10} + D_{01}(D_{20} + D_{01}D_{10}), \\
&= D_{30} + 3D_{11}D_{10} + D_{01}D_{20} + D_{01}^2D_{10}, \\
y^{(5)} &= D_{40} + 3D_{21}D_{10} + 3(D_{21} + D_{02}D_{10})D_{10} + 3D_{11}(D_{20}+D_{01}D_{10}), \\
&\quad +D_{11}D_{20}+D_{01}(D_{30}+2D_{11}D_{10})+2D_{01}D_{11}D_{10}+D_{01}^2(D_{20}+D_{01}D_{10}), \\
&= D_{40} + 6D_{21}D_{10} + 3D_{02}D_{10}D_{10} + 4D_{11}D_{20} + 7D_{11}D_{01}D_{10}, \\
&\quad + D_{01}D_{30} + D_{01}^2D_{20} + D_{01}^3D_{10}.
\end{aligned} \tag{5.2d}$$

To find the order conditions for a Runge–Kutta method, up to order 5, we need to systematically find the Taylor series for the stages, and finally for the output. In this analysis, we will assume that $\sum_{j=1}^{s} a_{ij} = c_i$ for all stages. For the stages it will be sufficient to work only to order 4 so that the scaled stage derivatives will include h^5 terms.

As a step towards finding the Taylor expansions of the stages and the output, we need to find the Taylor series for $hf(Y)$, for a given series $Y = y_0 + \cdots$. The following result does this for an arbitrary weighted series using the terms in (5.2 d).

Lemma 5.2B If

$$Y = y_0 + a_1 h\boldsymbol{D}_{00} + a_2 h^2 \boldsymbol{D}_{10} + a_3 h^3 \tfrac{1}{2}\boldsymbol{D}_{20} + a_4 h^3 \boldsymbol{D}_{01}\boldsymbol{D}_{10} + a_5 h^4 \tfrac{1}{6}\boldsymbol{D}_{30} + a_6 h^4 \boldsymbol{D}_{11}\boldsymbol{D}_{10} + a_7 h^4 \tfrac{1}{2}\boldsymbol{D}_{01}\boldsymbol{D}_{20} + a_8 h^4 \boldsymbol{D}_{01}^2\boldsymbol{D}_{10} + \mathscr{O}(h^5),$$

then

$$hf(x_0 + ha_1, Y) = hT_1 + h^2 T_2 + h^3 T_3 + h^4 T_4 + h^5 T_5 + \mathscr{O}(h^6),$$

where

$$\begin{aligned}
T_1 &= \boldsymbol{D}_{00},\\
T_2 &= a_1\boldsymbol{D}_{10},\\
T_3 &= \tfrac{1}{2}a_1^2\boldsymbol{D}_{20} + a_2\boldsymbol{D}_{01}\boldsymbol{D}_{10},\\
T_4 &= \tfrac{1}{6}a_1^3\boldsymbol{D}_{30} + a_1a_2\boldsymbol{D}_{11}\boldsymbol{D}_{10} + \tfrac{1}{2}a_3\boldsymbol{D}_{01}\boldsymbol{D}_{20} + a_4\boldsymbol{D}_{01}^2\boldsymbol{D}_{10},\\
T_5 &= \tfrac{1}{24}a_1^4\boldsymbol{D}_{40} + \tfrac{1}{2}a_1^2a_2\boldsymbol{D}_{21}\boldsymbol{D}_{10} + a_1a_3\boldsymbol{D}_{11}\boldsymbol{D}_{20} + \big(a_1a_4 + a_6\big)\boldsymbol{D}_{11}\boldsymbol{D}_{01}\boldsymbol{D}_{10}\\
&\quad + \tfrac{1}{2}a_2^2\boldsymbol{D}_{02}\boldsymbol{D}_{10}^2 + \tfrac{1}{6}a_5\boldsymbol{D}_{30}\boldsymbol{D}_{01} + \tfrac{1}{2}a_7\boldsymbol{D}_{01}^2\boldsymbol{D}_{20} + a_8\boldsymbol{D}_{01}^3\boldsymbol{D}_{10}.
\end{aligned}$$

Proof. Throughout this proof, an expression of the form $\partial_x^k \partial_y^m f$ is assumed to have been evaluated at (x_0, y_0). Evaluate T_1, T_2, T_3, T_4:

$$\begin{aligned}
T_1 &= f(x_0, y_0) = \boldsymbol{D}_{00},\\
T_2 &= a_1\,\partial_x f + a_1(\partial_y f)f = a_1\boldsymbol{D}_{10},\\
T_3 &= \tfrac{1}{2}a_1^2\,\partial_x^2 f + a_1^2(\partial_x\partial_y)\boldsymbol{D}_{00} + \tfrac{1}{2}a_1^2(\partial_y^2 f)\boldsymbol{D}_{00}^2 + a_2(\partial_y f)\boldsymbol{D}_{10}\\
&= \tfrac{1}{2}a_1^2\boldsymbol{D}_{20} + a_2\boldsymbol{D}_{01}\boldsymbol{D}_{10},\\
T_4 &= \tfrac{1}{6}a_1^3\,\partial_x^3 f + \tfrac{1}{2}a_1^3(\partial_x^2\partial_y f)\boldsymbol{D}_{10} + \tfrac{1}{2}a_1^3(\partial_x\partial_y^2 f)\boldsymbol{D}_{10}^2 + \tfrac{1}{6}a_1^3\partial_y^3 f\boldsymbol{D}_{10}^3\\
&\qquad + a_1a_2(\partial_x\partial_y f)\boldsymbol{D}_{10} + a_1a_2(\partial_y^2 f)\boldsymbol{D}_{10}\boldsymbol{D}_{01}\\
&\qquad + a_3(\partial_y f)\boldsymbol{D}_{20} + a_4(\partial_y f)\boldsymbol{D}_{01}\boldsymbol{D}_{10}\\
&= \tfrac{1}{6}a_1^3\boldsymbol{D}_{30} + a_1a_2\boldsymbol{D}_{11}\boldsymbol{D}_{10} + a_3\boldsymbol{D}_{01}\boldsymbol{D}_{20} + a_4\boldsymbol{D}_{01}^2\boldsymbol{D}_{10}.
\end{aligned}$$

The evaluation of the terms in T_5 is similar and is omitted except, as examples, the terms in a_1a_4 and a_6, which can be found from the simplified expression

$$hf(x_0 + a_1h, y_0 + ha_1\boldsymbol{D}_{00} + h^3a_4\boldsymbol{D}_{01}\boldsymbol{D}_{10} + h^5a_6\boldsymbol{D}_{11}\boldsymbol{D}_{10}).$$

The two example terms are

Table 16 Data for Theorem 5.2C

p	σ	$\boldsymbol{T}$	t	ϕ	$\boldsymbol{e}$
1	1	$\boldsymbol{D}_{00}$		$\sum b_i$	1
2	1	$\boldsymbol{D}_{10}$		$\sum b_i c_i$	$\frac{1}{2}$
3	2	$\boldsymbol{D}_{20}$		$\sum b_i c_i^2 i$	$\frac{1}{3}$
	1	$\boldsymbol{D}_{01}\boldsymbol{D}_{10}$		$\sum b_i a_{ij} c_j$	$\frac{1}{6}$
4	6	$\boldsymbol{D}_{30}$		$\sum b_i c_i^3$	$\frac{1}{4}$
	1	$\boldsymbol{D}_{11}\boldsymbol{D}_{10}$		$\sum b_i c_i a_{ij} c_j$	$\frac{1}{8}$
	2	$\boldsymbol{D}_{01}\boldsymbol{D}_{20}$		$\sum b_i a_{ij} c_j^2$	$\frac{1}{12}$
	1	$\boldsymbol{D}_{01}^2\boldsymbol{D}_{10}$		$\sum b_i a_{ij} a_{jk} c_k$	$\frac{1}{24}$
5	24	$\boldsymbol{D}_{40}$		$\sum b_i c_i^4$	$\frac{1}{5}$
	2	$\boldsymbol{D}_{21}\boldsymbol{D}_{10}$		$\sum b_i c_i^2 a_{ij} c_j$	$\frac{1}{10}$
	2	$\boldsymbol{D}_{11}\boldsymbol{D}_{20}$		$\sum b_i c_i a_{ij} c_j^2$	$\frac{1}{15}$
	1	$\boldsymbol{D}_{11}\boldsymbol{D}_{01}\boldsymbol{D}_{10}$		$\sum b_i (c_i + c_j) a_{ij} a_{jk} c_k$	$\frac{7}{120}$
	2	$\boldsymbol{D}_{02}\boldsymbol{D}_{10}^2$		$\sum b_i a_{ij} c_j a_{ik} c_k$	$\frac{1}{20}$
	6	$\boldsymbol{D}_{01}\boldsymbol{D}_{30}$		$\sum b_i a_{ij} c_3^3$	$\frac{1}{20}$
	2	$\boldsymbol{D}_{01}^2\boldsymbol{D}_{20}$		$\sum b_i a_{ij} a_{jk} c_k^2$	$\frac{1}{60}$
	1	$\boldsymbol{D}_{01}^3\boldsymbol{D}_{10}$		$\sum b_i a_{ij} a_{jk} a_{k\ell} c_\ell$	$\frac{1}{120}$

$$a_1 a_4 h^5 \left(\partial_x \partial_y f \boldsymbol{D}_{01}\boldsymbol{D}_{10} + \partial_y^2 f f \boldsymbol{D}_{01}\boldsymbol{D}_{10} \right) = h^5 a_1 a_4 \boldsymbol{D}_{10}\boldsymbol{D}_{11}\boldsymbol{D}_{01},$$

and

$$h^5 a_6 (\partial_y f \boldsymbol{D}_{11}\boldsymbol{D}_{10}) = h^5 a_6 \boldsymbol{D}_{10}\boldsymbol{D}_{11}\boldsymbol{D}_{01},$$

which combine to give the single term

$$h^5 (a_1 a_4 + a_6) \boldsymbol{D}_{10}\boldsymbol{D}_{11}\boldsymbol{D}_{01}. \qquad \square$$

For the stage values of a Runge–Kutta method, we have

$$\begin{aligned} Y_i &= y_0 + \sum_{j=1}^{s} a_{ij} h f(x_0 + hc_j, Y_j) \\ &= y_0 + hc_i \boldsymbol{D}_{00} + \mathscr{O}(h^2), \end{aligned}$$

and then, to one further order,

$$\begin{aligned} Y_i &= y_0 + \sum_{j=1}^{s} a_{ij} h f(x_0 + hc_j, y_0 + hc_j \boldsymbol{D}_{00}) + \mathscr{O}(h^3) \\ &= y_0 + hc_i \boldsymbol{D}_{00} + h^2 \sum_j a_{ij} c_j \boldsymbol{D}_{10} + \mathscr{O}(h^3). \end{aligned}$$

