Natural Continuous Extensions of Runge-Kutta Methods
Author(s): M. Zennaro
Source: *Mathematics of Computation*, Vol. 46, No. 173 (Jan., 1986), pp. 119-133
Published by: American Mathematical Society
Stable URL: https://www.jstor.org/stable/2008218
Accessed: 13-11-2019 14:58 UTC

REFERENCES
Linked references are available on JSTOR for this article:
https://www.jstor.org/stable/2008218?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
https://about.jstor.org/terms

American Mathematical Society is collaborating with JSTOR to digitize, preserve and extend access to *Mathematics of Computation*

MATHEMATICS OF COMPUTATION
VOLUME 46, NUMBER 173
JANUARY 1986, PAGES 119-133

Natural Continuous Extensions of Runge-Kutta Methods*

By M. Zennaro

Abstract. The present paper develops a theory of Natural Continuous Extensions (NCEs) for the discrete approximate solution of an ODE given by a Runge-Kutta process. These NCEs are defined in such a way that the continuous solutions furnished by the one-step collocation methods are included.

1. Introduction. We consider the following initial value problem (IVP) for ODEs:

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

where y_0, y and f are m-vectors and t is a real variable. Further, we suppose that f is as smooth as necessary both in t and in y.

A ν-stage Runge-Kutta (R-K) process is a means to find an approximation $\bar{y}$ to the value of y at the point $t_0 + h$.

Following the notation of Butcher [4], we write

$$g^{(i)} = f\left(t_0 + c_i h,\ y_0 + h \sum_{j=1}^{\nu} a_{ij} g^{(j)}\right), \qquad i = 1, \ldots, \nu, \tag{2}$$

$$\bar{y} = y_0 + h \sum_{i=1}^{\nu} b_i g^{(i)}, \tag{3}$$

where $g^{(i)}$, $i = 1, \ldots, \nu$, are m-vectors, the numbers a_{ij}, b_i characterize the method, and $c_i = \sum_{j=1}^{\nu} a_{ij}$, $i = 1, \ldots, \nu$. The integer ν is the number of stages.

If the R-K method is accurate of order p ($\geqslant 1$), then $|y(t_0 + h) - \bar{y}| = O(h^{p+1})$. The symbol $|\cdot|$ stands for any norm on $\mathbf{R}^m$. By iterating this process, one usually finds an approximate solution on a mesh $\Delta = \{t_0 < t_1 < \cdots < t_N = T\}$ of the interval $[t_0, T]$, such that

$$\max_{0 \leqslant n \leqslant N} |y_n - y(t_n)| = O(|\Delta|^p), \tag{4}$$

where $|\Delta| := \max_{1 \leqslant n \leqslant N} |t_n - t_{n-1}|$ and the y_n's are the approximate values.

In this way one gets information of the solution y on a discrete set of points. On the other hand, it is sometimes useful to have a continuous approximate solution available. For this purpose one could use interpolation at nodal values; that involves

Received June 11, 1984; revised April 1, 1985.
1980 *Mathematics Subject Classification*. Primary 65L05.

*Work performed within the Project Informatica, Subproject Pl, SOFMAT (of the Italian C.N.R.).

©1986 American Mathematical Society
0025-5718/86 $1.00 + $.25 per page

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

a number of mesh points and the resulting method has the characteristics of a multistep one. Nevertheless, it is well-known that some implicit R-K methods are equivalent to collocation methods (see Wright [13]). Therefore, they naturally give successively a continuous extension of the approximate solution, without any extra evaluations of the function f.

The following question then arises. Does there exist for each R-K process a continuous extension given successively by the method itself, without any extra evaluations of the function f?

A kind of answer has been recently given by Nørsett and Wanner [10], since they have proved that a large class of R-K processes can be considered as a somewhat perturbed collocation (PECO-method). Their definition includes ordinary collocation, as a particular case. However, if the number of stages is ν, they always find a polynomial of degree ν. This fact causes some troubles, particularly with explicit methods, since it often happens that the derivatives of the perturbed collocation solution are unbounded for $h \to 0$.

The aim of this paper is to provide Natural Continuous Extensions (NCEs) of the solution given by the R-K process (2)–(3), which, in case of equivalence with collocation, continue to include the collocation solution and, in a certain sense, behave like it in other cases. However, we do not look at the problem from Nørsett and Wanner's point of view.

