Derivation of Efficient Continuous Explicit Runge-Kutta Methods

by

Brynjulf Owren[1]
Computer Science department
University og Toronto
Toronto, ONT M5S 1A4 (Canada).

Marino Zennaro[2]
Dipartimento di Matematica Pura ed Applicada,
Universita' dell'Aquila,
I-67100 L'Aquila (Italy).

Abstract

Continuous Explicit Runge-Kutta methods with the minimal number of stages are considered. These methods are continuously differentiable if and only if one of the stages is the FSAL evaluation. A characterization of a subclass of these methods is developed for order 3,4 and 5. It is shown how the free parameters of these methods can be used either to minimize the continuous truncation error coefficients or to maximize the stability region. As a representative for these methods the 5th order method with minimized error coefficients is chosen, supplied with an error estimation method, and analysed by using the DETEST software. The results are compared with a similar implementation of the Dormand-Prince 5(4) pair with interpolant, showing a significant advantage to the new method for the chosen problems.

[1]Supported by the Information Technology Research Center of Ontario, the Natural Science and Engineering Research Council of Canada and by the norwegian research council, NAVF.
[2]Supported by the Italian M.P.I., funds 40 %

1 Introduction

We consider the first order system of differential equations

$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0, \quad x_0 \le x \le x_e \tag{1.1}$$

where $f : \mathbf{R} \times \mathbf{R}^{\mathrm{m}} \to \mathbf{R}^{\mathrm{m}}$ defines m generally non-linear equations, and y_0 is the m-vector of initial values. We are interested in a continuous approximation to the solution $y(x)$ for $x \in [x_0, x_e]$. A possible way to obtain such an approximation is to apply a continuous explicit Runge-Kutta method (CERK method) as proposed by the authors in [16]. One may expect such a method to be appropriate in the case when (1.1) is non-stiff. CERK methods may also be used for solving certain types of functional differential equations, like delay differential equations for which retarded arguments in the right hand side of (1.1) might be estimated by continuous extensions over previous subintervals. The continuous approximation $u(x)$ is obtained on $[x_0, x_e]$ by using a mesh $\{x_0 < x_1 < \cdots < x_N = x_e\}$ and computing polynomials $u_{n+1}(x_n + \theta h_n),\ n = 0, \ldots, N-1$ such that $u(x) = u_{n+1}(x_n + \theta h_n)$ for $x_n \le x = x_n + \theta h_n \le x_{n+1}$ where $h_n = x_{n+1} - x_n$. The continuity assumption on $u(x)$ requires that $u_n(x_n) = u_{n+1}(x_n) := y_n,\ n = 1, \ldots, N-1$. The general form of u_{n+1} is

$$K_i = f(x_n + c_i h_n, y_n + h_n \sum_{j=1}^{i-1} a_{ij} K_j), \quad i = 1, \ldots, s$$

$$u_{n+1}(x_n + \theta h_n) = y_n + h_n \sum_{i=1}^{s} b_i(\theta) K_i, \quad \theta \in [0, 1],$$

where $b_i(\theta),\ i = 1, \ldots, s$, are polynomials of degree $\le d$ for some positive integer d. We shall also require $c_i = \sum_{j=1}^{i-1} a_{ij}$, and $b_i(0) = 0$ for $i = 1, \ldots, s$, the last condition is necessary for continuity. Henceforth we shall denote by A the strictly lower triangular matrix defined by the coefficients a_{ij}. Observe that a conventional Runge-Kutta method is obtained by putting $y_{n+1} = u_{n+1}(x_n + h_n)$. In fact, a CERK method is equivalent to a Runge-Kutta method supplied with an interpolant. On surveying the literature one finds such interpolants for most of the commonly used Runge-Kutta formulas, e.g. Shampine [17, 18], Dormand and Prince [6, 7], Calvo et al. [4] and Horn [13]. Enright et al. [8] provides a general technique for constructing interpolants to a Runge-Kutta formula while Zennaro [20] discusses natural continuous extensions of Runge-Kutta methods which are especially suited for functional differential equations when certain restrictions are imposed on the mesh. In a recent paper Verner [19] elaborates differentiable interpolants of higher order.

An important issue in many of these papers has been whether or not the continuous approximation should yield the same order of consistency in the interior of the step as at the endpoints. Following [16] we shall define the uniform order, or simply the order of a CERK method as the greatest integer p for which

$$\max_{0 \le \theta \le 1} |y_{n+1}(x_n + \theta h_n) - u_{n+1}(x_n + \theta h_n)| = O(h_n^{p+1}) \tag{1.2}$$

where $y_{n+1}(x)$ is the local solution to the initial value problem $y'_{n+1}(x) = f(x, y_{n+1}(x))$, $y_{n+1}(x_n) = y_n$. Here $|\cdot|$ can be any norm on $\mathbf{R}^{\mathrm{m}}$. Of course the order q at the endpoints satisfies $q \geq p$ and in this paper we shall not impose the possible additional requirement that $q > p$ for any of our methods. We shall always require (see [16]) that the degree d of the polynomials $b_i(\theta)$, $i = 1, \ldots, s$, satisfies $d \leq p$. Again according to [16] one may for each order p define the order barrier as the smallest number of stages CEN(p) that a CERK method satisfying (1.2) can have. These order barriers were derived for $p \leq 5$, it was found that CEN(1)=1, CEN(2)=2, CEN(3)=4, CEN(4)=6 and CEN(5)=8. The first examples of 5th order CERK methods with only 8 stages were developed. It should be made clear that these new continuous methods with the minimal number of stages may only be expected to be cheap when the continuous approximation is required along the entire interval of integration. In some applications the continuous approximation is needed only occasionally, e.g. when it is used to locate discontinuities, see e.g. [9]. In such cases it is recommended that one uses a discrete method that can be supplied with a continuous extension, possibly at more than the minimal cost.

