Two continuous (4, 5) pairs of explicit 9-stage Runge–Kutta methods

MISHA STEPANOV†
Department of Mathematics and Program in Applied Mathematics,
University of Arizona, Tucson, AZ 85721, USA
[Received on 21 April 2022]

An 11-dimensional family of embedded (4, 5) pairs of explicit 9-stage Runge–Kutta methods with an interpolant of order 5 is derived. Two optimized for efficiency pairs are presented.

Keywords: embedded pair of Runge–Kutta methods, continuous formula, interpolant

1. Introduction

Runge–Kutta methods (see, e.g., (Butcher, 2008, sec. 23 and ch. 3), (Hairer et al., 2008, ch. II), (Ascher & Petzold, 1998, ch. 4), (Iserles, 2009, ch. 3)) are widely and successfully used to solve Ordinary Differential Equations (ODEs) numerically for over a century (Butcher & Wanner, 1996). Being applied to a system \( \frac{dx}{dt} = f(t, x) \), in order to propagate by the step size \( h \) and update the position, \( x(t) \mapsto x(t + h) \), an \( s \)-stage explicit Runge-Kutta method (which is determined by the coefficients \( a_{ij} \), weights \( b_j \), and nodes \( c_i \)) would compute intermediate vectors \( F_1, F_2, ..., F_s \), and then \( x(t + h) \):

\[
X_i = x(t) + h \sum_{j=1}^{i-1} a_{ij} F_j, \quad F_i = f(t + c_i h, X_i), \quad x(t + h) = x(t) + h \sum_{j=1}^{s} b_j F_j
\]

In the limit \( h \to 0 \) all the vectors \( F_i \), where \( 1 \leq i \leq s \), are the same, so it is natural and will be assumed that \( \sum_{j=1}^{i-1} a_{ij} = c_i \). For \( i = 1 \) the sum over \( j \) is empty, so \( c_1 = 0 \), \( X_1 = x(t) \), and \( F_1 = f(t, x(t)) \).

To obtain an accurate solution with less effort, various adaptive step size strategies were developed (see, e.g., (Butcher, 2008, secs. 271 and 33), (Hairer et al., 2008, sec. II.4), (Ascher & Petzold, 1998, sec. 4.5), (Iserles, 2009, ch. 6)). Typically a system of ODEs is solved in two different ways, and the step size is chosen so that the two solutions are sufficiently close. A computationally efficient procedure is to have two Runge–Kutta methods with different weights, but the same nodes and coefficients. The vectors \( F_1, F_2, ..., F_s \) are computed only once, and then are used in both methods, the latter are said to form an embedded pair. Two well known examples of such pairs are (Fehlberg, 1969, tab. III), (Fehlberg, 1970, tab. 1) and (Dormand & Prince, 1980, tab. 2).

There is no 5-stage explicit Runge–Kutta 5th order method (Butcher, 1964). The Fehlberg pair has 6 stages. The Dormand–Prince pair uses 7 stages, but has the so-called First Same As Last (FSAL) property (Fehlberg, 1969, p. 17), (Dormand & Prince, 1978): the vector \( F_1 \) at the current step is equal to the already computed \( F_u \) at the stage \( 1 < u \leq s \) of the previous step. An FSAL method requires \((s-1)\) evaluations of the r.h.s. function \( f \) per step, with the exception of the 1st step.

Continuous formulas or interpolants (see, e.g., (Horn, 1983), (Sarafyan, 1984), (Butcher, 2008, sec. 272), (Hairer et al., 2008, sec. II.6)) provide an inexpensive (i.e., with only a few if any additional evaluations of the r.h.s.) way to estimate the solution at anywhere within the integration interval.

† Email: stepanov@math.arizona.edu
Here the interpolant functions $\beta_\alpha$ \( \alpha \neq 1 \) are polynomials. For the approximation over several step intervals to be continuously differentiable a method should have the FSAL property, with the following conditions on the behavior of the row vector $\hat{\beta}(\theta) = [\beta_j(\theta)]$ at $\theta = 0$ and $\theta = 1$:

\[
\begin{align*}
  c_1 &= 0, \quad a_1 = 0 = \beta(0), \quad X_1 = x(t), \quad F_1 = f(t, x(t)), \quad \left. \frac{d\beta_j(\theta)}{d\theta} \right|_{\theta = 0} = \begin{cases} 1, & j = 1 \\ 0, & j \neq 1 \end{cases} \\
  c_u &= 1, \quad a_u = b = \beta(1), \quad X_u = x(t+h), \quad F_u = f(t+h, x(t+h)), \quad \left. \frac{d\beta_j(\theta)}{d\theta} \right|_{\theta = 1} = \begin{cases} 1, & j = u \\ 0, & j \neq u \end{cases}
\end{align*}
\]

Here $a_u = [a_{ij}]$ and $b = [b_j]$ are the row vectors of coefficients and weights, respectively. Below all interpolants are assumed to be continuously differentiable.