To formulate our definition, we need the following results.

Consider the one-step collocation method at ν Gaussian points (which is equivalent to the unique ν-stage R-K process of order $p = 2\nu$) to approximate the solution of (1) at $t_0 + h$. The collocation solution u, which is an m-vector polynomial of degree ν, satisfies the following error bounds (e.g., Guillou and Soulé [7], Hulme [8], Nørsett and Wanner [9] and [10]):

$$(5) \qquad \max_{t_0 \leqslant t \leqslant t_0+h} |y^{(k)}(t) - u^{(k)}(t)| = O(h^{\nu+1-k}), \qquad k = 0, 1, \ldots, \nu,$$

and

$$|y(t_0 + h) - u(t_0 + h)| = O(h^{2\nu+1}),$$

which is often referred to as the phenomenon of superconvergence at the nodes.

By using the same arguments of Nørsett and Wanner [9] and [10], the author [14] has proved the more general superconvergence result:

$$(6) \qquad \int_{t_0}^{t_0+h} G(t)[y'(t) - u'(t)]\,dt = O(h^{2\nu+1})$$

for every sufficiently smooth matrix-valued function G.

Equation (6) can be regarded as a sort of asymptotic orthogonality condition.

Our definition of NCE is given in such a way that conditions similar to (5) and (6) are required.

Definition 1. The ν-stage R-K process (2)–(3) of order p has a NCE u of degree d if there exist ν polynomials $b_i(\theta)$, $i = 1, \ldots, \nu$, of degree $\leqslant d$, independent of the function f, such that, by putting

$$(7) \qquad u(t_0 + \theta h) :\equiv y_0 + h \sum_{i=1}^{\nu} b_i(\theta) g^{(i)}, \qquad 0 \leqslant \theta \leqslant 1,$$

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

the following statements hold:

$$u(t_0) = y_0 \quad \text{and} \quad u(t_0 + h) = \bar{y}; \tag{8}$$

$$\max_{t_0 \leqslant t \leqslant t_0 + h} |y'(t) - u'(t)| = O(h^d); \tag{9}$$

$$\int_{t_0}^{t_0 + h} G(t)[y'(t) - u'(t)]\, dt = O(h^{p+1}) \tag{10}$$

for every sufficiently smooth matrix-valued function G.

It is easily seen that condition (9) implies the following error bounds for the higher derivatives of the NCEs:

$$\max_{t_0 \leqslant t \leqslant t_0 + h} |y^{(k)}(t) - u^{(k)}(t)| = O(h^{d-k+1}), \qquad k = 2, \dots, d, \tag{11}$$

and, obviously, $u^{(k)}(t) \equiv 0$ for $k \geqslant d + 1$.

Therefore all the derivatives of the NCE u are uniformly bounded as $h \to 0$.

Moreover, by integrating (9) and by (8), we have also

$$\max_{t_0 \leqslant t \leqslant t_0 + h} |y(t) - u(t)| = O(h^{d+1}), \tag{12}$$

while simple integration by parts, together with (10) and (8), yields

$$\int_{t_0}^{t_0 + h} G(t)[y(t) - u(t)]\, dt = O(h^{p+1}) \tag{13}$$

for every sufficiently smooth matrix-valued function G.

After proving the existence of NCEs for all R-K processes (Section 2), we find the relationships with collocation and with PECO-methods, and give the NCEs for some of the most popular explicit R-K processes (Section 3).

Section 4 is devoted to applications. First we prove a theorem which allows us to extend to all R-K processes some results of Bellen [2] and Vermiglio [12] concerning one-step collocation and one-step subregion methods, respectively, applied to DDEs (delay differential equations). Furthermore, some recent results of the author [14] concerning the order of uniform convergence and stepsize control for the one-step collocation method are extended to all R-K processes.

2. On the Existence of NCEs. In this section we prove the existence of NCEs for all R-K processes.

First of all,we recall some general results which were published by Butcher [4] and [5]. He considered, for the study of the ν-stage R-K process (2)–(3), the expansions

$$y(t_0 + h) = y_0 + \sum \alpha F \frac{h^r}{r!} \tag{14}$$

and

$$\bar{y} = y_0 + \sum \beta \Phi F \frac{h^r}{(r-1)!}. \tag{15}$$

The summations are over the different elementary differentials F for the function f, arranged in a sequence of nondecreasing order r. Φ is the corresponding elementary weight, while α and β are numerical coefficients independent of the function f.