In this paper, we consider CERK methods with CEN(p) stages for $p = 3, 4, 5$ with the additional property that they are $C^{(1)}$ continuous. We give a complete recipe for the construction of some if not all such methods, and we present pairs of formulas with optimized error constants and show how the regions of absolute stability may be optimized. Finally we present some results based on tests made with the DETEST package [10]. These tests are made with a 5th order representative of the CERK methods described in this paper, displaying the properties of the underlying discrete method.

2 Preliminary results

Henceforth we shall make extensive use of the theory of rooted trees and order conditions developed by Butcher [1, 2]. From [16] we find the continuous version of the order conditions

$$\sum_{j=1}^{s} b_j(\theta)\Phi_j(t) = \frac{\theta^{\rho(t)}}{\gamma(t)} \text{ for all trees } t \text{ such that } \rho(t) \leq p, \tag{2.3}$$

where $\Phi_j(t)$ is the jth elementary weight for the tree t, $\rho(t)$ is the order of t, and $\gamma(t)$ is a coefficient depending on the tree t. For each $r \geq 1$, let n_r be the number of trees such that $\rho(t) = r$. Thus, a CERK method of order p must satisfy N_p conditions where $N_p = \sum_{r=1}^{p} n_r$. We number the N_p trees t increasingly in terms of $\rho(t)$, such that $\rho(t_i) > \rho(t_j)$ only if $i > j$. This ordering is not unique. By putting $z_j(\theta) := b'_j(\theta)$, $j = 1, \ldots, s$, (2.3) can be written as

$$\sum_{j=1}^{s} \phi_{ij} z_j(\theta) = \frac{\rho(t_i)\theta^{\rho(t_i)-1}}{\gamma(t_i)}, \quad i = 1, \ldots, N_p, \tag{2.4}$$

where $\phi_{ij} = \Phi_j(t_i)$. Moreover, by writing

$$z_j(\theta) = \sum_{k=0}^{p-1} z_{jk}\theta^k \text{ and } \frac{\rho(t_i)\theta^{\rho(t_i)-1}}{\gamma(t_i)} = \sum_{l=0}^{p-1} q_{il}\theta^l,$$

and by defining the $N_p \times s$ matrix $\Phi := ((\phi_{ij}))$, the $s \times p$ matrix $Z := ((z_{jk}))$ and the $N_p \times p$ matrix $Q := ((q_{il}))$, (2.4) becomes

$$\Phi Z = Q.$$

The $N_p \times s$ matrix Φ depends on the $s \times s$ matrix A of the coefficients of the RK-method, whereas the $N_p \times p$ matrix Q is independent of A. For convenience we introduce the mappings

$$F_p : \bigcup_{s\geq 1} \mathcal{L}(\mathbf{R}^s, \mathbf{R}^s) \longrightarrow \bigcup_{s\geq 1} \mathcal{L}(\mathbf{R}^s, \mathbf{R}^{N_p}) \text{ such that } F_p(A) := \Phi,$$

and

$$G_p : \bigcup_{s\geq 1} \mathcal{L}(\mathbf{R}^s, \mathbf{R}^s) \longrightarrow \bigcup_{s\geq 1} \mathcal{L}(\mathbf{R}^{s+p}, \mathbf{R}^{N_p}) \text{ such that } G_p(A) := \Phi|Q,$$

where $\Phi|Q$ is the $N_p \times (s+p)$ matrix obtained by attaching the rows of Q to the rows of Φ. It was pointed out in [16] that a CERK method is of order p if and only if $\text{rank}(F_p(A)) = \text{rank}(G_p(A))$. These methods constitute a set

$$\mathcal{M}^p := \{A \in \textstyle\bigcup_{s\geq 1} \mathcal{L}(\mathbf{R}^s, \mathbf{R}^s) | \quad \begin{array}{l} A \text{ is strictly lower triangular} \\ \text{rank}(F_p(A)) = \text{rank}(G_p(A))\}. \end{array}$$

Consider the subset of $\mathcal{M}^p$

$$\mathcal{M}^p_* := \{A \in \mathcal{M}^p | \text{rank}(F_p(A)) = s\},$$

where s is the order of the matrix A. It was proved in [16] that a necessary condition for a CERK method of order p to have CEN(p) stages is that $A \in \mathcal{M}^p_*$. Thus, henceforth we shall only consider methods for which $A \in \mathcal{M}^p_*$.

3 Optimal $C^{(1)}$ approximations

A property possessed by all explicit Runge-Kutta methods is that the first stage of the step from x_{n+1} to x_{n+2} is given by $K_1 = f(x_{n+1}, y_{n+1})$. Many methods take advantage of this property by also using this stage in the previous step from x_n to x_{n+1}. In the literature this reusable stage is sometimes referred to as the FSAL (First Same As Last) evaluation. Because the methods we consider are explicit, the FSAL evaluation can not be involved in the end point approximation, but it may be used to obtain an error estimate or for the continuous approximation. It turns out that there is a close connection between the reusable stage being included in the CERK method and the uniform approximation being continuously differentiable. We shall consider CERK methods where the last stage K_s is the FSAL evaluation, i.e. we impose the *stage reuse conditions*

$$c_s = 1 \quad \text{and} \quad a_{sj} = b_j(1), \; j = 1, \ldots, s. \tag{3.5}$$

For our further discussion we need the following lemma

Lemma 3.1 *Let $A \in \mathcal{M}^p$ define a CERK method with stage reuse. Then $\Phi := F_{p+1}(A) := ((\phi_{ij}))$ is such that*

$$\phi_{is} = \frac{\rho(t_i)}{\gamma(t_i)}, \quad i = 1, \ldots, N_{p+1}.$$

Proof. The lemma is proved by induction on the row index. The result holds for $i = 1$ since the first row corresponds to the only tree, τ of order 1 and since $\gamma(\tau) = \rho(\tau) = \phi_{1s}(\tau) = 1$. Then assume that the lemma is true for all i such that $i \leq n - 1$. The nth condition corresponds to a tree t_n which either has the form $[t_{n'}]$ for some tree $t_{n'}$ of order $\rho(t_n) - 1$ or the form $[t_{\nu_1}, \ldots, t_{\nu_u}]$ for $u (\geq 2)$ trees t_{ν_i} where $1 \leq \rho(t_{\nu_i}) \leq \rho(t_n) - 2$ and $\rho(t_n) = 1 + \sum_{i=1}^{u} \rho(t_{\nu_i})$. In the latter case with $t_{n_i} = [t_{\nu_i}]$ we immediately obtain from the definition of γ [3, p.88] that

$$\gamma(t_n) = \rho(t_n) \prod_{i=1}^{u} \frac{\gamma(t_{n_i})}{\rho(t_{n_i})}$$

such that

$$\phi_{ns} = \prod_{i=1}^{u} \phi_{n_i s} = \prod_{i=1}^{u} \frac{\rho(t_{n_i})}{\gamma(t_{n_i})} = \frac{\rho(t_n)}{\gamma(t_n)}.$$

