myjournal manuscript No.
(will be inserted by the editor)

Strategies for deriving new explicit Runge–Kutta pairs*

J.H. Verner

Dedicated to Professor J. C.
Butcher in celebration of his sixtieth birthday

October 31, 2013

Abstract Different types of error-estimating pairs of explicit Runge–Kutta methods can be distinguished by the way the algebraic order-conditions are satisfied. Here, two new types of pairs are derived. The motivating theme develops algorithms for deriving new families of pairs of arbitrary orders that re-use the last stage. Additional restrictions on this design lead to subfamilies of conventionally propagated pairs which are more general than known families. Three examples are given. The new families contain some pairs now used in widely available general purpose software packages, and as well, new pairs which may be more effective.

Keywords Runge–Kutta pairs · Order conditions · local error estimation · stage-order

Mathematics Subject Classification (2000) 65L05 · 65L06 · 65L20

1 Introduction

To treat nonstiff vector differential equations of the form

$$y' = f(t, y), \qquad y(t_0) = y_0, \tag{1}$$

J.H. Verner
Department of Mathematics and Statistics, Queen's University at Kingston, K7L 3N6, Canada
URL: http://www.math.sfu.ca/∼jverner E-mail: jverner@pims.math.ca

The final publication of this manuscript appeared in Annals of Numerical Mathematics, Volume 1, pages 225-244. This Journal is no longer in print because J. C. Baltzer A.G. Science Publishers is now a defunct Netherlands publisher of scientific journals which was established in 1980 as J. C. Baltzer AG by Christoph and Marianne Baltzer. (See Wikipedia for more detail.)

pairs of explicit Runge–Kutta methods are often selected. From approximate derivative evaluations

$$f_{ni} = f(t_n + c_i h, y_n + h \sum_{j=1}^{i-1} a_{ij} f_{nj}), \qquad i = 1, \ldots, s. \tag{2}$$

two endpoint approximations

$$\begin{aligned} y_{n+1} &= y_n + h \sum_{i=1}^{s} b_i f_{ni}, \\ \hat{y}_{n+1} &= y_n + h \sum_{i=1}^{s} \hat{b}_i f_{ni}, \end{aligned} \qquad n = 0, 1, \ldots,$$

of orders p and $p-1$, respectively, are used to propagate an approximate solution and estimate the local error over a step of length h. The coefficients $\{b_i,\ \hat{b}_i,\ a_{ij}\}$ and nodes $\{c_i = \sum_{j=1}^{i-1} a_{ij}\}$ distinguish particular pairs of methods, and it is often convenient to denote each pair by a Butcher tableau where A is an $s \times s$ lower-triangular matrix, and each of $\mathbf{b},\ \hat{\mathbf{b}},\ \mathbf{c}$ are s-vectors. (In subsequent development, C will denote an $s \times s$ diagonal matrix with $\mathbf{c} = C\mathbf{e}$

Table 1 A Butcher Tableau.

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^t \\ & \hat{\mathbf{b}}^t \end{array} \tag{3}$$

where $\mathbf{e}$ is the s-vector with each element equal to 1.) The selection of these coefficients is in part a compromise between the reliability of the approximation and error estimate to be obtained, and the efficiency of the implemented algorithm. To obtain an approximation of a particular order of accuracy, an algorithm would be selected from one of several parametric families. In practice, the number of stages required for each step is often minimized to achieve a particular order, and some or all remaining arbitrary parameters are selected to optimize efficiency while maintaining adequate levels of stability and reliability.

For such a strategy to yield the best algorithms, it is desirable that all existing pairs be identified and preferably characterized parametrically. Although the minimum numbers of stages required for methods of orders at least $p \leq 8$ are known, the characterization of various *types* of methods of orders $p > 5$ remains incomplete. That is, there may exist different families or even different types of explicit Runge–Kutta methods which are yet undiscovered. In searching for new methods or pairs of methods, it is helpful to identify different *types* of methods by their general design, and within each type to characterize *families* for each order in terms of parameters which may be selected arbitrarily.

Algorithms for constructing pairs of orders $p-1$ and p requiring s stages are of interest, and Verner [10] has proposed a classification scheme for identifying potentially fruitful designs. Some known and new families are distinguished by their features in Table 2. For many pairs, only s *interior* stages are used to estimate both the propagated value and the error estimate. In contrast, in pairs of the FSAL type (for which the First stage of a new step is the Same As the Last stage of an immediately preceding successful step), the derivative evaluation of the propagated value is also utilized in the error estimator. In terms of their chronological development, pairs of both types are grouped as *traditional*, *recent* or *new.*

Table 2 Stages required for some high-order pairs.

Source of Derivation	Generic formula for number of stages	Stages for orders: 5:6	6:7	7:8	8:9
Traditional Pairs					
Fehlberg [4]	$(p^2 - 9p + 34)/2$	8	10	13	17
Verner [7]	$\leq (p^2 - 9p + 34)/2$	8	10	13	16
Prince & Dormand [5]	$(p^2 - 9p + 34)/2$	8		13	
Recent Pairs					
Sharp & Verner [6]	$\leq (p^2 - 7p + 24)/2$	9*	12*	15*	20*
Sharp & Verner [6]	$\leq (p^2 - 7p + 22)/2$	8	11	14	19
New Pairs					
New FSAL	$(p^2 - 9p + 36)/2$	9*			
New nonFSAL	$(p^2 - 9p + 34)/2$	8	10		

∗Indicates FSAL pairs (one stage may be reused after each successful step)

From Table 2, observe for $p > 6$ that each Sharp–Verner pair requires more stages than each traditional pair of the same orders. However, as Table 3 (below) reports, families of Sharp–Verner pairs have more arbitrary parameters in their parametric representation. Hence, the inefficiencies due to more stages may be offset if smaller coefficients in the local truncation error can be obtained by a careful choice among the larger set of arbitrary parameters.

The objective of this article is to derive pairs which preserve the smaller numbers of stages per step required by traditional pairs, but expand the set of arbitrary parameters available. To achieve this, we adapt the design of traditional pairs by removing the assumption that $c_{s-1} = 1$. In practice, to derive families of methods of this adapted design implies that new strategies for solving the order conditions are needed.

2 The order conditions and other notation

In the classification of [10], the interpretation of each stage of an explicit Runge–Kutta method as an internal approximation of some order $p_i \leq p$ allows for the distinction among different types of pairs and for the derivation of some particular types. This *stage-order* identifies the quadrature order of each stage together with a property which dissociates stages of lower order from those of higher order. Thus, for each $i = 1, \ldots, s$, stage i has stage-order p_i if

$$q_i^k \equiv \sum_{j=1}^{i-1} a_{ij} c_j^{k-1} - \frac{c_i^k}{k} = 0, \qquad k = 1, \ldots, p_i, \tag{4}$$

and

$$a_{ij} = 0, \qquad p_i > p_j + 1. \tag{5$'$}$$

This leads to the definition of the *stage-order vector* (SOV) of a method as $(p_1, \ldots, p_s)$, and the *dominant stage-order* (DSO) as the least value of p_i for which at least one of the corresponding weights b_i or $\hat{b}_i$ is nonzero. Generic values of the DSO for some families are given in Table 3. Furthermore, we form an *augmented* stage-order vector (ASOV) by appending to an SOV the stage-orders of stages used only for propagation of the solution or for error estimation. For example, a traditional eight-stage pair of orders 5 and 6 for which the DSO $= 2$ has an ASOV $= (6, 1, 2, 2, 2, 2, 2, 2 : 6, 5)$ while a new FSAL pair with the error estimate obtained using the first nine stages is characterized by the ASOV $= (6, 1, 2, 2, 2, 2, 2, 2, 6 : 5)$. For each positive k, the s-vector

$$\mathbf{q}^k \equiv AC^{k-1}\mathbf{e} - \frac{1}{k} C^k \mathbf{e}, \tag{5$''$}$$

with components given by the *subquadrature* expressions in (5), is used to specify the order conditions and to simplify proofs which follow.