A similar expression can be written down for the output from a step

$$y_1 = y_0 + h \sum_i b_i \boldsymbol{D}_{00} + h^2 \sum_i b_i c_i \boldsymbol{D}_{10} + \mathscr{O}(h^3).$$

A comparison with the exact solution, $y_0 + hy'(x_0) + \frac{1}{2}h^2 y''(x_0) + \mathscr{O}(h^3)$, evaluated using (5.2 d) gives, as second order conditions,

$$\begin{aligned} \sum_i b_i \boldsymbol{D}_{00} &= \boldsymbol{D}_{00}, \\ \sum_i b_i c_i \boldsymbol{D}_{10} &= \tfrac{1}{2} \boldsymbol{D}_{10}. \end{aligned}$$

Theorem 5.2C In the statement of this result, the quantities p, $\boldsymbol{T}$, σ, ϕ are given in Table 16

1. The Taylor expansion for the exact solution to the initial value problem

$$y'(x) = f(x,y), \qquad y(x_0) = y_0, \tag{5.2 e}$$

to within $\mathscr{O}(h^6)$, is y_0 plus the sum of terms of the form

$$\boldsymbol{e} h^p \sigma^{-1} \boldsymbol{T}.$$

2. The Taylor expansion for the numerical solution y_1 to (5.2 e), using a Runge–Kutta method $(A, b^{\intercal}, c)$, to within $\mathscr{O}(h^6)$, is y_0 plus the sum of terms of the form

$$\phi h^p \sigma^{-1} \boldsymbol{T}.$$

3. The conditions to order 5 for the solution of (5.2 e) using $(A, b^{\intercal}, c)$ are the equations of the form

$$\phi = \boldsymbol{e}.$$

This analysis can be taken further in a straightforward and systematic way and is summarized, as far as order 5, in Theorem 5.2C. This theorem, for which the detailed proof is omitted, has to be read together with Table 16 (p. 181). To obtain a convenient comparison with the non-scalar case, the corresponding t, or more than a single t, in Theorem 5.1A (p. 177), are also shown in this table.

Relation with isomeric trees

Isomeric trees, introduced in Section 2.7 (p. 77), involve quantities $\mathbf{s}_{mn}$ which correspond to $\boldsymbol{D}_{mn}$ in the present section. The isomers occur when the $\mathbf{s}$ factors are formally allowed to commute. Commutation of the $\boldsymbol{D}$ factors in the order analysis actually occurs because these are scalar quantities. However, if the same analysis were carried out in the $\mathbb{R}^N$ setting, commutation would not occur because the $\boldsymbol{D}$ factors then become vectors, linear operators and multilinear operators. Hence, the trees comprising the isomeric classes would yield independent order conditions. In particular, for order 5, the only non-trivial class is

$$\{\boldsymbol{D}_{11}\boldsymbol{D}_{01}\boldsymbol{D}_{10},\ \boldsymbol{D}_{01}\boldsymbol{D}_{11}\boldsymbol{D}_{10}\} = \{\mathbf{F}(\mathbf{t}_{12}),\ \mathbf{F}(\mathbf{t}_{15})\}. \tag{5.2 f}$$

These give separate order conditions in the vector case because $\boldsymbol{D}_{11}$ and $\boldsymbol{D}_{01}$ no longer commute. This phenomenon will be illustrated by the construction of a method with ambiguous order.

Derivation of an ambiguous method

We will now construct a method which has order 5 for a scalar problem but only order 4 for a vector based problem. This means that all the conditions $\Phi(\mathbf{t}_i) = 1/\mathbf{t}_i!$ are satisfied for $i = 1, 2, \ldots, 17$ *except* for $i = 12$ and $i = 15$, for which the corresponding order conditions are replaced by

$$\Phi(\mathbf{t}_{12}) + \Phi(\mathbf{t}_{15}) = \frac{1}{\mathbf{t}_{12}!} + \frac{1}{\mathbf{t}_{15}!} = \frac{7}{120}. \tag{5.2 g}$$

For convenience, we will refer to the order conditions as (O1), (O2), …, (O17), where (Oi) is the equation

$$\Phi(\mathbf{t}_i) = 1/\mathbf{t}_i!. \tag{Oi}$$

That is,

$$b^{\intercal}\mathbf{1} = 1, \tag{O1}$$

$$b^{\intercal}c = \tfrac{1}{2}, \tag{O2}$$

$$b^{\intercal}c^2 = \tfrac{1}{3}, \tag{O3}$$

$$\vdots \quad \vdots$$

$$b^{\intercal}A^3c = \tfrac{1}{120}. \tag{O17}$$

In construcing this method, it is convenient to introduce a vector $d^\intercal$ defined as

$$d^\intercal = b^\intercal A + b^\intercal C - b^\intercal,$$

where $C = \mathrm{diag}(c)$, which satisfies the property

$$d^\intercal c^{n-1} = 0, \qquad n = 1,2,3,4, \tag{5.2h}$$

because $d^\intercal c^{n-1} = b^\intercal A c^{n-1} + b^\intercal c^n - b^\intercal c^{n-1} = 1/n(n+1) + 1/(n+1) - 1/n = 0$. In the method to be constructed, some assumptions will be made. These are

$$\sum_{j=1}^{i-1} a_{ij} c_j = \tfrac{1}{2} c_i^2, \qquad i \neq 2,3, \tag{5.2i}$$

$$c_6 = 1, \tag{5.2j}$$

$$b_2 = b_3 = 0. \tag{5.2k}$$

From (5.2j), (5.2k), (O1), (O2), (O3), (O5), (O9), it follows that

$$\sum_{i=1}^{6} b_i c_i (c_i - c_4)(c_i - c_5)(1 - c_i) = 0,$$

$$\text{implying} \qquad \tfrac{1}{120}(20 c_4 c_5 - 10(c_4 + c_5) + 4) = 0$$

$$\text{and hence} \qquad (\tfrac{1}{2} - c_4)(c_5 - \tfrac{1}{2}) = \tfrac{1}{20}.$$

Choose the convenient values $c_4 = \frac{1}{4}$, $c_5 = \frac{7}{10}$ together with $c_2 = \frac{1}{2}$ and $c_3 = 1$. The value of b, from (O1), (O2), (O3), (O5), and d from (5.2h) are

$$b = \begin{bmatrix} \frac{1}{14} & 0 & 0 & \frac{32}{81} & \frac{250}{567} & \frac{5}{54} \end{bmatrix},$$

$$d = t \begin{bmatrix} 1 & 7 & \frac{7}{9} & -\frac{112}{27} & -\frac{125}{27} & 0 \end{bmatrix},$$

where t is a parameter, assumed to be non-zero. The third row of A can be found from

$$d_2(-\tfrac{1}{2} c_2^2) + d_3(a_{32} c_2 - \tfrac{1}{2} c_3^2) = 0, \tag{5.2l}$$

because, from (O3) – (O8),

$$\begin{aligned} d^\intercal(Ac - \tfrac{1}{2}c^2) &= b^\intercal A^2 c + b^\intercal C A c - b^\intercal A c - \tfrac{1}{2} b^\intercal A c^2 - \tfrac{1}{2} b^\intercal c^3 + \tfrac{1}{2} b^\intercal c^2 \\ &= \tfrac{1}{24} \quad + \tfrac{1}{8} \quad - \tfrac{1}{6} \quad - \tfrac{1}{24} \quad - \tfrac{1}{8} \quad + \tfrac{1}{6} \quad = 0. \end{aligned}$$

From (5.2l), it is found that $a_{32} = \frac{13}{4}$. The values of a_{42}, and a_{52} can be written in terms of the other elements of rows 4 and 5 of A and row 6 can be found in terms of the other rows. There are now four free parameters remaining: a_{43}, a_{53}, a_{54} and t, and four conditions that are not automatically satisfied. These (O10), O16), (O17) and (5.2g). The solutions are given in the complete tableau, with $t = -3/140$,

$$
\begin{array}{c|cccccc}
0 & & & & & & \\
\frac{1}{2} & \frac{1}{2} & & & & & \\
1 & -\frac{9}{4} & \frac{13}{4} & & & & \\
\frac{1}{4} & \frac{9}{64} & \frac{5}{32} & -\frac{3}{64} & & & \\
\frac{7}{10} & \frac{63}{625} & \frac{259}{2500} & \frac{231}{2500} & \frac{252}{625} & & \\
1 & -\frac{27}{50} & -\frac{139}{50} & -\frac{21}{50} & \frac{56}{25} & \frac{5}{2} & \\
\hline
 & \frac{1}{14} & 0 & 0 & \frac{32}{81} & \frac{250}{567} & \frac{5}{54}
\end{array}
\tag{5.2 m}
$$

Numerical tests on the ambiguous method

For these tests we use the test problem (1.3 c) (p. 11), written in two alternative formulations, one scalar and one vector-valued, using matching initial and final values. Let

$$
t_0 = \exp(\tfrac{1}{10}\pi), \quad x_0 = t_0 \sin\big(\ln(t_0)\big), \quad y_0 = t_0 \cos\big(\ln(t_0)\big), \quad z_0 = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix},
$$

$$
t_1 = \exp(\tfrac{1}{2}\pi), \quad x_1 = t_1 \sin\big(\ln(t_1)\big), \quad y_1 = t_1 \cos\big(\ln(t_1)\big), \quad z_1 = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}.
$$

The scalar formulation, as an initial value problem with a specified output value is

$$
\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{y-x}{y+x}, \quad y(x_0) = y_0, \qquad y_1 = y(x_1)
$$

and the vector-valued formulation is

$$
\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} z^1 \\ z^2 \end{bmatrix} = \|z\|^{-1}\begin{bmatrix} z^2 + z^1 \\ z^2 - z^1 \end{bmatrix}, \quad z(t_0) = z_0, \qquad z_1 = z(t_1).
$$

Numerical tests were made for each problem on the intervals $[x_0, x_1]$ and $[t_0, t_1]$ respectively, using a sequence of stepsizes based on a total of $n = 5, 10, 20, \ldots, 5 \times 2^6$ steps, in each case.

Shown below are the absolute value, or the norm in the vector case, of the error at the output point. These are denoted by err_n. Also shown are $\mathrm{err}_{n/2}/\mathrm{err}_n$. For fourth order behaviour, these ratios should be approximately 16 and for fifth order, they should be approximately 32. The results are given in the display

n	err_n	$\mathrm{err}_{n/2}/\mathrm{err}_n$	err_n	$\mathrm{err}_{n/2}/\mathrm{err}_n$
5×2^0	4.3170×10^{-4}		9.4865×10^{-4}	
5×2^1	1.0906×10^{-5}	39.583	5.2577×10^{-5}	18.043
5×2^3	2.8486×10^{-7}	38.286	3.4454×10^{-6}	15.260
5×2^4	8.3007×10^{-9}	34.318	2.3100×10^{-7}	14.915
5×2^5	2.5422×10^{-10}	32.651	1.5117×10^{-8}	15.281
5×2^6	7.8960×10^{-12}	32.198	9.6908×10^{-10}	15.599

As we see, the predictions are confirmed by the computed error ratios.