In the former case we get

$$\phi_{ns} = \sum_{j=1}^{s-1} a_{sj} \phi_{n'j} = \sum_{j=1}^{s-1} b_j(1) \phi_{n'j} = 1/\gamma(t_{n'}) = \frac{\rho(t_n)}{\gamma(t_n)},$$

where we have applied the stage reuse conditions along with the order conditions at $\theta = 1$. Finally, observe that the induction works for all n such that $\rho(t_n) \leq p + 1$ since we only used the order condition corresponding to $t_{n'}$ and $\rho(t_{n'}) = \rho(t_n) - 1$. $\square$

With this result, the following theorem is now easy to prove.

Theorem 3.2 *Let $A \in \mathcal{M}_*^p$ be a CERK method with stage reuse. Then the global continuous approximation is continuously differentiable.*

Proof. It is sufficient to prove that $u'(x_0) = K_1$ and that $u'(x_0+h) = K_s$. By Lemma 3.1 the last column of $F_p(A)$ is equal to the right hand side of (2.4) evaluated at $\theta = 1$. Moreover, the first column of $F_p(A)$ equals the right hand side of (2.4) evaluated at $\theta = 0$. Since, by assumption, the columns of $F_p(A)$ are linearly independent we obtain $z_i(1) = \delta_{is}$ and $z_i(0) = \delta_{i1}$ such that $u'(x_0) = K_1$ and $u'(x_0 + h) = K_s$. □

It is of interest to know whether there exist CERK methods with stage reuse having a total of CEN(p) stages. It is easy to prove that such methods cannot exist for $p \leq 2$ under the assumption that the degree d of the polynomial weights does not exceed p. We have not been able to answer this question for general $p > 2$, but we shall see that such methods exist for $p = 3, 4, 5$. During this discussion we shall sometimes impose some additional conditions which we shall refer to as the simplifying assumptions, see e.g. [3, p.195],

$$\sum_{j=1}^{i-1} a_{ij} c_j = \frac{1}{2} c_i^2, \quad i = 3, \ldots, s. \tag{3.6}$$

We begin by considering the case $p = 3$. By combining the 4 continuous order conditions with (3.5) we find that a third order CERK method with 4 stages and stage reuse can be constructed in the following way

- Choose c_2 and c_3 with $c_2 \neq 0$, $c_3 \neq 0$ $c_2 \neq c_3$ and $c_2 \neq \frac{2}{3}$ in order that $A \in \mathcal{M}_*^3$.
- Put $c_4 = 1$ and compute a_{32}, a_{42} and a_{43} from the formulas

$$a_{32} = \tfrac{c_3(c_3-c_2)}{c_2(2-3c_2)},$$
$$a_{42} = \tfrac{3c_3-2}{6c_2(c_3-c_2)}, \quad a_{43} = \tfrac{2-3c_2}{6c_3(c_3-c_2)}.$$

- The continuous weights are then given by

$$\begin{aligned}
b_1(\theta) &= (1 - 2a_{41})\theta^3 + (3a_{41} - 2)\theta^2 + \theta, \\
b_2(\theta) &= -a_{42}(2\theta^3 - 3\theta^2), \\
b_3(\theta) &= -a_{43}(2\theta^3 - 3\theta^2), \\
b_4(\theta) &= \theta^3 - \theta^2.
\end{aligned}$$

Observe that this continuous extension of the 3-stage discrete method above is nothing but the cubic Hermite interpolant based on the endpoints x and $x + h$.

Considering order $p = 4$ we shall restrict ourselves to the methods derived in [16] having CEN(4)=6 stages. These methods satisfy the simplifying assumptions (3.6) with $s = 6$ and, using the notation of [3], we let the row space of $G_4(A)$ be spanned by the rows corresponding to the order conditions arising from the trees

$$\tau,\ [\tau],\ [\tau^2],\ [_2\tau]_2,\ [\tau^3],\ [_3\tau]_3. \tag{3.7}$$

Combining these conditions with (3.5) we find that one may choose $c_2 \neq 0$ and c_3, c_4, c_5 nonzero and distinct. Then if $6c_3c_4 - 4(c_3 + c_4) + 3 \neq 0$ the remaining coefficients are uniquely determined. If $c_3 = \frac{1}{2}$ and $c_4 = 1$, there exists a one-parameter family of methods (with a_{54} arbitrary). Given c_2, c_3, c_4, c_5 one may use the following procedure to obtain a 4th order CERK method with stage reuse having 6 stages

- Put $c_6 = 1$, $a_{62} = 0$ and compute a_{63}, a_{64}, a_{65} from the linear system

  $$\sum_{j=3}^{5} a_{6j} c_j^{k-1} = \frac{1}{k}, \;\; k = 2, 3, 4.$$

  By assumption there exists a unique solution.

- Consider the equation

  $$12 a_{65} a_{54} c_4 (c_4 - c_3) = 1 - 2c_3,$$

  arising from the condition corresponding to the tree $[_3\tau]_3$. If $a_{65} \neq 0$ this equation can be satisfied in any case and thereby a_{54} is uniquely determined. This requires that $6c_3c_4 - 4(c_3 + c_4) + 3 \neq 0$ If $a_{65} = 0$ then $c_3 = \frac{1}{2}$ leading to $c_4 = 1$, in which case a_{54} is arbitrary.

