MATHEMATICS OF COMPUTATION
VOLUME 56, NUMBER 194
APRIL 1991, PAGES 645–661

ORDER BARRIERS FOR CONTINUOUS EXPLICIT RUNGE-KUTTA METHODS

BRYNJULF OWREN AND MARINO ZENNARO

ABSTRACT. In this paper we deal with continuous numerical methods for solving initial value problems for ordinary differential equations, the need for which occurs frequently in applications. Whereas most of the commonly used multistep methods provide continuous extensions by means of an interpolant which is available without making extra function evaluations, this is not always the case for one-step methods. We consider the class of explicit Runge-Kutta methods and provide theorems used to obtain lower bounds for the number of stages required to construct methods of a given uniform order p. These bounds are similar to the Butcher barriers known for the discrete case, and are derived up to order $p = 5$. As far as we know, the examples we present of 8-stage continuous Runge-Kutta methods of uniform order 5 are the first of their kind.

1. INTRODUCTION

Consider the initial value problem (IVP) for ordinary differential equations (ODE's)

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

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

The large variety of methods available nowadays allows almost any problem of the kind (1.1) to be handled efficiently. Most of these methods are designed to furnish the solution at a discrete set of points, say a mesh $\Delta := \{x_0 < x_1 < \cdots < x_N := x_f\}$. However, many applications require a continuous approximation to $y(x)$ in the entire interval $[x_0, x_f]$. These include differential equations with deviating arguments, problems where dense output is required, and problems where discontinuities are present. At first sight, it seems that *multistep* methods would be appropriate for these cases, as they provide a continuous extension by means of an interpolant which is available without making extra function evaluations. But their poor ability to handle problems with discontinuities, and

Received July 3, 1989; revised May 11, 1990.
1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 65L05.
The work of the first author was supported in part by the Norwegian Research Council for Science and Humanities.
The work of the second author was supported in part by the Italian M.P.I.

the fact that they cannot combine high order of accuracy with good stability properties are serious deficiencies that have to be taken into account. For these reasons, many authors have recently investigated continuous extensions of *one-step* methods (see, e.g., Bellen and Zennaro [1], Enright et al. [5], Horn [7], Nørsett and Wanner [8], Shampine [10], and Zennaro [11, 12, 13], as well as the book by Hairer et al. [6]). It appears that one can either construct a continuous one-step method directly, giving a discrete method as a by-product, or one can extend an already existing discrete method, possibly by including some additional function evaluations.

In this paper we follow the former approach in order to construct *continuous explicit Runge-Kutta methods* (CERK methods) of the form

$$K_i = f\left(x_0 + c_i h,\, y_0 + h\sum_{j=1}^{i-1} a_{ij}K_j\right), \qquad i = 1, \ldots, \nu, \tag{1.2a}$$

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

$u(x_0 + \theta h)$ is a *continuous approximation* to $y(x)$ in the interval $[x_0, x_0 + h]$ and $b_i(\theta)$, $i = 1, \ldots, \nu$, 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 $i = 1, \ldots, \nu$. Note that $c_1 = 0$, which implies that the first stage reduces to $K_1 = f(x_0, y_0)$. Moreover, the coefficients a_{ij} define a strictly lower triangular $\nu \times \nu$-matrix A.

As in the discrete case, which is obtained by setting $y_1 := u(x_0 + h)$, ν is the number of *stages*, whereas the *uniform order* (which we shall simply refer to as the *order*) is defined as the greatest integer p for which

$$\max_{0\leq\theta\leq 1} |y(x_0 + \theta h) - u(x_0 + \theta h)| = O(h^{p+1}). \tag{1.3}$$

Here, $|\cdot|$ stands for any norm on $\mathbf{R}^m$.

It is well known that efficiency, viewed as the ratio between the accuracy of the computed approximations and computational effort, is a very important parameter to be considered when designing new numerical methods. Therefore, the main goal of this paper is to find methods which use the lowest possible number of stages to attain a fixed prescribed order. Section 2 is devoted to providing theorems which can be used to determine 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 with order p. The numbers $CEN(p)$ are similar to the famous Butcher barriers $EN(p)$ for the discrete case.

It is well known (see, for example, the papers quoted above) that for implicit Runge-Kutta methods the minimal number of stages, say $N(p)$ and $CN(p)$, necessary to get order p for the discrete and continuous case, respectively, are

easy to find in general, and are attained by collocation methods. They are

$$N(p) = [(p+1)/2] \quad \text{and} \quad CN(p) = p.$$

For CERK methods things are, as for the discrete case, considerably more complicated. One has the obvious result

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

and from the literature referred to above, one can extract the following bounds:

$$\begin{gathered} CEN(1) = 1, \qquad CEN(2) = 2, \qquad 3 \leq CEN(3) \leq 4, \\ 5 \leq CEN(4) \leq 6, \qquad 6 \leq CEN(5) \leq 9, \end{gathered} \tag{1.4}$$

where the upper bounds are determined by known CERK methods. Although in §3 we solve the problem completely up to $p = 5$, we will not be able to derive a general formula for $CEN(p)$, and we suspect that, as for $EN(p)$, this is a very hard task.

2. Lower bounds on $CEN(p)$

In this section we shall use extensively the theory developed by Butcher [2, 3] without giving specific references, as we shall assume that the reader is acquainted with trees, order conditions, and related topics. We recommend the books by Butcher [4] and Hairer et al. [6] for background material, and we will use the notation of the latter.

It is easy to see that in order to fulfill (1.3), the degree d of the polynomials $b_i(\theta)$ must satisfy $d \geq p$. On the other hand, allowing $d > p$ can lead to approximate solutions $u(x)$ whose derivatives are unbounded as $h \to 0$ (see [8]). Therefore, we always choose $d = p$, so that, according to [12, Theorem 5], the polynomials $b_i'(\theta)$ span the space Π_{p-1} of polynomials of degree $p-1$. With reference to (1.2a–b) it is necessary that the number ν^* of distinct c_i's satisfies

$$\nu^* \geq p. \tag{2.1}$$

Consider the continuous version of the *order conditions*, which becomes

$$\sum_{j=1}^{\nu} b_j(\theta)\Phi_j(t) = \frac{\theta^{\rho(t)}}{\gamma(t)} \quad \text{for all trees } t \text{ such that } \rho(t) \leq p, \tag{2.2}$$

where $\Phi_j(t)$ is the jth elementary weight of the tree t, $\rho(t)$ is the order of t, and $\gamma(t)$ is a coefficient depending on the tree t. Now, putting

$$z_j(\theta) := b_j'(\theta), \qquad j = 1, \ldots, \nu, \tag{2.3}$$

(2.2) becomes

$$\sum_{j=1}^{\nu} z_j(\theta)\Phi_j(t) = \frac{\rho(t)\theta^{\rho(t)-1}}{\gamma(t)} \quad \text{for all trees } t \text{ such that } \rho(t) \leq p. \tag{2.4}$$