The first sixth order method

In [59, 60] (Hut'a, 1966, 1967), the detailed conditions for a method with 8 stages to have order six were derived. The very intricate analysis in this work was combined with stage order conditions and other simplifying assumptions to yield methods with the required properties. In [8] (Butcher, 1963) it was shown that the simplifications had the effect of forcing the 31 conditions assumed by Hut'a to actually hold for the full set of 37 conditions required for applications to vector-valued problems. We will review this result starting with Table 17, which generalizes the pairing of two trees because they share an isomeric class. This generalization is taken to order 6 trees. Using Table 17, we can write down the scalar order conditions in the case of these isomeric classes.

$$\begin{aligned}
\Phi(\mathbf{t}_{12})+\Phi(\mathbf{t}_{15}) &= \frac{1}{\mathbf{t}_{12}!}+\frac{1}{\mathbf{t}_{15}!},\\
\Phi(\mathbf{t}_{21})+\Phi(\mathbf{t}_{29}) &= \frac{1}{\mathbf{t}_{21}!}+\frac{1}{\mathbf{t}_{29}!},\\
\Phi(\mathbf{t}_{25})+\Phi(\mathbf{t}_{31}) &= \frac{1}{\mathbf{t}_{25}!}+\frac{1}{\mathbf{t}_{31}!},\\
\Phi(\mathbf{t}_{28})+\Phi(\mathbf{t}_{33}) &= \frac{1}{\mathbf{t}_{28}!}+\frac{1}{\mathbf{t}_{33}!},\\
\Phi(\mathbf{t}_{26})+\Phi(\mathbf{t}_{32})+\Phi(\mathbf{t}_{35}) &= \frac{1}{\mathbf{t}_{26}!}+\frac{1}{\mathbf{t}_{32}!}+\frac{1}{\mathbf{t}_{35}!}.
\end{aligned}$$

The single-tree isomeric classes provide 26 order conditions which, together with the 5 listed above, constitute the 31 conditions to obtain order 6 for scalar problems, as in [59] (Hut'a, 1966).

Remarkably, Hut'a's methods are actually order 6, even for high-dimensional problems. To verify this, it is only necessary to show that

$$\Phi(\mathbf{t}_i)=\frac{1}{\mathbf{t}_i!},\qquad i=15,29,31,32,33,35. \tag{5.2 n}$$

But each of these trees is of the form $\mathrm{t}=[\mathrm{t}']$ so that, according to the $D(1)$ simplifying assumption, which holds for the Hut'a methods,

Table 17 Trees arranged in isomeric classes, with corresponding order conditions

$\boldsymbol{D}_{11}\boldsymbol{D}_{01}\boldsymbol{D}_{10}$		$\Phi(\mathrm{t}_{12}) = \frac{1}{\mathrm{t}_{12}!}$
$\boldsymbol{D}_{01}\boldsymbol{D}_{11}\boldsymbol{D}_{10}$		$\Phi(\mathrm{t}_{15}) = \frac{1}{\mathrm{t}_{15}!}$
$\boldsymbol{D}_{21}\boldsymbol{D}_{01}\boldsymbol{D}_{10}$		$\Phi(\mathrm{t}_{21}) = \frac{1}{\mathrm{t}_{21}!}$
$\boldsymbol{D}_{01}\boldsymbol{D}_{21}\boldsymbol{D}_{10}$		$\Phi(\mathrm{t}_{29}) = \frac{1}{\mathrm{t}_{29}!}$
$\boldsymbol{D}_{11}\boldsymbol{D}_{01}\boldsymbol{D}_{20}$		$\Phi(\mathrm{t}_{25}) = \frac{1}{\mathrm{t}_{25}!}$
$\boldsymbol{D}_{01}\boldsymbol{D}_{11}\boldsymbol{D}_{20}$		$\Phi(\mathrm{t}_{31}) = \frac{1}{\mathrm{t}_{31}!}$
$\boldsymbol{D}_{02}\boldsymbol{D}_{10}\boldsymbol{D}_{01}\boldsymbol{D}_{10}$		$\Phi(\mathrm{t}_{28}) = \frac{1}{\mathrm{t}_{28}!}$
$\boldsymbol{D}_{01}\boldsymbol{D}_{02}\boldsymbol{D}_{10}\boldsymbol{D}_{10}$		$\Phi(\mathrm{t}_{33}) = \frac{1}{\mathrm{t}_{33}!}$
$\boldsymbol{D}_{11}\boldsymbol{D}_{01}\boldsymbol{D}_{01}\boldsymbol{D}_{10}$		$\Phi(\mathrm{t}_{26}) = \frac{1}{\mathrm{t}_{26}!}$
$\boldsymbol{D}_{01}\boldsymbol{D}_{11}\boldsymbol{D}_{01}\boldsymbol{D}_{10}$		$\Phi(\mathrm{t}_{32}) = \frac{1}{\mathrm{t}_{32}!}$
$\boldsymbol{D}_{01}\boldsymbol{D}_{01}\boldsymbol{D}_{11}\boldsymbol{D}_{10}$		$\Phi(\mathrm{t}_{35}) = \frac{1}{\mathrm{t}_{35}!}$

$$\Phi(\mathrm{t}) - \Phi(\mathrm{t}') + \Phi(\mathrm{t}' * \tau) = \frac{1}{\mathrm{t}!} - \frac{1}{\mathrm{t}'!} + \frac{1}{(\mathrm{t}' * \tau)!}.$$

Consequently, $\Phi(\mathrm{t}) = 1/\mathrm{t}!$ because $\Phi(\mathrm{t}') = 1/\mathrm{t}'!$ and $\Phi(\mathrm{t}' * \tau) = 1/(\mathrm{t}' * \tau)!$, in each of these cases listed in (5.2 n).

5.3 Stability of Runge–Kutta methods

The stability function

Given a method (A, b^{T}, c), consider the result computed for the linear problem, $y' = qy$, where q is a (possibly complex) constant. If $z = hq$, the output after a single step is $y_1 = R(z)y_0$, where $R(z)$ is the "stability function", defined by

$$Y = y_0\mathbf{1} + zAY, \tag{5.3 a}$$

$$R(z)y_0 = y_0 + zb^{\mathsf{T}}Y. \tag{5.3 b}$$

From (5.3 a), $Y = y_0(I - zA)^{-1}$ and from (5.3 b),

$$R(z) = 1 + zb^{\intercal}(I - zA)^{-1}\mathbf{1}. \tag{5.3 c}$$

For an explicit s-stage method, it further follows that

$$R(z) = 1 + \sum_{n=1}^{s} \Phi([_n 1]_n) z^n. \tag{5.3 d}$$

Exercise 44 Verify (5.3 d) for an explicit s stage method.

If an explicit method has order $p = s$, it further follows that

$$R(z) = 1 + \sum_{n=1}^{p} \frac{z^n}{n!}. \tag{5.3 e}$$

Exercise 45 Verify (5.3 e) for an explicit method with $p = s$.

Finally, we find a convenient general formula for the stability function. See for example [53] (Hairer, Wanner, 1996).

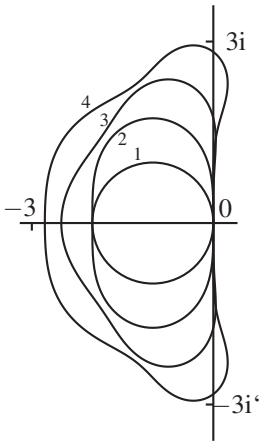
Lemma 5.3A The stability function for a Runge–Kutta method $(A, b^{\intercal}, c)$ is equal to

$$R(z) = \frac{\det\left(I + z(\mathbf{1}b^{\intercal} - A)\right)}{\det(I - zA)}. \tag{5.3 f}$$

Proof. If a square non-singular matrix M is perturbed by a rank 1 matrix $uv^{\intercal}$, the determinant is modified according to $\det(M + uv^{\intercal}) = \det(M) + v^{\intercal}\,\mathrm{adj}(M)u$. It follows that $\det(M + uv^{\intercal})/\det(M) = 1 + v^{\intercal}M^{-1}u$. Substitute $M = I - zA$, $u^{\intercal} = zb^{\intercal}$, $v = \mathbf{1}$ and the result follows from (5.3 c). □

The stability region and the stability interval

The set of points in the complex plane for which $|R(z)| \leq 1$ is the "stability region". The interval $I = [-r, 0]$, with maximal r, such that I lies in the stability region, is the stability "interval". In the case of the explicit methods for which $1 \leq p = s \leq 4$, the boundaries of these finite regions are as shown in the diagram:



The stability intervals are

$$p = 1: \quad I = [-2.000, 0]$$
$$p = 2: \quad I = [-2.000, 0]$$
$$p = 2: \quad I = [-2.513, 0]$$
$$p = 4: \quad I = [-2.785, 0]$$

The stability interval is an important attribute of a numerical method because, for a decaying exponential component of a problem, we want to avoid exponential growth of the corresponding numerical approximation.

Exercise 46 Find the stability function for the implicit method

$$\begin{array}{c|cc} \frac{1}{4} & \frac{7}{24} & -\frac{1}{24} \\ 1 & \frac{2}{3} & \frac{1}{3} \\ \hline & \frac{2}{3} & \frac{1}{3} \end{array}$$

5.4 Explicit Runge–Kutta methods

In this section we will review the derivation of the classical explicit methods in the full generality of multi-dimensional autonomous problems. That is, we will define the order of a method as given by Theorem 5.1A (p. 177)).

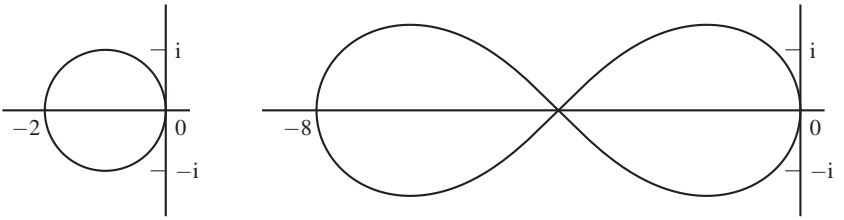
Low orders

Order 1

For a single stage there is only the Euler method but for $s > 1$ other options are possible such as

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{7}{8} & \frac{1}{8} \end{array} \tag{5.4 a}$$

This so-called "Runge–Kutta–Chebyshev" method [90] (van der Houwen, Sommeijer, 1980) is characterized by its extended real interval of stability; that is, a high value of r for its stability interval $I = [-r, 0]$. For this method the stability function is $R(z) = 1 + z + \frac{1}{8}z^2$ compared with $R_E(z) = 1 + z$ for the Euler method. The corresponding stability regions are shown in the following diagrams, with Euler on the left and (5.4 a) on the right.