- We obtain the remaining coefficients from the formulas

  $$\begin{array}{ll} a_{32} = \frac{c_3^2}{2c_2}, & \\ a_{42} = \frac{(3c_3 - 2c_4)c_4^2}{2c_2c_3}, & a_{43} = \frac{(c_4 - c_3)c_4^2}{c_3^2}, \\ a_{52} = \frac{(3c_3 - 2c_5)c_5^2 + 6a_{54}c_4(c_4 - c_3)}{2c_2c_3}, & \\ a_{53} = \frac{(c_5 - c_3)c_5^2 - a_{54}c_4(3c_4 - 2c_3)}{c_3^2}. & \end{array}$$

The continuous weights are found by solving the linear 6×6 system of equations arising from the order conditions corresponding to the trees (3.7).

Also for the order 5 case we shall impose the simplifying assumptions (3.6) with $s = \mathrm{CEN}(5) = 8$ and the row space of $G_5(A)$ is assumed to be spanned by the rows corresponding to the trees

$$\tau,\ [\tau],\ [\tau^2],\ [_2\tau]_2,\ [\tau^3],\ [_3\tau]_3,\ [\tau^4],\ [_4\tau]_4. \tag{3.8}$$

For a detailed discussion of such 5th order CERK methods with 8 stages, see [16]. Now combining these conditions with the stage reuse conditions (3.5) and the assumption $A \in \mathcal{M}_*^5$ one finds after some long but straightforward algebra that $c_2 \neq 0$, $c_3 \neq 0$, $c_6 \neq 0$, $c_7 \neq 0$ and $a_{54} \neq 0$ can be chosen subject to the constraints $c_6 \neq c_3$, $c_6 \neq 2c_3$, $c_7 \neq c_3$, $c_7 \neq 2c_3$, $c_7 \neq c_6$, $c_8 \neq 2c_3$, i.e. $c_3 \neq \frac{1}{2}$ and $\frac{2}{3}c_3^2 - \frac{3}{4}c_3 + \frac{1}{5} \neq c_6(c_3 - \frac{1}{2})^2$. The following procedure provides the remaining coefficients

- Put $c_5 = c_4 = 2c_3$.

- Put $c_8 = 1$ and $a_{82} = 0$. Compute a_{86} and a_{87} from the formulas

$$a_{86} = -\frac{\frac{2}{3}c_3^2 - \frac{3}{4}c_3 + \frac{1}{5} - c_7(c_3 - \frac{1}{2})^2}{c_6(c_6 - c_3)(c_6 - 2c_3)(c_7 - c_6)}, \quad a_{87} = \frac{\frac{2}{3}c_3^2 - \frac{3}{4}c_3 + \frac{1}{5} - c_6(c_3 - \frac{1}{2})^2}{c_7(c_7 - c_3)(c_7 - 2c_3)(c_7 - c_6)}.$$

Observe that $a_{87} = 0$ would violate the assumption $\frac{2}{3}c_3^2 - \frac{3}{4}c_3 + \frac{1}{5} \neq c_6(c_3 - \frac{1}{2})^2$.

- We may now compute a_{76} from the formula

$$a_{76} = \frac{\frac{1}{3}c_3^2 - \frac{1}{4}c_3 + \frac{1}{20}}{a_{87}c_6(c_6 - c_3)(c_6 - 2c_3)}.$$

- Now we turn to the 6th stage which is independent of the stage reuse conditions

$$\begin{aligned} a_{62} &= \tfrac{c_6^2(2c_3 - c_6)}{2c_2c_3}, & a_{63} &= \tfrac{c_6^2(c_6 - c_3)}{3c_3^2}, \\ a_{64} &= \tfrac{c_6^2(c_6 - c_3)(2c_3 + a_{54} - c_6)}{12a_{54}c_3^2}, & a_{65} &= \tfrac{c_6^2(c_6 - c_3)(c_6 - 2c_3)}{12a_{54}c_3^2}. \end{aligned}$$

- The remaining coefficients of the 7th stage are given by

$$a_{75} = \frac{c_7^2(c_7 - c_3)(c_7 - 2c_3) - a_{76}c_6(c_6 - c_3)(3c_6 - 4c_3)}{12a_{54}c_3^2},$$

$$a_{74} = \frac{\frac{1}{12}c_7^2(c_7 - c_3)(c_7 - 2c_3) + a_{76}c_6(c_3(\frac{1}{2}c_7 - \frac{1}{3}c_3) - c_6(\frac{1}{2}c_7 - \frac{1}{3}c_6))}{c_3^2(c_7 - 2c_3)} - a_{75},$$

$$a_{73} = \frac{\frac{2}{3}c_7^2(c_7 - c_3)(c_7 - 2c_3) - a_{76}c_6(\frac{4}{3}(c_6 - 4c_3)(c_6 - \frac{1}{2}c_3) + c_6c_7)}{c_3^2(c_7 - 2c_3)} - 4(a_{74} + a_{75}).$$

a_{72} is obtained from (3.6).

- We next give formulas for the remaining coefficients of the 8th stage

$$a_{85} = \frac{(1 - c_3)(1 - 2c_3) - a_{86}f_6 - a_{87}f_7}{12a_{54}c_3^2},$$

where $f_i = 4c_i^3 - 9c_i^2c_3 + 4c_ic_3^2 + 2c_2c_3a_{i2}, \ i = 6, 7$.

$$a_{84} = \frac{\frac{1}{4}(1 - c_3)(1 - 2c_3) - (1 - 2c_3 + 2a_{54})a_{85}c_3^2 - a_{86}g_6 - a_{87}g_7}{(1 - 2c_3)c_3^2},$$

where $g_i = \frac{1}{4}c_i(3c_i - 2c_3)(1 - 2c_3) - \frac{1}{4}c_i^2 + \frac{1}{3}(c_i^3 + a_{i2}c_2c_3), \ i = 6, 7$.

$$a_{83} = \frac{1 - c_3 - a_{86}c_6(3c_6 - 2c_3) - a_{87}c_7(3c_7 - 2c_3)}{c_3^2} - 8(a_{84} + a_{85}).$$

- The remaining coefficients are independent of the stage reuse conditions

$$a_{32} = \frac{c_3^2}{2c_2}, \quad a_{42} = -\frac{2c_3^2}{c_2}, \quad a_{43} = 4c_3,$$

$$a_{52} = \frac{2c_3(3a_{54} - c_3)}{c_2}, \quad a_{53} = 4c_3 - 8a_{54}.$$

The polynomials $b_1(\theta), \ldots, b_8(\theta)$ are found by solving the linear 8×8 system of equations arising from the order conditions corresponding to the trees (3.8).