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

PROPOSITION 2. *Each elementary weight Φ has the form*

$$\Phi = \sum_{i=1}^{\nu} b_i \Phi^{(i)}$$

where $\Phi^{(i)}$ is independent of b_j, $j = 1, \dots, \nu$.

PROPOSITION 3. *The R-K process* (2)–(3) *is of order p if and only if*

$$\Phi = \alpha/\beta r$$

for all elementary weights Φ corresponding to elementary differentials F of order $r \leqslant p$.

PROPOSITION 4. *Let the R-K process* (2)–(3) *be of order p. If $\Phi = \sum_{i=1}^{\nu} b_i \Phi^{(i)}$ is an elementary weight corresponding to an elementary differential F of order $r \leqslant p - 1$ and s is an integer such that $1 \leqslant s \leqslant p - r$, then, by putting*

$$\tilde{\Phi} := \sum_{i=1}^{\nu} b_i c_i^s \Phi^{(i)},$$

we get an elementary weight corresponding to an elementary differential $\tilde{F}$ of order $r + s$ and, moreover, $(r + s)\tilde{\Phi} = r\Phi$.

In view of (7), it is quite easy to verify that the following continuous version of (15) holds:

$$u(t_0 + \theta h) \equiv y_0 + \sum \beta \Phi(\theta) F \frac{h^r}{(r-1)!}, \qquad 0 \leqslant \theta \leqslant 1, \tag{16}$$

where, by Proposition 2,

$$\Phi(\theta) \equiv \sum_{i=1}^{\nu} b_i(\theta) \Phi^{(i)}, \qquad 0 \leqslant \theta \leqslant 1, \tag{17}$$

are continuous elementary weights.

Moreover, we have the continuous version of (14)

$$y(t_0 + \theta h) \equiv y_0 + \sum \alpha F \frac{\theta^r h^r}{r!}, \qquad 0 \leqslant \theta \leqslant 1. \tag{18}$$

Therefore, by comparing (16) and (18), one can easily see that the two conditions (9) and (10) are equivalent to

$$\Phi'(\theta) \equiv \alpha \theta^{r-1}/\beta \quad \text{for all } \Phi \text{ corresponding to elementary differentials } F \text{ of order } r \leqslant d; \tag{19}$$

together with

$$\int_0^1 \theta^s \Phi'(\theta)\, d\theta = \frac{\alpha}{\beta(r+s)} \quad \text{for all } \Phi \text{ corresponding to elementary differentials } F \text{ of order } r = d+1, \dots, p \text{ and for every } s = 0, \dots, p - r. \tag{20}$$

Observe that, if $d = p$, condition (20) must not be considered.

As far as condition (8) is concerned, it is clear that it is equivalent to

$$b_i(0) = 0 \quad \text{and} \quad b_i(1) = b_i, \qquad i = 1, \dots, \nu. \tag{21}$$

Hence, by (17), we have

$$\Phi(0) = 0 \quad \text{and} \quad \Phi(1) = \Phi \quad \text{for all } \Phi. \tag{22}$$

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

Observe that, if $d = p - 1$, then (22) and (19) imply also (20).

If p is the order of the R-K method (2)–(3), then we define

$$q := [(p+1)/2],$$

where $[\cdot]$ stands for "the integer part of".

Moreover, we define ν^* as the number of distinct values of the numerical coefficients c_i, $i = 1, \dots, \nu$. One obviously has $\nu^* \leqslant \nu$.

THEOREM 5. *If u is a NCE of the ν-stage R-K method* (2)–(3) *of order p, then its degree d must satisfy*

$$q \leqslant d \leqslant \min\{\nu^*, p\}. \tag{23}$$

Further, the polynomials $b_i'(\theta)$, $i = 1, \dots, \nu$, span the whole space Π_{d-1} of polynomials of degree $\leqslant d-1$.

Proof. Butcher [4] has shown that, for every $r \geqslant 1$, there exists an elementary weight of the form

$$\Phi = \sum_{i=1}^{\nu} b_i c_i^{r-1} \quad \bigl(\text{if } c_{i^*} = 0, \text{ then } c_{i^*}^0 := 1\bigr).$$

Hence, by Proposition 2, and by (17) and (19), we can state

$$\sum_{i=1}^{\nu} b_i'(\theta) c_i^{r-1} \equiv \theta^{r-1}, \qquad 1 \leqslant r \leqslant d, \tag{24}$$

since in this case $\alpha = \beta = 1$.