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 (2.4), where $N_p = \sum_{r=1}^{p} n_r$. It is well known that $n_1 = 1$, $n_2 = 1$, $n_3 = 2$, $n_4 = 4$, $n_5 = 9$, so that $N_1 = 1$, $N_2 = 2$, $N_3 = 4$, $N_4 = 8$, $N_5 = 17$. In general, we can number the N_p trees t increasingly in terms of $\rho(t)$, such that $i > j$ if $\rho(t_i) > \rho(t_j)$. We then rewrite the conditions (2.4) as

$$\sum_{j=1}^{\nu} \phi_{ij} z_j(\theta) = \frac{\rho(t_i)\theta^{\rho(t_i)-1}}{\gamma(t_i)}, \qquad i = 1, \ldots, N_p, \tag{2.5}$$

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

$$z_j(\theta) = \sum_{k=0}^{p-1} z_{jk}\theta^k, \tag{2.6a}$$

$$\frac{\rho(t_i)\theta^{\rho(t_i)-1}}{\gamma(t_i)} = \sum_{l=0}^{p-1} q_{il}\theta^l, \tag{2.6b}$$

and by defining the $N_p \times \nu$ matrix $\Phi := (\phi_{ij})$, the $\nu \times p$ matrix $Z := (z_{jk})$, and the $N_p \times p$ matrix $Q := (q_{il})$, (2.5) becomes

$$\Phi Z = Q. \tag{2.7}$$

The $N_p \times \nu$ matrix Φ depends on the $\nu \times \nu$ matrix A of the coefficients of the RK method, whereas the $N_p \times p$ matrix Q is independent of A. Incidentally, observe that (2.6b) implies

$$q_{i,\rho(t_i)-1} = \frac{\rho(t_i)}{\gamma(t_i)} \quad \text{and} \quad q_{il} = 0 \quad \text{for } l \neq \rho(t_i) - 1. \tag{2.8}$$

So we can define the maps

$$F_p\colon \bigcup_{\nu \geq 1} \mathscr{L}(\mathbf{R}^\nu, \mathbf{R}^\nu) \to \bigcup_{\nu \geq 1} \mathscr{L}(\mathbf{R}^\nu, \mathbf{R}^{N_p}) \quad \text{such that } F_p(A) := \Phi,$$

and

$$G_p\colon \bigcup_{\nu \geq 1} \mathscr{L}(\mathbf{R}^\nu, \mathbf{R}^\nu) \to \bigcup_{\nu \geq 1} \mathscr{L}(\mathbf{R}^{\nu+p}, \mathbf{R}^{N_p}) \quad \text{such that } G_p(A) := \Phi|Q,$$

where $\Phi|Q$ is the $N_p \times \nu$ matrix obtained by attaching the rows of Q to the rows of Φ.

Proposition 2.1. *A strictly lower triangular $\nu \times \nu$ matrix A defines a ν-stage CERK method of order p if and only if* $\operatorname{rank}(F_p(A)) = \operatorname{rank}(G_p(A))$.

Proof. From (2.7) it is obvious that any CERK method of order p satisfies $\operatorname{rank}(F_p(A)) = \operatorname{rank}(G_p(A))$. Vice versa, if a $\nu \times \nu$ matrix A is such that $\operatorname{rank}(F_p(A)) = \operatorname{rank}(G_p(A))$, then the system $F_p(A)Z = Q$ has at least one solution Z. By (2.3) and (2.6a), this matrix Z defines uniquely ν polynomials $b_i(\theta)$, $i = 1, \ldots, \nu$, of degree $\leq p$ such that $b_i(0) = 0$, and hence, the matrix A defines the stages of some CERK method of order p. □

In view of the result above, we are only interested in matrices belonging to the set

$$\mathcal{M}^p := \left\{ A \in \bigcup_{\nu \geq 1} \mathcal{L}(\mathbf{R}^\nu, \mathbf{R}^\nu) \mid A \text{ is strictly lower triangular}, \right.$$
$$\left. \operatorname{rank}(F_p(A)) = \operatorname{rank}(G_p(A)) \right\}.$$

From this point on, we shall say that ν is the *dimension* of A if A is a $\nu \times \nu$ matrix, and we write $\dim(A) = \nu$. It is clear that in general we have $\dim(A) \geq \operatorname{rank}(F_p(A))$.

Definition 2.2. A matrix $A \in \mathcal{M}^p$ is called *p-minimal* if $\dim(A) = \operatorname{rank}(F_p(A))$. Moreover, we define

$$\mathcal{M}_*^p := \{A \in \mathcal{M}^p \mid A \text{ is } p\text{-minimal}\}.$$

Proposition 2.3. *If the matrix $A \in \mathcal{M}_*^p$, then it cannot have two equal rows; in particular, we must have $c_2 \neq 0$. Moreover, $\dim(A) \leq N_p$.*

Proof. The first part follows easily from the fact that two equal rows in A imply two equal columns in $F_p(A)$. To see that $\dim(A) \leq N_p$, it is sufficient to observe that $\operatorname{rank}(F_p(A)) \leq N_p$. □

The following theorem represents a basic result for our theory, since it allows us to restrict consideration to p-minimal matrices.

Theorem 2.4. *Let $A \in \mathcal{M}^p$ be such that $\rho := \operatorname{rank}(F_p(A)) < \dim(A)$. Then there exists a matrix $A^* \in \mathcal{M}_*^p$ such that $\dim(A^*) = \rho$.*

Proof. It is sufficient to prove that there exists a matrix $A' \in \mathcal{M}^p$ such that $\dim(A') = \dim(A) - 1$ and $\operatorname{rank}(F_p(A')) = \rho$. In fact, this procedure can be applied $(\dim(A) - \rho)$ times in order to get the desired result. Let $\nu := \dim(A)$ and $\Phi := F_p(A)$. By hypothesis, we can find a column, say the kth column $(\phi_{1k}, \ldots, \phi_{N_p k})^T$, which is a linear combination of the preceding $k-1$ columns, that is

$$(\phi_{1k}, \ldots, \phi_{N_p k})^T = \sum_{j=1}^{k-1} \lambda_j (\phi_{1j}, \ldots, \phi_{N_p j})^T \tag{2.9}$$

for some $\lambda_j \in \mathbf{R}$. Now, define the $\nu \times \nu$ matrix A'' as follows:

$$a''_{kj} := 0 \quad \forall j = 1, \ldots, \nu, \tag{2.10a}$$

$$a''_{ik} := 0 \quad \forall i = 1, \ldots, \nu, \tag{2.10b}$$

$$a''_{ij} := a_{ij} + \lambda_j a_{ik} \quad \forall i = 1, \ldots, \nu, \ i \neq k; \ j = 1, \ldots, k-1, \tag{2.10c}$$

$$a''_{ij} := a_{ij} \quad \forall i = 1, \ldots, \nu, \ i \neq k; \ j = k+1, \ldots, \nu. \tag{2.10d}$$

