

Continuous Explicit Runge-Kutta methods

Brynjulf Owren and Marino Zennaro

Norwegian Institute of Technology and University of L'Aquila

1 Introduction

Consider the initial value problem (IVP) for ordinary differential equations (ODEs)

$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0 \quad (1.1)$$

where $y_0, y(x) \in R^m$, x is a real variable and $f : R \times R^m \rightarrow R^m$. The solution y is assumed to be as smooth as necessary and is sought in the interval $[x_0, x_f]$.

Among the large variety of methods available for the numerical solution of (1.1), one can distinguish between *continuous* and *discrete* methods [1]. The class of *explicit Runge-Kutta methods* has traditionally belonged to the latter group. Recently, however, many authors have investigated continuous extensions of *one step* methods (see e.g. Bellen and Zennaro[2], Enright et al.[1], Horn[3], Nørsett and Wanner[4], Shampine[5], Zennaro[6, 7], as well as the book by Hairer et al.[8]). Some of these authors use the strategy of supplying an already existing discrete Runge-Kutta method with an interpolant (possibly by adding some stages) in order to obtain a continuous method with the desired accuracy. Another possibility is of course to construct *continuous explicit Runge-Kutta methods* (CERK methods) without regard to any existing discrete method. In this paper, we use both the above strategies, considering methods of the general form

$$K_i = f(x_0 + c_i h, y_0 + h \sum_{j=1}^{i-1} a_{ij} K_j), \quad i = 1, \dots, \nu \quad (1.2a)$$

$$u(x_0 + \theta h) = y_0 + h \sum_{i=1}^{\nu} b_i(\theta) K_i, \quad \theta \in [0, 1] \quad (1.2b)$$

where $u(x_0 + \theta h)$ is a *continuous approximation* to $y(x)$ in the interval $[x_0, x_0 + h]$ and $b_i(\theta)$, $i = 1, \dots, \nu$, are polynomials of degree $\leq d$ where d is a positive integer. We shall also require $c_i = \sum_{j=1}^{i-1} a_{ij}$, and $b_i(0) = 0$ for $i = 1, \dots, \nu$. The coefficients a_{ij} define a strictly lower triangular

$\nu \times \nu$ -matrix A . It follows that one obtains a discrete method simply by putting $\theta = 1$ in (1.2b) leading to $y_1 := u(x_0 + h)$. Clearly, using the value y_1 to advance integration implies global continuity of the continuous approximation.

We define the *uniform order* (which we shall simply refer to as the *order*) as the greatest integer p for which

$$\max_{0 \leq \theta \leq 1} |y(x_0 + \theta h) - u(x_0 + \theta h)| = O(h^{p+1}).$$

Here $|\cdot|$ stands for any norm on R^m .

In Section 2, we shall be concerned with finding lower bounds for the numbers

$$CEN(p) := \min_{m(\nu) \in M_p} \nu$$

where $m(\nu)$ is a CERK method with ν stages, and M_p is the set of all CERK methods of order p . The numbers $CEN(p)$ are similar to the famous Butcher barriers $EN(p)$ for the discrete case. We provide some theorems that can be used to find such lower bounds in general, and we solve the problem completely up to and including order $p = 5$. For proofs and further details, we refer to [9].

In Section 3, we consider continuous extensions of the well-known Dormand-Prince (4,5) discrete embedded pair.

2 Lower bounds for CEN(p)

Letting $CEN(p)$ and $EN(p)$ be the minimum number of stages required for p th order continuous and discrete explicit Runge-Kutta methods respectively, the following result is obvious,

$$CEN(p) \geq EN(p).$$

Moreover, from the literature quoted in Section 1, one can extract the following bounds

$$\begin{aligned} CEN(1) &= 1, & CEN(2) &= 2, & 3 &\leq CEN(3) \leq 4, \\ 5 &\leq CEN(4) \leq 6, & 6 &\leq CEN(5) \leq 9. \end{aligned}$$

Subsequently, we shall make extensive use of the theory developed by Butcher [10, 11] without giving specific references. In doing this we rely on the reader's acquaintance with trees, order conditions and related topics. We recommend the books by Butcher [12] and Hairer et al.[8] as background material, and we will use the notation of the latter.