It is clear that the matrix $C := ((c_i^{r-1}))$ satisfies

$$\operatorname{rank}(C) = \min\{\nu^*, d\}.$$

Moreover, if we choose d distinct values θ_j, $0 \leqslant \theta_j \leqslant 1$ ($0^0 := 1$), we can consider the matrices $B := ((b_i'(\theta_j)))^T$ and $\Theta := ((\theta_j^{r-1}))$. We obviously have

$$\operatorname{rank}(B) \leqslant d \quad \text{and} \quad \operatorname{rank}(\Theta) = d.$$

On the other hand, (24) gives

$$BC = \Theta \tag{25}$$

which, by well-known arguments of linear algebra, implies

$$d \leqslant \min\bigl\{\min\{\nu^*, d\}, \operatorname{rank}(B)\bigr\}.$$

This yields both $d \leqslant \nu^*$ and $\operatorname{rank}(B) = d$.

The order of the R-K method being p, the inequality $d \leqslant \nu^*$, by (8) and (12), implies $d \leqslant \min\{\nu^*, p\}$. Moreover, $\operatorname{rank}(B) = d$ states that at least d polynomials among the $b_i'(\theta)$'s are linearly independent. Since they all have degree $\leqslant d-1$, they span the whole space Π_{d-1}.

In order to prove $q \leqslant d$, we observe that, for $r = d+1$, condition (20) must be satisfied for every $s = 0, \dots, p-d-1$. This imposes $p - d$ linearly independent conditions to the polynomial $\Phi'(\theta)$, the degree of which is $\leqslant d-1$. Thus, we must have $p - d \leqslant d$, and then $q \leqslant d$. □

This theorem gives a range for the possible degree d of a NCE. However it is not true that there exists a NCE for every d satisfying (23).

As far as the existence of NCEs is concerned, we have the following theorems.

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

THEOREM 6. *Let $p \geqslant 2$. If the R-K process* (2)–(3) *has a NCE u of degree $d > q$, then it has also a NCE $\tilde{u}$ of degree d' for every $d' = q, \ldots, d-1$.*

THEOREM 7. *Every R-K process* (2)–(3) *has a NCE u of minimal degree q.*

Before proving these theorems, we remark that Theorem 7 gives a positive answer to the question we have posed in Section 1.

We shall need the following lemma to prove Theorem 6.

LEMMA 8. *If n is an integer such that $q \leqslant n \leqslant p - 1$ (this can be true for $p \geqslant 2$), then there exists a linear projector P_n of the space C^0 of the continuous functions in* $[0,1]$ *onto the space Π_{n-1} of polynomials of degree $\leqslant n-1$, such that*

$$\int_0^1 \theta^s P_n f(\theta)\, d\theta = \int_0^1 \theta^s f(\theta)\, d\theta, \qquad s = 0, \ldots, p-n-1, \tag{26}$$

for every $f \in C^0$.

Proof. Since $[(p+1)/2] \leqslant n \leqslant p-1$, we have $1 \leqslant p - n \leqslant [p/2]$. On the other hand $[p/2] \leqslant [(p+1)/2]$ and hence $1 \leqslant p - n \leqslant n$. This is sufficient to prove the lemma, provided other $2n - p$ linear conditions (if $2n - p > 0$), linearly independent of those defined by (26) on the space Π_{n-1}, are chosen. □

Proof of Theorem 6. Since $q \leqslant d' < d \leqslant p$, by Lemma 8 we get the existence of a linear projector $P_{d'}$ of C^0 onto $\Pi_{d'-1}$ such that (26) holds with $n = d'$. Put

$$\tilde{b}_i'(\theta) :\equiv P_{d'} b_i'(\theta), \qquad \tilde{b}_i(0) := 0, \qquad i = 1, \ldots, \nu,$$

and consider the polynomial (see (7))

$$\tilde{u}(t_0 + \theta h) :\equiv y_0 + h \sum_{i=1}^{\nu} \tilde{b}_i(\theta) g^{(i)}, \qquad 0 \leqslant \theta \leqslant 1.$$

By (17), we can state that the corresponding continuous elementary weights are

$$\tilde{\Phi}'(\theta) \equiv P_{d'} \Phi'(\theta), \qquad \tilde{\Phi}(0) = 0. \tag{27}$$

Since (19) holds for u and since $P_{d'}$ reduces to the identity map on the space $\Pi_{d'-1}$, by (27) we can state also

$$\tilde{\Phi}'(\theta) \equiv \frac{\alpha \theta^{r-1}}{\beta}$$

for all $\tilde{\Phi}$ corresponding to elementary differentials $\tilde{F}$ of order $r \leqslant d'$.