Without altering the strategy of step size choice, this can be used in applications that require values of the solution $x(t)$ at specific points $t_1, t_2, \ldots$ (dense output) or the place (time or position $x(t))$ where the solution crosses a hypersurface $g(t, x(t)) = 0$ (event location). The continuous approximation to the solution in the interval $[t, t+h]$ is typically of the form

\[
x(t + \theta h) = x(t) + h \sum_{j=1}^{s} \beta_j(\theta) F_j
\]

where the interpolant functions $\beta_j(\theta) = \sum \beta_{ij} \theta^k$ are polynomials. For the approximation over several step intervals to be continuously differentiable a method should have the FSAL property, with the following conditions on the behavior of the row vector $\hat{\beta}(\theta) = [\beta_j(\theta)]$ at $\theta = 0$ and $\theta = 1$:

\[
\begin{align*}
  c_1 &= 0, \quad a_1 = 0 = \beta(0), \quad X_1 = x(t), \quad F_1 = f(t, x(t)), \quad \left. \frac{d\beta_j(\theta)}{d\theta} \right|_{\theta = 0} = \begin{cases} 1, & j = 1 \\ 0, & j \neq 1 \end{cases} \\
  c_u &= 1, \quad a_u = b = \beta(1), \quad X_u = x(t+h), \quad F_u = f(t+h, x(t+h)), \quad \left. \frac{d\beta_j(\theta)}{d\theta} \right|_{\theta = 1} = \begin{cases} 1, & j = u \\ 0, & j \neq u \end{cases}
\end{align*}
\]

Here $a_u = [a_{ij}]$ and $b = [b_j]$ are the row vectors of coefficients and weights, respectively. Below all interpolants are assumed to be continuously differentiable.
The product of column vectors $x, y$ is to be understood element-wise: $(xy)_i = x_i y_i$. Let $1$ be the $s$-dimensional column vector with all components being equal to 1; $c = [c_i]$ be the nodes vector; and $A = [a_{ij}]$ be the $s \times s$ matrix with $a_{ij}$ as its matrix element in the $i^{\text{th}}$ row and $j^{\text{th}}$ column (for an explicit method $a_{ij} = 0$ if $i \neq j$). Let $q_0 = A c^s - \frac{1}{\pi^2} e^{\sigma^s+1}$. The condition $\sum_j a_{ij} = c_i$, $A 1 = c$, or $q_0 = 0$ is assumed. The following quantities will be used for the estimation of the local error:

$$T_p^2 (x, \theta) = \sum_{\text{rooted trees } t \text{ of order } p} \tau^2 (t, x, \theta), \quad \tau (t, x, \theta) = \frac{1}{\sigma (t)} \left( x \Phi (t) - \frac{\theta^p}{\gamma (t)} \right)$$

Here $\sigma (t)$ is the order of the symmetry group of the tree $t$ (see, e.g., (Butcher, 2008, p. 140)). Whenever the argument $x$ or $\theta$ is omitted, its value is meant to be equal to $b (\theta)$ and 1, respectively. Note that $b (1) = b$. An interpolant of order $p$ should satisfy the conditions $\tau (t, \theta) \equiv 0$ for all rooted trees $t$ with up to $p$ vertices. For $p = 5$, with the assumption that the intermediate positions $X_i$ are at least 2nd order accurate for all $i > 2$, these conditions are listed in Table 1, see also (Butcher, 2008, sec. 31), (Hairer et al., 2008, sec. II.2), (Dormand & Prince, 1980, tab. 1).

There are certain properties one would expect from a practical Runge–Kutta method (see, e.g., a list in (Verner, 1978, p. 785)). The natural or desirable behavior of the interpolant function $\beta_i (\theta)$, where $1 \leq i \leq s$, is that it has a notably positive slope around $\theta \approx c_j$, while it is hardly changing anywhere else. Deviations from this behavior can be divided into two categories: non-positivity, when an interpolant function’s derivative is negative; and non-locality, when an interpolant function has substantial slope (positive or negative) far from the position of the corresponding node.

An explicit Runge–Kutta method with continuously differentiable interpolant of order 5 has at least 8 stages (Owren & Zennaro, 1991, sec. 3.3). Such methods were completely classified in (Verner & Zennaro, 1995). In (Owren & Zennaro, 1992) a 5-dimensional family of continuous (4, 5) pairs of 8-stage Runge–Kutta methods was constructed, and an optimized for efficiency (Owren & Zennaro, 1992, fig. 3) pair was suggested, which is also shown in Figure 1. All the pairs in this 5-dimensional family satisfy $2 c_3 = c_4 = c_5$. This is an indication that the family lacks sufficient flexibility or is stressed by numerous imposed conditions. Looking at the interpolant functions in Figure 1, $b_5 (\theta)$ goes down for $\theta > 0.8$, $b_6 (\theta)$ goes down for $\theta < 0.5$, and both $b_7 (\theta)$ and $b_8 (\theta)$ express non-locality for $\theta < 0.3$.

Any Runge–Kutta method could be equipped with an interpolant by adding, if needed, additional stages (Enright et al., 1986), (Verner, 1993). Several interpolants were constructed for the (Dormand & Prince, 1980, tab. 2) pair, see, e.g., (Shampine, 1986, p. 149) and (Calvo et al., 1990). With $u = 7$ both interpolants use $s = 9$ stages (see also (Owren & Zennaro, 1991, corollary 2.13)). The second interpolant has somewhat smaller local error, see (Calvo et al., 1990, fig. 1).

Increasing the number of stages (and thus the amount of computation per step) provides additional flexibility in choosing the nodes, coefficients, and weights, which may be exploited to construct viable pairs that produce an accurate solution in fewer steps. In (Sharp & Smart, 1993, sec. 3.1) and (Bogacki & Shampine, 1996) non-FSAL embedded (4, 5) pairs of 7-stage Runge–Kutta methods were suggested. The objective in the construction of the latter pair was an improvement of the (Dormand & Prince, 1980, tab. 2) pair (see Bogacki & Shampine, 1996, p. 19)). The pair was also equipped with an interpolant of order 5. The minimal number of stages would be 9, but in (Bogacki & Shampine, 1996) the suggested interpolant (with the local error, to the leading order, being a problem-independent function of the local error at the end of the step) is using 11 stages.