We begin by considering the continuous order conditions

$$\sum_{j=1}^{\nu} b_j(\theta) \Phi_j(t) = \frac{\theta^{\rho(t)}}{\gamma(t)} \text{ for all trees } t \text{ such that } \rho(t) \leq p, \quad (2.1)$$

where $\Phi_j(t)$ is the j th elementary weight of the tree t , $\rho(t)$ is the order of t , and $\gamma(t)$ is a coefficient depending on the tree t . Furthermore, define $z_j(\theta) := b_j'(\theta)$, $j = 1 \dots \nu$, let N_p be the number of trees such that $\rho(t_i) \leq p$ and supply subscripts to the trees such that $i > j$ if $\rho(t_i) > \rho(t_j)$. Then (2.1) becomes

$$\sum_{j=1}^{\nu} \phi_{ij} z_j(\theta) = \frac{\rho(t_i) \theta^{\rho(t_i)-1}}{\gamma(t_i)}, \quad i = 1 \dots N_p, \quad (2.2)$$

where $\phi_{ij} := \Phi_j(t_i)$. Now, write $z_j(\theta)$ as $\sum_{k=0}^{p-1} z_{jk} \theta^k$, the right hand side of (2.2) as $\sum_{l=0}^{p-1} q_{il} \theta^l$ and define the matrices $\Phi := ((\phi_{ij})) \in R^{N_p \times \nu}$, $Z := ((z_{jk})) \in R^{\nu \times p}$ and $Q := ((q_{il})) \in R^{N_p \times p}$. The order conditions can now be written as a matrix equation

$$\Phi Z = Q. \quad (2.3)$$

Notice that the matrix Φ depends solely on the lower-triangular $\nu \times \nu$ matrix A and the order p . The following theorem allows us to exclude a considerable number of candidates among the CERK methods in our search for $CEN(p)$.

Theorem 1 *If there exists $A \in R^{\nu \times \nu}$ and $Z \in R^{\nu \times p}$ such that (2.3) is satisfied for some order p with $\text{rank}(\Phi) = \rho < \nu$, then there exists $A^* \in R^{\rho \times \rho}$ with a corresponding $\Phi^* \in R^{N_p \times \rho}$ and $Z^* \in R^{\rho \times p}$ satisfying $\Phi^* Z^* = Q$.*

Proof. See [9].

Consequently, when investigating $CEN(p)$, we need only consider $\nu \times \nu$ matrices A such that the corresponding matrix Φ satisfies $\text{rank}(\Phi) = \nu$. We shall denote by \mathcal{M}_*^p the set of these matrices that also cause (2.3) to be satisfied for some Z .

In order to provide tools for determining the lower bounds for $CEN(p)$, we must take a closer look at the conditions (2.2) and introduce a few more definitions. In (2.2) we shall distinguish between the p *primary conditions*

$$\sum_{j=1}^{\nu} c_j^{r-1} z_j(\theta) = \theta^{r-1}, \quad r = 1, \dots, p$$

and the remaining $N_p - p$ conditions which we shall call the *secondary conditions*. For a $\nu \times \nu$ matrix $A \in \mathcal{M}_*^p$ with $p \geq 3$, we introduce the following equivalence relation on the set of indices $\{1, \dots, \nu\}$:

$$i \equiv j \quad \text{if and only if} \quad c_i = c_j.$$

There are ν^* equivalence classes S_1, \dots, S_{ν^*} , and we assume without restrictions that $1 \in S_1$ (i.e. $c_i = 0 \Leftrightarrow i \in S_1$) and that $2 \in S_2$. We shall

call a *good index set* for A either the empty set \emptyset or any non-empty subset of $\{3, \dots, \nu\}$ which elements do not belong to more than $p - 3$ equivalence classes among S_3, \dots, S_ν . The following theorem has proved useful for obtaining lower bounds for $CEN(p)$.

Theorem 2 *Let $A \in \mathcal{M}_*^p$ with $p \geq 3$, and let N secondary conditions from (2.2) be linearly independent. Let S be the set formed by the indices $j \geq 3$ of the polynomials $z_j(\theta)$ which are not explicitly involved by these N conditions (possibly $S = \emptyset$). Then*

$$\dim(A) \geq N + s + 2,$$

where s is the cardinality of S . Moreover, if S is a good index set for A , then

$$\dim(A) \geq N + s + 3.$$

Proof. Consult [9] for a proof and further details.