The extended stability interval $[-8, 0]$ is regarded as an advantage for the solution of mildly stiff problems.

Order 2

From the order conditions

$$\begin{aligned} b_1 + b_2 &= 1, \\ b_2 c_2 &= \tfrac{1}{2}, \end{aligned}$$

the family of methods is found, where $c_2 \neq 0$,

$$\begin{array}{c|cc} 0 & & \\ c_2 & c_2 & \\ \hline & 1 - \frac{1}{2c_2} & \frac{1}{2c_2} \end{array}.$$

This family, particularly the special cases $c_2 = \frac{1}{2}$ and $c_2 = 1$, were made famous in the pioneering paper by Runge [82].

Exercise 47 Derive an explicit Runge–Kutta method with $s = p = 2$ and $b_2 = 1$.

Order 3

These methods are associated with the paper by Heun [56]. For a three stage method, with $p = 3$, we need to satisfy

$$\begin{aligned} b_1 + b_2 + b_3 &= 1, \\ b_2 c_2 + b_3 c_2 &= \tfrac{1}{2}, \\ b_2 c_2^2 + b_3 c_2^2 &= \tfrac{1}{3}, \\ b_3 a_{32} c_2 &= \tfrac{1}{6}. \end{aligned}$$

There are four cases to consider
(i) $c_2 \neq c_3,\ c_2 \neq 0 \neq c_3,\ c_2 \neq \frac{2}{3} \neq c_3$,
(ii) $c_3 = \frac{2}{3},\ 0 \neq c_2 \neq \frac{2}{3}$,
(iii) $c_2 = \frac{2}{3},\ c_3 = 0,\ b_3 \neq 0$,
(iv) $c_2 = c_3 = \frac{2}{3}$.
These are

(i)
$$\begin{array}{c|ccc} 0 & & & \\ c_2 & c_2 & & \\ c_3 & \frac{(3c_2^2-3c_2+c3)c_3}{(3c_2-2)c_2} & \frac{c_3(c_2-c_3)}{(3c_2-2)c_2} & \\ \hline & \frac{6c_2c_3-3c_2-3c_3+2}{6c_2c_3} & \frac{3c_3-2}{6c_2(c_3-c_2)} & \frac{2-3c_2}{6c_3(c_3-c_2)} \end{array},$$

(ii)
$$\begin{array}{c|ccc} 0 & & & \\ c_2 & c_2 & & \\ \frac{2}{3} & \frac{6c_2-2}{9c_2} & \frac{2}{9c_2} & \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array},$$

(iii)
$$\begin{array}{c|ccc} 0 & & & \\ \frac{2}{3} & \frac{2}{3} & & \\ 0 & -\frac{1}{4b_3} & \frac{1}{4b_3} & \\ \hline & \frac{1}{4} - b_3 & \frac{3}{4} & b_3 \end{array},$$

(iv)
$$\begin{array}{c|ccc} 0 & & & \\ \frac{2}{3} & \frac{2}{3} & & \\ \frac{2}{3} & \frac{2}{3} - \frac{1}{4b_3} & \frac{1}{4b_3} & \\ \hline & \frac{1}{4} & \frac{3}{4} - b_3 & b_3 \end{array}.$$

Examples from each case are

(i)
$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 1 & -1 & 2 & \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array},$$

(ii)
$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{3} & \frac{1}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array},$$

(iii)
$$\begin{array}{c|ccc} 0 & & & \\ \frac{2}{3} & \frac{2}{3} & & \\ 0 & -1 & 1 & \\ \hline & 0 & \frac{3}{4} & \frac{1}{4} \end{array},$$

(iv)
$$\begin{array}{c|ccc} 0 & & & \\ \frac{2}{3} & \frac{2}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & \frac{3}{8} & \frac{3}{8} \end{array}.$$

Exercise 48 Derive an explicit Runge–Kutta method with $s = p = 3$, $c_2 = c_3$, $b_2 = 0$.

Order 4

To obtain an order 4 method, with $s = 4$ stages, the equations $\Phi(\mathrm{t}) = 1/\mathrm{t}!$, $|\mathrm{t}| \leq 4$, must be satisfied. Write these as $u_i = v_i$, $i = 1, 2, \ldots, 8$, where the vectors u and v are given by

$$u := \begin{bmatrix} b_1 + b_2 + b_3 + b_4 \\ b_2c_2 + b_3c_3 + b_4c_4 \\ b_2c_2^2 + b_3c_3^2 + b_4c_4^2 \\ b_3a_{32}c_2 + b_4a_{42}c_2 + b_4a_{43}c_3 \\ b_2c_2^3 + b_3c_3^3 + b_4c_4^3 \\ b_3c_3a_{32}c_2 + b_4c_4a_{42}c_2 + b_4c_4a_{43}c_3 \\ b_3a_{32}c_2^2 + b_4a_{42}c_2^2 + b_4a_{43}c_3^2 \\ b_4a_{43}a_{32}c_2 \end{bmatrix}, \quad v := \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{3} \\ \frac{1}{6} \\ \frac{1}{4} \\ \frac{1}{8} \\ \frac{1}{12} \\ \frac{1}{24} \end{bmatrix}. \tag{5.4 b}$$

The value of c_4

If c_2, c_3, c_4 are parameters we might attempt to solve the system in three steps: (i) solve the linear system $u_2 = v_2$, $u_3 = v_3$, $u_5 = v_5$, to obtain b_2, b_3, b_3, (ii) solve $u_4 = v_4$, $u_6 = v_6$, $u_7 = v_7$ to obtain a_{32}, a_{42}, a_{43}, (iii) substitute the solutions to steps (ii) and (iii) into $u_8 = v_8$. This will give a condition on the c values. A short-circuit to this analysis is given in the following:

Lemma 5.4A For any explicit Runge–Kutta method, with $p = s = 4$, $c_4 = 1$.

Proof. From the equations

$$\begin{aligned}(u_7 - c_2u_4) &= \quad b_4a_{43} \cdot (c_3 - c_2)c_3,\\ (u_6 - c_4u_4) &= b_3(c_3 - c_4) \cdot a_{32}c_2,\\ u_5 - (c_2 + c_4)u_3 + c_2c_4u_2 &= b_3(c_3 - c_4) \cdot (c_3 - c_2)c_3,\\ u_8 &= \quad b_4a_{43} \cdot a_{32}c_2,\end{aligned}$$

it follows that

$$(u_7 - c_2u_4)(u_6 - c_4u_4) - \big(u_5 - (c_2 + c_4)u_3 + c_2c_4u_2\big)u_8 = 0.$$

Substitute $u_i = v_i$, with the result

$$(v_7 - c_2v_4)(v_6 - c_4v_4) - \big(v_5 - (c_2 + c_4)v_3 + c_2c_4v_2\big)v_8 = 0,$$

which simplifies to

$$c_2(c_4 - 1) = 0.$$

If $c_2 = 0$ we deduce the contradiction $u_8 = 0 \neq v_8$. Hence, $c_4 = 1$. □

Lemma 5.4B For any explict Runge–Kutta method, with $p = s = 4$, $D(1)$ holds. That is, $d_i = 0$, where $d_j = \sum_{i>j} b_i a_{ij} - b_j(1 - c_j)$, $j = 1, 2, 3, 4$.

Proof. From Lemma 5.4A, $d_4 = 0$; from $\sum_{j=1}^4 d_j(c_j - c_2)c_j = 0$, $d_3 = 0$; from $\sum_{j=1}^4 d_j(c_j = 0, d_2 = 0$; and from $\sum_{j=1}^4 d_j = 0$, $d_1 = 0$. □

Reduced tableaux for order 4

For Runge–Kutta methods in which $D(1)$ holds, the first column and last row of A can be omitted from the analysis and restored later when the derivation of the method is completed in other respects. Let $\widetilde{b}^{\mathsf{T}} = b^{\mathsf{T}}A$ and consider the reduced tableau

$$\begin{array}{c|cc} c_2 & & \\ c_3 & a_{32} & \\ \hline & \widetilde{b}_2 & \widetilde{b}_3 \end{array},$$

satisfying the reduced order conditions

$$\begin{aligned}\widetilde{b}_2c_2 + \widetilde{b}_3c_3 &= \frac{1}{6},\\ \widetilde{b}_2c_2^2 + \widetilde{b}_3c_3^2 &= \frac{1}{12},\end{aligned}$$

$$\widetilde{b}a_{32}c_2 = \frac{1}{24}.$$

Assuming $c_2, c_3 \notin \{0,1\}$ and $c_2 \neq \frac{1}{2}$ (to avoid $\widetilde{b}_3 = 0$), we find that

$$\widetilde{b}_2 = \frac{2c_3 - 1}{12c_2(c_3 - c_2)},$$
$$\widetilde{b}_3 = \frac{(1-2c_2)}{12c_3(c_3 - c_2)},$$
$$a_{32} = \frac{c_3(c_3 - c_2)}{2c_2(1-2c_2)},$$

leading to the full tableau:

$$\begin{array}{c|cccc} 0 & & & & \\ c_2 & c_2 & & & \\ c_3 & \frac{c_3(3c_2-4c_2^2-c_3)}{2c_2(1-2c_2)} & \frac{c_3(c_3-c_2)}{2c_2(1-2c_2)} & & \\ 1 & a_{41} & a_{42} & a_{43} & \\ \hline & \frac{6c_2c_3-2c_2-2c_3+1}{12c_2c_3)} & \frac{2c_3-1}{12c_2(1-c_2)(c_3-c_2)} & \frac{(1-2c_2)}{12c_3(1-c_3)(c_3-c_2)} & \frac{6c_2c_3-4c_2-4c_3+3}{12(1-c_3)(1-c_2))} \end{array} \quad , \tag{5.4 c}$$

where

$$a_{41} = \frac{12c_2^2c_3^2 - 12c_2^2c_3 - 12c_2c_3^2 + 4c_2^2 + 15c_2c_3 + 4c_3^2 - 6c_2 - 5c_3 + 2}{2c_3c_2(6c_2c_3 - 4c_2 - 4c_3 + 3)},$$
$$a_{42} = \frac{(1-c_2)(-4c_3^2 + c_2 + 5c_3 - 2)}{2c_2(c_3 - c_2)(6c_2c_3 - 4c_2 - 4c_3 + 3)},$$
$$a_{43} = \frac{(1-2c_2)(1-c_2)(1-c_3)}{c_3(c_3 - c_2)(6c_2c_3 - 4c_2 - 4c_3 + 3)}.$$

This "general" case takes a simpler form if c_2 and c_3 are symmetrically placed in $[0,1]$. Write $c_2 = 1 - c_3$ and (5.4 c) becomes

$$\begin{array}{c|cccc} 0 & & & & \\ 1-c_3 & 1-c_3 & & & \\ c_3 & \frac{c_3(1-2c_3)}{2(1-c_3)} & \frac{c_3}{2(1-c_3)} & & \\ 1 & \frac{(1-2c_3)(4-9c_3+6c_3^2)}{2(1-c_3)(-1+6c_3-6c_3^2)} & \frac{c_3(1-2c_3)}{2(1-c_3)(-1+6c_3-6c_3^2)} & \frac{1-c-3}{1-6c_3+6c3^2} & \\ \hline & \frac{-1+6c_3-6c_3^2}{12c_3(1-c_3)} & \frac{1}{12c_3(1-c_3)} & \frac{1}{12c_3(1-c_3)} & \frac{-1+6c_3-6c_3^2}{12c_3(1-c_3)} \end{array} \tag{5.4 d}$$

Exercise 49 Derive an explicit Runge–Kutta method with $s = p = 4$, $c_2 = \frac{1}{3}$, $c_3 = \frac{3}{4}$.