The interpolant (Calvo et al., 1990) for the (Dormand & Prince, 1980, tab. 2) pair and the interpolant for the (Bogacki & Shampine, 1996) pair are depicted in Figure 2. In the case of Dormand–Prince pair, $b_5 < 0$, plus the 1st order condition $\sum_j b_j (\theta) = 0$ is obtained through the cancellation of wiggles in $\beta_3$,
\begin{align*}
\begin{array}{c|cc}
\theta & b_1 & b_2 \\
0 & 0 & 0.5 \\
1/5 & 1/5 & 3/10 \\
1/4 & 1/4 & -3/4 \\
1/2 & -3/4 & 15/4 \\
1 & -3 & 3 \\
\end{array}
\end{align*}

\[ s = u = 8 \]

\[ s = u = 8 \]

\[
\begin{align*}
\begin{array}{ccccccccc}
\theta & 0 & 1/5 & 1/4 & 1/2 & 1 & 3/4 & 15/4 & 1/2 & 1/5 \\
0 & 945 & 83 & 0.5 & 825 & 0.5 & 180 & 36 & 0.5 & 180 \\
0.5 & 1372 & 133 & 248 & 0 & 6016 & 7945 & 12845 & 315 & 156065 \\
1 & 369 & -4 & 243 & 0 & 343 & 9904 & 1485 & 297 & 4802 \\
1.5 & 1113 & 12845 & 343 & 1485 & 297 & 7945 & 156065 & 315 & 156065 \\
2 & 7945 & 156065 & 7945 & 156065 & 315 & 156065 & 315 & 156065 & 315 \\
2.5 & 12845 & 315 & 156065 & 315 & 156065 & 315 & 156065 & 315 & 156065 \\
3 & 343 & 9904 & 1485 & 297 & 4802 & 7945 & 156065 & 315 & 156065 \\
3.5 & 825 & 0.5 & 180 & 36 & 0.5 & 180 & 36 & 0.5 & 180 \\
4 & 0 & 0.5 & 825 & 0.5 & 180 & 36 & 0.5 & 180 \\
\end{array}
\end{align*}
\]

---

**FIG. 1.** The continuous (4, 5) pair (Owren & Zennaro, 1992, fig. 3): the Butcher tableau (upper panel), the weights vector \( b \) repeats the last row of \( A \) and is not shown; the interpolant functions \( \beta_j(\theta) \), with \( j = 1, 3, 4, \ldots, 8 \) (the panels on the left, the function \( \beta_2(\theta) = 0 \) is not shown); the local error \( T_6(\theta) \) (middle right panel); and the region of absolute stability (lower right panel). (The thin dashed and solid lines correspond to the pairs on Figure 3 and Figure 5, respectively. The three regions of absolute stability are scaled for equal cost, i.e., the regions where \( |R(s-1)z| \leq 1 \) are depicted, here \( R(z) \) is the stability function and \( s \) is the number of stages.)
Fig. 2. The interpolant functions $\beta_j(\theta)$, with $j = 1, 3, 4, ..., 9$ (Calvo et al., 1990) for the embedded $(4,5)$ pair (Dormand & Prince, 1980, tab. 2) (panels on the left, $s = 9, u = 7$) and functions $\beta_j(\theta)$, with $j = 1, 3, 4, ..., 11$, for the (Bogacki & Shampine, 1996) pair (panels on the right, $s = 11, u = 8$).
of slopes in $\beta_1$ and $\beta_5$ for $\theta > 0.6$. In the case of Bogacki–Shampine pair, there is some non-locality in $\beta_1$, $\beta_5$, $\beta_9$, and $\beta_0$, which is compensated by $\beta_0$, $\beta_{10}$, and $\beta_{11}$. If an interpolant is obtained by adding stages to an already formed pair, the interpolant function for an added stage has a negative slope somewhere, as it should have zero values (and derivatives) at $\theta = 0$ and $\theta = 1$.

In this work embedded (4,5) pairs are constructed, like the (Owren & Zennaro, 1992, fig. 3) pair, so that they have an interpolant right away, although not the minimal number of stages is used. The family of such pairs is constructed in Section 2. How the values of the free parameters are chosen is discussed in Section 3. The performance of built pairs is demonstrated in Section 4.

2. A family of continuous (4,5) pairs

There are two different ways to generate an embedded pair with an interpolant: to construct a pair with no interpolant, and then add one or several stages in order to build one; or to design both a pair and an interpolant at once. The latter approach is used here. As for continuous (4,5) pairs 8 stages do not provide enough flexibility (Owren & Zennaro, 1992), here $s = 9$ stages are used. For the update of position, $x(t) \rightarrow x(t + h)$, to be as accurate as possible, the FSAL stage is the last one: $u = 9$.

To increase the similarity with collocation methods (see, e.g., (Hairer et al., 2008, p. 211), (Ascher & Petzold, 1998, sec. 4.7.1), (Iserles, 2009, sec. 3.4)), the Dominant Stage Order (DSO) (see, e.g., (Verner, 2010, eq. (5))) is chosen to be equal to 3. This goes against the observation (Verner, 2010, p. 386) that most efficient for computation pairs of order $p$ have the DSO being equal to $(p - 4)$ or $(p - 3)$. Increasing the DSO makes order conditions more redundant, and may not reduce richness or flexibility of the set of pairs much.