In order to prove that $A'' \in \mathcal{M}^p$, first observe that the strictly lower triangular form of A is inherited by A''. Now, define $\Phi'' := F_p(A'')$. In view of (2.10a), we can easily conclude that the kth column $(\phi''_{1k}, \dots, \phi''_{N_p k})^T$ of Φ'' is equal to $(1, 0, \dots, 0)^T$. Moreover, since A and A'' are strictly lower triangular, the first column of both Φ and Φ'' are equal to $(1, 0, \dots, 0)^T$. As for the remaining columns of Φ'', we prove by induction on the row index i that they are all equal to the corresponding columns of Φ. This is clearly true for $i = 1$, since the first row of $F_p(A)$ is equal to $(1, \dots, 1)^T$ for any matrix A, and corresponds to the only condition (2.5) of order $r = 1$. We assume that the property is true for all $i \le n-1$ and prove it for $i = n$. Select the nth condition of (2.5) which corresponds to the tree t_n, where $\rho(t_n) \ge 2$. This tree can either have the form $[t_{n'}]$ for some tree $t_{n'}$ of order $\rho(t_n) - 1$, or the form $[t_{\nu_1}, \dots, t_{\nu_s}]$ for s (≥ 2) trees t_{ν_i}, where $1 \le \rho(t_{\nu_i}) \le \rho(t_n) - 2$ and $\rho(t_n) = 1 + \sum_{i=1}^{s} \rho(t_{\nu_i})$. In the former case, since $a_{jl} = a''_{jl} = 0$ for $l \ge j$, we have

$$\phi_{nj} = \sum_{l=1}^{j-1} a_{jl}\phi_{n'l}, \qquad j = 2, \dots, \nu, \ j \ne k, \tag{2.11a}$$

and

$$\phi''_{nj} = \sum_{l=1}^{j-1} a''_{jl}\phi''_{n'l}, \qquad j = 2, \dots, \nu, \ j \ne k, \tag{2.11b}$$

whereas in the latter case, with $t_{n_i} := [t_{\nu_i}]$, $i = 1, \dots, s$ (with $\rho(t_{n_i}) = 1 + \rho(t_{\nu_i}) \le \rho(t_n) - 1$), we have

$$\phi_{nj} = \prod_{i=1}^{s} \phi_{n_i j} \quad \text{and} \quad \phi''_{nj} = \prod_{i=1}^{s} \phi''_{n_i j}, \qquad j = 2, \dots, \nu, \ j \ne k. \tag{2.12}$$

Since we have numbered the conditions increasingly in terms of their order, we get in either case $n', n_1, \dots, n_s \le n-1$ and hence, by the inductive hypothesis,

$$\phi''_{n'j} = \phi_{n'j} \quad \forall j = 2, \dots, \nu, \ j \ne k, \tag{2.13}$$

and

$$\phi''_{n_i j} = \phi_{n_i j} \quad \forall i = 1, \dots, s \quad \text{and} \quad j = 2, \dots, \nu, \ j \ne k. \tag{2.14}$$

Therefore, in the latter case, by (2.12) and (2.14) we immediately get $\phi''_{nj} = \phi_{nj}$, $j = 2, \dots, \nu$, $j \ne k$. As for the former case, by (2.11b), (2.13), and (2.10c–d), and since $\phi''_{nk} = 0$, we have

$$\phi''_{nj} = \sum_{l=1}^{j-1} a''_{jl}\phi''_{n'l} = \sum_{\substack{l=1 \\ l\ne k}}^{j-1} a''_{jl}\phi_{n'l} = \sum_{\substack{l=1 \\ l\ne k}}^{j-1} a_{jl}\phi_{n'l} + a_{jk} \sum_{\substack{l=1 \\ l<k}}^{j-1} \lambda_l \phi_{n'l},$$

and hence, by (2.9) and since $a_{jk} = 0$ for $j \leq k$,

$$\phi''_{nj} = \sum_{l=1}^{j-1} a_{jl}\phi_{n'l},$$

which by (2.11a) yields $\phi''_{nj} = \phi_{nj}$, $j = 2, \ldots, \nu$, $j \neq k$. So the induction works. By (2.9), and since

$$(\phi''_{1k}, \ldots, \phi''_{N_p k})^T = (\phi''_{11}, \ldots, \phi''_{N_p 1})^T = (1, 0, \ldots, 0)^T,$$

we can conclude that the range of Φ is equal to the range of Φ'' and that $A'' \in \mathscr{M}^p$ with $\operatorname{rank}(\Phi'') = \rho$. Moreover, it is clear that the $\nu \times p$ matrix Z satisfying $\Phi'' Z = Q$ (see (2.7)) can be chosen with the kth column equal to the zero-vector, which means that $b_k(\theta) = 0$ in (1.2b). Furthermore, by (2.10b), the kth stage in (1.2a) is completely useless for the CERK method defined by A'', as it is not involved in the computations of the following stages. Consequently, the $(\nu - 1) \times (\nu - 1)$ matrix A' obtained by suppressing the kth row and the kth column of A'' defines the same CERK method (without the useless kth stage) and the matrix $F_p(A')$ is obtained by suppressing the kth column of Φ''. So A' is the desired matrix, satisfying $\dim(A') = \dim(A) - 1$ and $\operatorname{rank}(F_p(A')) = \rho$. □

By virtue of the theorem above, the following result is now obvious.

Corollary 2.5. *The set $\mathscr{M}^p_*$ is nonempty for all $p \geq 1$, and the minimum number of stages $CEN(p)$ required for a CERK method of order p is*

$$CEN(p) = \min_{A \in \mathscr{M}^p_*} \dim(A).$$

Moreover, if $A \in \mathscr{M}^p$ and $A \notin \mathscr{M}^p_*$, then $\dim(A) > CEN(p)$.

The problem of finding $CEN(p)$ can be slightly simplified by isolating the following p conditions of (2.5), which we shall call the *primary conditions*:

$$\sum_{j=1}^{\nu} c_j^{r-1} z_j(\theta) = \theta^{r-1}, \qquad r = 1, \ldots, p. \tag{2.15}$$

These conditions correspond to the trees defined recursively by $\tau^r := [\tau, \tau^{r-2}]$, where $\tau := [\,]$ and $\tau^2 := [\tau]$. Since the matrices A are strictly lower triangular, the remaining $N_p - p$ conditions of (2.5), which we shall call *secondary conditions*, do not explicitly involve the polynomials $z_1(\theta)$ and $z_2(\theta)$, as they satisfy $\phi_{i1} = \phi_{i2} = 0$. Roughly speaking, the dimension of the problem is, in some sense, reduced by two units.

Remark 2.6. Since all the secondary conditions correspond to trees of order $r \geq 3$, they always yield $q_{i1} = 0$ in (2.6b).

Now, for a $\nu \times \nu$ matrix $A \in \mathscr{M}^p_*$ with $p \geq 3$, we introduce the following equivalence relation on the set of indices $\{1, \ldots, \nu\}$:

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

There are ν^* equivalence classes $S_1, \ldots, S_{\nu^*}$, and we assume without restrictions that $1 \in S_1$ (i.e., $c_i = 0$ if and only if $i \in S_1$) and that $2 \in S_2$ (recall that $c_2 \neq 0$ by Proposition 2.3).

Definition 2.7. For a $\nu \times \nu$ matrix $A \in \mathcal{M}_*^p$ with $p \geq 3$ we shall call a *good index set* either the empty set $\varnothing$ or any nonempty subset of $\{3, \ldots, \nu\}$ whose elements do not belong to more than $p-3$ equivalence classes among $S_3, \ldots, S_{\nu^*}$.

Remark 2.8. If $p = 3$, then $S \subset S_1 \cup S_2$ for any good index set S.