Remark: An interesting question is whether there exist methods among this fifth order class such that their first 6 stages define a discrete RK formula of order 5. This would clearly require $a_{87} = b_7(1) = 0$, a case which is excluded in the formulas above because of the assumption $\frac{2}{3}c_3^2 - \frac{3}{4}c_3 + \frac{1}{5} \neq c_6(c_3 - \frac{1}{2})^2$. However, it can be shown that this condition is also necessary for the existence of 8 stage 5th order CERK methods with stage reuse based on the simplifying assumptions and the choice of linearly independent rows of $G_5(A)$ made above.

4 Error estimation

Variable step Runge-Kutta methods are usually supplied with an error estimation device. We shall be concerned with the strategy based on embedded formulas (see [12] for details) and we shall see how this strategy can be adapted to pairs of CERK methods. Such a pair will be denoted by CERK(p,q) which means that the integration is proceeded by a continuous method of order p as described in the previous section, while a discrete formula of order q is used to obtain an error estimate at the end point of each step. It is obvious that we must require $p \neq q$ and it is customary to assume that $|p - q| = 1$, but it is not clear whether one should have $q > p$ or $p > q$. Most implementations of the discrete pairs proposed by Fehlberg [11] are of the former type while in the methods of Dormand and Prince [5] the intention is to impose $p = q + 1$ (local extrapolation). We will pay most of our attention to the latter type of pairs, mainly for two reasons. Firstly, we have the following negative result for the typical Fehlberg type implementations.

Proposition 4.1 *There exist no CERK($p, p+1$) pairs with $s =$CEN(p) stages.*

Proof. Assume that $s =$CEN(p) for a method given by the $s \times s$ matrix A. Then, since rank$(F_p(A)) = s$ it is impossible to find two distinct sets of weights that satisfy the first N_p discrete order conditions. □

The second reason is based on a recent result by Jackiewicz and Zennaro [14]. They find that given a CERK method of order p with s=CEN(p) stages, it is possible to obtain a two-step Runge-Kutta method of order $p+1$ at no additional cost. Hence, one may use this two-step approximation of order $p + 1$ to obtain an error estimate. The authors have no experience with practical use of such estimates.

Having constructed a CERK method of order p, several possibilities remain for the construction of a $(p-1)$st order discrete formula. We suggest that the final reusable stage is omitted from the error estimation formula, as this will cause rejected steps to cost only $s-2$ function evaluations if properly implemented. We shall take advantage of the fact that all our pth order methods with CEN(p) stages, $p = 3, 4, 5$, described in the previous section turns out to have an imbedded CERK method of order $p-1$ with CEN$(p-1)$ stages. We shall use this method evaluated at $\theta = 1$ as the error estimation method.

5 Minimization of the error constant

For a CERK method of order p, the local truncation error is given by

$$|y(x_0+\theta h) - u(x_0+\theta h)| = \frac{h^{p+1}}{(p+1)!} \left| \sum_{i=N_p+1}^{N_{p+1}} e_i(\theta)\alpha_i F_i(x_0, y_0) \right| + \mathcal{O}(h^{p+2})$$

where $F_i := F(t_i)$ is the elementary differential corresponding to the tree t_i, $\alpha_i := \alpha(t_i)$ is a positive integer weight corresponding to the tree t_i and $e_i(\theta)$, $i = N_p+1, \ldots, N_{p+1}$ are the *error polynomials* given by

$$e_i(\theta) = \theta^{\rho(t_i)} - \gamma(t_i) \sum_{j=1}^{s} \phi_{ij} b_j(\theta), \quad i = N_p+1, \ldots, N_{p+1}.$$

By considering the structure of the matrix $\Phi = ((\phi_{ij}))$, it is easy to see that all the error polynomials have vanishing zeroth and first degree coefficients such that both $e_i(0)$ and $e_i'(0)$ are zero. Moreover, for the CERK methods of the previous section the derivatives of the error polynomials also vanish at $\theta = 1$. We have

Proposition 5.1 *Let $A \in \mathcal{M}_*^p$ be a CERK method with stage reuse. Then all the error polynomials $e_i(\theta)$, $i = N_p+1, \ldots, N_{p+1}$ satisfy $e_i'(1) = 0$.*

Proof. The derivatives of the error polynomials are given by

$$e_i'(\theta) = \rho(t_i)\theta^{\rho(t_i)-1} - \gamma(t_i) \sum_{j=1}^{s} \phi_{ij} z_j(\theta), \; i = N_p+1, \ldots, N_{p+1}.$$

As in the proof of Theorem 3.2 we must have $z_s(1) = 1$ and $z_1(1) = \cdots = z_{s-1}(1) = 0$ and by Lemma 3.1 it follows that $e_i'(1) = 0$. □

Following e.g. Calvo et al.[4] we shall attempt to make the error polynomials small in some sense. The literature is not consistent with regard to what norm should be used on this n_{p+1}-vector of polynomials. Calvo et al. [4] minimize the quantity

$$g^* = \int_0^1 g(\theta) d\theta$$

over the free parameters, where

$$g(\theta) = \sqrt{\frac{\sum_{i=N_p+1}^{N_{p+1}} [e_i(\theta)\alpha(t_i)]^2}{\sum_{i=N_p+1}^{N_{p+1}} [e_i(1)\alpha(t_i)]^2}}. \tag{5.9}$$