Kutta's classification of fourth order methods

In the famous 1901 paper by Kutta [66] the classical theory of Runge–Kutta methods, up to order 4, was effectively completed. Given the value $c_4 = 1$, 5 families of methods were formulated, in addition to the general case. These are as follows, where the reduced tableau is shown in the first column and the full tableau in the second. In each case λ is a non-zero parameter.

$$\begin{array}{c|cc} \lambda & & \\ \frac{1}{2} & \frac{1}{8\lambda} & \\ \hline & 0 & \frac{1}{3} \end{array}, \qquad \begin{array}{c|ccc} 0 & & & \\ \lambda & \lambda & & \\ \frac{1}{2} & \frac{1}{2}-\frac{1}{8\lambda} & \frac{1}{8\lambda} & \\ 1 & -1+\frac{1}{2\lambda} & -\frac{1}{2\lambda} & 2 \\ \hline & \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6} \end{array}, \tag{5.4e}$$

$$\begin{array}{c|cc} \frac{1}{2} & & \\ \frac{1}{2} & \frac{1}{12\lambda} & \\ \hline & \frac{1}{3}-\lambda & \lambda \end{array}, \qquad \begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ \frac{1}{2} & \frac{1}{2}-\frac{1}{12\lambda} & \frac{1}{12\lambda} & \\ 1 & 0 & 1-6\lambda & 6\lambda \\ \hline & \frac{1}{6} & \frac{2}{3}-2\lambda & 2\lambda & \frac{1}{6} \end{array}, \tag{5.4f}$$

$$\begin{array}{c|cc} 1 & & \\ \frac{1}{2} & \frac{1}{8} & \\ \hline & 0 & \frac{1}{3} \end{array}, \qquad \begin{array}{c|ccc} 0 & & & \\ 1 & 1 & & \\ \frac{1}{2} & \frac{3}{8} & \frac{1}{8} & \\ 1 & 1-\frac{1}{4\lambda} & -\frac{1}{12\lambda} & \frac{1}{3\lambda} \\ \hline & \frac{1}{6} & \frac{1}{6}-\lambda & \frac{2}{3} & \lambda \end{array}, \tag{5.4g}$$

$$\begin{array}{c|cc} \frac{1}{2} & & \\ 0 & \frac{1}{12\lambda} & \\ \hline & \frac{1}{3} & \lambda \end{array}, \qquad \begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 0 & -\frac{1}{12\lambda} & \frac{1}{12\lambda} & \\ 1 & -\frac{1}{2}-6\lambda & \frac{3}{2} & 6\lambda \\ \hline & \frac{1}{6}-\lambda & \frac{2}{3} & \lambda & \frac{1}{6} \end{array}. \tag{5.4h}$$

The contributions of S. Gill

A four stage method usually requires a memory of size $4N + C$ to carry out a single step. In [45] (Gill, 1951), a new fourth order method was derived in which the

memory requirements were reduced to $3N+C$. However, the work of S. Gill in this paper has a wider significance.

An important feature of Gill's analysis, and derivation of the new method, was the use of elementary differentials written in tensor form

$$f^i, \quad f^i_j f^j, \quad f^i_{jk} f^j f^k, \quad f^i_j f^j_k f^k, \quad \ldots,$$

represented by the trees

The Gill Runge–Kutta method

To motivate this discussion, we need to ask how many vectors are generated in each step of the calculation process, as stage values are evaluated and stage derivatives are then calculated. The calculations, and a list of the vectors needed at the end of each stage derivative calculation, are:

$$\begin{aligned}
Y_1 &= y_0, \\
hF_1 &= hf(Y_1), \left[\, Y_1 \quad hF_1 \,\right], \\
Y_2 &= y_0 + a_{21}hF_1, \\
hF_2 &= hf(Y_2), \left[\, Y_2 \quad hF_1 \quad hF_2 \,\right], \\
Y_3 &= y_0 + a_{31}hF_1 + a_{32}hF_2, \\
hF_3 &= hf(Y_3), \left[\, Y_3 \quad y_0 + a_{41}hF_1 + a_{42}hF_2 \quad y_0 + b_1hF_1 + b_2hF_2 \quad hF_3 \,\right], \\
Y_4 &= y_0 + a_{41}hF_1 + a_{42}hF_2 + a_{43}hF_3, \\
hF_4 &= hf(Y_4), \left[\, Y_4 \quad y_0 + b_1hF_1 + b_2hF_2 + b_3hF_3 \quad hF_4 \,\right].
\end{aligned}$$

Note that, at the time $hf(Y_3)$ is about to be computed, hF_1 and hF_2 are still needed for the eventual calculation of Y_4 and the output value y_1 and these are shown, in the list of required vectors, as partial formulae for these quantities, which can be updated as soon as hF_3 and hF_4 become available.

At the point in the process, immediately before hF_3 is computed, we need to have values of the three vectors, $Y_3 = y_0 + a_{31}hF_1 + a_{32}hF_2$, $y_0 + a_{41}hF_1 + a_{42}hF_2$ and $y_0 + b_1hF_1 + b_2hF_2$. This information can be stored as just two vectors if

$$\det\left(\begin{bmatrix} 1 & a_{31} & a_{32} \\ 1 & a_{41} & a_{42} \\ 1 & b_1 & b_2 \end{bmatrix}\right) = 0. \tag{5.4 i}$$

Gill's derivation based on the Kutta class (5.4 f), for which (5.4 i) holds gives

$$72\lambda^2 - 24\lambda + 1 = 0,$$

leading to

$$\lambda = \tfrac{1}{6} \pm \tfrac{1}{12}\sqrt{2}.$$

Gill recommends $\lambda = \frac{1}{6} - \frac{1}{12}\sqrt{2}$ based on the magnitude of the error coefficients.

An alternative solution satisfying Gill's criterion

If the assumption $c_2 = 1 - c_3$ is made, as in (5.4 d), Gill's criterion (5.4 i) gives

$$3c_3^3 - 3c_3 + 1 = 0,$$

with solutions

$$c_3 = -\frac{2}{\sqrt{3}} \cos\left(\tfrac{\pi}{18}\right), \tag{5.4 j}$$

$$c_3 = \frac{2}{\sqrt{3}} \cos\left(\tfrac{5\pi}{18}\right), \tag{5.4 k}$$

$$c_3 = \frac{2}{\sqrt{3}} \cos\left(\tfrac{7\pi}{18}\right). \tag{5.4 l}$$

The first case (5.4 j) gives $c_2 > 1$, $c_3 < 0$ and this should be rejected. The second case (5.4 k) has elements of rather large magnitude in A and seems, on this basis, less desirable than the third case (5.4 l), for which the tableau is, to 10 decimal places,

0				
0.6050691568	0.6050691568			
0.3949308432	0.0685790216	0.3263518216		
1	−0.5530334218	0.1581025768	1.3949308450	
	0.1512672890	0.3487327110	0.3487327110	0.1512672890

.

Fifth order and higher order methods

The pattern which holds up to order 4, in which methods with order p exist with $s = p$ stages, does not hold above order 4. It will be shown in Theorem 5.5A (p. 200) that, for $p > 4$, $s > p$ is necessary. We will complete this survey of explicit Runge–Kutta methods by presenting a single fifth order method with 6 stages and referring to a famous method with $p = 10$.

The role of simplifying assumptions

The $D(1)$ condition was necessary in the case of $s = p = 4$ and, at the same time, simplified the order requirements. The $C(2)$ condition is not really possible because it would imply $\frac{1}{2}c_2^2 = a_{21}c_1 = 0$. This would mean that $c_2 = c_1$ and the first two stages compute the same value and could be combined into a single stage. Taking this argument further, we conclude that *all* stages are equivalent to a single stage and only order 1 is possible.

But for order at least 5, it becomes very difficult to construct methods without assuming something related in some way to $C(2)$. We can see this by evaluating the following expression on the assumption that the order 5 order conditions are satisfied.

$$\begin{aligned}\sum_{i=1}^{s} b_i \Big(\sum_{j=1}^{s-1} a_{ij}c_j - \tfrac{1}{2}c_i^2 \Big)^2 &= \sum_{i=1}^{s} b_i \sum_{j=1}^{s-1} a_{ij}c_j a_{ik}c_k - \sum_{i=1}^{s}\sum_{j=1}^{s-1} b_i c_i^2 a_{ij}c_j + \tfrac{1}{4}\sum_{i=1}^{s} b_i c_i^4 \\ &= \tfrac{1}{20} - \tfrac{1}{10} + \tfrac{1}{20} \\ &= 0.\end{aligned}$$

If for example the $C(2)$ requirement were satisfied for each stage *except the second stage*, it would be necessary that $b_2 = 0$. It would also be necessary that

$$\sum b_i a_{i2} = 0, \quad \sum b_i c_i a_{i2} = 0, \quad \sum b_i a_{ij} a_{j2} = 0, \tag{5.4 m}$$

otherwise it would be impossible for the following three pairs of conditions to hold simultaneously.

$$\begin{aligned} \sum b_i a_{ij} a_{jk} c_k &= \tfrac{1}{24}, & \sum b_i a_{ij} c_j^2 &= \tfrac{1}{12}, \\ \sum b_i c_i a_{ij} a_{jk} c_k &= \tfrac{1}{30}, & \sum b_i c_i a_{ij} c_j^2 &= \tfrac{1}{15}, \\ \sum b_i a_{ij} a_{jk} a_{k\ell} c_\ell &= \tfrac{1}{120}, & \sum b_i a_{ij} a_{jk} c_k^2 &= \tfrac{1}{60}. \end{aligned}$$

If $D(1)$ holds, a suitably modified form of $C(2)$ does not require (5.4 m) but only $\sum b_i(1-c_i)a_{i2} = 0$, in addition to $b_2 = 0$. These assumptions open a path to the construction of fifth order methods. Rewrite (5.4 b) with u_i replaced by $\widehat{u}_i$, where the b_i are replaced by

$$\widehat{b}_i = \sum_{j>i} b_j a_{ji}, \qquad i = 1,2,3,4,$$

and the v_i are replaced by the elements of the vector

$$\widehat{v}^{\mathsf{T}} = \left[\begin{array}{cccccccc} \frac{1}{2} & \frac{1}{6} & \frac{1}{12} & \frac{1}{24} & \frac{1}{20} & \frac{1}{40} & \frac{1}{60} & \frac{1}{120} \end{array} \right].$$