The parameters of the 11-dimensional family of continuous (4,5) pairs described below are $c_2$, $c_4$, $c_5$, $c_6$, $c_7$, $c_8$, $a_{65}$, $a_{75}$, $a_{76}$, $a_{86}$, and $a_{87}$. Their values are arbitrary, except for some degenerate cases, e.g., $c_5 = c_4$ for which the matrix $A$ ends up being infinite. Other nodes $c_1$, $c_3$, $c_9$, and the first 8 rows of $A$ are expressed through the 11 parameters as follows:

\[
\begin{align*}
    c_1 &= 0, \quad c_3 = 2c_4/3, \quad c_9 = 1 \\
    h_{ij} &= a_{ij}c_j(c_j - c_4)/\prod_{k \in \{1,4,5,6,7,8\}}(c_i - c_k), \quad Y_i = 3 - 5c_4 - 5c_j + 10c_4c_j \\
    Z_m &= 12 - 15c_4 - 15c_5 - 15c_m + 20c_4c_5 + 20c_4c_m + 20c_5c_m - 30c_4c_5c_m \\
    \sum_{ij \in \mathscr{C}} Y_i h_{ij} &= \sum_{ij \in \mathscr{C}} \sum_{kl \in \mathscr{C}} (c_i - c_k)(c_j - c_l)Z_{21-i-k}h_{ij}h_{kl} \\
    a_{i4} &= \frac{1}{c_4^2} \left( c_i^2(c_i - c_4) - 3 \sum_{j=5}^{i-1} a_{ij}c_j(c_j - c_3) \right), \quad i \geq 4 \\
    a_{i3} &= \frac{1}{c_3^2} \left( c_i^2(c_4 - 2c_3) + 2 \sum_{j=5}^{i-1} a_{ij}c_j(c_j - c_4) \right), \quad i \geq 4 \\
    a_{i2} &= \left\{ \begin{array}{ll} 
        c_i^2/2c_2, & i = 3 \\
        0, & i \neq 3 
    \end{array} \right. \\
    a_{i1} &= c_i - \sum_{j=2}^{i-1} a_{ij} \quad \text{for all } i
\end{align*}
\]

\[
\begin{align*}
    \text{determines } a_{85} \\
    \text{ensures } b_9 = 0 \\
    \{q_0 = 0 \} \\
    \{q_1_i = 0 \text{ for all } i \neq 2 \} \\
    \{Aq_2_i = 0 \text{ for all } i \neq 3 \} \\
    \{q_2 \} = 0 \text{ for all } i \neq 2, 3
\end{align*}
\]
where \( \mathcal{S} = \{65, 75, 85, 76, 86, 87\} \). In particular, \( a_{44} = c_4/4 \) and \( a_{43} = 3c_4/4 \).

The vectors \( q_3 \) and \( q_3 c \) are linear combinations of \( q_1 \) and \( Aq_1 \), also \((Aq_1)c = c_1 Aq_1\). The four vectors \( q_1, Aq_1, A^2q_1, \) and \( q_3 \) should be in the null space of the matrix \( B = [\beta_{ij}] \), see Table 1. The interpolant matrix \( B \) is generated as

\[
B = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{5}{6} & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\begin{bmatrix}
1 & c & c^2 & c^3 & c^4 & q_1 & Aq_1 & A^2q_1 & q_3 \\
\end{bmatrix}^{-1}
\]

Here the last components of the four vectors \( q_1, Aq_1, A^2q_1, \) and \( q_3 \) are set to 0, which is compatible with the order conditions. The interpolant \( \beta(\theta) = [\theta \ \theta^2 \ \theta^3 \ \theta^4 \ \theta^5]B \) is continuously differentiable if

\[
B = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 0 \\
\end{bmatrix}
\]

The 1st and 3rd rows of this equation are satisfied automatically due to how the first and last rows of the 9 \( \times \) 9 matrix in eq. (2.1) look like. The 2nd row is used to determine \( a_0 = b \). The interpolant functions \( \beta_2(\theta) = 0 \) and \( \beta_3(\theta) = 0 \) as \( (q_1)_i = 0 \) and \( (Aq_1)_i = 0 \) for all \( i \neq 2 \) and \( i \neq 3 \), respectively.

The 5-dimensional (could be parameterized by \( c_2, c_4, c_5, c_6, \) and \( c_7 \) ) family of (5,6) pairs (see (Dormand et al., 1989, tab. 5), (Verner, 1991, tab. 2), (Sharp & Verner, 1994, tab. 4), (Verner, 2010, tab. 3)) is constructed in a similar fashion. It has DSO \( = 3, c_8 = 1, \) and \( a_2 = 0 \) for all \( i \neq 3 \). The 9 \( \times \) 9 matrix in the equation (2.1) is singular. With just 9 stages, in order for the 5th order interpolant to exist, the last 4 columns of the matrix should be linearly dependent, which happens if \( c_4 = c_5 \). Then the 5th stage repeats the 4th one, while some of the coefficients become infinite. To equip such an embedded (5,6) pair with an interpolant of order 5, one needs to add at least one more stage.

3. Choice of the degrees of freedom

A measure of the amount of non-positivity present in an interpolant is its total variation:

\[
V_0^1(B) = \sum_{j=1}^{4} \int_{0}^{1} \frac{d}{d\theta} |d\beta_i(\theta)|
\]

As \( \sum_{j} \beta_j(\theta) = \theta \), the minimal possible value of \( V_0^1(B) \) is equal to 1. If the integration over \( \theta \) is restricted to the region where \( \beta_j(\theta) \) has a negative slope: \( N_1^1(B) = \sum_{j=1}^{4} \int_{0}^{1} d\theta |\beta_j(\theta)|H(-\beta_j(\theta)) \), then \( V_0^1(B) = 1 + 2N_1^1(B) \). Here \( H \) is the Heaviside step function.