We have applied Theorem 2 to obtain lower bounds for $CEN(p)$, and by comparing these results to already existing methods (see the papers quoted in Section 1), we have been able to determine the following values for $CEN(p)$, $p = 1, \dots, 5$.

Order p	CEN(p)
1	1
2	2
3	4
4	6
5	8

Since we did not find any 8-stage CERK methods of order 5 in the literature, we had to construct some ourselves. One example is the following. ($c|A$)

0							
1/4	1/4						
1/4	1/8	1/8					
1/2	0	-1/2	1				
1/2	1/12	0	1/3	1/12			
3/4	0	-9/8	3/2	-3/4	9/8		
1	0	4/5	0	-3/5	0	4/5	
1	1/6	0	0	4/15	2/5	0	1/6

where the continuous weights are given by

$$\begin{aligned}
 b_1(\theta) &= \frac{32}{15}\theta^5 - \frac{20}{3}\theta^4 + \frac{70}{9}\theta^3 - \frac{25}{6}\theta^2 + \theta \\
 b_2(\theta) &= 0 \\
 b_3(\theta) &= -\frac{128}{15}\theta^5 + 24\theta^4 - \frac{208}{9}\theta^3 + 8\theta^2 \\
 b_4(\theta) &= \frac{32}{5}\theta^5 - 12\theta^4 + \frac{16}{3}\theta^3 \\
 b_5(\theta) &= \frac{32}{5}\theta^5 - 20\theta^4 + 20\theta^3 - 6\theta^2 \\
 b_6(\theta) &= -\frac{128}{15}\theta^5 + \frac{56}{3}\theta^4 - \frac{112}{9}\theta^3 + \frac{8}{3}\theta^2 \\
 b_7(\theta) &= -\frac{20}{3}\theta^5 + 15\theta^4 - \frac{95}{9}\theta^3 + \frac{5}{2}\theta^2 \\
 b_8(\theta) &= \frac{44}{5}\theta^5 - 19\theta^4 + 13\theta^3 - 3\theta^2
 \end{aligned}$$

3 Continuous extensions of the Dormand-Prince (4,5) discrete embedded pair

The Dormand-Prince embedded pairs are among the most popular discrete explicit Runge-Kutta methods. We shall discuss the various continuous extensions to which one can supply the Dormand-Prince (4,5) (DP45) [13] embedded pairs, using an approach different from the ones in [14] and [5, 15]. The coefficients of DP45 are as follows.

0								
$\frac{1}{5}$	$\frac{1}{5}$							
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$						
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$					
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$				
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$			
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$		
y_1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0	
\hat{y}_1	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$	

(3.1)

The method y_1 is the one of order 5. From [8] we find a fourth order continuous extension which is obtained by using the first six stages of (3.1) and which is important for the fifth order interpolant constructed later.

$$u_0(x_0 + \theta h) = y_0 + h \sum_{i=1}^6 b_i(\theta) K_i$$

$$\begin{aligned}
b_1(\theta) &= \theta - \frac{1337}{480}\theta^2 + \frac{1039}{360}\theta^3 - \frac{1163}{1152}\theta^4 \\
b_2(\theta) &\equiv 0 \\
b_3(\theta) &= \frac{4216}{1113}\theta^2 - \frac{18728}{3339}\theta^3 + \frac{7580}{3339}\theta^4 \\
b_4(\theta) &= -\frac{27}{16}\theta^2 + \frac{9}{2}\theta^3 - \frac{415}{192}\theta^4 \\
b_5(\theta) &= -\frac{2187}{8480}\theta^2 + \frac{2673}{2120}\theta^3 - \frac{8991}{6784}\theta^4 \\
b_6(\theta) &= \frac{33}{35}\theta^2 - \frac{319}{105}\theta^3 + \frac{187}{84}\theta^4
\end{aligned}$$

If we use also the 7th stage, it turns out that there is a 3-parameter family of 4th order continuous extensions of DP45. As for the 5th order case, we can use the theory from the previous section to deduce the following result.