Lemma 2.9. *Let $A \in \mathcal{M}_*^p$ with $p \geq 3$, and let S be a good index set for A. Then, with reference to* (2.5) *and* (2.6a), *in the set of polynomials* $\{z_j(\theta) | j \geq 3, j \notin S\}$ *(which is nonempty by* (2.1)*) there exists at least one, say $z_J(\theta)$, such that $z_{J1} \neq 0$.*

Proof. Choose an index $j_k \in S_k$ for any $k = 3, \ldots, \nu^*$, and assume, without restrictions, that $S \subset S_1 \cup S_2 \cup \cdots \cup S_r$ for some $r \leq p-1 \leq \nu^* - 1$. Since the polynomials $z_j(\theta)$, $j = 1, \ldots, \nu$, satisfy the primary conditions (2.15) for any polynomial $\pi(\theta) \in \Pi_{p-1}$, we easily get

$$\pi(\theta) = \sum_{j=1}^{\nu} \pi(c_j) z_j(\theta) = \sum_{k=1}^{\nu^*} \pi(c_{j_k}) \sum_{j \in S_k} z_j(\theta). \tag{2.16}$$

So, if we define $\pi(\theta) := \theta(\theta - c_2)(\theta - c_{j_3}) \cdots (\theta - c_{j_r})$, we get $\pi(c_j) = 0$ for all $j \in S_1 \cup S_2 \cup \cdots \cup S_r$, so that (2.16) becomes

$$\pi(\theta) = \sum_{k=r+1}^{\nu^*} \pi(c_{j_k}) \sum_{j \in S_k} z_j(\theta).$$

Since the coefficient of θ in $\pi(\theta)$ is $(-1)^{r-1} c_2 c_{j_3} \cdots c_{j_r} \neq 0$, and since $\pi(c_{j_k}) \neq 0$ for all $k = r+1, \ldots, \nu^*$, the proof is complete. □

We shall say that N (≥ 1) conditions (2.5) are linearly independent if and only if the corresponding N rows of the matrix $G_p(A)$ are linearly independent.

Lemma 2.10. *Let $A \in \mathcal{M}^p$, and let N conditions* (2.5) *be linearly independent. Then they explicitly involve N polynomials $z_j(\theta)$.*

Proof. It is sufficient to observe that, since $\operatorname{rank}(F_p(A)) = \operatorname{rank}(G_p(A))$, if N rows of the matrix $G_p(A)$ are linearly independent, then the same N rows of the matrix $F_p(A)$ must also be linearly independent. □

Now we are in a position to state the main result of this section, which is a tool for finding lower bounds for $CEN(p)$.

Theorem 2.11. *Let $A \in \mathcal{M}_*^p$ with $p \geq 3$, and let N secondary conditions from* (2.5) *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 in these* N *conditions (possibly* $S = \varnothing$*). 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. Since the N conditions we consider are secondary, they involve neither $z_1(\theta)$ nor $z_2(\theta)$. Thus by Lemma 2.10, we have in any case that $\dim(A) \geq N + s + 2$. To prove the stronger inequality, assume that S is a good index set for A, and that $\dim(A) = N + s + 2$. Then, again by Lemma 2.10, there are exactly N polynomials $z_j(\theta)$, the ones with $j \geq 3$ and $j \notin S$, which are involved in these N conditions. Moreover, the $N \times N$ matrix Φ^*, obtained by suppressing the $s + 2$ vanishing elements relevant to the missing polynomials $z_j(\theta)$ in each of the corresponding N rows of $F_p(A)$, is nonsingular. Thus, solving the subsystem (2.7) defined by these N conditions yields $z_{j1} = 0$ for all $j \geq 3$, $j \notin S$, which contradicts Lemma 2.9. □

We close this section by considering the minimal number of stages that must be added to a discrete Runge-Kutta method to extend it to a CERK method of the same (uniform) order. On the basis of the theory of this section, the following result is easy to prove.

Theorem 2.12. *Let the* $\nu \times \nu$ *matrix* A *define a discrete explicit Runge-Kutta method of order* p*. Consider a continuous extension of this method with* $\hat{\nu}$ *stages and uniform order* p*. Then* $\hat{\nu} - \nu \geq \delta$*, where* $\delta = \operatorname{rank}(G_p(A)) - \operatorname{rank}(F_p(A))$.

Proof. Let the $\hat{\nu} \times \hat{\nu}$ matrix $\hat{A}$ define the extended method. Since the first ν columns of $F_p(A)$ and $F_p(\hat{A})$ are identical, it follows that $\operatorname{rank}(F_p(\hat{A})) - \operatorname{rank}(F_p(A)) \leq \hat{\nu} - \nu$. On the other hand, the $\hat{\nu} - \nu$ columns arising from the additional stages cannot decrease the rank of $G_p(A)$, so that we must have $\operatorname{rank}(G_p(\hat{A})) \geq \operatorname{rank}(G_p(A))$. Thus, using Proposition 2.1, we have

$$\begin{aligned}\delta &= \operatorname{rank}(G_p(A)) - \operatorname{rank}(F_p(A)) \leq \operatorname{rank}(G_p(\hat{A})) - \operatorname{rank}(F_p(A)) \\ &= \operatorname{rank}(F_p(\hat{A})) - \operatorname{rank}(F_p(A)) \leq \hat{\nu} - \nu,\end{aligned}$$

and the proof is complete. □

For a given discrete Runge-Kutta method, this bound is easy to calculate and in many of the cases we have considered it is sharp. By comparing the above theorem with known interpolants we get

Corollary 2.13. *The minimal (total) number of stages of any* 5*th-order continuous extension of the Dormand-Prince*(4,5) *pair, or the Runge-Kutta-Fehlberg*(4,5) *pair, is* 9.

3. Finding $CEN(p)$ for $p \leq 5$

Now we shall apply the results from §2 in order to find the minimum number of stages $CEN(p)$ for $p = 3, 4, 5$. Our strategy will always be the following:

(i) In view of Corollary 2.5 we consider a $\nu \times \nu$ matrix $A \in \mathcal{M}_*^p$. We assume the maximum number of linearly independent secondary conditions to be in turn $1, 2, \ldots, N_p - p$, so that we either obtain an absurdity or, by Theorem 2.11, a lower bound for $\nu = \dim(A)$.

(ii) We will compare these lower bounds to the upper bound given by some existing method, already known for $p = 3, 4$ (see (1.4)) and new for $p = 5$.