Butcher [1] has shown that the set of conditions which must be satisfied for a method to be of order p may be identified by a one-to-one mapping with the set of rooted trees on no more than p nodes. Some knowledge of this theory is assumed, although an equivalent modified set of order conditions, described formally in [6] and briefly here, is also used in the development. The tree $t = \tau$ consisting of one node which is its root has height $H(t) = 1$. This tree together with trees $t = [\tau^{r-1}],\ r = 2, \ldots, p,$ of height $H(t) = 2$, each defined by attaching $r - 1$ nodes to a root, are identified with the quadrature order-conditions

$$\sum_{i=1}^{s} b_i c_i^{r-1} = \frac{1}{r}, \qquad r = 1, \ldots, p, \tag{5}$$

and for each stage $i,\ i = 1, \ldots, s$, with the *compound weights*,

$$\psi_i(\tau) = c_i, \qquad \psi_i([\tau^{r-1}]) = q_i^r, \quad r = 2, \ldots, p. \tag{6$'$}$$

Then, for a suitable ordering of the rooted trees, the remaining $N_p - p$ order conditions are formulated recursively. For increasing values of $r(t),\ 3 \le r(t) \le p$, each tree $t = [t_1, \ldots, t_k]$ with $r(t) = 1 + r(t_1) + \cdots + r(t_k)$ nodes and height $H(t) \ge 3$, and formed by attaching the k subtrees $t_1, \ldots, t_k$, to a root, corresponds to the subquadrature order-condition

$$\psi(t) \equiv \sum_{j=1}^{s} b_j \prod_{l=1}^{k} \psi_j(t_l) = 0, \tag{6}$$

and the compound weights,

$$\psi_i(t) = \sum_{j=1}^{s} a_{ij} \prod_{l=1}^{k} \psi_j(t_l), \qquad i = 1, \ldots, s. \tag{7$'$}$$

The remainder of the paper focuses on finding coefficients which satisfy (5)–(7) together with a second vector of weights $\widehat{\mathbf{b}}^t$ for which (6) and (7) are valid when $\widehat{\mathbf{b}}^t$ replaces $\mathbf{b}^t$ and $p-1$ replaces p. Four complementary devices for solving these order conditions are described in §3. Section 4 weaves an intricate path through the order conditions using these devices to identify parametric families of new methods of order p requiring $s = (p^2 - 9p + 36)/2$ stages. While the development motivates the choices made and indicates a variety of choices that remain to be studied, a reader wishing only to construct new methods may do so by implementing the equations of Lemma 4. Section 5 indicates how further constraints yield embedded pairs, and two types are given in Lemmas 6 and 7. So far, only some families for $p = 6$ and $p = 7$ have been constructed, and details for constructing three of these with particular examples are given in §6.

Table 3 tabulates the dominant stage orders and the numbers of arbitrary nodes and other arbitrary coefficients (separated by a slash) of both known and new families of pairs. Furthermore, one additional arbitrary parameter is available for each family since any nontrival convex combination of approximations of orders $p-1$ and p has order $p-1$.

Table 3 Stages (s) and Arbitrary Parameters (AP) required.

Orders $p-1:p$		5:6		6:7		7:8		8:9	
Type	DSO	s	AP	s	AP	s	AP	s	AP
Verner [7]	$p-4$	8	4	10	5	13	7/1	16	7/1
Prince & Dormand [5]	$p-4$	8	4/2			13	7/3		
Sharp & Verner [6]	$p-3$	9*	5	12*	7/1	15*	7/1	20*	8/3
Sharp & Verner [6]	$p-3$	8	4	11	6/1	14	6/1	19	7/3
New FSAL	$p-4$	9*	6/2						
New nonFSAL	$p-4$	8	6/1	10	7/1				

3 Devices for solving the order conditions

3.1 STAGE ORDER SELECTION

Both types of new pairs are derived from a single method of order p which uses $s = (p^2 - 9p + 34)/2$ stages. The first of four devices for deriving this basic method is the choice of an augmented stage-order vector: for $p = 6$, ASOV $= (6, 1, 2, 2, 2, 2, 2, 2 : 6)$, and for $p = 7$, ASOV $= (7, 1, 2, 3, 3, 3, 3, 3, 3, 3 : 7)$ are appropriate choices. In general, the appropriate ASOV is determined by selecting certain subsets of nodes to be distinct, and then sequentially constraining the coefficients for successive stages.

LEMMA 1

For each $p \geq 6$, there exist methods with $s = (p^2 - 9p + 34)/2$ stages, an ASOV $=$ $(p, 1, 2, 3, 3, \ldots,\ \ p-4, \ldots, p-4 : p)$ where these stage-orders are

$$
\begin{aligned}
&p_1 = p, &&(7)\\
&p_2 = 1,\\
&p_i = q, \quad i = \rho_q + 1, \ldots, \rho_q + q - 1, \quad \rho_q = (q^2 - 3q + 6)/2, \quad 2 \leq q \leq p - 5,\\
&p_i = p - 4, \quad i = \rho + 1, \ldots, \rho + p, \quad \rho \equiv \rho_{p-4} = (p^2 - 11p + 34)/2 \equiv s - p,
\end{aligned}
$$

and with weights which satisfy (6) and

$$b_i = 0, \qquad i = 2, \ldots, \rho. \tag{6''}$$

Proof

For each $q,\ 2 \leq q \leq p - 5$, $\rho_q + q - 1 = \rho_{q+1}$, so that (8) uniquely defines the entries of an augmented stage-order vector. The general proof is an analog of the proof of Lemma 1 in [6] which shows how to restrict the coefficients and weights to satisfy (5), (5′) (6), and (6″). For illustration, the result is proved for the case $p = 7$ and $s = 10$ only. For $c_1 = 0$ and $a_{1i} = 0,\ 1 \leq i \leq s,\ q_1^r = 0$ for all r, so that (5) is valid for *any* choice of p_1. For $a_{21} = c_2 \neq 0$ (to make stage 2 different from stage 1), $p_2 = 1$, and for $a_{32} = c_3^2/2c_2 \neq 0$, $p_3 = 2$, so that (5) is valid for stages 2 and 3. For stages 4 to 10, then (5′) as well as (5) must be satisfied, so that $a_{i2} = 0,\ i = 4, \ldots, 10$, is imposed. Now if $c_1 = 0,\ c_3,\ c_4$ are distinct, for each stage $i \geq 5$, $\{a_{ij},\ j = 1, 3, \ldots, i-1\}$ may be chosen with $i - 5$ arbitrary choices to satisfy (5). Otherwise, for stage $i = 4$, we need to choose $a_{41},\ a_{43}$ to satisfy three conditions of (5). This is possible if and only if

$$\int_0^{c_4} c(c - c_3)dc = 0, \tag{8}$$

or $c_3 = 2c_4/3$. Finally, by choosing $b_2 = b_3 = 0$ to satisfy (6″), and b_4 arbitrary, the remaining seven weights $\{b_i,\ i = 1, 5, \ldots, 10\}$ can be chosen to satisfy (6) if the seven corresponding nodes are distinct. □

Observe that $\rho = s - p$ so that (6″) implies $b_i = 0$ if $p_i \leq p - 5$. Hence, (6″) will be interpreted as a modified analog of (5′) when the approximation of order p is specified as stage $s + 1$. The arbitrary choices identified in the proof will be exploited to solve remaining order conditions, or else will remain arbitrary to characterize a parametric family.

3.2 LEFT HOMOGENEOUS POLYNOMIALS

For the second device, define $\boldsymbol{\Lambda}_r$, the set of *left homogeneous polynomials* of degree r, recursively by

$$
\begin{aligned}
\boldsymbol{\Lambda}_1 &= \{\mathbf{b}^t\},\\
\boldsymbol{\Lambda}_r &= \boldsymbol{\Lambda}_{r-1}A \cup \boldsymbol{\Lambda}_{r-1}C, \qquad r = 2, \ldots,
\end{aligned} \tag{9}
$$

where $\boldsymbol{\Lambda}_{r-1}A$ and $\boldsymbol{\Lambda}_{r-1}C$ denote sets obtained by post-multiplying elements of $\boldsymbol{\Lambda}_{r-1}$ by A and C, respectively. For example, $\boldsymbol{\Lambda}_3 = \{\mathbf{b}^tA^2,\ \mathbf{b}^tAC,\ \mathbf{b}^tCA,\ \mathbf{b}^tC^2\}$. Now denote the first vector of each set by $\mathbf{B}_r \equiv \mathbf{b}^tA^{r-1}$, so for example $B_{3k} = \sum_{i>j=k+1}^{s} b_i a_{ij} a_{jk},\ k = 1,\ldots,s-2$. Then for each positive r, the traditional order condition (arising from linear differential equations) corresponding to the tree $_r[\tau^q]_r$ requires that

$$\begin{aligned}\mathbf{b}^tA^{r-1}C^{q-1}\mathbf{e} \equiv \sum_{i=1}^{s} B_{ri}c_i^{q-1} &= \frac{1}{q(q+1)\cdots(q+r-1)} \qquad (10)\\ &\equiv \int_0^1\int_0^{x_1}\cdots\int_0^{x_{r-1}} x_r^{q-1}dx_r\cdots dx_2dx_1, \quad q = 1,\ldots,p-r+1.\end{aligned}$$