To enable us to focus on the parts of the tableau that are most significant, we use a reduced tableau of the form

$$\begin{array}{c|ccc} c_3 & & & \\ c_4 & a_{43} & & \\ c_5 & a_{53} & a_{54} & \\ \hline & \widehat{b}_3 & \widehat{b}_4 & \widehat{b}_5 \end{array}.$$

We need to solve

$$\begin{aligned} \widehat{b}_3 c_3 + \widehat{b}_4 c_4 + \widehat{b}_5 c_5 &= \tfrac{1}{6}, \\ \widehat{b}_3 c_3^2 + \widehat{b}_4 c_4^2 + \widehat{b}_5 c_5^2 &= \tfrac{1}{12}, \\ \widehat{b}_3 c_3^3 + \widehat{b}_4 c_4^3 + \widehat{b}_5 c_5^3 &= \tfrac{1}{20}, \\ \widehat{b}_5 a_{54} c_4 (c_4 - c_3) &= \tfrac{1}{60} - \tfrac{1}{24} c_3. \end{aligned}$$

These conditions give no information about a_{43} and a_{53}. We also need to take into account the relations based on the $C(2)$ condition. These are

$$\begin{aligned} a_{32} c_2 &= \tfrac{1}{2} c_3^2, \\ a_{42} c_2 &= \tfrac{1}{2} c_4^2 - a_{43} c_3, \\ \widehat{b}_5 a_{52} &= -\widehat{b}_3 a_{32} - \widehat{b}_4 a_{42}, \\ a_{53} c_3 &= \tfrac{1}{2} c_5^2 - a_{52} c_2 - a_{54} c_4. \end{aligned}$$

We can solve these equations sequentially for a_{32}, a_{42}, a_{52}, and a_{53}, with a_{43} chosen arbitrarily, to complete the reduced tableau. In the special case

$$c = \begin{bmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \frac{3}{4} & 1 \end{bmatrix}^{\mathsf{T}},$$

with the chosen value $a_{43} = 1$ we find the reduced tableau to be

$$\begin{array}{c|ccc} \frac{1}{4} & 1 & & \\ \frac{1}{2} & 0 & \frac{9}{16} & \\ \hline & \frac{4}{15} & \frac{1}{15} & \frac{4}{45} \end{array},$$

leading to the full tableau

$$\begin{array}{c|cccccc} 0 & & & & & & \\ \frac{1}{4} & \frac{1}{4} & & & & & \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & & & & \\ \frac{1}{2} & 0 & -\frac{1}{2} & 1 & & & \\ \frac{3}{4} & \frac{3}{16} & 0 & 0 & \frac{9}{16} & & \\ 1 & -\frac{3}{7} & \frac{2}{7} & \frac{12}{7} & -\frac{12}{7} & \frac{8}{7} & \\ \hline & \frac{7}{90} & 0 & \frac{16}{45} & \frac{2}{15} & \frac{16}{45} & \frac{7}{90} \end{array}.$$

A tenth order method

Using a combination of simplifying assumptions and variants of these assumptions, a 17 stage method of order 10 [47] (Hairer, 1978) has been constructed. It is not known if methods of this order exist with $s < 17$.

5.5 Attainable order of explicit methods

As we have seen, it is possible to obtain order p with $s = p$ stages for $p \leq 4$. However, for $p \geq 5$, methods only exist if $s \geq p+1$. Furthermore, for $p \geq 7$, methods only exist if $s \geq p+2$.

Before presenting these results, we make some preliminary remarks.

Remarks

It can always be assumed that $c_2 \neq 0$

If a method existed with $c_2 = 0$, the second stage will give a result identical to the first stage so that the second stage can be effectively eliminated. That is, the tableau

$$\begin{array}{c|ccc} 0 & & & \\ 0 & 0 & & \\ c_3 & a_{21} & & \\ c_4 & a_{31} & a_{32} & \\ \vdots & \vdots & \vdots & \ddots \\ \hline & b_1 & b_2 & \cdots \end{array},$$

can be replaced by

$$\begin{array}{c|ccc} 0 & & & \\ c_3 & a_{21} & & \\ c_4 & a_{31}+a_{32} & & \\ \vdots & \vdots & & \ddots \\ \hline & b_1+b_2 & & \cdots \end{array},$$

with one less stage but the same order.

Some low rank matrix products

In the proofs given in this section, products of matrices U and V occur in which many terms cancel from UV because of zero elements in the final columns of U and the initial rows of V. Typically the rows of U have the form $b^{\intercal}A^m$, with some of the A factors replaced by some other strictly lower triangular matrices, and with the columns of V of the form $A^n c$, again with some of the A factors replaced by strictly lower triangular matrices. From the specific structure of U and V, an upper bound on the rank of UV can be given.

Order bounds

In this section, $C = \mathrm{diag}(c)$, $d_I = 1 - c_i$, $D = I - C$.

Theorem 5.5A No (explicit) Runge–Kutta s-stage method exists with order $p = s > 4$.

Proof. We will assume a method of this type exists and obtain a contradiction. Let

$$U = \begin{bmatrix} b^{\intercal}A^{p-3} \\ b^{\intercal}A^{p-4}(D-d_4I) \end{bmatrix},$$
$$V = \begin{bmatrix} Ac & (C-c_2I)c \end{bmatrix}.$$

Since each of these matrices has rank 1, their product is singular. However, the product is given by

$$UV = \begin{bmatrix} \frac{1}{p!} & \frac{2}{p!} - \frac{c_2}{(p-1)!} \\ \frac{p-3}{s!} - \frac{c_4}{(p-1)!} & \frac{2(p-3)}{p!} - \frac{2(p-3)c_2}{(p-1)!} - \frac{2c_4}{(p-1)!} + \frac{c_2c_4}{(p-2)!} \end{bmatrix}$$
$$= \begin{bmatrix} 1 & 0 \\ p-3 & -d_4 \end{bmatrix} \begin{bmatrix} \frac{1}{p!} & \frac{1}{(p-1)!} \\ \frac{1}{(p-1)!} & \frac{1}{(p-2)!} \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & -c_2 \end{bmatrix}.$$

Since the last two factors are non-singular, it follows that $d_4 = 0$ so that $c_4 = 1$. Repeat this argument with V unchanged but

$$U = \begin{bmatrix} b^{\intercal}A^{p-3} \\ b^{\intercal}A^{p-5}(D-d_5I)A \end{bmatrix},$$

and it follows that $c_5 = 1$. From $c_4 = c_5 = 1$, it follows that $b^{\intercal}A^{p-5}(I - \mathrm{diag}(c))A^2c$ is zero. However, by the order conditions,

$$b^{\intercal}A^{p-5}(I - \mathrm{diag}(c))A^2c = \frac{p-4}{p!} \neq 0. \qquad \square$$

For further results on attainable order, see [11] (Butcher 1965), [16] (Butcher 1985), [20] (Butcher 2016).

5.6 Implicit Runge–Kutta methods

The classical methods in which A is strictly lower triangular can be implemented explicitly. That is the stages can be evaluated in sequence, using only information already available so that the stage values, Y_i, and the corresponding B-series, which will be denoted by η_i, are computed by

$$Y_i = y_0 + h\sum_{j<i} a_{ij}f(Y_j), \qquad i = 1,2,\ldots,s,$$
$$\eta_i = 1 + \sum_{j<i} a_{ij}\eta_j \mathrm{D}, \qquad i = 1,2,\ldots,s.$$

In this section, we consider the consequences of allowing A to have non-zero elements on or above the diagonal. Four examples are the methods with tableaux

$$\begin{array}{c|cc} \frac{1}{2}-\frac{1}{6}\sqrt{3} & \frac{1}{4}-\frac{1}{6}\sqrt{3} & \frac{1}{4} \\ \frac{1}{2}+\frac{1}{6}\sqrt{3} & \frac{1}{4} & \frac{1}{4}+\frac{1}{6}\sqrt{3} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}, \tag{5.6a}$$

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}, \tag{5.6b}$$

$$\begin{array}{c|cc} 1-\frac{1}{2}\sqrt{2} & 1-\frac{1}{2}\sqrt{2} & 0 \\ 1 & \frac{1}{2}\sqrt{2} & 1-\frac{1}{2}\sqrt{2} \\ \hline & \frac{1}{2}\sqrt{2} & 1-\frac{1}{2}\sqrt{2} \end{array}, \tag{5.6c}$$

$$\begin{array}{c|cc} 3-2\sqrt{2} & \frac{5}{4}-\frac{3}{4}\sqrt{2} & \frac{7}{4}-\frac{5}{4}\sqrt{2} \\ 1 & \frac{1}{4}+\frac{1}{4}\sqrt{2} & \frac{3}{4}-\frac{1}{4}\sqrt{2} \\ \hline & \frac{1}{4}+\frac{1}{4}\sqrt{2} & \frac{3}{4}-\frac{1}{4}\sqrt{2} \end{array}, \tag{5.6d}$$

The fourth order method (5.6a), due to Hammer and Hollingsworth [54] (1955), is a member of the important Gauss family. These methods have maximal order $p = 2s$ amongst methods with s stages. They are A-stable and symplectic but they are *fully* implicit with high implementation costs, which increase as s increases. The method (5.6b) is an example of the Radau IIA family. The order is only $p = 2s - 1$, for the method with s stages, but it has the perceived advantage that $b^{\mathsf{T}} = e_s^{\mathsf{T}} A$. The effect of this restriction is a stronger stability property and improved behaviour for differential-algebraic problems.

In contrast to these fully-implicit methods, (5.6c) has order two, but the implementation cost is only twice that of the implicit Euler method. Like (5.6a), it is A-stable. The method (5.6d), while having a full A matrix, has comparable implementation cost to (5.6d) because $\sigma(A)$ contains only a single eigenvalue. That is, it is a "singly-implicit" method.

Competing attributes of implicit methods

The order, the stability and the implementability — meaning the existence of a low-cost implementation algorithm — are all attributes that make an implicit method capable of yielding results for a stiff problem efficiently and accurately. In this survey, some additional attributes are also considered.