Enright et al. [8] consider bounds for the principal error term on the intervals $0 \le \theta \le 1$ and $0 \le \theta \le 2$. They use the norm

$$\max_{N_p+1 \le i \le N_{p+1}} \{ \max_{\theta \in [0,\theta_e]} |e_i(\theta)| \}$$

where $\theta_e = 1$ or $\theta_e = 2$. This latter case is of interest when the continuous method is to be used beyond the current step e.g. for the purpose of handling discontinuities. We present

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{12}{23} & \frac{12}{23} & & & \\ \frac{4}{5} & -\frac{68}{375} & \frac{368}{375} & & \\ 1 & \frac{31}{144} & \frac{529}{1152} & \frac{125}{384} & \\ \hline \hat{y}_{n+1} & \frac{1}{24} & \frac{23}{24} & 0 & 0 \end{array} , \qquad \begin{aligned} b_1(\theta) &= \tfrac{41}{72}\theta^3 - \tfrac{65}{48}\theta^2 + \theta \\ b_2(\theta) &= -\tfrac{529}{576}\theta^3 + \tfrac{529}{384}\theta^2 \\ b_3(\theta) &= -\tfrac{125}{192}\theta^3 + \tfrac{125}{128}\theta^2 \\ b_4(\theta) &= \theta^3 - \theta^2 \end{aligned}$$

Table 1: Optimal 3rd order CERK method with stage reuse.

here method pairs of order 3,4 and 5 where the free parameters have been chosen to minimize the numerator of (5.9). The optimization was done numerically, and the values found for the free parameters were approximated by rational numbers. The weights $\hat{y}_{n+1}$ in Table 1-3 are those of the error estimation method and have nothing to do with the underlying discrete method of the CERK method. Several numerical investigations conducted with the methods derived in this paper show that the vector of error polynomials tends to have one dominating component which is the polynomial that corresponds to the tree $[_p\tau]_p$. Moreover, this error polynomial turns out to be independent or only weakly dependent on some of the free parameters of our optimal CERK methods. We illustrate this point by considering the optimal third order CERK methods of the previous section. The error polynomial $e_8(\theta)$ corresponding to the tree $[_3\tau]_3$ of order 4 is given by

$$e_8(\theta) = \theta^2(\theta - 2)^2$$

and hence completely independent of the free parameters c_2 and c_3. For $0 \le \theta \le 1$ it attains the maximum value 1 at $\theta = 1$ and it can be shown that c_2 and c_3 can be chosen such that the three remaining error polynomials satisfy $|e_i(\theta)| < 0.05, \ \theta \in [0,1]$. Consequently, one may suspect that the minima of the various error measures are flat and if the max-norm is used the minimum is not likely to be unique.

$$
\begin{array}{c|cccccc}
0 & & & & & & \\
\frac{1}{6} & \frac{1}{6} & & & & & \\
\frac{11}{37} & \frac{44}{1369} & \frac{363}{1369} & & & & \\
\frac{11}{17} & \frac{3388}{4913} & -\frac{8349}{4913} & \frac{8140}{4913} & & & \\
\frac{13}{15} & -\frac{36764}{408375} & \frac{767}{1125} & -\frac{32708}{136125} & \frac{210392}{408375} & & \\
1 & \frac{1697}{18876} & 0 & \frac{50653}{116160} & \frac{299693}{1626240} & \frac{3375}{11648} & \\
\hline
\hat{y}_{n+1} & \frac{101}{363} & 0 & -\frac{1369}{14520} & \frac{11849}{14520} & 0 & 0
\end{array}
$$

$$
\begin{aligned}
b_1(\theta) &= -\tfrac{866577}{824252}\theta^4 + \tfrac{1806901}{618189}\theta^3 - \tfrac{104217}{37466}\theta^2 + \theta \\
b_2(\theta) &= 0 \\
b_3(\theta) &= \tfrac{12308679}{5072320}\theta^4 - \tfrac{2178079}{380424}\theta^3 + \tfrac{861101}{230560}\theta^2 \\
b_4(\theta) &= -\tfrac{7816583}{10144640}\theta^4 + \tfrac{6244423}{5325936}\theta^3 - \tfrac{63869}{293440}\theta^2 \\
b_5(\theta) &= -\tfrac{624375}{217984}\theta^4 + \tfrac{982125}{190736}\theta^3 - \tfrac{1522125}{762944}\theta^2 \\
b_6(\theta) &= \tfrac{296}{131}\theta^4 - \tfrac{461}{131}\theta^3 + \tfrac{165}{131}\theta^2
\end{aligned}
$$

Table 2: Optimal 4th order CERK method with stage reuse.

REMARK: Dormand and Prince attempt to minimize the denominator of (5.9) in their discrete pair RK5(4)7M [5]. If we use the same error measure for our discrete underlying formulas of the 5th order methods, it turns out that we can only obtain an error at the end point about three times the size of that of RK5(4)7M. The corresponding error for the optimized 5th order method above is about 4 times that of RK5(4)7M.

6 Stability

When applied to ODEs, the methods of the previous sections will obviously have the stability characteristics of the underlying discrete method. The region of absolute stability of the methods of order four and five is influenced by the choice of the free parameters. We write the stability polynomials of pth order CERK methods with CEN(p) stages and stage reuse in the form

$$p(z) = \sum_{k=0}^{CEN(p)-1} \alpha_k \frac{z^k}{k!}$$

where $\alpha_0 = \cdots = \alpha_p = 1$. It can be shown that for our fourth order methods we have $\alpha_5 = 5(1-2c_3)c_4$. Following [15] the largest disk of stability is obtained by choosing $\alpha_5 \approx 0.5806$;