Now select values for $\mathbf{B}_r$ in an analogous way to the selection of weights in Lemma 1. Stages of lower order are suppressed by requiring that

$$B_{ri} = 0, \qquad i = 2,\ldots,\rho, \quad \rho \equiv s-p. \qquad (11)$$

Next, choose one more value arbitrarily, selected without loss of generality as $B_{r,\rho+1}$. Then with nodes $\{c_i,\ i = 1,\rho+2,\ldots,s\}$ distinct, for $r = 1,2,\ldots,$ define polynomials

$$\pi^{s-r+1}(c) = (c-c_1)(c-c_{\rho+2})\cdots(c-c_{s-r+1}) \qquad (12)$$

of degree $p-r+1$, and

$$\pi_j^{s-r+1}(c) = \frac{\pi^{s-r+1}(c)}{(c-c_j)}, \qquad j = 1,\rho+2,\ldots,s-r+1, \qquad (13')$$

of degree $p-r$. Then Lagrange interpolation implies that values chosen by (12) and

$$B_{ri} = \left\{B_{r,\rho+1}\pi_i^{s-r+1}(c_{\rho+1}) - \int_0^1\cdots\int_0^{x_{r-1}}\pi_i^{s-r+1}(x_r)dx_r\cdots dx_1\right\}\Big/\pi_i^{s-r+1}(c_i),$$
$$i = 1,\rho+2,\ldots,s-r+1, \qquad (12')$$

satisfy (11). For $r = 1$, (12) and (12′) are equivalent to the choice of weights $\{b_i,\ i = 1,\ldots,s\}$ in Lemma 1.

Although alternative arbitrary choices are possible, (12) serves to satisfy other order conditions as well. The derivation of traditional pairs in [7] uses (12) and an analog of (12′) explicitly for $r = 1,2,3,4$. This article imposes these constraints explicitly *only* for $r = 1,2,3$. Thus, from coefficients $\{a_{ij},\ i \le s-2\}$ computed to satisfy stage-order conditions (5) and (5′) and other constraints yet to be determined, and nodes selected so that $B_{2,s-1} \equiv b_sa_{s,s-1}$ evaluated by (12′) is nonzero, coefficients for the final two stages will be obtained by the *back substitution*

$$a_{s-r+2,j} = \frac{B_{rj} - \sum_{i=j+1}^{s-r+1} B_{r-1,i}a_{ij}}{B_{r-1,s-r+2}}, \qquad r = 2,3, \quad j = s-r+1,\ldots,1. \quad (13)$$

3.3 BUTCHER'S ROW SIMPLIFYING ASSUMPTIONS

The third device required is the set of Butcher's simplifying conditions,

$$B_{2i} = b_i(1 - c_i), \qquad i = 1, \ldots, s. \tag{14}$$

By (12), this is valid trivially for $i = 2, \ldots, \rho$, and can be established for all other values of i if the coefficients are restricted to satisfy only two conditions of (15) explicitly. Specifically, we require $c_s = 1$ (which implies (15) for $i = s$), and $B_{2,\rho+1} \equiv \sum_{i=\rho+2}^{s} b_i a_{i,\rho+1} = b_{\rho+1}(1 - c_{\rho+1})$. With (12′) for $r = 1$, these with (13′) give for each $i = 1, \rho + 2, \ldots, s - 1$,

$$b_i(1 - c_i) = \left\{ b_{\rho+1}\pi_i^s(c_{\rho+1}) - \int_0^1 \pi_i^s(x)dx \right\} (1 - c_i) \Big/ \pi_i^s(c_i) \equiv B_{2i}, \tag{15′}$$

where the final equality is obtained by reversing the order of integration in (12′) for $r = 2$.

Observe to use (14) that $b_{s-1}(1 - c_{s-1}) \neq 0$ which constrains the arbitrary choices slightly (and precludes the traditional methods in [7] at least formally). Subject to $c_s = 1$ and this requirement, $b_{\rho+1}$ and $B_{3,\rho+1}$ remain arbitrary while $B_{2,\rho+1}$ is now determined by (15).

3.4 RIGHT HOMOGENEOUS POLYNOMIALS

The final device is motivated by observing from (7) that both $\mathbf{b}^t$ and $\mathbf{b}^t C$ must be orthogonal to certain vectors of $\mathbb{R}^s$. Interpret (7) as a requirement that $\mathbf{b}^t$ be orthogonal to columns of the $(N_p - p) \times s$ matrix Ψ_p'. Each column, a vector of compound weights, is a polynomial of degree $r - 1$ in A and C, which is uniquely determined by $t = [t_1, \ldots, t_k]$, a tree of height H(t)≥ 3 and $r \leq p$ nodes, post-multiplied by $\mathbf{e}$. Next, partition these columns of Ψ_p' into subsets $\boldsymbol{\Theta}_r,\ 2 \leq r < p$, of *right homogeneous polynomials* of degree r. For example,

$$\begin{aligned} \boldsymbol{\Theta}_2 &= \{\mathbf{q}^2\}, \\ \boldsymbol{\Theta}_3 &= \{\mathbf{q}^3,\ A\mathbf{q}^2,\ C\mathbf{q}^2\}, \\ \boldsymbol{\Theta}_4 &= \{\mathbf{q}^4\} \cup A\boldsymbol{\Theta}_3 \cup C\boldsymbol{\Theta}_3 \cup \{(\mathbf{q}^2)^2\}. \end{aligned} \tag{15}$$

where $A\boldsymbol{\Theta}_r$ and $C\boldsymbol{\Theta}_r$ designate the sets obtained on pre-multiplication of each element of $\boldsymbol{\Theta}_r$ by A and C, respectively. Furthermore, a recursive algorithm is evident from (16). In particular,

$$\boldsymbol{\Theta}_r = \{\mathbf{q}^r\} \cup A\boldsymbol{\Theta}_{r-1} \cup C\boldsymbol{\Theta}_{r-1} \cup \tilde{\boldsymbol{\Theta}}_r, \qquad 2 < r < p, \tag{16′}$$

where $\tilde{\boldsymbol{\Theta}}_r$ is the set of all componentwise products of two vectors, one taken from each of $\boldsymbol{\Theta}_{\bar{r}}$ and $\boldsymbol{\Theta}_{r-\bar{r}}$ for each $\bar{r}$ with $2 \leq \bar{r} \leq r/2$.

4 Derivation of the propagating method

Conditions (6) imply that $\mathbf{b}^t$ and $\mathbf{b}^tC$ are orthogonal to other vectors as well.

LEMMA 2

(a) Conditions (6) are valid if and only if

$$\mathbf{b}^t\mathbf{e} = 2\mathbf{b}^tC\mathbf{e} = 1, \tag{16}$$

and both $\mathbf{b}^t$ and $\mathbf{b}^tC$ are orthogonal to each vector

$$\boldsymbol{\omega}^r \equiv \frac{d^2}{dC^2}\left\{C^r(I-C)^2\right\}\mathbf{e}, \qquad r = 2,\ldots,p-2. \tag{17}$$

(b) Suppose that the weights and coefficients of an s-stage method satisfy (11) for $r = 1,2,3$, and (15). Then $\mathbf{b}^t$ and $\mathbf{b}^tC$ are orthogonal to each of $\mathbf{q}^r$, $r = 2,\ldots,p-2$, and $\mathbf{b}^t$ is orthogonal to $\mathbf{q}^{p-1}$.

(c) Suppose that coefficients of stages 1 to $s-2$ are chosen to satisfy (5) and (5′) for the ASOV of Lemma 1, that $c_s = 1$ and $B_{2,\rho+1}$ satisfies (15), and that the weights and coefficients of stages $s-1$ and s satisfy (12)–(14). Then (5) and (5′) are valid for $p_{s-1} = p_s = p-4$.

Proof

(a) If (6) holds, the orthogonality of $\mathbf{b}^t$ and $\mathbf{b}^tC$ to (18) can be verified by substitution. Otherwise (17), the orthogonality of $\mathbf{b}^t$ to each of (18), and of $\mathbf{b}^tC$ to $\boldsymbol{\omega}^{p-2}$, forms a system of p linearly independent conditions which imply (6).