Theorem 3 *There are no 5th order 8-stage continuous extensions of DP45.*

However, in DP45 the last stage of step k is identical to the first stage of step $k + 1$, which means that the effective cost for each step corresponds to 8 stages. Moreover, DP45 supports facilities for step size control. So, if it is possible to construct 9-stage 5th order continuous extensions of DP45, we will obtain CERK methods which are competitive with the 8-stage methods of the previous section. Indeed, by [14] and [5, 15] such 9-stage extensions exist: Also following [1], with u_0 as above we choose two additional points $c_8, c_9 \in [0, 1]$ and construct a 5th degree polynomial u_1 satisfying

$$\begin{aligned}
u_1(x_0) &= y_0, \quad u_1'(x_0) = K_1, \\
u_1(x_0 + h) &= y_1, \quad u_1'(x_0 + h) = K_7, \\
u_1'(x_0 + c_8h) &= f(x_0 + c_8h, u_0(x_0 + c_8h)) = K_8, \\
u_1'(x_0 + c_9h) &= f(x_0 + c_9h, u_0(x_0 + c_9h)) = K_9.
\end{aligned}$$

The coefficients c_8 and c_9 must be chosen with some care as $u_1(x_0 + \theta h)$ is an Hermite-Birkhoff interpolant. The following result is easily obtained.

Theorem 4 *The polynomial $u_1(x_0 + \theta h)$ defines a 5th order CERK method if*

$$\begin{aligned}
c_8, c_9 &\neq 0, & c_8, c_9 &\neq 1, \\
c_8 &\neq c_9, & c_8 &\neq (c_9 - \frac{3}{5})/(2c_9 - 1).
\end{aligned} \tag{3.1}$$

It is now natural to ask how one should choose c_8 and c_9 in order to minimize the local truncation error. u_1 can be written in the form

$$u_1(x_0 + \theta h) = y_0 + h \sum_{i=1}^9 d_i(c_8, c_9; \theta) K_i.$$

There are 20 trees, $t_i, i = 18, \dots, 37$, of order 6, and thereby 20 corresponding elementary differentials, $F(t_i)$ with elementary weights $\phi(t_i)$. We obtain the following expression for the principal error term :

$$y(x_0 + \theta h) - u_1(x_0 + \theta h) = \frac{h^6}{6!} \sum_{i=18}^{37} \alpha(t_i) \epsilon(t_i) F(t_i)(y_0)$$

where

$$\epsilon(t_i) = \theta^6 - \gamma(t_i) \sum_{j=1}^9 d_j \phi_j(t_i).$$

$\epsilon(t_i)$ are called the *error coefficients*. An appropriate choice for the coefficients c_8 and c_9 can be obtained by minimizing the error coefficients in some norm on R^{20} , e.g. find c_8 and c_9 that minimizes the functional

$$K(c_8, c_9) = \max_{18 \leq i \leq 37} \{ \max_{\theta \in [0,1]} |e(t_i)(c_8, c_9; \theta)| \}.$$

Numerical experiments show that apart from values of c_8, c_9 lying very close to the singularities (see (3.1)), $K(c_8, c_9)$ is constant. In fact, for most values of c_8 and c_9 we have

$$K(c_8, c_9) = \max_{\theta \in [0,1]} |e(t_{37})(c_8, c_9; \theta)|$$

where t_{37} is the tree corresponding to $\phi(t_{37}) = A^5 u$ with $u = (1, \dots, 1)^T$. It can be shown that

$$e(t_{37})(c_8, c_9; \theta) = \theta^6 + \frac{186}{25} \theta^5 - \frac{432}{25} \theta^4 + \frac{216}{25} \theta^3$$

which is indeed independent of c_8 and c_9 . The interpolant above can be rewritten into the following canonical form. For convenience, the parameters c_8 and c_9 have been replaced by s and t respectively.

$$u_1(\theta) = \phi_0(\theta) y_0 + \phi_1(\theta) y_1 + \psi_0(\theta) h k_1 + \psi_s(\theta) h k_8 + \psi_t(\theta) h k_9 + \psi_1(\theta) h k_7$$