The continuous (4,5) pair shown in Figure 3 was obtained by minimizing the following function:

\[
\max_{0 \leq \theta \leq 1} T_6(\theta) + 10^{-4} V_0^1(B) + 10^{-7} \sum_{ij} (4|a_{ij}|^2 + |a_{ij}|^4)
\]

The term \( \max_{0 \leq \theta \leq 1} T_6(\theta) \) was motivated by the discussion in (Bogacki & Shampine, 1996, p. 24). In the resulted pair the local error \( T_6 \) is only slightly smaller than \( \max_{0 \leq \theta \leq 1} T_0(\theta) \), see Table 2. The term
\[ R(z) = 1 + z + \frac{1}{2} z^2 + \frac{1}{6} z^3 + \frac{1}{24} z^4 + \frac{1}{120} z^5 + \frac{1}{720} r_6 z^6 + \frac{1}{5040} r_7 z^7 + \frac{1}{40320} r_8 z^8 \]

\[ d_1 = \begin{bmatrix} 3 \\ 43 \\ 80 \\ 0 \end{bmatrix},
\quad d_2 = \begin{bmatrix} -3 \\ -4 \\ -5 \\ 0 \end{bmatrix},
\quad d_3 = \begin{bmatrix} 0 \\ 1 \\ -1 \\ -1 \end{bmatrix} \]

**Fig. 3.** A continuous \((4, 5)\) pair: the Butcher tableau and the region of absolute stability \(|R(z)| \leq 1\). The weights vector \(b\) repeats the last row of \(A\) and is not shown. The difference between the weights vectors of the 4th and the 5th order methods within the pair could be any linear combination of \(d_1, d_2, \) and \(d_3\). Regions of absolute stability of the (Dormand & Prince, 1980, tab. 2) pair (with \(r_7 = 0\)) and the (Bogacki & Shampine, 1996) pair (with \(r_7 = 2421/2464\)) are shown for comparison. The three regions of absolute stability shown on the small right panels are scaled for equal cost, i.e., the regions where \(|R((s-1)z)| \leq 1\) are depicted. For the (Dormand & Prince, 1980, tab. 2), (Bogacki & Shampine, 1996) pairs and the pair presented in this figure the number of stages \(s\) is equal to 7, 8, 9 on upper panel (just position is updated, \(x(t) \rightarrow x(t+h)\)) and 9, 11, 9 on lower panel (interpolant of order 5 with \(\max_{0 \leq \theta \leq 1} T_6(\theta) \approx T_6\) is computed), respectively.
Fig. 4. Columns of the matrix $B$ and interpolant functions $\beta_j(\theta)$, where $j = 1, 4, 5, ..., 9$, for the pair shown in Figure 3.
Fig. 5. A continuous (4,5) pair: the Butcher tableau and the region of absolute stability $|R(z)| \leq 1$. The weights vector $b$ repeats the last row of $A$ and is not shown. The difference between the weights vectors of the 4th and the 5th order methods within the pair could be any linear combination of $d_1$, $d_2$, and $d_3$. Regions of absolute stability of the (Dormand & Prince, 1980, tab. 2) pair (with $r_7 = 0$) and the (Bogacki & Shampine, 1996) pair (with $r_7 = 2421$) are shown for comparison. The three regions of absolute stability shown on the small right panels are scaled for equal cost, i.e., the regions where $|R((s-1)z)| \leq 1$ are depicted. For the (Dormand & Prince, 1980, tab. 2), (Bogacki & Shampine, 1996) pairs and the pair presented in this figure the number of stages $s$ is equal to 7, 8, 9 on upper panel (just position is updated, $x(t) \rightarrow x(t+h)$, or interpolant of order 4, 4, 5 is computed) and 7, 9, 9 on lower panel (low-cost interpolant of order 4, 5, 5 is computed), respectively.
FIG. 6. Columns of the matrix $B$ and interpolant functions $\beta_j(\theta)$, where $j = 1, 4, 5, ..., 9$, for the pair shown in Figure 5.
not even uncommon to use interpolants of order 4.

With the pair brings the local error $T_6$ to zero, see Table 4. The value of $\max_{0 \leq \theta \leq 1} T_6(\theta)$ is much larger than $T_6$, see Table 2. This by itself is not necessarily problematic, as less accurate interpolant is connecting endpoints of the integration step that contain an error accumulated in many steps. With (4,5) pairs it is not even uncommon to use interpolants of order 4.

| Owren–Zennaro      | $10^5 \times T_6$ | $10^5 \times T_7$ | $10^5 \times \max_\theta T_6(\theta)$ | $\max_{ij} |a_{ij}|$ | $V_0^1(B)$ |
|---------------------|-------------------|-------------------|--------------------------------------|-----------------|------------|
| 108.62...           | 154.05...         | 108.62...         | 3.75                                 | 1.6496...       |
| 39.908...           | 395.57...         | 39.908...         | 11.595...                            | 3.2490...       |
| 2.2169...           | 21.260...         | 2.2169...         | 1.1637...                            | 3.3092...       |
| Figure 3            | 9.2847...         | 19.904...         | 9.7178...                            | 2.0803...       | 1.4857...  |
| Figure 5            | 0.59809...        | 5.9203...         | 16.134...                            | 2.7916...       | 1.6424...  |

Table 2. A comparison of five continuous (4,5) pairs. The first three are from the literature: (Owren & Zennaro, 1992, fig. 3), (Dormand & Prince, 1980, tab. 2), and (Bogacki & Shampine, 1996).