(b) Using definition (5″), (11) for $r = 1,2$, imply that

$$\sum_{i=1}^{s} b_i q_i^k \equiv \mathbf{b}^t\mathbf{q}^k = \mathbf{b}^t(AC^{k-1}\mathbf{e} - \frac{1}{k}C^k\mathbf{e}) = 0, \qquad k = 1,\ldots,p-1, \tag{18}$$

and for $r = 2,3$, that

$$\sum_{i=1}^{s} B_{2i} q_i^k \equiv \mathbf{b}^tA\mathbf{q}^k = \mathbf{b}^tA(AC^{k-1}\mathbf{e} - \frac{1}{k}C^k\mathbf{e}) = 0, \qquad k = 1,\ldots,p-2. \tag{19}$$

Since (15) holds for all stages, $\mathbf{b}^tC = \mathbf{b}^t - \mathbf{b}^tA$, and then (19) and (20) imply that

$$\mathbf{b}^tC\mathbf{q}^k = \mathbf{b}^t\mathbf{q}^k - \mathbf{b}^tA\mathbf{q}^k = 0, \qquad k = 1,\ldots,p-2, \tag{20}$$

yielding the stated result.

(c) First, (12)–(14) imply that (11) holds for $r = 1,2,3$. The hypotheses imply (15) holds for $i = s$ (trivially), and for $i = \rho+1 = s-p+1$, so that with (12), (15′) implies (15) is valid for all i. That is, the hypotheses in (b) are valid, and so the conclusions of (b) hold. Since $a_{s-1,i}$ is computed by (14),

$B_{2,s-1} = b_s a_{s,s-1} \neq 0$. Now if $p_i < p-4$, (12) implies $B_{2i} = 0$, and if $p_i = p-4$, the ASOV implies that $q_i^k = 0$ for each $k \leq p-4$ and $i \leq s-2$, so that

$$B_{2,s-1} q_{s-1}^k \equiv \sum_{i=1}^{s-1} B_{2i} q_i^k = 0, \qquad k = 1, \ldots, p-4, \tag{20$'$}$$

by (20). Hence, $q_{s-1}^k = 0$ establishing (5) for $i = s-1$. Also, if $p_i < p-5$, $B_{2,s-1} a_{s-1,i} \equiv B_{3i} = 0$, so that $a_{s-1,i} = 0$ establishing (5$'$) for $i = s-1$, and so $p_{s-1} = p-4$. Similarly, (19) implies that $b_s q_s^k = 0$ for $k = 1, \ldots, p-4$, and so $b_s a_{si} = 0$ when $p_i < p-5$, implying that $p_s = p-4$ since $b_s \neq 0$. □

To understand why $\boldsymbol{\omega}_r$ is represented in (18) as a formal derivative, observe for $r \geq 2$ that the second derivative $\{c^r(1-c)^2\}''$ is orthogonal to both $p(c) = 1$ and $p(c) = c$ under the inner product defined by integration on [0,1].

THEOREM 1

Suppose $\{\mathbf{b},\ A,\ \mathbf{c}\}$ for an s-stage method satisfy (11) for $r = 1, 2, 3$, and (15). Then the method is of order p if in addition both $\mathbf{b}^t$ and $\mathbf{b}^t C$ are orthogonal to $\{\boldsymbol{\Theta}_r,\ r = 2, \ldots, p-2\}$, and $\mathbf{b}^t$ is orthogonal to $\tilde{\boldsymbol{\Theta}}_{p-1}$.

Proof

Conditions (11) for $r = 1$ imply that the quadrature conditions (6) are satisfied. It suffices to establish that $\mathbf{b}^t$ is orthogonal to Ψ_p', or equivalently to $\{\boldsymbol{\Theta}_r,\ r = 2, \ldots, p-1\}$. By the hypotheses, only the orthogonality to $\boldsymbol{\Theta}_{p-1}$ is in doubt. Lemma 2(b) establishes that $\mathbf{b}^t \mathbf{q}^{p-1} = 0$. By (15), $\mathbf{b}^t A = \mathbf{b}^t (I - C) \equiv \mathbf{b}^t - \mathbf{b}^t C$, and then it follows from the hypotheses that $\mathbf{b}^t A$ also is orthogonal to $\boldsymbol{\Theta}_{p-2}$. Together with the hypotheses, these imply that $\mathbf{b}^t$ is orthogonal to the set $\boldsymbol{\Theta}_{p-1} = \{\mathbf{q}^{p-1}\} \cup A\boldsymbol{\Theta}_{p-2} \cup C\boldsymbol{\Theta}_{p-2} \cup \tilde{\boldsymbol{\Theta}}_{p-1}$, and so the method is of order p. □

On one hand, Lemma 2(c) gives equations for computing coefficients of a method, while on the other, Theorem 1 gives conditions additional to those of Lemma 2(b) which are sufficient for the method to have order p. Next, we condense this latter set of sufficient conditions towards those of Lemma 2(c). All conditions other than (17) are equivalent to ensuring that certain vectors are in the nullspace of $\{\mathbf{b}^t,\ \mathbf{b}^t C\}$, and so we focus on minimizing its dimension. This is achieved by refining the identification of order conditions with vectors in $\boldsymbol{\Lambda}_r$ and $\boldsymbol{\Theta}_d$ for $1 \leq r + d \leq p$.

LEMMA 3

Suppose that $\{\mathbf{b}^t,\ A,\ \mathbf{c}\}$ satisfy the conditions of Lemma 2(c).

(a) If, in addition, the coefficients satisfy

$$(\mathbf{b}^t A^3)_j = (\mathbf{b}^t C^2 A)_j = 0, \qquad p_j = p-5, \tag{21}$$

then for every $\tilde{\mathbf{B}}_r \in \boldsymbol{\Lambda}_r$, $\tilde{B}_{rj} = 0$ whenever $p_j < p - max(4, r)$.

(b) For every $\mathbf{R}_d \in \boldsymbol{\Theta}_d$, $R_{jd} = 0$ whenever $d \le p_j \le p - 4$.

Proof

By Lemma 2(c), (5) and (5′) hold for the ASOV of Lemma 1.

(a) Observe $(\mathbf{b}^t A^3)_j = \sum_{i=j+1}^{s-2} (\mathbf{b}^t A^2)_i a_{ij}$ and $(\mathbf{b}^t C^2 A)_j = \sum_{i=j+1}^{s} (\mathbf{b}^t C^2)_i a_{ij}$. For $p_j < p - 5$, either $p_i \le p_j + 1 < p - 4$ and $(\mathbf{b}^t A^2)_i = (\mathbf{b}^t C^2)_i = 0$ by (12), or else $p_i > p_j + 1$ and $a_{ij} = 0$ by (5′). Hence, (22) is valid for all $p_j \le p - 5$.

Now, we establish the result for $r \le 4$. For $B_{rj} = (\mathbf{b}^t A^{r-1})_j$, $p_j < p-4$, the result follows by (12) for $r = 1, 2, 3$, and by the first expression of (22) for $r = 4$. Then, (15) implies that $(\mathbf{b}^t CA)_j = B_{2,j} - B_{3,j} = 0$, and that $(\mathbf{b}^t CA^2)_j = B_{3,j} - B_{4,j} = 0$. Next, (15) implies $(\mathbf{b}^t ACA)_j = (\mathbf{b}^t CA)_j - (\mathbf{b}^t C^2 A)_j = 0$ by the previous case and the second expression of (22). Since the result holds for $\boldsymbol{\Lambda}_1 = \{\mathbf{b}^t\}$, and each remaining vector of $\boldsymbol{\Lambda}_r$, $r \le 4$, is obtained by post-multiplication of one from $\boldsymbol{\Lambda}_{r-1}$ by C, the result holds by finite induction on $r \le 4$.

The result is established for $r > 4$ by induction. For each $r > 4$ and $\tilde{\mathbf{B}}_r \in \boldsymbol{\Lambda}_r$, $\tilde{B}_{r,j}$ is either $\sum_{i=j+1}^{s} \tilde{B}_{r-1,i} a_{ij}$ or $\tilde{B}_{r-1,j} c_j$. Using the stage suppression conditions (5′) with the result for $\boldsymbol{\Lambda}_{r-1}$ shows that the first expression is zero when $p_j \le min(p_i - 1, p_i) < min\,(p - max(4, r-1) - 1, p - max(4, r-1)) = p - max(4, r)$, giving the stated result. The argument for the second expression is similar but easier.