$$
\begin{array}{c|cccccccc}
0 & & & & & & & & \\
\frac{1}{6} & \frac{1}{6} & & & & & & & \\
\frac{1}{4} & \frac{1}{16} & \frac{3}{16} & & & & & & \\
\frac{1}{2} & \frac{1}{4} & -\frac{3}{4} & 1 & & & & & \\
\frac{1}{2} & -\frac{3}{4} & \frac{15}{4} & -3 & \frac{1}{2} & & & & \\
\frac{9}{14} & \frac{369}{1372} & -\frac{243}{343} & \frac{297}{343} & \frac{1485}{9604} & \frac{297}{4802} & & & \\
\frac{7}{8} & -\frac{133}{4512} & \frac{1113}{6016} & \frac{7945}{16544} & -\frac{12845}{24064} & -\frac{315}{24064} & \frac{156065}{198528} & & \\
1 & \frac{83}{945} & 0 & \frac{248}{825} & \frac{41}{180} & \frac{1}{36} & \frac{2401}{38610} & \frac{6016}{20475} & \\
\hline
\hat{y}_{n+1} & -\frac{1}{9} & 0 & \frac{40}{33} & -\frac{7}{4} & -\frac{1}{12} & \frac{343}{198} & 0 & 0
\end{array}
$$

$$
\begin{aligned}
b_1(\theta) &= \tfrac{596}{315}\,\theta^5 - \tfrac{4969}{819}\,\theta^4 + \tfrac{17893}{2457}\,\theta^3 - \tfrac{3292}{819}\,\theta^2 + \theta \\
b_2(\theta) &= 0 \\
b_3(\theta) &= -\tfrac{1984}{275}\,\theta^5 + \tfrac{1344}{65}\,\theta^4 - \tfrac{43568}{2145}\,\theta^3 + \tfrac{5112}{715}\,\theta^2 \\
b_4(\theta) &= \tfrac{118}{15}\,\theta^5 - \tfrac{1465}{78}\,\theta^4 + \tfrac{3161}{234}\,\theta^3 - \tfrac{123}{52}\,\theta^2 \\
b_5(\theta) &= 2\,\theta^5 - \tfrac{413}{78}\,\theta^4 + \tfrac{1061}{234}\,\theta^3 - \tfrac{63}{52}\,\theta^2 \\
b_6(\theta) &= -\tfrac{9604}{6435}\,\theta^5 + \tfrac{2401}{1521}\,\theta^4 + \tfrac{60025}{50193}\,\theta^3 - \tfrac{40817}{33462}\,\theta^2 \\
b_7(\theta) &= -\tfrac{48128}{6825}\,\theta^5 + \tfrac{96256}{5915}\,\theta^4 - \tfrac{637696}{53235}\,\theta^3 + \tfrac{18048}{5915}\,\theta^2 \\
b_8(\theta) &= 4\,\theta^5 - \tfrac{109}{13}\,\theta^4 + \tfrac{75}{13}\,\theta^3 - \tfrac{18}{13}\,\theta^2
\end{aligned}
$$

Table 3: Optimal 5th order CERK method with stage reuse.

E.g. the choice $c_3 = \frac{2}{5}$ and $c_4 = \frac{4}{7}$ leads to a nearly optimal stability region in the sense of [15]. The fifth order methods yield

$$\alpha_6 = 6(-5c_3^2 + 2c_3) - 2c_6\beta, \quad \alpha_7 = 14c_3c_6\beta$$

with $\beta = 20c_3^2 - 15c_3 + 3$.
Thus, for any $(\mu, \nu) \in \mathbf{R} \times \mathbf{R}$ there exist real c_3 and c_6 such that $\alpha_6 = \mu$ and $\alpha_7 = \nu$. So the stability polynomial is only restricted by the assumptions made in the construction of the CERK-formulas. In particular one may choose c_3 and c_6 such that $\alpha_6 \approx 0.7956$ and $\alpha_7 \approx 0.3305$ which, again according to [15], maximizes the region of absolute stability.

7 Numerical results with DETEST

We have tested the 5th order pair, henceforth denoted CM54, on some selected problems using the DETEST package [10]. We used absolute error control, attempting to control the local error at the end point of each step. As reference we have performed the same tests with an identical implementation of the Dormand-Prince RK5(4)7M method [5], henceforth called DP54, with a continuous extension obtained at the additional cost of 2 stages per step, see e.g. [4]. Hence, the effective cost per step is 8 stages for the Dormand-Prince extension and 7 stages for our 5th order CERK-method. In order to compare these two pairs, we have found it natural to use normalised efficiency, a feature of the DETEST software. Thus, instead of comparing the cost of the two pairs for a given tolerance, we make comparisons for a given *expected global accuracy* at the end point of integration. This expected accuracy is obtained in terms of the tolerance by assuming a relation of the form

$$\text{global error} \approx C \cdot \text{TOL}^E$$

where C and E are found by a least squares fit to the computed data. Piecewise linear interpolation then yields continuous extensions of the tabulated efficiency statistics. See [10] for a more detailed explanation.

We believe that the test problems chosen give a good idea of how the two methods perform with DETEST. On some of the omitted problems the discrepancy between equivalent tolerances (see tables) for the two methods were substantial or the least squares fit was too poor to be reliable.

Expected	Equiv. TOL		STEPS		FCALLS	
Accuracy	DP54	CM54	DP54	CM54	DP54	CM54
10**-3	10**-2.50	10**-1.35	4	2	33	19
10**-4	10**-3.55	10**-2.38	4	3	37	31
10**-5	10**-4.61	10**-3.42	6	4	50	41
10**-6	10**-5.66	10**-4.45	9	7	83	59
10**-7	10**-6.72	10**-5.49	14	11	137	91
10**-8	10**-7.77	10**-6.52	22	17	208	138

Table 4: Efficiency statistics for DETEST problem A4, $y' = \frac{y}{4}(1 - \frac{y}{20}),\ y(0) = 1.$

Expected	Equiv. TOL		STEPS		FCALLS	
Accuracy	DP54	CM54	DP54	CM54	DP54	CM54
10**-2	10**-3.16	10**-1.15	5	3	42	23
10**-3	10**-4.13	10**-2.31	6	4	52	33
10**-4	10**-5.10	10**-3.47	9	6	77	49
10**-5	10**-6.07	10**-4.63	14	10	117	74
10**-6	10**-7.05	10**-5.79	21	17	173	123
10**-7	10**-8.02	10**-6.95	33	29	266	206

Table 5: Efficiency statistics for DETEST problem C5, the five body problem.