That is, (19) is verified also for $\tilde{u}$.

Moreover, since (19) and (20) hold for u, we have that

$$\int_0^1 \theta^s \Phi'(\theta)\, d\theta = \frac{\alpha}{\beta(r+s)}$$

for all Φ corresponding to elementary differentials F of order $r = d' + 1, \ldots, p$ and for every $s = 0, \ldots, p - r$.

Since $p - r \leqslant p - d' - 1$, by Lemma 8 and by (27) we have that (20) is satisfied also by $\tilde{u}$.

To complete the proof, observe that $p - d' - 1 \geqslant 0$. Thus, for $s = 0$, Lemma 8, together with (21) for u, yields

$$\tilde{b}_i(1) = \int_0^1 \tilde{b}_i'(\theta)\, d\theta = \int_0^1 b_i'(\theta)\, d\theta = b_i(1) = b_i, \qquad i = 1, \ldots, \nu,$$

that is, (21) is satisfied also by $\tilde{u}$. □

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

Proof of Theorem 7. For every $i = 1, \dots, \nu$, consider the unique polynomial $b_i(\theta)$ of degree $\leqslant q$ which satisfies

$$b_i(0) = 0 \tag{28}$$

and

$$\int_0^1 \theta^s b_i'(\theta)\, d\theta = b_i c_i^s, \qquad s = 0, \dots, q-1. \tag{29}$$

Then, by using (7), we get a polynomial u of degree $\leqslant q$.

By (29) for $s = 0$ we get $b_i(1) = b_i$, so that, by (28), Eqs. (21) and (22) hold.

Let $1 \leqslant r \leqslant p$ and let $\Phi(\theta)$ be a continuous elementary weight corresponding to an elementary differential F of order r. Then (17), together with (29), yields for $r + s \leqslant p$,

$$\int_0^1 \theta^s \Phi'(\theta)\, d\theta = \sum_{i=1}^{\nu} \left(\int_0^1 \theta^s b_i'(\theta)\, d\theta \right) \Phi^{(i)} = \sum_{i=1}^{\nu} b_i c_i^s \Phi^{(i)},$$

which, by Proposition 4, is an elementary weight $\tilde{\Phi}$ corresponding to an elementary differential $\tilde{F}$ of order $r + s \leqslant p$. $\tilde{\Phi}$ is such that $(r+s)\tilde{\Phi} = r\Phi$, where, by (22), $\Phi = \Phi(1)$. Hence, since $\Phi = \alpha/\beta r$ by Proposition 3, we get

$$\int_0^1 \theta^s \Phi'(\theta)\, d\theta = \frac{\alpha}{\beta(r+s)} \quad \text{if } r + s \leqslant p \text{ and } s = 0, \dots, q-1. \tag{30}$$

Since $p - q - 1 \leqslant q - 1$, we can conclude that (20) holds with $d = q$.

On the other hand, for $1 \leqslant r \leqslant q$, we have $q - 1 \leqslant p - r$. Thus (30) implies that the polynomial $\Phi'(\theta) - \alpha\theta^{r-1}/\beta$, the degree of which is $\leqslant q - 1$, lies in the kernel of q linearly independent linear functionals; this implies (19) with $d = q$.

The existence of a NCE u of degree q is proved. □

Observe that both proofs are of a constructive type.

In general, nothing can be said about the uniqueness of the NCEs and, in fact, there are examples of nonuniqueness. The only uniqueness result is included in Theorem 9 (next section).

3. Collocation, Perturbed Collocation and NCEs. Further Examples. In order to simplify the discussion, we briefly recall the definition of PECO-method given by Nørsett and Wanner [10].