(b) By (5), the result is valid for $\mathbf{R}_d = \mathbf{q}^d$, $d \le p - 4$, and hence trivially for all $\mathbf{R}_d \in \boldsymbol{\Theta}_2$. The result is proved by induction on $d \ge 2$, so for each $2 \le c < d \le p - 4$, we assume the result is valid for $\mathbf{R}_c \in \boldsymbol{\Theta}_c$. For $d \le p - 4$, $\mathbf{R}_d$ is one of $\mathbf{q}^d$, $A\mathbf{R}_{d-1}$, $C\mathbf{R}_{d-1}$, or $\mathbf{R}_b \cdot \mathbf{R}_c$ where $b + c = d$ for $b, c \ge 2$. The result has been established for the first case. In the second case, consider $(A\mathbf{R}_{d-1})_j = \sum_{k=1}^{j-1} a_{jk} R_{k,d-1}$ for $d \le p_j \le p - 4$. If $p_k < p_j - 1$ then $a_{jk} = 0$. Otherwise, if $p_j - 1 \le p_k \le p_j$ then $d - 1 \le p_k \le p - 4$, and in this case, the inductive assumption implies $R_{k,d-1} = 0$. Hence, each term in the summation is zero, and the result holds. For each of the two remaining alternatives, the result holds directly because it holds for each $\mathbf{R}_c$, $c < d$. This completes the inductive step, so the result holds for all $d \le p - 4$. □

THEOREM 2

Suppose the coefficients are selected to satisfy the conditions of Lemma 2(c), and (22), and that $\mathbf{b}^t$ and $\mathbf{b}^t C$ are orthogonal to each of $A\mathbf{q}^{p-3}$ and $C\mathbf{q}^{p-3}$. Then the method has order p.

Proof

The conditions of Lemma 2(c) imply that (11) is valid for $r = 1, 2, 3$, and in particular the quadrature conditions (6) hold. Each remaining condition (7) is interpreted as requiring that $\tilde{\mathbf{B}}_r \in \boldsymbol{\Lambda}_r$, $r \ge 1$ is orthogonal to some $\mathbf{R}_d \in \boldsymbol{\Theta}_d$, $r + d \le p$.

First assume that $d \le p - 4$. If $p_j < p - max(r, 4)$, Lemma 3(a) establishes that $\tilde{B}_{rj} = 0$. Otherwise, if $p_j \ge p - max(r, 4) = min(p - r, p - 4) \ge min(p -$

$r, d) = d$ (since $d \leq p - r$), then by Lemma 3(b), $R_{jd} = 0$. Hence, $\tilde{\mathbf{B}}_r \cdot \mathbf{R}_d = \sum_{j-1}^{s} \tilde{B}_{rj} R_{jd} = 0$, and all order conditions with $d \leq p - 4$ hold.

Now we proceed by induction on decreasing r. Observe for each $r \geq 4$, that $r + d \leq p$ implies that $d \leq p - 4$, and for these values of r and d, the first part establishes that $\tilde{\mathbf{B}}_r \cdot \mathbf{R}_d = 0$. We use this to establish the same result for each of $r = 3, 2, 1$ inductively whenever $p - r \geq d > p - 4$. Recall that $\mathbf{R}_d$ takes one of four forms implied by (16′). If $\mathbf{R}_d \in A\boldsymbol{\Theta}_{d-1} \cup C\boldsymbol{\Theta}_{d-1}$, the inductive assumption for $d - 1$ implies that $\tilde{\mathbf{B}}_r \cdot \mathbf{R}_d \equiv \tilde{\mathbf{B}}_{r+1} \cdot \mathbf{R}_{d-1} = 0$. Next, for $\mathbf{R}_d \in \tilde{\boldsymbol{\Theta}}$, $\mathbf{R}_d = \mathbf{R}_b \cdot \mathbf{R}_c$ with $b + c = d$. In this case, $min(b, c) \leq [(b + c)/2] = [d/2] \leq [(p - r)/2]$, where $[x]$ is the integer part of x. Since $p \geq 6$, it may be verified by computation for each $r \geq 1$ that $[(p - r)/2] \leq p - 4$ with equality holding only if $p = 6$ or $p = 7$. Hence, by Lemma 3(a) if $p_j = p - 4$, $R_{jb} R_{jc} = 0$. Also for $r < 4$, $\tilde{B}_{rj} = 0$ for $p_j < p - 4$ by Lemma 3(b), and these imply that $\tilde{\mathbf{B}}_r \cdot \mathbf{R}_d \equiv \sum_{j=1}^{s} \tilde{B}_{rj} R_{jb} R_{jc} = 0$. It remains to show that each vector of $\boldsymbol{\Lambda}_r$ is orthogonal to $\mathbf{q}^d,\ p - r \geq d > p - 4$.

Observe that the results of Lemma 2(b) hold, so that $\mathbf{q}^d$ is orthogonal to $\mathbf{b}^t$ for $d = p - 3, p - 2, p - 1$, and to $\mathbf{b}^t C$ and then by (15) to $\mathbf{b}^t A$ for $d = p - 3, p - 2$. Now continuing the induction, for $r = 3$, the hypotheses imply that each of $\mathbf{b}^t CA$ and $\mathbf{b}^t C^2$ are orthogonal to $\mathbf{q}^{p-3}$. Also, with the results that $\mathbf{b}^t C \mathbf{q}^{p-3} = \mathbf{b}^t A \mathbf{q}^{p-3} = 0$, (15) implies that $\mathbf{b}^t A^2 = \mathbf{b}^t A - \mathbf{b}^t CA$ and $\mathbf{b}^t AC = \mathbf{b}^t C - \mathbf{b}^t C^2$, so that these vectors also are orthogonal to q^{p-3}, and this completes the case for $r = 3$. Now, the induction for $r = 2$ and then for $r = 1$ can be completed without further difficulty. □

The result shows that only a few constraints must be imposed in addition to the design identified in Lemma 2(c) in order to obtain a method of order p. Of several alternatives, we attempt to make each of $A\mathbf{q}^{p-3}$ and $C\mathbf{q}^{p-3}$ linear combinations of vectors in $\boldsymbol{\Omega}_{p-2} \cup \mathbf{Q}_{p-2}$ where $\boldsymbol{\Omega}_\rho = \cup_{r=2}^{\rho} \{\boldsymbol{\omega}^r\}$ and $\mathbf{Q}_\rho = \cup_{r=2}^{\rho} \{\mathbf{q}^r\}$ since Lemma 2(a) and (b) imply each subset is orthogonal to both $\mathbf{b}^t$ and $\mathbf{b}^t C$. The proof is facilitated by representing vectors spanned by $\boldsymbol{\Omega}_{p-2}$ as linear combinations of $\boldsymbol{\omega}^2$ and

$$\bar{\boldsymbol{\omega}}^r = \frac{d^3}{dC^3} \left\{ C^r (I - C)^3 \right\} \mathbf{e} \equiv r\boldsymbol{\omega}^{r-1} - (r + 3)\boldsymbol{\omega}^r, \quad r = 3, 4, \ldots, p - 2. \tag{18′}$$

If the weights satisfy (6), the argument used to prove Lemma 3(a) can be extended to show that each of $\mathbf{b}^t,\ \mathbf{b}^t C,\ \mathbf{b}^t C^2$ is orthogonal to $\bar{\boldsymbol{\omega}}^r$ if $r \leq p - 3$.

LEMMA 4

Suppose that coefficients of stages 1 to $s - 2$ are chosen to satisfy (5) and (5′) for the ASOV of Lemma 1, and in addition so that

$$A\mathbf{q}^{p-4} = \sum_{r=3}^{p-3} J_r \bar{\boldsymbol{\omega}}^r + \sum_{r=2}^{p-3} \tilde{J}_r \mathbf{q}^r, \tag{22}$$

$$A\mathbf{q}^{p-3} = K_2 \boldsymbol{\omega}^2 + \sum_{r=3}^{p-2} K_r \bar{\boldsymbol{\omega}}^r + \sum_{r=2}^{p-2} \tilde{K}_r \mathbf{q}^r, \tag{23}$$

and

$$C\mathbf{q}^{p-3} = L_2\boldsymbol{\omega}^2 + \sum_{r=3}^{p-2} L_r\bar{\boldsymbol{\omega}}^r + \sum_{r=2}^{p-2} \tilde{L}_r\mathbf{q}^r \tag{24}$$

are valid for components $i = 1, \ldots, s-2$, with $\tilde{J}_{p-3} \neq 0$. Also suppose that the homogeneous polynomials $\mathbf{B}_r,\ r = 1, 2, 3$, can be chosen to satisfy (12), (13), (15) and (22), and that the weights and coefficients of stages $s-1$ and s are computed by (14). Then, (23)–(25) also hold for $i = s-1$ and s.