Tables 4-7 show the efficiency of DP54 versus CM54. The first column contains the expected accuracy, while columns 2 and 3 predict the value of the tolerance that corresponds to this

Expected	Equiv. TOL		STEPS		FCALLS	
Accuracy	DP54	CM54	DP54	CM54	DP54	CM54
10**-1	10**-3.40	10**-1.76	52	40	553	384
10**-2	10**-4.27	10**-2.63	70	57	741	529
10**-3	10**-5.14	10**-3.51	99	80	1017	735
10**-4	10**-6.01	10**-4.39	142	115	1427	1031
10**-5	10**-6.88	10**-5.26	205	165	1699	1370
10**-6	10**-7.75	10**-6.14	307	237	2464	1713

Table 6: Efficiency statistics for DETEST problem D4, an orbit problem.

Expected	Equiv. TOL		STEPS		FCALLS	
Accuracy	DP54	CM54	DP54	CM54	DP54	CM54
10**-2	10**-3.04	10**-1.36	42	33	488	317
10**-3	10**-3.97	10**-2.29	61	48	651	430
10**-4	10**-4.90	10**-3.23	91	68	944	582
10**-5	10**-5.83	10**-4.16	135	101	1346	861
10**-6	10**-6.75	10**-5.09	201	150	1768	1242
10**-7	10**-7.68	10**-6.02	305	224	2534	1717

Table 7: Efficiency statistics for DETEST problem E2, the van der Pol oscillator.

accuracy for the two methods. Notice that CM54 is more pessimistic in estimating the error than DP54. Columns 4-7 show the expected number of steps and function evaluations for the two methods respectively. Note that since the number of steps and the number of function calls are obtained by interpolation, they may disagree with the number of stages per step that each of the methods actually have.

DETEST itself does not support any feature for testing interpolants. But by carefully modifying the program code, we found that the maximum of the uniform global error of OZ54 rarely exceeds the maximum global error of the underlying discrete method for any problem in the DETEST package. This can be explained by the fact that the underlying discrete method has the same local order as the continuous method. When comparing the error at the end point of the first step (i.e. the local error) to the maximum of the uniform error over this step, we found a maximum ratio of about 6.14. The ratio was equal to 1.0 in 87% of the cases, and in the range 1.0-1.32 in 99 % of the tests.

In this paper we have used the framework of [16] to characterize differentiable continuous explicit Runge-Kutta methods with the minimal number of stages for order 3,4 and 5. Of particular interest are the 5th order methods with only 7 effective stages since no such methods have been presented before. It is by no means obvious that the number of stages per step is an appropriate efficiency measure for continuous Runge-Kutta methods of a given order. However, the numerical results presented in this section may indicate that the new 5th order methods are at least comparable with those previously known.

Acknowledgement

We would like to express our thanks to the referees who gave us many useful suggestions and pointed out several misprints in the manuscript.

References

[1] J.C. Butcher, *Coefficients for the study of Runge-Kutta integration processes*, J. Austral. Math. Soc., 3 (1963), pp. 185–201.

[2] ——, *Implicit Runge-Kutta processes*, Math. Comp., 18 (1964), pp. 50–64.

[3] ——, *The numerical analysis of ordinary differential equations*, J. Wiley & Sons, 1987.

[4] M. Calvo, J.I. Montijano, and L. Rández, *A fifth order interpolant for the Dormand and Prince Runge-Kutta method*, J. Comput. Appl. Math., (1990), pp. 91–100.

[5] J.R. Dormand and P.J. Prince, *A family of embedded Runge-Kutta formulae*, J. Comput. Appl. Math., 6 (1980), pp. 19–26.

[6] ——, *Runge-Kutta triples*, Comput. Math. Appl., 12A (1986), pp. 1007–1017.

[7] ——, *Practical Runge-Kutta processes*, SIAM J. Sci. Stat. Comput., 5 (1989), pp. 977–989.

[8] W.H. Enright, K.R. Jackson, S.P. Nørsett, and P.G. Thomsen, *Interpolants for Runge-Kutta formulas*, ACM Transactions on Mathematical Software, 12 (1986), pp. 193–218.

[9] ——, *Effective Solution of Discontinuous IVPs Using a Runge-Kutta Formula Pair with Interpolants*, Appl. Math. and Comput., 27 (1988), pp. 313–335.

[10] W.H. Enright and J.D. Pryce, *Two Fortran Packages for Assessing Initial Value Problems*, TOMS, 13 (1987), pp. 1–27.

[11] E. Fehlberg, *Classical fifth-,sixth-,seventh- and eighth order Runge-Kutta formulas with step size control*, Tech. Rep. 287, NASA, 1968. Extract published in Computing 4 (1969) 93–106.

[12] E. Hairer, S.P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems*, Springer Verlag, 1987.

[13] M.K. Horn, *Fourth- and fifth-order, scaled Runge-Kutta algorithms for treating dense output*, SIAM J. Numer. Anal., 20 (1983), pp. 558–568.

[14] Z. Jackiewicz and M. Zennaro, *Variable step explicit two-step Runge-Kutta methods*, Tech. Rep. 125, Arizona State University, May 1990.

[15] B. Owren and K. Seip, *Some stability results for explicit Runge-Kutta methods*, BIT, 30 (1990), pp. 700–706.

[16] B. Owren and M. Zennaro, *Order barriers for continuous explicit Runge-Kutta methods*, Math. Comp., 56 (1991), pp. 645–661.

[17] L.F. Shampine, *Interpolation for Runge-Kutta methods*, SIAM J. Numer. Anal., 22 (1985), pp. 1014–1027.

[18] ——, *Some Practical Runge-Kutta Formulas*, Math. Comp., 173 (1986), pp. 135–150.

[19] J. Verner, *Differentiable Interpolants for High-Order Runge-Kutta Methods*, Tech. Rep. 1990–9, Queens's University, 1990.

[20] M. Zennaro, *Natural Continuous Extensions of Runge Kutta Methods*, Math. Comp., 46 (1986), pp. 119–133.