We shall denote by $\phi^{(i)}$ the ith row $(\phi_{i1}, \ldots, \phi_{i\nu})$ of the matrix $F_p(A)$ and by $\psi^{(i)}$ the ith row $(\phi_{i1}, \ldots, \phi_{i\nu}, q_{i0}, \ldots, q_{ip-1})$ of the matrix $G_p(A)$. Moreover, for each $r \geq 1$, let R_r be the set of rows of the matrix $F_p(A)$ which correspond to conditions of order r. In the proof of Theorem 2.4 we saw that, if $\phi^{(i)} \in R_r$, then either $\phi^{(i)} = \phi^{(i')}A^T$, where $\phi^{(i')} \in R_{r-1}$ (and we shall say that $\phi^{(i)}$, as well as the corresponding condition, is an *A-transformation*) or $\phi_{ij} = \prod_{n=1}^{s} \phi_{i_n j}$, $j = 1, \ldots, \nu$, where $\phi^{(i_n)} \in R_{r_n}$ with $r_n \leq r-1$. In particular, in the latter case we may have $\phi^{(i)} = \phi^{(i')}C$, where $\phi^{(i')} \in R_{r-1}$ and $C := \mathrm{diag}(0, c_2, \ldots, c_\nu)$ (and we shall say that $\phi^{(i)}$, as well as the corresponding condition, is a *C-transformation*). Note that the primary condition of order r is the C-transformation of the primary condition of order $r-1$. In view of this, and since we have decided to number the conditions increasingly in terms of their order, each set of n_r conditions of order r will be numbered as follows: First the C-transformations, then the A-transformations, and finally the remaining conditions if any. It turns out that the primary condition will always be the first in each set of n_r conditions.

3.1. **Order** $p = 3$. There is only one secondary condition for the case $p = 3$. This condition corresponds to $\psi^{(4)} = (\phi^{(4)}, 0, 0, 1/2)$, where $\phi^{(4)} = \phi^{(2)}A^T$. Since $\psi^{(4)} \neq 0$, Theorem 2.11 yields in every case $\nu \geq 4$. Thus, the existence of the 4-stage CERK method of order 3 associated with the RKN(3,4) embedded pair (see Enright et al. [5]) implies

$$CEN(3) = 4.$$

Alternatively, this upper bound also follows from Proposition 2.3 since $N_3 = 4$. Moreover, it is clear that every 4×4 matrix A that results in a nonsingular $F_3(A)$ (which is a 4×4 matrix as well) determines a 4-stage CERK method of order 3.

3.2. **Order** $p = 4$. There are four secondary conditions: one of order 3, corresponding to $\psi^{(4)} = (\phi^{(4)}, 0, 0, 1/2, 0)$ and three of order 4, corresponding to $\psi^{(6)} = (\phi^{(6)}, 0, 0, 0, 1/2)$, $\psi^{(7)} = (\phi^{(7)}, 0, 0, 0, 1/3)$, and $\psi^{(8)} = (\phi^{(8)}, 0, 0, 0, 1/6)$. Moreover, $\phi^{(6)} = \phi^{(4)}C$, $\phi^{(7)} = \phi^{(3)}A^T$, and $\phi^{(8)} = \phi^{(4)}A^T$.

To begin with, we observe that the maximum number N of linearly independent conditions clearly obeys $N \geq 2$. Then assume $N = 2$. In this case $\psi^{(6)}$, $\psi^{(7)}$, and $\psi^{(8)}$ are proportional and therefore, since $\phi_{83} = 0$, we get $\phi_{63} = 0$ and $\phi_{73} = a_{32}c_2^2 = 0$ as well. So $a_{32} = 0$ because $c_2 \neq 0$, and hence $\phi_{43} = a_{32}c_2 = 0$, which means that $z_3(\theta)$ is not involved in the secondary conditions. Consequently, with reference to Theorem 2.11, we have $S \supset \{3\}$, which is a good index set for A, and hence, in every case we obtain $\nu \geq 6$. $N \geq 3$ implies in every case $\nu \geq 6$ by Theorem 2.11. The 6-stage CERK method of order 4 associated with the Dormand-Prince(4,5) embedded pair (see Hairer et al. [6]) provides an upper bound for $CEN(4)$, so we have

$$CEN(4) = 6.$$

Now, consider the conditions to be imposed on a 6×6 matrix A, necessarily 4-minimal (see Corollary 2.5), in order that it determines a CERK method of order 4. First, observe that, in view of the case $p = 3$ above, the rows $\psi^{(1)}$, $\psi^{(2)}$, $\psi^{(3)}$, $\psi^{(4)}$, and $\psi^{(5)}$ must be linearly independent. Note that $\psi^{(1)} = (1, 1, 1, 1, 1, 1, 1, 0, 0, 0)$, $\psi^{(2)} = (0, c_2, c_3, c_4, c_5, c_6, 0, 1, 0, 0)$, $\psi^{(3)} = (0, c_2^2, c_3^2, c_4^2, c_5^2, c_6^2, 0, 0, 1, 0)$, and $\psi^{(5)} = (0, c_2^3, c_3^3, c_4^3, c_5^3, c_6^3, 0, 0, 0, 1)$ correspond to the primary conditions. We can conclude that in order to have $\operatorname{rank}(G_4(A)) = 6$, it is necessary and sufficient that at least one of the following conditions be satisfied, where $S := \operatorname{span}\{\psi^{(1)}, \psi^{(2)}, \psi^{(3)}, \psi^{(4)}, \psi^{(5)}\}$:

(i) $\psi^{(6)} \notin S$ and $\psi^{(7)}, \psi^{(8)} \in \operatorname{span}\{S, \psi^{(6)}\}$,
(ii) $\psi^{(7)} \notin S$ and $\psi^{(6)}, \psi^{(8)} \in \operatorname{span}\{S, \psi^{(7)}\}$,
(iii) $\psi^{(8)} \notin S$ and $\psi^{(6)}, \psi^{(7)} \in \operatorname{span}\{S, \psi^{(8)}\}$.

Of course, we must also require $\operatorname{rank}(F_4(A)) = 6$. Now consider the following three groups of conditions for the coefficients of A:

$$c_2(c_4^2 - 2\phi_{44}) = c_4^3 + \lambda c_4\phi_{44} + \mu(a_{42}c_2^2 + a_{43}c_3^2), \tag{3.1a}$$

$$c_2(c_5^2 - 2\phi_{45}) = c_5^3 + \lambda c_5\phi_{45} + \mu(a_{52}c_2^2 + a_{53}c_3^2 + a_{54}c_4^2), \tag{3.1b}$$

$$c_2(c_6^2 - 2\phi_{46}) = c_6^3 + \lambda c_6\phi_{46} + \mu(a_{62}c_2^2 + a_{63}c_3^2 + a_{64}c_4^2 + a_{65}c_5^2), \tag{3.1c}$$

where

$$\lambda = 2\frac{a_{32}c_2^2 + c_3^2(c_2 - c_3)}{a_{32}c_2(2c_3 - 3c_2)} \quad \text{and} \quad \mu = \frac{3(c_2 - c_3)(2a_{32}c_2 - c_3^2)}{a_{32}c_2(2c_3 - 3c_2)}; \tag{3.1d}$$

$$c_2(c_4^2 - 2\phi_{44}) = c_4^3 + \lambda c_4\phi_{44} + \mu a_{43}\phi_{43}, \tag{3.2a}$$

$$c_2(c_5^2 - 2\phi_{45}) = c_5^3 + \lambda c_5\phi_{45} + \mu(a_{53}\phi_{43} + a_{54}\phi_{44}), \tag{3.2b}$$

$$c_2(c_6^2 - 2\phi_{46}) = c_6^3 + \lambda c_6\phi_{46} + \mu(a_{63}\phi_{43} + a_{64}\phi_{44} + a_{65}\phi_{45}), \tag{3.2c}$$

where

$$\lambda = \frac{c_3^2(c_2 - c_3) - 2a_{32}c_2^2}{a_{32}c_2c_3} \quad \text{and} \quad \mu = \frac{3(c_2 - c_3)(2a_{32}c_2 - c_3^2)}{a_{32}c_2c_3}; \tag{3.2d}$$