Proof

The proof, which establishes and uses the fact that $\mathbf{b}^t$, $\mathbf{b}^t C$, and $\mathbf{b}^t A$ are each orthogonal to each of $A\bar{\boldsymbol{\omega}}^r$ and $C\bar{\boldsymbol{\omega}}^r$ for $3 \leq r \leq p-3$, is omitted. □

COROLLARY

Suppose that coefficients of stages 1 to $s-2$ are chosen to satisfy (5) and (5′) for the ASOV of Lemma 1 and (23)–(25). Furthermore, let $c_s = 1$ and $B_{2,\rho+1}$ satisfy (15), and suppose that the weights and coefficients of stages $s-1$ and s satisfy (12)–(14) and (22). Then the method has order p.

Proof

Lemmas 2(c) and 4 establish that the hypotheses of Theorem 2 are valid. □

This completes the design for the method of order p. It will be seen later that the conditions of Lemma 4 are not easy to satisfy, so that not all of the target families have been constructed. Possibly other alternatives may be more fruitful.

5 Embedded and imbedded error estimators

A strategy for obtaining a Runge–Kutta *pair* is based on the development in §4. For each $p \geq 6$, we apply Theorem 2 to obtain a method of order p by selecting the arbitrary parameters remaining from Lemma 1 together with $b_{\rho+1}$ and $B_{3,\rho+1}$ to satisfy (22)–(25). In this section, we consider the possibility of adding to this a related, but different method of order $p-1$. This method uses the s stages of the first method together with the additional stage defined with $c_{s+1} = 1$ and $a_{s+1,i} = b_i,\ i = 1, \ldots, s$. For this development, it is convenient to append the coefficients defining this stage to the matrices A and C, and the vector $\mathbf{c}$, and to append 0 to the vector $\mathbf{b}^t$ to define new matrices and vectors of size $s+1$. Furthermore, this extension naturally induces an expansion of the vectors in $\boldsymbol{\Lambda}_r$, $\boldsymbol{\Theta}_d$ and $\boldsymbol{\Omega}_r$, which are henceforth interpreted to be subsets of $\mathbb{R}^{s+1}$. When required, restriction to only the first s components will be denoted by a prime, so for example, $\boldsymbol{\Omega}_r'$ is a subset of $\mathbb{R}^s$. To satisfy conditions of order $p-1$ for the second method, we have the $s+1$ weights of the vector

$\widehat{\mathbf{b}}^t$ and perhaps some arbitrary coefficients remaining from Lemma 1 available. To motivate their choice, we examine the order conditions that have to be satisfied. For this, we denote $\mathcal{N}_r = span(\boldsymbol{\Omega}_r \cup \boldsymbol{\Theta}_2 \cup \cdots \cup \boldsymbol{\Theta}_r)$, so that $\mathbf{b}^t$ is a vector which is orthogonal to $(I-2C)\mathbf{e}$ and $\mathcal{N}_{p-1}$, and for which $\mathbf{b}^t\mathbf{e} = 1$. In some respects, the proofs which follow resemble those in [8].

LEMMA 5

(a) If $(\mathbf{b}^t,\ A,\ \mathbf{c})$ represents a method of order p, and $\widehat{\mathbf{b}}^t - \mathbf{b}^t$ is a nonzero vector that is orthogonal to each vector of $\left(\cup_{r=0}^{p-2}\{C^r\mathbf{e}\}\right) \cup \left(\cup_{r=2}^{p-2}\boldsymbol{\Theta}_r,\right)$, then $(\widehat{\mathbf{b}}^t,\ A,\ \mathbf{c})$ represents a different method of order $p-1$.

(b) Suppose for a method of order p that $span(\{(I-2C)\mathbf{e}\} \cup \mathcal{N}_{p-2})$ is a proper subspace of the $span(\{(I-2C)\mathbf{e}\} \cup \mathcal{N}_{p-1})$. Then a different method of order $p-1$ which uses the $s+1$ stages can be constructed.

Proof

(a) Since the method with weights $\mathbf{b}^t$ is of order p, the hypotheses imply that $\widehat{\mathbf{b}}^t$ satisfies each of the quadrature conditions (6) for $r = 1, \ldots, p-1$, with $\widehat{\mathbf{b}}^t$ replacing $\mathbf{b}^t$, and that $\widehat{\mathbf{b}}^t$ is orthogonal to each vector of $\cup_{r=2}^{p-2}\boldsymbol{\Theta}_r$. These are precisely the conditions that the method with weights $\widehat{\mathbf{b}}^t$ be of order $p-1$. Since $\widehat{\mathbf{b}}^t - \mathbf{b}^t \neq 0$, the new method is different from that with weights $\mathbf{b}^t$.

(b) Since $\mathbf{b}^t \in \mathbb{R}^{s+1}$ and $\mathbf{b}^t\mathbf{e} = 1$, the nullspace of $\mathbf{b}^t$ satisfies

$$dim(span\,(\{(I-2C)\mathbf{e}\} \cup \mathcal{N}_{p-1})) \leq s.$$

Therefore, $\mathbf{b}^t\mathbf{e} \neq 0$ and the assumption of proper inclusion implies that

$$span\,(\{\mathbf{e}\} \cup \{(I-2C)\mathbf{e}\} \cup \mathcal{N}_{p-2})$$

is a proper subspace of $\mathbb{R}^{s+1}$. Hence, there exists a nonzero (s+1)-vector orthogonal to this subspace, and we denote it by $\widehat{\mathbf{b}}^t - \mathbf{b}^t$. Furthermore, this subspace contains each of the vectors specified by part (a), and so $\widehat{\mathbf{b}}^t$ determines the required weights of a different method of order $p-1$. □

If a Runge–Kutta pair exists, then the conditions of Lemma 5(a) hold, and these imply that each of $\mathbf{b}^t,\ \mathbf{b}^tC,\ \widehat{\mathbf{b}}^t$ is orthogonal to $\mathcal{N}_{p-2}$. For the two methods to be different, these three vectors must be linearly independent, and hence, $dim(\mathcal{N}_{p-2}) \leq$ number of stages-3 (i.e. $s-3$ for conventional pairs and $s-2$ for FSAL pairs). Lemma 5(b) indicates what difficulties must be overcome to obtain an embedded method of order $p-1$. For a FSAL pair, it becomes necessary to make components for stage $s+1$ in *each* vector of $\{(I-2C)\mathbf{e}\} \cup \mathcal{N}_{p-2}$ consistent with choosing the vector $\widehat{\mathbf{b}}^t$ to have $\hat{b}_{s+1} \neq 0$. Alternatively, a pair which uses only the original s stages could be obtained if the subspace of $\mathbb{R}^s$ spanned by $\{(I-2C)\mathbf{e}'\} \cup \mathcal{N}'_{p-1}$ has dimension $s-2$. It is not clear that either alternative can be achieved for all values of p of interest, although some direction is possible.

LEMMA 6 (FSAL pairs)

Suppose for a method determined by the Corollary to Lemma 4, that the vector space spanned by $\{(I-2C)\mathbf{e}'\} \cup \mathcal{N}'_{p-2}$ is a nullspace of $\mathbf{b}'$ of dimension $s-1$, and that conditions (23)–(25) are valid for stage $s+1$. Then a different method of order $p-1$ which uses the $s+1$ stages can be constructed.

Proof

The distinctness of the nodes $c_1, c_{s-p+2}, \ldots, c_s$ required by Lemma 1 is assumed, and this is sufficient to imply that the p vectors in the set $\{\mathbf{e}'\} \cup \{(I-2C)\mathbf{e}'\} \cup \boldsymbol{\Omega}'_{p-1}$ are linearly independent. Now, choose the basis of the nullspace *span* $\left\{(I-2C)\mathbf{e}'\} \cup \mathcal{N}'_{p-2}\right)$ of $\mathbf{b}'^t$ by appending to $\{(I-2C)\mathbf{e}'\} \cup \boldsymbol{\Omega}'_{p-2}$, a basis of the remaining unspanned vectors of $\cup_{r=2}^{p-2} \boldsymbol{\Theta}'_r$. Let $\mathcal{B}'_{p-2}$ denote the subset of this basis which excludes $(I-2C)\mathbf{e}'$, and let $\mathcal{B}_{p-2}$ denote the corresponding vectors of $\mathbb{R}^{s+1}$. Then the hypothesis implies that each of $\mathcal{B}'_{p-2}$ and $\mathcal{B}_{p-2}$ contain exactly $s-1$ linearly independent vectors.

Now for each nonzero value of $\hat{b}_{s+1}$, the elements of an s-vector $\hat{\mathbf{b}}'^t$ may be uniquely determined so that $\widehat{\mathbf{b}}^t\mathbf{e} = 1$ and $\widehat{\mathbf{b}}^t$ is orthogonal to the $s-1$ linearly independent vectors of $\{(I-2C)\mathbf{e}\} \cup \mathcal{B}_{p-2}$.