| Dormand–Prince, $x = b + $ order | $10^5 \times T_5(x)$ | $10^5 \times T_6(x)$ | $10^5 \times T_7(x)$ |
|-----------------------------------|-----------------------|-----------------------|-----------------------|
| $x = \frac{1}{3}b + \frac{2}{3}b$ | 118.29...             | 182.37...             | 414.05...             |
| Bogacki–Shampine, $x = b + E$     | 10.595...              | 12.204...              | 24.114...              |
| $x = B$                           | 10.615...              | 10.992...              | 20.562...              |
| Figure 3, $x = b + \frac{1}{3}d_1$ | 19.765...             | 14.406...             | 18.948...             |
| $x = b - \frac{1}{323}d_2$       | 19.465...              | 29.174...              | 33.473...              |
| $x = b + \frac{1}{132}d_3$       | 19.511...              | 27.858...              | 27.751...              |
| Figure 5, $x = b + \frac{1}{2430}d_1$ | 0.99808...         | 1.0189...             | 6.0611...             |
| $x = b + \frac{1}{2377}d_2$      | 0.99189...             | 1.4234...             | 6.4042...             |
| $x = b - \frac{1}{995}d_3$       | 0.99041...             | 1.3645...             | 6.4662...             |

Table 3. Local errors for the lower order methods of four embedded pairs. The second line of the (Dormand & Prince, 1980, tab. 2) entry is the modification suggested in (Shampine, 1986, p. 141). The notation $E$ and $B$ in the (Bogacki & Shampine, 1996) entry is taken from the rksuite.f code (Brankin et al., 1993).
TWO CONTINUOUS (4, 5) PAIRS OF RUNGE–KUTTA METHODS

4. Numerical tests

The performance of the pairs constructed in the previous section is demonstrated on test problems A3, D5, and E2 from (Hull et al., 1972) in Figure 7; and on new suggested test problems U1, U2, and U4 in Figure 8. The error, that is the difference between the exact and numerical solutions, was computed only at the ends of integration steps. The adaptive step size scheme $h \leftarrow 0.9h(\text{ATOL}/E)^{1/5}$ was used. (The starting step size $h_0 = 10^{-3}$ was swiftly corrected by the adaptive step size control.) Here ATOL is the absolute error tolerance, and $E$ is the $l^2$-norm of the difference vector between the two solutions within a pair. The steps with $E > \text{ATOL}$ were rejected, but they were still contributing to the number of the r.h.s. evaluations. For the pairs with multiple difference vectors between the weights of the higher and lower order methods (i.e., the (Bogacki & Shampine, 1996) pair and the pairs in Figures 3 and 5, see Table 3), the difference vectors that use smaller number of stages were tried first. Once a step was rejected, no further difference vectors were tried. The size of the next step, whether the previous step was rejected or not, was chosen according to the maximal value of $E$ between the tried vectors.

The pairs with higher order 6 performed better on test problems A3 and D5. On other problems their performance was similar to the (Bogacki & Shampine, 1996) pair. The (Dormand & Prince, 1980, tab. 2) pair performed well on problem A3, while on problems E2, U2, and U4 it performed the worst. The pair shown in Figure 3 performed the worst on problems A3 and D5. The pair from Figure 5 seems to be at least as efficient as the (Bogacki & Shampine, 1996) pair. Note that the efficiency curves in Figures 7 and 8 show the cost of obtaining the numerical solution without computing interpolants. In case of the (Bogacki & Shampine, 1996) pair, no additional stages are needed to use the interpolant of order 4 (Bogacki, 1990). One or three additional r.h.s. evaluations per step are needed to compute less and more accurate interpolant of order 5, respectively (Bogacki & Shampine, 1996, p. 24), which corresponds to the increase of the cost by factors 8/7 and 10/7.

Figure 9 shows the performance of five pairs and their interpolants on the system $dx/dt = -y, dy/dt = x$ with initial condition $(x(0), y(0)) = (1, 0)$. Just one step $h = \pi/2$ is made, with no error control. The exact solution is $x(t) = \cos t, y(t) = \sin t$ with $(x(\pi/2), y(\pi/2)) = (0, 1)$. In cases of the (Dormand & Prince, 1980, tab. 2) and (Owren & Zennaro, 1992, fig. 3) pairs the intermediate positions $X_i, 1 \leq i \leq s$, with $s = u = 9$.

| 0 | 1 |
|---|---|
| 1/14 | 1/14 |
| 1/7 | 0 | 1/7 |
| 3/56 | 0 | 9/56 |
| 2/7 | 29/77 | -25/77 | 14/9 |
| 9/14 | -17/56 | 93/56 | -8/7 | 3 |
| 6/7 | 199/357 | -195/357 | 1259/357 | -3855/357 | 45 |
| 1 | 4903/9772 | 2847/357 | 102384/70285 | 33847/70285 | 53 |
| 1/14 | 16807/335480 | 53 | 2401 | 2401 | 79 |

Table 4. The Butcher tableau of the embedded (4, 6) pair with FSAL property. The weights vector $b$, which produces the 6th order update, repeats the last row of $A$ and is not shown. The $9 \times 9$ matrix in eq. (2.1) is non-singular, and no weights vector other than $b$ would give a method of order 5. The 5th order interpolant is constructed as in eq. (2.1). The differences $d_{1,2,3}$ between the weights vectors of the 4th and the 6th order methods, that could be used for error control, are the same as in Figure 5.
test problem A3: $x' = x \cos t, \quad x(0) = 1$

test problem D5: $x'_1 = x_3, x'_2 = x_4, x'_3 = -x_1/(x_1^2 + x_2^2)^{3/2}, x'_4 = -x_2/(x_1^2 + x_2^2)^{3/2}$