$$(3.3a)\quad c_2(c_4^2 - 2\phi_{44}) = c_4^3 + \lambda(a_{42}c_2^2 + a_{43}c_3^2) + \mu a_{43}\phi_{43},$$

$$(3.3b)\quad c_2(c_5^2 - 2\phi_{45}) = c_5^3 + \lambda(a_{52}c_2^2 + a_{53}c_3^2 + a_{54}c_4^2) + \mu(a_{53}\phi_{43} + a_{54}\phi_{44}),$$

$$(3.3c)\quad c_2(c_6^2 - 2\phi_{46}) = c_6^3 + \lambda(a_{62}c_2^2 + a_{63}c_3^2 + a_{64}c_4^2 + a_{65}c_5^2) + \mu(a_{63}\phi_{43} + a_{64}\phi_{44} + a_{65}\phi_{45}),$$

where

$$(3.3d)\qquad \lambda = \frac{c_3^2(c_2 - c_3) - 2a_{32}c_2^2}{a_{32}c_2^2} \quad \text{and} \quad \mu = -2\frac{a_{32}c_2^2 + c_3^2(c_2 - c_3)}{a_{32}c_2^2}.$$

Recall that $\phi_{43} = a_{32}c_2$, $\phi_{44} = a_{42}c_2 + a_{43}c_3$, $\phi_{45} = a_{52}c_2 + a_{53}c_3 + a_{54}c_4$, and $\phi_{46} = a_{62}c_2 + a_{63}c_3 + a_{64}c_4 + a_{65}c_5$.

Simple, but tedious calculations lead to the fact that, for all cases (i), (ii), and (iii), we must have $a_{32} \neq 0$ (otherwise, the matrix A would not be 4-minimal) and that:

(i) is equivalent to (3.1a–d) and (3.2a–d), where $c_2 \neq c_3$, $c_3 \neq 0$, $3c_2 - 2c_3 \neq 0$, and $2a_{32}c_2 - c_3^2 \neq 0$;

(ii) is equivalent to (3.1a–d) and (3.3a–d), where $3c_2 - 2c_3 \neq 0$ and $a_{32}c_2^2 + c_3^2(c_2 - c_3) \neq 0$;

(iii) is equivalent to (3.2a–d) and (3.3a–d), where $c_3 \neq 0$ and $c_3^2(c_2 - c_3) - 2a_{32}c_2^2 \neq 0$.

Moreover, it is also easy to see that, for all cases (i), (ii), and (iii), the following conditions, expressing that $\psi^{(6)}$, $\psi^{(7)}$, and $\psi^{(8)}$ are linearly dependent, can equivalently replace either of the two corresponding groups of conditions among (3.1a–d), (3.2a–d), and (3.3a–d):

$$(3.4a)\qquad a_{42}c_2^2(c_3 - c_4) + a_{43}c_3(c_3^2 - c_2c_4) = (2c_3 - 3c_2)a_{43}\phi_{43},$$

$$(3.4b)\qquad a_{52}c_2^2(c_3 - c_5) + a_{53}c_3(c_3^2 - c_2c_5) + a_{54}c_4(c_3c_4 - c_2c_5) = (2c_3 - 3c_2)(a_{53}\phi_{43} + a_{54}\phi_{44}),$$

$$(3.4c)\qquad a_{62}c_2^2(c_3 - c_6) + a_{63}c_3(c_3^2 - c_2c_6) + a_{64}c_4(c_3c_4 - c_2c_6) + a_{65}c_5(c_3c_5 - c_2c_6) = (2c_3 - 3c_2)(a_{63}\phi_{43} + a_{64}\phi_{44} + a_{65}\phi_{45}).$$

The method associated with the Dormand-Prince(5,4) pair obeys $2c_3 - 3c_2 = 0$ (i.e., $\psi^{(6)}$ and $\psi^{(7)}$ are proportional) and (iii) holds, but (i) and (ii) do not. This method is included in the class of 6-stage CERK methods of order 4, which is obtained by imposing $c_3 \neq 0$ and the following two groups of conditions:

$$(3.5a)\qquad 2\phi_{43} = c_3^2,$$

$$(3.5b)\qquad 2\phi_{44} = c_4^2,$$

$$(3.5c)\qquad 2\phi_{45} = c_5^2,$$

$$(3.5d)\qquad 2\phi_{46} = c_6^2$$

and

$$2a_{42}c_2c_3 + 3a_{43}c_3^2 = c_4^3, \tag{3.6a}$$

$$2a_{52}c_2c_3 + 3(a_{53}c_3^2 + a_{54}c_4^2) = c_5^3, \tag{3.6b}$$

$$2a_{62}c_2c_3 + 3(a_{63}c_3^2 + a_{64}c_4^2 + a_{65}c_5^2) = c_6^3. \tag{3.6c}$$

Indeed, $c_3 \neq 0$ and (3.5) imply (3.2) with $\lambda = -2$ and $\mu = 0$. Moreover, since the relations (3.2) are satisfied, conditions (3.6) are equivalent to conditions (3.3) with $\lambda = -2c_3/c_2$ and $\mu = 2(2c_3 - 3c_2)/c_2$. Since $c_3^2(c_2 - c_3) - 2a_{32}c_2^2 = -c_3^3 \neq 0$, it therefore follows that (iii) holds. Summarizing, we can choose arbitrary c_2, c_3, and c_4, subject to the only restrictions $c_2 \neq 0$, $c_3 \neq 0$, $c_3 \neq c_4$, and we get

$$a_{32} = \frac{c_3^2}{2c_2}, \tag{3.7b}$$

$$a_{42} = \frac{(3c_3 - 2c_4)c_4^2}{2c_2c_3}, \qquad a_{43} = \frac{(c_4 - c_3)c_4^2}{c_3^2}. \tag{3.7b}$$

Furthermore, we can choose arbitrary c_5, a_{54}, c_6, a_{64}, a_{65} (subject to A being 4-minimal) leading to

$$\begin{aligned} a_{52} &= \frac{(3c_3 - 2c_5)c_5^2 + 6a_{54}c_4(c_4 - c_3)}{2c_2c_3}, \\ a_{53} &= \frac{(c_5 - c_3)c_5^2 - a_{54}c_4(3c_4 - 2c_3)}{c_3^2}, \end{aligned} \tag{3.7c}$$

$$\begin{aligned} a_{62} &= \frac{(3c_3 - 2c_6)c_6^2 + 6a_{64}c_4(c_4 - c_3) + 6a_{65}c_5(c_5 - c_3)}{2c_2c_3}, \\ a_{63} &= \frac{(c_6 - c_3)c_6^2 - a_{64}c_4(3c_4 - 2c_3) - a_{65}c_5(3c_5 - 2c_3)}{c_3^2}. \end{aligned} \tag{3.7d}$$