It remains to show that this choice of $\widehat{\mathbf{b}}^t$ satisfies all of the conditions of order $p-1$. We need to show that $\widehat{\mathbf{b}}^t$ is orthogonal to *span* $(\{(I-2C)\mathbf{e}\} \cup \mathcal{N}_{p-2})$ and it is sufficient to show that $\{(I-2C)\mathbf{e}\} \cup \mathcal{B}_{p-2}$ is a basis for this set. This is established by considering that each vector of *span* $(\{(I-2C)\mathbf{e}\} \cup \mathcal{N}_{p-2})$ is either in *span* $(\{(I-2C)\mathbf{e}\} \cup \boldsymbol{\Omega}_{p-2})$, or in *span* $\left(\cup_{r=2}^{p-2}\{\boldsymbol{\Theta}_r\}\right)$, or else it is a linear combination of vectors in these two sets. In the first case, it is spanned by vectors in the basis, by the choice of $\mathcal{B}'_{p-2}$. Next, observe for every vector in *span* $\left(\cup_{r=2}^{p-2}\{\boldsymbol{\Theta}_r\}\right)$, element $s+1$ is equal to zero because stage $s+1$ has order p, and this implies the result for the second case. On reviewing the conditions defining the method, it is found that the only order conditions that are satisfied specifically because a vector in the nullspace of $\mathbf{b}^t$ is a linear combination of the two sets are identified by (23)–(25). As these conditions hold for stage $s+1$ by hypothesis, this completes the proof. □

LEMMA 7 (Conventional pairs)

Suppose for a method determined by the Corollary to Lemma 4, that the vector space spanned by $\{(I-2C)\mathbf{e}'\} \cup \mathcal{N}'_{p-2}$ is a nullspace of $\mathbf{b}'$ of dimension $\leq s-2$. Then a different method of order $p-1$ which uses only the original s stages can be constructed.

Proof

With notation used in the proof of Lemma 6, $\mathbf{b}'^t\mathbf{e}' = 1$ and the hypothesis implies that $\{\mathbf{e}'\} \cup \{(I-2C)\mathbf{e}'\} \cup \mathcal{B}'_{p-2}$ is a set of not more than $s-1$ linearly independent vectors of $\mathbb{R}^s$. Hence, there is a nonzero vector $\hat{\mathbf{b}}'^t - \mathbf{b}'^t$ orthogonal to this set. Since this set spans $\{(I-2C)\mathbf{e}'\} \cup \boldsymbol{\Omega}'_{p-2} \cup \left(\cup_{r=2}^{p-2}\boldsymbol{\Theta}'_r\right)$, it follows that $\hat{\mathbf{b}}'^t$ satisfies all of the conditions of order $p-1$, and is different from $\mathbf{b}'^t$. □

For $p=6$ and $p=7$, reducing the size of the basis suggested by Lemma 7 has been achieved by choosing parameters to make the set $\boldsymbol{\Omega}_{p-2} \cup \{\mathbf{q}^{p-3}\} \cup \{\mathbf{q}^{p-2}\}$ linearly dependent. Possibly other choices may achieve the same result.

We conclude this section with a general algorithm which may be used to obtain weights for many of the embedded methods proposed. For $c_{s-1} \neq 1$, define

$$\begin{aligned} \hat{b}_i &= \frac{2B_{3i} - b_i(1-c_i)(c_{s-1}-c_i)}{(1-c_i)(1-c_{s-1})}, \qquad i = 1, \ldots, s-2, \\ \hat{b}_{s-1} &= 0, \end{aligned} \tag{25}$$

and then define $\hat{b}_i,\ i = s, s+1$, to satisfy

$$\widehat{\mathbf{b}}^t \mathbf{e} = 1, \qquad \widehat{\mathbf{b}}^t \mathbf{q}^{p-3} = 0. \tag{26$'$}$$

It is easily verified that the solution of these equations satisfies many of the conditions of order $p-1$.

6 Algorithms for three families

So far only families with $p=6$ and $p=7$ have been constructed. For some of these, algorithms for finding the nodes and coefficients explicitly, as solutions of linear systems, and occasionally as roots of polynomials follow. For actual computation, these have been implemented using the MAPLE programming language. Often, the solutions will involve radicals (if the equations are solvable by a direct method), although in the pairs displayed, radicals have been avoided. Furthermore, for each displayed pair, the 2-norms of the vectors of leading error coefficients for the lower-order and higher-order methods are recorded as $\hat{A}_{p-1,2}$ and A_{p2}, respectively. See [5] and [9] for some comparison values.

EXAMPLE 1

There exists a family of FSAL pairs of orders 5 and 6 with seven arbitrary parameters, $c_2, \ldots, c_7$, and a_{52}.

ALGORITHM

Select eight nodes so that $c_1 = 0, c_4, c_5, c_6, c_7, c_8 = 1$ are distinct, $c_2 \neq 0$ and c_3 is arbitrary. For the ASOV=(6,1,2,2,2,2,2,2,6:5), select the coefficients of stages 1 to 6 to satisfy (5), (5$'$), and

$$\begin{aligned} A\mathbf{q}^2 &= \tilde{J}_3\mathbf{q}^3, \\ A\mathbf{q}^3 &= \hat{K}_3 C(I-C)(I-5C+5C^2)\mathbf{e} + \tilde{K}_2\mathbf{q}^2 + \tilde{K}_3\mathbf{q}^3 + \tilde{K}_4\mathbf{q}^4, \\ C\mathbf{q}^3 &= \hat{L}_3 C(I-C)(I-5C+5C^2)\mathbf{e} + \tilde{L}_2\mathbf{q}^2 + \tilde{L}_3\mathbf{q}^3 + \tilde{L}_4\mathbf{q}^4. \end{aligned} \tag{26}$$

To accommodate stage 1 easily, and as well anticipate the derivation of an embedded method, the linear combinations of $\{\boldsymbol{\omega}^2,\ \bar{\boldsymbol{\omega}}^3,\ \bar{\boldsymbol{\omega}}^4\}$ proposed in (24) and (25) have been replaced by a multiple of a single polynomial, which is orthogonal to each of $\mathbf{b}^t,\ \mathbf{b}^tC$, and is zero for each of $c_1 = 0$ and $c_{s+1} = 1$. The coefficients on the right sides of (27) are determined by stages 2 to 5, and one coefficient of stage 5 remains arbitrary (chosen to be a_{52} without loss of generality). Observe that each equation of (27) is trivially valid for stage 9. Next, select $b_2 = B_{22} = B_{32} = 0$, and b_3, B_{23} and B_{33} to satisfy $(\mathbf{b}^t(I - C)(c_7I - C)A)_2 = (\mathbf{B}_2 - \mathbf{b}^t(I - C))_3 = (\mathbf{B}_3A)_2 = 0$ so that (15) and (22) are valid. Then choose left homogeneous polynomials to satisfy (12) and (13), and compute the coefficients of stages 7 and 8 by (14). Finally, select the weights $\widehat{\mathbf{b}}^t$ to satisfy (26) and (26′). This yields a family which satisfies Lemmas 4 and 6. □

EXAMPLE 2

There exist two families of conventional pairs of orders 5 and 6 with six arbitrary nodes $c_2, \ldots, c_7$.

ALGORITHM

Select eight nodes so that $c_1 = 0, c_4, c_5, c_6, c_7, c_8 = 1$ are distinct, $c_2 \neq 0$ and c_3 is arbitrary. For the ASOV=(6,1,2,2,2,2,2,2:6,5), select the coefficients of stages 1 to 6 to satisfy (5), (5′), and (27). It is again convenient to confine the representation of each of (24) and (25) to a single vector of $\boldsymbol{\Omega}_4$, again to accommodate stage 1, and in this case, the need to restrict a_{52} so that $\{\boldsymbol{\omega}^2,\ \bar{\boldsymbol{\omega}}^3,\ \bar{\boldsymbol{\omega}}^4,\ \mathbf{q}^2,\ \mathbf{q}^3,\ \mathbf{q}^4\}$ is a linearly dependent set. After obtaining the coefficients in (27), the linear dependence of the last set yields a quadratic in a_{52}, and each choice leads to a pair. The remaining coefficients are obtained exactly as in Example 1. This yields a family which satisfies Lemmas 4 and 7, and such a pair appears in Table 4. □

In addition to these two families, some other special families for $p = 6$ may be obtained for certain values of the nodes. One of these pairs is displayed in [10].

For $p = 7$, attempts to obtain a family of FSAL pairs have been unsuccessful. Accordingly, the following result is rather surprising.