$x_1(0) = 1/10, x_2(0) = 0, x_3(0) = 0, x_4(0) = \sqrt{19}$

test problem E2: $x'_1 = x_2, x'_2 = (1 - x_1^2)x_2 - x_1, \quad x_1(0) = 2, x_2(0) = 0$

Fig. 7. Efficiency curves for problems A3 (Hull et al., 1972, p. 617), D5 (Hull et al., 1972, p. 620), and E2 (Hull et al., 1972, p. 621); the (Dormand & Prince, 1980, tab. 2) pair (dash-dotted curve), the (Bogacki & Shampine, 1996) pair (dotted curve), the pair in Figure 3 (dashed curve), the pair in Figure 5 (solid curve), the (Verner, 2010, tab. 3) pair (thin dotted curve), and the (Verner, 2010, tab. 3) pair (thin solid curve).
The five test problems Un, where \( n = 1, 2, 3, 4, \) and \( 5 \), are based on the initial value problem describing the motion of a particle with unit mass in a potential \( U(x,y) = \frac{1}{2 + \cos(2\pi x) + \cos(2\pi y)} \), with the initial conditions \( x(0) = y(0) = 0, \ p(0) = \frac{7}{2}, \ q(0) = -2 \). The potential has spikes at semi-integer values of \( x \) and \( y \), off which the particle scatters. The problem \( U_n \) corresponds to the computation of \( x(n) \) and \( y(n) \).

\[
\begin{bmatrix}
x(1) \\
y(1)
\end{bmatrix} = \begin{bmatrix} 2.45719163557503409569 \\
0.75988615298279252162 \\
\end{bmatrix},
\begin{bmatrix}
x(2) \\
y(2)
\end{bmatrix} = \begin{bmatrix} 4.35443562594961881563 \\
2.39389146204407615117 \\
\end{bmatrix},
\begin{bmatrix}
x(3) \\
y(3)
\end{bmatrix} = \begin{bmatrix} 2.11505288065117566506 \\
0.52937555595567336001 \\
\end{bmatrix},
\begin{bmatrix}
x(4) \\
y(4)
\end{bmatrix} = \begin{bmatrix} 2.29431416810009081225 \\
1.33175191382089012750 \\
\end{bmatrix},
\begin{bmatrix}
x(5) \\
y(5)
\end{bmatrix} = \begin{bmatrix} 1.85902085285052227134 \\
4.21660738720576932899 \\
\end{bmatrix}.
\]

Fig. 8. The test problems Ur, where \( r = 1, 2, 3, 4, \) and \( 5 \); accurate up to \( 10^{-20} \) values of \( (x(n), y(n)) \), \( n = 1, 2, 3, 4, \) and \( 5 \) (middle left); the trajectory \( (x(t), y(t)) \) for \( 0 < t < 5 \) (middle right); and efficiency curves for problems U1, U2, and U4 (bottom panel): the (Dormand & Prince, 1980, tab. 2) pair (dash-dotted curve), the (Bogacki & Shampine, 1996) pair (dotted curve), the pair in Figure 3 (dashed curve), the pair in Figure 5 (solid curve), the (Verner, 2010, tab. 3) pair (thin dotted curve).
FIG. 9. One step \( h = \frac{\pi}{4} \) for the system \( \frac{dx}{dt} = -y, \frac{dy}{dt} = x \) with initial conditions \( x(0) = 1, y(0) = 0 \): the intermediate positions \( X_i, 1 \leq i \leq s \), and the error \( (x(\frac{\pi}{2} \theta) - \cos(\frac{\pi}{2} \theta), y(\frac{\pi}{2} \theta) - \sin(\frac{\pi}{2} \theta)) \), where \( 0 \leq \theta \leq 1 \), made by an interpolant. Nodes are shown by the radial ticks. Open circles correspond to additional stages that are used to construct an interpolant. Ticks on the error curves correspond to \( \theta = \frac{1}{12}, \frac{1}{6}, \frac{1}{4}, ..., \frac{11}{12} \). In case of the (Dormand & Prince, 1980, tab. 2) pair, the 5th order interpolant is the (Calvo et al., 1990) one, while the 4th order interpolant with \( u = s = 7 \) is from \texttt{ntrp45.m} code that is a part of MATLAB® software.
are also shown for the step size $h = \frac{\pi}{4}$. For the (Dormand & Prince, 1980, tab. 2) pair, $(q^2)_i = 0$ for all $i \neq 2$. In the leading order the deviation of $X_i$ from $x(t + c_i h)$ is controlled by $T_3(a_i, c_i) = |a_i A c_i c_i^3 / 6|$ that for $i = 3, 4, 5, 6$ is equal to $\frac{9}{2000} = 0.0045$, $\frac{28}{375} = 0.0746...$, $\frac{2536}{1935} = 0.231...$, and $\frac{71}{300} = 0.231...$, respectively. Relatively large values of $T_3(a_5, c_5)$ and $T_3(a_6, c_6)$ explain the observed deviation of $X_5$ and $X_6$ from $(\cos \frac{8h}{9}, \sin \frac{8h}{9})$ and $(\cos h, \sin h)$. Even for $h = \frac{\pi}{2}$ the position update $x(t + h) = X_7$ is remarkably close to the exact value, though.