3.3. **Order** $p = 5$. There are 12 secondary conditions, corresponding to

$$\begin{aligned} &\psi^{(4)} = (\phi^{(4)}, 0, 0, 1/2, 0, 0) \quad \text{of order 3}, \\ &\psi^{(6)} = (\phi^{(6)}, 0, 0, 0, 1/2, 0), \qquad \psi^{(7)} = (\phi^{(7)}, 0, 0, 0, 1/3, 0), \\ &\psi^{(8)} = (\phi^{(8)}, 0, 0, 0, 1/6, 0) \quad \text{of order 4}, \end{aligned}$$

and

$$\begin{aligned} &\psi^{(10)} = (\phi^{(10)}, 0, 0, 0, 0, 1/2), \qquad \psi^{(11)} = (\phi^{(11)}, 0, 0, 0, 0, 1/3), \\ &\psi^{(12)} = (\phi^{(12)}, 0, 0, 0, 0, 1/6), \qquad \psi^{(13)} = (\phi^{(13)}, 0, 0, 0, 0, 1/4), \\ &\psi^{(14)} = (\phi^{(14)}, 0, 0, 0, 0, 1/8), \qquad \psi^{(15)} = (\phi^{(15)}, 0, 0, 0, 0, 1/12), \\ &\psi^{(16)} = (\phi^{(16)}, 0, 0, 0, 0, 1/24), \qquad \psi^{(17)} = (\phi^{(17)}, 0, 0, 0, 0, 1/4) \end{aligned}$$

of order 5. Moreover, $\phi^{(10)} = \phi^{(6)}C$, $\phi^{(11)} = \phi^{(7)}C$, $\phi^{(12)} = \phi^{(8)}C$, $\phi^{(13)} = \phi^{(5)}A^T$, $\phi^{(14)} = \phi^{(6)}A^T$, $\phi^{(15)} = \phi^{(7)}A^T$, $\phi^{(16)} = \phi^{(8)}A^T$, and $\phi_{17,j} = (\phi_{4,j})^2$, $j = 1, \dots, \nu$.

To begin with, we observe that the maximum number N of linearly independent secondary conditions clearly obeys $N \geq 3$. If we assume $N = 3$, then $\psi^{(16)}$ and $\psi^{(17)}$ are proportional and hence, since $\phi^{(16)} = \phi^{(4)}(A^T)^2$ and $\phi_{17,j} = (\phi_{4,j})^2$, $j = 1, \dots, \nu$, we easily obtain the absurdity $\phi^{(16)} = \phi^{(17)} = 0$. Therefore, $\psi^{(16)}$ and $\psi^{(17)}$ must be linearly independent, and we can conclude that $N \geq 4$. So we assume $N = 4$. From above, we know that the dimension of $\text{span}\{\psi^{(10)}, \psi^{(11)}, \psi^{(12)}, \psi^{(13)}, \psi^{(14)}, \psi^{(15)}, \psi^{(16)}, \psi^{(17)}\}$ equals 2. First assume that $\psi^{(17)}$ is a linear combination of $\psi^{(12)}$, $\psi^{(14)}$, $\psi^{(15)}$, and $\psi^{(16)}$. In this case, since $\phi_{12,3} = \phi_{14,3} = \phi_{15,3} = \phi_{16,3} = 0$, we also have $\phi_{17,3} = (a_{32}c_2)^2 = 0$. Again, because $c_2 \neq 0$, we must have $a_{32} = 0$ and hence, $\phi_{43} = a_{32}c_2 = 0$, $\phi_{63} = a_{32}c_2c_3 = 0$, $\phi_{73} = a_{32}c_2^2 = 0$, $\phi_{10,3} = a_{32}c_2c_3^2 = 0$, $\phi_{11,3} = a_{32}c_2^2c_3 = 0$, and $\phi_{13,3} = a_{32}c_2^3 = 0$. Since also $\phi_{83} = 0$, the polynomial $z_3(\theta)$ is not involved in the secondary conditions, and therefore, with reference to Theorem 2.11 we get $S \supset \{3\}$, which is a good index set for A, and $\nu \geq 8$. Then assume $a_{32} \neq 0$ and that $\psi^{(17)}$ is linearly independent of $\psi^{(12)}$, $\psi^{(14)}$, $\psi^{(15)}$, and $\psi^{(16)}$. In this case, the rows $\psi^{(12)}$, $\psi^{(14)}$, $\psi^{(15)}$, and $\psi^{(16)}$ must be proportional, and consequently, since $\phi_{16,4} = 0$, we must also have $\phi_{12,4} = \phi_{14,4} = 0$ and $\phi_{15,4} = a_{43}a_{32}c_2^2 = 0$. Further, $a_{32}c_2^2 \neq 0$ implies $a_{43} = 0$ and $\phi_{84} = a_{43}a_{32}c_2c_4 = 0$. Moreover, since $\psi^{(6)}$, $\psi^{(7)}$, and $\psi^{(8)}$ cannot be linearly independent (this would imply $N \geq 5$), (3.4a) holds, and reduces to $a_{42}c_2^2(c_3 - c_4) = 0$. So only two cases are possible, either $a_{42} \neq 0$ and $c_3 = c_4$, or $a_{42} = 0$. Indeed, the former case cannot hold, since it contradicts the fact that A is 5-minimal. In fact, since $\psi^{(13)}$ must depend linearly on $\psi^{(16)}$ and $\psi^{(17)}$, and since $\phi_{16,3} = \phi_{16,4} = 0$, there should exist $\lambda \neq 0$ such that $\phi_{13,3} = \lambda\phi_{17,3}$ and $\phi_{13,4} = \lambda\phi_{17,4}$, leading to $a_{32}c_2^3 = \lambda(a_{32}c_2)^2$ and $a_{42}c_2^3 = \lambda(a_{42}c_2)^2$, so that $a_{32} = a_{42}$. Thus, the matrix A has two equal rows, contradicting Proposition 2.3. So we are left with $a_{42} = 0$, which implies $\phi_{44} = a_{42}c_2 + a_{43}c_3 = 0$, $\phi_{64} = c_4(a_{42}c_2 + a_{43}c_3) = 0$, $\phi_{74} = a_{42}c_2^2 + a_{43}c_3^2 = 0$, $\phi_{10,4} = c_4^2(a_{42}c_2 + a_{43}c_3) = 0$, $\phi_{11,4} = c_4(a_{42}c_2^2 + a_{43}c_3^2) = 0$, $\phi_{13,4} = a_{42}c_2^3 + a_{43}c_3^3 = 0$, and $\phi_{17,4} = (a_{42}c_2 + a_{43}c_3)^2 = 0$. In conclusion, $z_4(\theta)$ is not involved in the secondary conditions, so with reference to Theorem 2.11 we have $S \supset \{4\}$, which is a good index set for A, and hence, $\nu \geq 8$. If $N \geq 5$, then Theorem 2.11 yields again, in every case, $\nu \geq 8$.