EXAMPLE 3

There exist two families of conventional pairs of orders 6 and 7 with seven arbitrary nodes $c_2, c_4, \ldots, c_9$.

ALGORITHM

Select ten nodes so that $c_1 = 0, c_5, c_6, c_7, c_8, c_9, c_{10} = 1$ are distinct, $c_2 \neq 0$, $c_3 = 2c_4/3$ and c_4 is arbitrary. For the ASOV=(7,1,2,3,3,3,3,3,3,3:7,6), select

Table 4 A Conventional (8, 5:6) Pair: Six nodes arbitrary, $\mathrm{A}_{72} \approx .000105$, $\hat{\mathrm{A}}_{62} \approx .00808$.

0								
$\frac{1}{5}$	$\frac{1}{5}$							
$\frac{1}{6}$	$\frac{7}{72}$	$\frac{5}{72}$						
$\frac{1}{3}$	$-\frac{1}{18}$	$-\frac{5}{18}$	$\frac{2}{3}$					
$\frac{1}{2}$	$\frac{13}{344}$	$-\frac{75}{344}$	$\frac{15}{43}$	$\frac{57}{172}$				
$\frac{3}{4}$	$-\frac{287}{1216}$	$\frac{765}{1216}$	$\frac{1239}{1216}$	$-\frac{1155}{608}$	$\frac{1505}{1216}$			
$\frac{11}{12}$	$\frac{19723}{59904}$	$-\frac{275}{1872}$	$-\frac{5335}{6656}$	$\frac{12397}{6656}$	$-\frac{16555}{19968}$	$\frac{209}{416}$		
1	$-\frac{1409}{726}$	$-\frac{45}{22}$	$\frac{915}{77}$	$-\frac{1480}{77}$	$\frac{559}{33}$	$-\frac{1520}{231}$	$\frac{1664}{847}$	
$\mathbf{b}^t$	$\frac{113}{2475}$	0	$\frac{239}{875}$	$\frac{9}{3500}$	$\frac{43}{125}$	$\frac{1216}{7875}$	$\frac{1664}{9625}$	$\frac{11}{1500}$
$\widehat{\mathbf{b}}^t$	$\frac{7}{90}$	0	$\frac{18}{175}$	$\frac{9}{25}$	0	$\frac{608}{1575}$	0	$\frac{11}{150}$

the coefficients of stages 1 to 8 to satisfy (5), (5′), and

$$\begin{aligned} A\mathbf{q}^3 &= \hat{J}_4 C(4I - 30C + 65C^2 - 35C^3)\mathbf{e} + \sum_{r=2}^{4} \tilde{J}_r \mathbf{q}^r, \\ A\mathbf{q}^4 &= (\hat{K}_4 + \hat{K}_5 C)C(4I - 30C + 65C^2 - 35C^3)\mathbf{e} + \sum_{r=2}^{5} \tilde{K}_r \mathbf{q}^r, \\ C\mathbf{q}^4 &= (\hat{L}_4 + \hat{L}_5 C)C(4I - 30C + 65C^2 - 35C^3)\mathbf{e} + \sum_{r=2}^{5} \tilde{L}_r \mathbf{q}^r. \end{aligned} \tag{27}$$

Table 5 A Conventional (10,6:7) Pair: Seven nodes arbitrary, $\mathrm{A}_{82} \approx .000172$, $\hat{\mathrm{A}}_{72} \approx .00885$.

0										
$\frac{2}{3}$	$\frac{2}{3}$									
$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{12}$								
$\frac{1}{2}$	$\frac{1}{8}$	0	$\frac{3}{8}$							
$\frac{1}{4}$	$\frac{1}{8}$	0	$\frac{3}{16}$	$-\frac{1}{16}$						
$\frac{1}{3}$	$\frac{217}{1458}$	0	$\frac{73}{162}$	$-\frac{82}{729}$	$-\frac{112}{729}$					
$\frac{1}{6}$	$\frac{277}{5832}$	0	$-\frac{119}{648}$	$\frac{59}{729}$	$\frac{344}{729}$	$-\frac{1}{4}$				
$\frac{5}{8}$	$-\frac{3787}{165888}$	0	$-\frac{1745}{9216}$	$\frac{54149}{165888}$	$-\frac{2359}{10368}$	$\frac{357}{2048}$	$\frac{1155}{2048}$			
$\frac{3}{4}$	$-\frac{823}{4080}$	0	$-\frac{399}{544}$	$\frac{11}{48}$	$-\frac{4}{3}$	$\frac{567}{544}$	$\frac{567}{374}$	$\frac{216}{935}$		
1	$\frac{151}{180}$	0	2	$\frac{157}{18}$	$\frac{107}{9}$	$-\frac{351}{28}$	$-\frac{432}{77}$	$-\frac{2592}{385}$	$\frac{17}{7}$	
$\mathbf{b}^t$	$\frac{31}{630}$	0	0	$\frac{104}{105}$	$\frac{16}{45}$	$-\frac{243}{490}$	$\frac{486}{2695}$	$-\frac{2048}{2695}$	$\frac{272}{441}$	$\frac{4}{63}$
$\widehat{\mathbf{b}}^t$	$\frac{17}{210}$	0	0	$-\frac{146}{105}$	$-\frac{32}{35}$	$\frac{837}{490}$	$\frac{108}{385}$	$\frac{3072}{2695}$	0	$\frac{2}{21}$

Again the basis vectors of $\boldsymbol{\Omega}_5$ are conveniently selected to accommodate the first stage and the necessity to restrict a_{72} so that $\boldsymbol{\Omega}_5 \cup \{\mathbf{q}^4\} \cup \{q^5\}$ is a linearly dependent set. When coefficients from (28) and those of stages 2–7 are represented in terms of a_{72}, this linear dependence is quadratic in a_{72} and each choice leads to a pair. Next, select $b_i = B_{2i} = B_{3i} = 0,\ i = 2, 3$, and $b_4,\ B_{24}$ and B_{34} to satisfy $(\mathbf{b}^t(I - C)(c_9 I - C)A)_3 = (\mathbf{B}_2 - \mathbf{b}^t(I - C))_4 = (\mathbf{B}_3 A)_3 = 0$ so that (15) and (22) are valid. Then choose left homogeneous polynomials to satisfy (12) and (13), and compute the coefficients of stages 9 and 10 by (14). Finally, select the weights $\widehat{\mathbf{b}}^t$ to satisfy (26) and (26′). This yields a family which satisfies Lemmas 4 and 7. One pair of this type is displayed in Table 5. □

Acknowledgments

The author thanks P.W. Sharp for discussions and continuing collaboration which have contributed to the development of new algorithms. This research was completed in part while the author visited the Department of Mathematics and Statistics at the University of Auckland, New Zealand, and was supported by the Natural Sciences and Engineering Research Council of Canada under grant A8147, and a grant from the Information Technology Research Centre of Ontario, Canada.

References

1. J.C. Butcher, *The Numerical Analysis of Ordinary Differential Equations* (Wiley, Toronto, 1987).
2. M. Calvo, J.I. Montijano, L. Rández, A new embedded pair of Runge–Kutta formulas or orders 5 and 6, Computers Math. Applic. 20 (1990) 15-24.
3. J.R. Dormand, M.R. Lockyer, N.E. McCorrigan and P.J. Prince, Global error estimation with Runge–Kutta triples, J. Comput. Appl. Math. 7 (1989) 835-846.
4. E. Fehlberg, Classical fifth-, sixth-, seventh-, and eighth-order Runge–Kutta formulas with stepsize control, NASA Technical Report NASA TR R-287 (1968) 82 pages.
5. P.J. Prince and J.R. Dormand, High-order embedded Runge–Kutta formulae, J. Comput. Appl. Math. 7 (1981) 67-76.
6. P.W. Sharp and J.H. Verner, Completely imbedded Runge–Kutta formula pairs, SIAM J. Numer. Anal. 31 (1994) to appear.
7. J.H. Verner, Explicit Runge–Kutta methods with estimates of the local truncation error, SIAM J. Numer. Anal. 15 (1978) 772-790.
8. J.H. Verner, Families of imbedded Runge–Kutta methods, SIAM J. Numer. Anal. 16 (1979) 857-875.
9. J.H. Verner, Some Runge–Kutta formula pairs, SIAM J. Numer. Anal. 28 (1991) 496-511.
10. J.H. Verner, A classification scheme for studying explicit Runge–Kutta pairs, Dept. Mathematics and Statistics, Queen's University, Kingston, Preprint 1992-04 (1992) 25 pages.