with

$$\begin{aligned} \phi_0(\theta) &= (\theta - 1)^2(1 + 2\theta) + \frac{3}{5} \xi(\theta - 1)^2 \theta^2 (4\theta - 5(s + \hat{t})) \\ \phi_1(\theta) &= \theta^2(3 - 2\theta) - \frac{3}{5} \xi(\theta - 1)^2 \theta^2 (4\theta - 5(s + \hat{t})) \\ \psi_0(\theta) &= \theta(\theta - 1)^2 [1 + \frac{\xi \theta}{s \hat{t}} (\frac{2}{5} ((3t - 1)s + \frac{1}{2} - t)) \theta + \frac{1}{2} (s(s + \hat{t})(1 - 3t) + \hat{t}t)] \\ \psi_1(\theta) &= \theta^2(\theta - 1) [1 + \frac{\xi(\theta - 1)}{(s - 1)(\hat{t} - 1)} (\frac{2}{5} ((3t - 2)s + \frac{3}{2} - 2t)) \theta \\ &\quad + \frac{1}{2} (s(2 - 3t)(s + \hat{t} - 1) + 2\hat{t}(t - 1))] \\ \psi_s(\theta) &= \theta^2(\theta - 1)^2 \frac{\xi}{s(s - 1)(s - \hat{t})} (\frac{2}{5} (t - \frac{1}{2}) \theta - \frac{1}{2} \hat{t}t) \\ \psi_t(\theta) &= -\theta^2(\theta - 1)^2 \frac{\xi}{\hat{t}(\hat{t} - 1)(s - \hat{t})} (\frac{2}{5} (s - \frac{1}{2}) \theta - \frac{1}{2} \hat{s}s) \end{aligned}$$

where

$$\xi = [(2t - 1)s - \hat{t}]^{-1}$$

and

$$\begin{aligned} \hat{t} &= t - \frac{3}{5} \\ \hat{s} &= s - \frac{3}{5} \end{aligned}$$

The error has been minimized with respect to c_8 and c_9 for several test problems, using a least square criterion, and it is found that the choices $c_8 = 0.2$ and $c_9 = 0.5$ is almost optimal for most of these problems. However, as one would expect from the error analysis above, this optimum is very flat.

References

- [1] Enright, W.H., Jackson, K.R., Nørsett S.P. and Thomsen P.G. (1986). Interpolants for Runge-Kutta formulas. *ACM Transactions on Mathematical Software*, **12**, 193-218.
- [2] Bellen, A. and Zennaro, M. (1988). Stability of interpolants for Runge-Kutta methods. *SIAM J. Numer. Anal.* **25**, 411-432.
- [3] Horn, M.K. (1983) Fourth- and fifth-order, scaled Runge-Kutta algorithms for treating dense output. *SIAM J. Numer. Anal.*, **20**, 558-568.
- [4] Nørsett, S.P. and Wanner, G. (1981). Perturbed Collocation and Runge-Kutta methods. *Numer. Math.*, **38**, 193-208.
- [5] Shampine, L.F. (1985). Interpolation for Runge-Kutta methods. *SIAM J. Numer. Anal.*, **22**, 1014-1027.
- [6] Zennaro, M. (1986). Natural Continuous Extensions of Runge-Kutta methods. *Math. Comp.*, **46**, 119-133.
- [7] Zennaro, M. (1988) Natural Runge-Kutta and projection methods. *Numer. Math.*, **53**, 423-438.
- [8] Hairer, E., Nørsett, S.P. and Wanner, G. (1987). *Solving Ordinary Differential Equations I, Nonstiff Problems*. Springer Verlag.
- [9] Owren, B. and Zennaro M. (1989). *Order barriers for Continuous Explicit Runge-Kutta Methods*. Report no. 2, Division of Mathematical Sciences, Norwegian Institute of Technology, Trondheim, Norway.
- [10] Butcher, J.C. (1964). Coefficients for the study of Runge-Kutta integration processes. *J. Austral. Math. Soc.*, **3**, 233-243.
- [11] Butcher, J.C. (1964). Implicit Runge-Kutta processes. *Math. Comp.* **18**, 50-64

- [12] Butcher, J.C. (1987). *The numerical analysis of ordinary differential equations*, J. Wiley & Sons.
- [13] Dormand, J.R. and Prince, P.J. (1980). A family of embedded Runge-Kutta formulae. *J. Comput. Appl. Math.*, **6**.
- [14] Calvo, M., Montijano, J.I. and Rández, L. A fifth order interpolant for the Dormand and Prince Runge-Kutta method. *J. Comput. Appl. Math.*, To appear.
- [15] Shampine, L.F. (1986). Some Practical Runge-Kutta Formulas. *Math. of Comp.*, **173**, 135-150.