Gauss methods

Let $\widetilde{P}_n$, $n = 0, 1, 2, \ldots$ denote the shifted Legendre polynomials, orthogonal on $[0, 1]$, and normalized by $\widetilde{P}_n(1) = 1$. They satisfy the recursion

$$\begin{aligned}\widetilde{P}_0(x) &= 1,\\ \widetilde{P}_1(x) &= 2x-1,\\ n\widetilde{P}_n(x) &= (2n-1)(2x-1)\widetilde{P}_{n-1}(x) - (n-1)\widetilde{P}_{n-2}(x).\end{aligned}$$

The zeros of $\widetilde{P}_n$ are real and distinct and lie in $(0, 1)$. The first few polynomials and their zeros are equal to

$$\begin{aligned}&\widehat{P}_0(x) = 1, && \text{zeros: } \{\},\\ &\widehat{P}_1(x) = 2x-1, && \text{zeros: } \{\tfrac{1}{2}\},\\ &\widehat{P}_2(x) = 6x^2-6x+1, && \text{zeros: } \{\tfrac{1}{2}-\tfrac{1}{6}\sqrt{3},\ \tfrac{1}{2}+\tfrac{1}{6}\sqrt{3}\},\\ &\widehat{P}_3(x) = 20x^3-30x^2+12x-1, && \text{zeros: } \{\tfrac{1}{2}-\tfrac{1}{10}\sqrt{15},\ \tfrac{1}{2},\ \tfrac{1}{2}+\tfrac{1}{10}\sqrt{15}\}.\end{aligned}$$

To construct the Gauss method of order $2s$, choose c, b and A as follows

1. Choose the components of c as the zeros of $\widetilde{P}_s$.
2. Choose $b^\intercal$ as the solution to the linear system $b^\intercal c^{k-1} = 1/k, \quad k = 1, 2, \ldots, s$.
3. Choose A as the solution to the linear system $A^\intercal c^{k-1} = c^k/k, \quad k = 1, 2, \ldots, s$.

Exercise 50 Find the tableau for the Gauss method with $s = 3$.

Radau IIA methods

These methods are usually preferred to Gauss methods for the solution of stiff problems and differential-algebraic equations. The components of c are the zeros of $\widetilde{P}_s - \widetilde{P}_{s-1}$. Also $b^\intercal$ and A are the solutions of $b^\intercal c^{k-1} = 1/k$, $k = 1, 2, \ldots, s$ and $A^\intercal c^{k-1} = c^k/k$, $k = 1, 2, \ldots, s$.

Exercise 51 Show that 1 is a component of c for a Radau IIA method.

Exercise 52 Find the tableau for the Radau IIA method with $s = 3$.

Implementability

We will discuss the valuation of the solution of the implicit equations in a single step, using a simplified form of Newton's method in which the Jacobian matrices $\mathbf{f}'(Y_i)$, $i = 1, 2, \ldots, s$, are approximated by a single matrix J. Each iteration takes the form $Y \mapsto Y - D$, where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix}, \quad F = \begin{bmatrix} f(Y_1) \\ f(Y_2) \\ \vdots \\ f(Y_s) \end{bmatrix}, \quad D = \begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_s \end{bmatrix},$$

$$\Phi(Y) = Y - h(A \otimes I_N)F,$$
$$\Phi'(Y) = I_{sN} - h(A \otimes J),$$
$$D = \Phi'(Y)^{-1}\Phi(Y).$$

Write $V = \Phi(Y)$, $M = \Phi'(Y)$, and consider the solution of $MD = V$, where we see that

$$M = \begin{bmatrix} I_N - a_{11}J & -a_{12}J & \cdots & -a_{1s}J \\ -a_{21}J & I_N - a_{22}J & \cdots & -a_{2s}J \\ \vdots & \vdots & & \vdots \\ -a_{s1}J & -a_{s2}J & \cdots & I_N - a_{ss}J \end{bmatrix}.$$

The cost of carrying out an LU factorization, in preparation for a sequence of iterations in a single step, or in a run of steps for slowly varying values of $\mathbf{f}'(Y_i)$, is const s^3N^3, where const is a small constant. However, if A is lower triangular, with constant diagonal λ, it is only necessary to prepare, by LU factorization, the single $N \times N$ matrix $I_N - h\lambda J$.

Transformations and practical implementable methods

Let T be a non-singular $s \times s$ matrix and define transformed vectors $\overline{Y} = TY$, $\overline{F} = TF$, $\overline{V} = TV$ and $\overline{D} = TD$. Also write $\overline{A} = TAT^{-1}$, $\overline{M} = TMT^{-1}$. The iterations can now be carried out using the solution of $\overline{MD} = \overline{V}$. It was proposed [5] (Burrage, 1978) to use "singly-implicit" methods, for which $\sigma(A) = \{\lambda\}$; with T chosen so that $\overline{A}$ is lower triangular, the preparation costs reduce to that of $I_N - h\lambda J$.

To obtain singly-implicit methods with stage-order s we must satisfy the two conditions

$$(A - \lambda I)^s = 0, \tag{5.6e}$$
$$Ac^{k-1} = \frac{1}{k}c^k, \quad k = 1, 2, \ldots, s. \tag{5.6f}$$

From (5.6f), $A^k\mathbf{1} = (1/k!)c^k$, and from (5.6e) we then have

$$(A - \lambda I)^s\mathbf{1} = \sum_{i=0}^{s} \binom{s}{i} (-\lambda)^{s-i} A^i \mathbf{1} \tag{5.6g}$$
$$= (-\lambda)^s \frac{s!}{(s-i)!(i!)^2} (-\lambda^{-1}c)^i. \tag{5.6h}$$

The Laguerre polynomials are given by

$$L_n(x) = \sum_{i=0}^{n} \frac{n!}{(n-i)!(i!)^2}(-x)^i, \tag{5.6 i}$$

and it follows that

$$c = \lambda \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_s \end{bmatrix}, \tag{5.6 j}$$

where $\xi_1, \xi_2, \ldots, \xi_s$ are the zeros of L_s.

5.7 Effective order methods

In [13] it was pointed out that the fifth-order barrier can be overcome by weakening the order conditions slightly to obtain methods which are *conjugate* to a fifth order method but have only 5 stages. The methods can be used efficiently, at least with constant stepsize, by carrying out pre- and post- processing steps at the beginning and end of a sequence of integration steps.

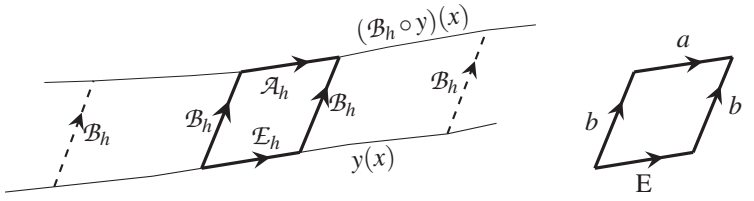
If the main steps correspond to a mapping $\mathcal{A}_h$ and the perturbation step, used as the pre-processor, corresponds to the mapping $\mathcal{B}_h$ then "effective order p" means that ${\mathcal{B}_h}^{-1} \circ \mathcal{A}_h \circ \mathcal{B}_h$ has order p. If $\mathcal{A}_h$ corresponds to $\mathbf{B}_h a$ and $\mathcal{B}_h$ to $\mathbf{B}_h b$ then we replace the usual order conditions by

$$bab^{-1} = \mathrm{E} + \mathsf{O}_{p+1},$$

or, what is equivalent,

$$ba = \mathrm{E}b + \mathsf{O}_{p+1}. \tag{5.7 a}$$

The diagram on the left of (5.7 b) illustrates the relationship between $\mathcal{A}_h$, $\mathcal{B}_h$ and $\mathcal{E}_h$. The diagram on the right represents the corresponding relationships between B-series.



(5.7 b)

In this section we will start by finding explicit conditions on the elementary weights required to satisfy (5.7 a), in terms of the free parameters made available from the elementary weights for b.

Complexity of effective order conditions

Using the notation introduced in Section 2.4, write a_n for the number of trees with order n. This means that, for methods with classical order p, the number of order conditions is

$$a_1 + a_2 + \cdots + a_p,$$

but for effective order this is replaced by $1 + a_p$. This is shown in [26]. Although there are $a_1 + a_2 + \cdots + a_p$ parameters from b, these are not all *free* parameters. The equations in (5.7 a) do not involve $b(\mathrm{t})$ when $|\mathrm{t}| = p$ because these terms cancel from the sides of $(ba)(\mathrm{t}) = (\mathrm{E}b)(\mathrm{t})$. Furthermore, there is no loss in generality in assuming that $b(\tau) = 0$ because, for any parameter θ, b can be replaced by $E^{(\theta)}b$, because, if (5.7 a) holds, then

$$(\mathrm{E}^{(\theta)}b)a = \mathrm{E}^{(\theta)}ba = \mathrm{E}^{(\theta)}\mathrm{E}b = \mathrm{E}^{(\theta+1)}b = \mathrm{E}(\mathrm{E}^{(\theta)}b).$$

We will focus on the case $p = 5$. For simplicity we will consider methods in the intersection of $C(2)$ and $D(1)$.

Methods satisfying $D(1)$ and $C(2)$

For the special methods in $D(1)$, we can substitute $\tau * \mathrm{t} = \tau\mathrm{t} - \mathrm{t} * \tau$ as a step towards writing $a(\tau * \mathrm{t}) = a(\tau)a(\mathrm{t}) - a(\mathrm{t} * \tau)$, for $|\mathrm{t}| \leq 5$. In the case of $C(2)$, $[\tau] = \frac{1}{2}\tau^2$ and for certain pairs $(\mathrm{t}, \mathrm{t}')$, we have $\mathrm{t}' = \frac{1}{2}\mathrm{t}$. These assumptions are shown together in the following tabulation

$C(2)$	$D(1)$
$\mathbf{t}_2 = \frac{1}{2}\mathbf{t}_1^2,$	$\mathbf{t}_2 = \frac{1}{2}\mathbf{t}_1$
$\mathbf{t}_4 = \frac{1}{2}\mathbf{t}_3,$	$\mathbf{t}_4 = \mathbf{t}_1\mathbf{t}_2 - \mathbf{t}_3 = \frac{1}{2}\mathbf{t}_1^3 - \mathbf{t}_3,$
$\mathbf{t}_6 = \frac{1}{2}\mathbf{t}_5,$	
	$\mathbf{t}_7 = \mathbf{t}_1\mathbf{t}_3 - \mathbf{t}_5,$
$\mathbf{t}_8 = \frac{1}{2}\mathbf{t}_7,$	$\mathbf{t}_8 = \mathbf{t}_1\mathbf{t}_4 - \mathbf{t}_6 = \frac{1}{2}\mathbf{t}_1^4 - \mathbf{t}_1\mathbf{t}_3 - \mathbf{t}_6,$
$\mathbf{t}_{10} = \frac{1}{2}\mathbf{t}_9,$	
$\mathbf{t}_{12} = \frac{1}{2}\mathbf{t}_{11},$	
$\mathbf{t}_{13} = \frac{1}{2}\mathbf{t}_{10} = \frac{1}{4}\mathbf{t}_9,$	
	$\mathbf{t}_{14} = \mathbf{t}_1\mathbf{t}_5 - \mathbf{t}_9,$
$\mathbf{t}_{15} = \frac{1}{2}\mathbf{t}_{14},$	$\mathbf{t}_{15} = \mathbf{t}_1\mathbf{t}_6 - \mathbf{t}_{10},$
	$\mathbf{t}_{16} = \mathbf{t}_1\mathbf{t}_7 - \mathbf{t}_{11} = \mathbf{t}_1^2\mathbf{t}_3 - \mathbf{t}_1\mathbf{t}_5 - \mathbf{t}_{11},$
$\mathbf{t}_{17} = \frac{1}{2}\mathbf{t}_{16},$	$\mathbf{t}_{17} = \mathbf{t}_1\mathbf{t}_8 - \mathbf{t}_{12} = \frac{1}{2}\mathbf{t}_1^5 - \mathbf{t}_1^2\mathbf{t}_3 - \mathbf{t}_1\mathbf{t}_6 - \mathbf{t}_{12}.$

By comparing the two lists, we see that $\mathbf{t}_3 = \frac{1}{3}\mathbf{t}_1^3$ and $\mathbf{t}_4 = \frac{1}{6}\mathbf{t}_1^3$. From the combination of these conditions, the set of trees necessary to analyse the effective order conditions reduces to $\{\mathbf{t}_1, \mathbf{t}_5, \mathbf{t}_9, \mathbf{t}_{11}\}$.