As far as we know, the cheapest known CERK methods of order 5 require nine stages. As an example, we quote the 9-stage CERK method of order 5 associated with the RKV(5, 6) embedded pair (see Enright et al. [5]). However, since we shall find examples of 8-stage CERK methods of order 5, we can

conclude that

$$CEN(5) = 8.$$

Now we want to find conditions to be imposed on an 8×8 matrix A, necessarily 5-minimal, such that it determines a CERK method of order 5. In view of the previous case, $p = 4$, we observe that at least six rows among $\psi^{(i)}$, $i = 1, \ldots, 8$, must be linearly independent. So we can, for example, assume that $\psi^{(1)}$, $\psi^{(2)}$, $\psi^{(3)}$, $\psi^{(4)}$, $\psi^{(5)}$, and $\psi^{(8)}$ are linearly independent, and then impose condition (iii). Then we must have $a_{32} \neq 0$, $c_3 \neq 0$, and $c_3^2(c_2 - c_3) - 2a_{32}c_2^2 \neq 0$. Moreover, conditions (3.2a–d) and (3.3a–d) (or, equivalently, (3.2a–d) and 3.4a–d)) must be satisfied. However, since we now have $\dim(A) = 8$, we must supply both (3.2a–d) and (3.3a–d) with the obvious two conditions corresponding to the last two stages (see [9]).

Now, observe that, if $\phi^{(i)} \in R_r$ is a C-transformation of $\phi^{(i')} \in R_{r-1}$, then by (2.8) we get $q_{i,r-1} = q_{i',r-2}$, whereas if $\phi^{(i)} \in R_r$ is an A-transformation of $\phi^{(i')} \in R_{r-1}$, then $q_{i,r-1} = q_{i',r-2}/(r-1)$. So we can conclude that condition (iii) automatically implies

(iv) $\psi^{(10)}, \psi^{(11)} \in \text{span}\{\psi^{(2)}, \psi^{(3)}, \psi^{(5)}, \psi^{(6)}, \psi^{(9)}, \psi^{(12)}\}$ and

(v) $\psi^{(14)}, \psi^{(15)} \in \text{span}\{\psi^{(2)}, \psi^{(4)}, \psi^{(7)}, \psi^{(8)}, \psi^{(13)}, \psi^{(16)}\}$. Therefore, we assume that $\psi^{(16)} \notin \text{span}\{S, \psi^{(8)}, \psi^{(9)}\}$ and impose that

(vi) $\psi^{(12)}, \psi^{(13)}, \psi^{(17)} \in \text{span}\{S, \psi^{(8)}, \psi^{(9)}, \psi^{(16)}\}$ in order to get $\text{rank}(G_5(A)) = 8$.

A particular class of these methods is obtained by imposing $c_3 \neq 0$ and the conditions (3.5) and (3.6) together with their counterparts for the last two stages. It can be shown (see [9]) that this requires $2c_3 = c_4 = c_5$.

Moreover, like for $p = 4$, a_{54} is a free parameter. However, for the matrix A to be 5-minimal we must require $a_{54} \neq 0$. In conclusion, we can choose arbitrary c_2, c_3, a_{54} subject to the restrictions $c_2 \neq 0$, $c_3 \neq 0$, and $a_{54} \neq 0$. By using $c_5 = c_4 = 2c_3$ in (3.7a–c) we get

$$a_{32} = \frac{c_3^2}{2c_2},$$

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

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

Furthermore, we can choose arbitrary c_6, subject to the only restrictions $c_6 \neq c_3$ and $c_6 \neq 2c_3$, and we get

$$a_{62} = \frac{c_6^2(2c_3 - c_6)}{2c_2c_3}, \qquad a_{63} = \frac{c_6^2(c_6 - c_3)}{3c_3^2},$$

$$a_{64} = \frac{c_6^2(c_6 - c_3)(2c_3 + a_{54} - c_6)}{12a_{54}c_3^2}, \qquad a_{65} = \frac{c_6^2(c_6 - c_3)(c_6 - 2c_3)}{12a_{54}c_3^2}.$$

Finally, we can choose arbitrary c_7, a_{76}, c_8, a_{86}, a_{87} (apart from combinations leading to A not being 5-minimal), and solve for the remaining coefficients $a_{72}, \ldots, a_{75}$ and $a_{82}, \ldots, a_{85}$. Their general expressions are quite complicated, so we prefer to present the Butcher tableau $\mathbf{c}|A$ as an example of such methods, together with the continuous weights $b_i(\theta)$. The choices for the free parameters are not motivated by stability considerations or error constant minimization, but rather by our desire to obtain simple coefficients.

$$
\begin{array}{c|ccccccc}
0 & & & & & & & \\
\frac{1}{4} & \frac{1}{4} & & & & & & \\
\frac{1}{4} & \frac{1}{8} & \frac{1}{8} & & & & & \\
\frac{1}{2} & 0 & -\frac{1}{2} & 1 & & & & \\
\frac{1}{2} & \frac{1}{12} & 0 & \frac{1}{3} & \frac{1}{12} & & & \\
\frac{3}{4} & 0 & -\frac{9}{8} & \frac{3}{2} & -\frac{3}{4} & \frac{9}{8} & & \\
1 & 0 & \frac{4}{5} & 0 & -\frac{3}{5} & 0 & \frac{4}{5} & \\
1 & \frac{1}{6} & 0 & 0 & \frac{4}{15} & \frac{2}{5} & 0 & \frac{1}{6}
\end{array}
$$

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

BIBLIOGRAPHY

1. A. Bellen and M. Zennaro, *Stability of interpolants for Runge Kutta methods*, SIAM J. Numer. Anal. **25** (1988), 411–432.
2. John C. Butcher, *Coefficients for the study of Runge-Kutta integration processes*, J. Austral. Math. Soc. **3** (1963), 185–201.
3. ____, *Implicit Runge-Kutta processes*, Math. Comp. **18** (1964), 50–64.
4. ____, *The numerical analysis of ordinary differential equations*, Wiley, Chichester, 1987.
5. W. H. Enright, K. R. Jackson, S. P. Nørsett, and P. G. Thomsen, *Interpolants for Runge-Kutta formulas*, ACM Trans. Math. Software **12** (1986), 193–218.

6. E. Hairer, S. P. Nørsett, and G. Wanner, *Solving ordinary differential equations.* I, *Nonstiff problems*, Springer, Berlin, 1987.
7. M. K. Horn, *Fourth- and fifth-order, scaled Runge-Kutta algorithms for treating dense output*, SIAM J. Numer. Anal. **20** (1983), 558–568.
8. S. P. Nørsett and G. Wanner, *Perturbed collocation and Runge-Kutta methods*, Numer. Math. **38** (1981), 193–208.
9. B. Owren and M. Zennaro, *Order barriers for continuous explicit Runge-Kutta methods*, Mathematics and Computation no. 2/89, The University of Trondheim, Norway.
10. L. F. Shampine, *Interpolation for Runge-Kutta methods*, SIAM J. Numer. Anal. **22** (1985), 1014–1027.
11. M. Zennaro, *One-step collocation: uniform superconvergence, predictor-corrector method, local error estimate*, SIAM J. Numer. Anal. **22** (1985), 1135–1152.
12. ——, *Natural continuous extensions of Runge Kutta methods*, Math. Comp. **46** (1986), 119–133.
13. ——, *Natural Runge-Kutta and projection methods*, Numer. Math. **53** (1988), 423–438.

DIVISION OF MATHEMATICAL SCIENCES, NORWEGIAN INSTITUTE OF TECHNOLOGY, N-7034 TRONDHEIM-NTH, NORWAY

DIPARTIMENTO DI MATEMATICA E INFORMATICA, UNIVERSITÀ DI UDINE I-33100 UDINE, ITALY