5. Conclusion

Utilizing 9 stages, it is possible to construct embedded (4,5) pairs of explicit Runge–Kutta methods with FSAL property that are as cost-efficient as the best known conventional (i.e., interpolant is either of order 4 or would require extra stages) pairs (e.g., the Dormand–Prince and Bogacki–Shampine ones), but with the benefit of having continuous formulae or interpolants of order 5 available at no additional cost.

References

ASCHER, U. M. & PETZOLD, L. R. (1998) Computer methods for ordinary differential equations and differential-algebraic equations, SIAM.

BOGACKI, P. (1990) Efficient Runge–Kutta pairs and their interpolants. Ph.D. thesis, Department of Mathematics, Southern Methodist University, Dallas, TX, USA.

BOGACKI, P. & SHAMPINE, L. F. (1996) An efficient Runge–Kutta (4,5) pair. Computers & Mathematics with Applications, 32 (6) 15–28.

BRANKIN, R. W., GLADWELL, I. & SHAMPINE, L. F. (1993) RKSUITE: A suite of explicit Runge-Kutta codes. Contributions in numerical mathematics, ed. R. P. Agarwal, World Scientific, pp. 41–53.

BUTCHER, J. C. (1964) On Runge–Kutta processes of high order. Journal of the Australian Mathematical Society, 4 (2) 179–194.

BUTCHER, J. C. (2008) Numerical methods for ordinary differential equations, 2nd ed., John Wiley & Sons Ltd.

BUTCHER, J. C. & WANNER, G. (1996) Runge–Kutta methods: some historical notes. Applied Numerical Mathematics, 22 (1–3) 113–151.

CALVO, M., MONTIJANO, J. I. & RANDEZ, L. (1990) A fifth-order interpolant for the Dormand and Prince Runge–Kutta method. Journal of Computational and Applied Mathematics, 29 (1) 91–100.

DORMAND, J. R., LOCKYER, M. A., McGORRIGAN, N. E. & PRINCE, P. J. (1989) Global error estimation with Runge–Kutta triples. Computers & Mathematics with Applications, 18 (9) 835–846.

DORMAND, J. R. & PRINCE, P. J. (1978) New Runge–Kutta algorithms for numerical simulation in dynamical astronomy. Celestial Mechanics, 18 (3) 223–232.

DORMAND, J. R. & PRINCE, P. J. (1980) A family of embedded Runge-Kutta formulae. Journal of Computational and Applied Mathematics, 6 (1) 19–26.
REFERENCES

ENRIGHT, W.H., JACKSON, K. R., NØRSETT, S. P. & THOMSEN, P. G. (1986) Interpolants for Runge–Kutta formulas. ACM Transactions on Mathematical Software, 12 (3) 193–218.

FEHLBERG, E. (1969) Low-order classical Runge–Kutta formulas with stepsize control and their application to some heat transfer problems. NASA Technical Report R-315.

FEHLBERG, E. (1970) Klassische Runge–Kutta-Formeln vierten und niedrigerer Ordnung mit Schrittweiten-Kontrolle und ihre Anwendung auf Wärmeleitungsprobleme. Computing 6, 61–71.

Hairer, E., Norsett, S. P. & Wanner, G. (2008) Solving ordinary differential equations I: nonstiff problems, 2nd ed., Springer.

Horn, M. K. (1983) Fourth- and fifth-order, scaled Runge–Kutta algorithms for treating dense output. SIAM Journal on Numerical Analysis, 20 (3) 558–568.

Hull, T. E., Enright, W.H., Fellen, B. M. & Sedwick, A. E. (1972) Comparing numerical methods for ordinary differential equations. SIAM Journal on Numerical Analysis, 9 (4) 603–637.

Iserles, A. (2009) A first course in the numerical analysis of differential equations, 2nd ed., Cambridge University Press.

Owren, B. & Zennaro, M. (1991) Order barriers for continuous explicit Runge–Kutta methods. Mathematics of Computation, 56 (194) 645–661.

Owren, B. & Zennaro, M. (1992) Derivation of efficient, continuous, explicit Runge–Kutta methods. SIAM Journal on Scientific and Statistical Computing, 13 (6) 1488–1501.

Sarafyan, D. (1984) Continuous approximate solution of ordinary differential equations and their systems. Computers & Mathematics with Applications, 10 (2) 139–159.

Shampine, L. F. (1986) Some practical Runge–Kutta formulas. Mathematics of Computation, 46 (173) 135–150.

Sharp, P. W. & Smart, E. (1993) Explicit Runge–Kutta pairs with one more derivative evaluation than the minimum. SIAM Journal on Scientific Computing, 14 (2) 338–348.

Sharp, P. W. & Verner, J. H. (1994) Completely imbedded Runge–Kutta pairs. SIAM Journal on Numerical Analysis, 31 (4) 1169–1190.

Verner, J. H. (1978) Explicit Runge–Kutta methods with estimates of the local truncation error. SIAM Journal on Numerical Analysis, 15 (4) 772–790.

Verner, J. H. (1991) Some Runge–Kutta formula pairs. SIAM Journal on Numerical Analysis, 28 (2) 496–511.

Verner, J. H. (1993) Differentiable interpolants for high-order Runge–Kutta methods. SIAM Journal on Numerical Analysis, 30 (5) 1446–1466.

Verner, J. H. (2010) Numerically optimal Runge–Kutta pairs with interpolants. Numerical Algorithms, 53 (3) 383–396.

Verner, J. H. & Zennaro, M. (1995) Continuous explicit Runge–Kutta methods of order 5. Mathematics of Computation, 64 (211) 1123–1146.