To facilitate evaluations of group products calculate the bi-product in Sweedler notation

$$\begin{aligned}
\Delta(\mathbf{t}_1) &= \mathbf{t}_1 \otimes \varnothing + 1 \otimes \mathbf{t}_1,\\
\Delta(\mathbf{t}_5) &= \mathbf{t}_5 \otimes \varnothing + \mathbf{t}_1^3 \otimes \mathbf{t}_1 + 3\mathbf{t}_1^2 \otimes \mathbf{t}_2 + \mathbf{t}_1 \otimes \mathbf{t}_1^3 + 1 \otimes \mathbf{t}_5\\
&= \mathbf{t}_5 \otimes \varnothing + \mathbf{t}_1^3 \otimes \mathbf{t}_1 + \tfrac{3}{2}\mathbf{t}_1^2 \otimes \mathbf{t}_1^2 + \mathbf{t}_1 \otimes \mathbf{t}_1^3 + 1 \otimes \mathbf{t}_5,\\
\Delta(\mathbf{t}_9) &= \mathbf{t}_9 \otimes \varnothing + \mathbf{t}_1^4 \otimes \mathbf{t}_1 + 4\mathbf{t}_1^3 \otimes \mathbf{t}_2 + 6\mathbf{t}_1^2 \otimes \mathbf{t}_3 + 4\mathbf{t}_1 \otimes \mathbf{t}_5 + 1 \otimes \mathbf{t}_9\\
&= \mathbf{t}_9 \otimes \varnothing + \mathbf{t}_1^4 \otimes \mathbf{t}_1 + 2\mathbf{t}_1^3 \otimes \mathbf{t}_1^2 + 2\mathbf{t}_1^2 \otimes \mathbf{t}_1^3 + 4\mathbf{t}_1 \otimes \mathbf{t}_5 + 1 \otimes \mathbf{t}_9,\\
\Delta(\mathbf{t}_{11}) &= \mathbf{t}_{11} \otimes \varnothing + \mathbf{t}_1\mathbf{t}_3 \otimes \mathbf{t}_1 + (\mathbf{t}_1^3 + \mathbf{t}_3) \otimes \mathbf{t}_2 + \mathbf{t}_1^2 \otimes \mathbf{t}_3\\
&\qquad + 2\mathbf{t}_1^2 \otimes \mathbf{t}_4 + 2\mathbf{t}_1 \otimes \mathbf{t}_6 + \mathbf{t}_1 \otimes \mathbf{t}_7 + 1 \otimes \mathbf{t}_{11}\\
&= \mathbf{t}_{11} \otimes \varnothing + \tfrac{1}{3}\mathbf{t}_1^4 \otimes \mathbf{t}_1 + \tfrac{2}{3}\mathbf{t}_1^3 \otimes \mathbf{t}_1^2 + \tfrac{2}{3}\mathbf{t}_1^2 \otimes \mathbf{t}_1^3 + \tfrac{1}{4}\mathbf{t}_1 \otimes \mathbf{t}_1^4 + 1 \otimes \mathbf{t}_{11}.
\end{aligned}$$

To find the formula for $(ba)(\mathrm{t})$, replace each term in $(\mathrm{f} \otimes \mathrm{f}')$ in $\Delta(\mathrm{t})$ by $b(\mathrm{f})a(\mathrm{f}')$ and similarly for $\mathrm{E}(\mathrm{f})b(\mathrm{f}')$. As has been remarked, we will, without loss of generality, assume that $b(\mathrm{t}) = 0$ for $\mathrm{t} = \tau$ and whenever $|\mathrm{t}| = 5$. The results are

$$\begin{aligned}
(ba)(\mathbf{t}_1) &= a(\mathbf{t}_1), & (\mathrm{E}b)(\mathbf{t}_1) &= 1,\\
(ba)(\mathbf{t}_5) &= b(\mathbf{t}_5) + a(\mathbf{t}_5), & (\mathrm{E}b)(\mathbf{t}_3) &= \tfrac{1}{4} + b(\mathbf{t}_5),\\
(ba)(\mathbf{t}_9) &= b(\mathbf{t}_9) + a(\mathbf{t}_9), & (\mathrm{E}b)(\mathbf{t}_9) &= \tfrac{1}{5} + 4b(\mathbf{t}_5) + b(\mathbf{t}_9),\\
(ba)(\mathbf{t}_{11}) &= a(\mathbf{t}_{11}), & (\mathrm{E}b)(\mathbf{t}_{11}) &= \tfrac{1}{15}.
\end{aligned}$$

These imply the usual order conditions for $\mathrm{t} = \mathbf{t}_1, \mathbf{t}_5, \mathbf{t}_{11}$, whereas the $\mathbf{t}_9$ equation can be satisfied by choosing $b(\mathbf{t}_5) = \frac{1}{4}\big(a(\mathbf{t}_9) - \frac{1}{5}\big)$. With this proviso, we can construct a method

$$\begin{array}{c|cccc}
0 & & & & \\
c_2 & c_2 & & & \\
c_3 & c_3 - a_{32} & a_{32} & & \\
1 & 1 - a_{42} - a_{43} & a_{42} & a_{43} & \\
\hline
 & b_1 & 0 & b_3 & b_4
\end{array},$$

satisfying order as well as the $C(2)$ and $D(1)$ conditions

$$\begin{aligned}
\textstyle\sum_j a_{ij}c_j &= \tfrac{1}{2}c_i^2, & i &= 3,4,\\
\textstyle\sum_i b_i a_{ij} &= b_j(1 - c_j), & j &= 1,2,3,4,\\
b_1 + b_3 + b_4 + b_5 &= 1, & & \\
b_3 c_3^{k-1} + b_4 c_4^{k-1} + b_5 &= \tfrac{1}{k}, & k &= 2,3,4,\\
\textstyle\sum b_i c_i a_{ij} c_j^2 &= \tfrac{1}{15}. &&
\end{aligned} \tag{5.7 c}$$

Exercise 53 Show that, for a solution of (5.7 c), $c_3 = \frac{2}{5}$.

An example solution of (5.7 c) is given by the tableau

$$\begin{array}{c|ccccc} 0 & & & & & \\ \frac{2}{5} & \frac{2}{5} & & & & \\ \frac{2}{5} & \frac{1}{5} & \frac{1}{5} & & & \\ \frac{1}{2} & \frac{3}{16} & 0 & \frac{5}{16} & & \\ 1 & \frac{1}{4} & -\frac{5}{4} & 0 & 2 & \\ \hline & \frac{1}{6} & 0 & 0 & \frac{2}{3} & \frac{1}{6} \end{array} . \tag{5.7 d}$$

Summary of Chapter 5 and the way forward

Summary

The early Runge–Kutta methods were built around the aim of obtaining successively higher orders for a generic scalar problem. However, in modern computing there is no interest in numerical methods which are applicable only to scalar problems and, above order 4, an analysis based on B-series is more appropriate. Even for order 5 there exist scalar methods with reduced order when applied to non-scalar problems.

Explicit methods to order 5 are derived. Order barriers are introduced through the simplest case (that order $p = s$ is impossible for explicit methods with $p > 4$). It is shown that this barrier can be circumvented through the use of effective order (conjugate order). Implicit methods, intended for the solution of stiff problems, are analysed and derived. Effective order is introduced and a new method of effective order 5 is derived.

The way forward

Not discussed in this chapter are Runge–Kutta methods with error estimators and methods of Nystöm type. In Chapter 6, Runge–Kutta methods are used in the construction of starting methods for general linear methods. In Chapter 7 symplectic methods are introduced for the solution of Hamiltonian problems.

Teaching and study notes

The following items are suggested reading associated with Runge–Kutta methods.
Butcher, J.C. *Coefficients for the study of Runge–Kutta integration processes* (1963) [7]

Butcher, J.C. *A stability property of implicit Runge–Kutta methods* (1975) [15]

Butcher, J.C. *The Numerical Analysis of Ordinary Differential Equations, Runge–Kutta and General Linear Methods* (1987) [17]

Cooper, G.J. and Verner, J.H. *Some explicit Runge-Kutta methods of high order* (1972) [33]

Gill, S. *A process for the step-by-step integration of differential equations in an automatic computing machine* (1951) [45]

Hairer, E. *A Runge–Kutta method of order* 10 (1978) [47]

Hammer, P.C. and Hollingsworth, J.W. *Trapezoidal methods of approximating solutions of differential equations* (1955) [54]

Merson, R.H. *An operational method for the study of integration processes* (1957) [72]

Nyström, E.J. *Über die numerische Integration von Differentialgleichungen* (1925) [77]

Prince, P.J. and Dormand, J.R. *High order embedded Runge–Kutta formulae* (1981) [78]

Verner, J.H. *Some Runge–Kutta formula pairs* (1991) [91]

Projects

Project 14 The stability function of a Runge–Kuta method with $s = 6$, $p = 5$ has the form $1+z+z^2/2+z^3/6+z^4/24+z^5/120+Cz^6$, where C depends on free parameters in the method tableau. Explore the dependence of the stability region on C.

Project 15 Show that a Runge–Kutta method with $s = 6$, and order 5 for scalar problems, also has the same order for high-dimensional problems if $D(1)$ holds.

Project 16 Explore the family of Runge–Kutta methods with $s = 7$, $p = 6$.

Project 17 Explore the family of Runge–Kutta methods with $s = 11$, $p = 8$.

Project 18 Study the theory of Order Stars and its application to the relationship between A-stability and order of Runge–Kutta methods.

Project 19 Study the theory of Order Arrows.

Project 20 Learn about Runge–Kutta pairs such as those in [78] and [91].

Project 21 Learn about the Radau family of numerical codes developed by Ernst Hairer.