They define a perturbation operator P: $\Pi_\nu \to \Pi_\nu$ by

$$Pu(t) :\equiv u(t) + \sum_{j=1}^{\nu} N_j(\theta) u^{(j)}(t_0) h^j, \qquad 0 \leqslant \theta := \frac{t - t_0}{h} \leqslant 1,$$

where the $N_j(\theta)$'s are polynomials of degree $\leqslant \nu$. Then the PECO-method is

$$\begin{cases} u(t_0) := y_0, & u \in \Pi_\nu, \\ u'(t_0 + c_i h) = f(t_0 + c_i h,\ Pu(t_0 + c_i h)), & i = 1, \dots, \nu, \\ \bar{y} := u(t_0 + h), \end{cases} \tag{31}$$

where the c_i's are assumed to be distinct. They define the polynomial $M(\theta) :\equiv \prod_{i=1}^{\nu}(\theta - c_i)$, which is of degree ν.

The PECO-method (31) is equivalent to a suitable R-K process (2)–(3) which has the same c_i's (hence $\nu^* = \nu$) and is such that $g^{(i)} = u'(t_0 + c_i h)$, $i = 1, \dots, \nu$.

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

In particular, for $N_j(\theta) \equiv 0$, $j = 1, \ldots, \nu$, that is, for $P = I$ (identity map), ordinary collocation is obtained.

THEOREM 9. *Let the ν-stage R-K process* (2)–(3) *of order p be such that $\nu^* = \nu \leqslant p$. Then it has a NCE u of degree ν if and only if it is equivalent to a PECO-method* (31) *such that $N_j(\theta) \equiv 0$, $j = 1, \ldots, \nu - 1$, and, if $\nu < p$, such that*

$$\int_0^1 \theta^s M(\theta)\, d\theta = \int_0^1 \theta^s N_\nu(\theta)\, d\theta = 0, \qquad s = 0, \ldots, p - \nu - 1.$$

Moreover, in this case, the NCE u of degree ν is unique and it is just the perturbed collocation solution.

Proof. First of all, we observe that $d = \nu$ satisfies the necessary condition (23), since (as is well-known) $\nu \geqslant q$.

Consider, like Nørsett and Wanner [10], the Gröbner-Alekseev nonlinear variation-of-constants formula. If u is a function such that $u(t_0) = y_0$, then

$$y(t) - u(t) = \int_{t_0}^{t} K(t, x)\big[f(x, u(x)) - u'(x)\big]\, dx, \tag{32}$$

where $K(t, x)$ is a variational matrix depending on u such that $K(t, t) \equiv I$.

By differentiating (32) we get

$$\begin{aligned} y'(t) - u'(t) = \int_{t_0}^{t} \frac{\partial}{\partial t} K(t, x)\big[f(x, u(x)) - u'(x)\big]\, dx \\ + f(t, u(t)) - u'(t). \end{aligned} \tag{33}$$

Then, by multiplying (33) by a sufficiently smooth matrix-valued function G, after a small calculation we obtain

$$\begin{aligned} &\int_{t_0}^{t_0+h} G(t)\big[y'(t) - u'(t)\big]\, dt \\ &\qquad = \int_{t_0}^{t_0+h} H(t_0 + h, x)\big[f(x, u(x)) - u'(x)\big]\, dx, \end{aligned} \tag{34}$$

where

$$H(t, x) :\equiv G(x) + \int_x^t G(\xi) \circ \frac{\partial}{\partial t} K(\xi, x)\, d\xi.$$

If) By the equivalence of the methods, (8) follows for the perturbed collocation solution u.

Like Nørsett and Wanner (Theorem 10 in [10]), we split (34) in two parts:

$$\int_{t_0}^{t_0+h} G(t)\big[y'(t) - u'(t)\big]\, dt = (\mathrm{I}) + (\mathrm{II}),$$

where

$$(\mathrm{I}) := \int_{t_0}^{t_0+h} H(t_0 + h, x)\big[f(x, Pu(x)) - u'(x)\big]\, dx$$

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

and

$$\begin{aligned}(\mathrm{II}) &:= \int_{t_0}^{t_0+h} H(t_0+h,x)[f(x,u(x)) - f(x,Pu(x))]\,dx \\ &= -\int_{t_0}^{t_0+h} H(t_0+h,x)\circ\frac{\partial}{\partial y}f(x,u(x))N_\nu\left(\frac{x-t_0}{h}\right)u^{(\nu)}(t_0)h^\nu\,dx \\ &\quad + O(h^{2\nu+1}).\end{aligned}$$

Since (as is well-known) $p \leqslant 2\nu$ and since the hypotheses of Theorem 10 in [10] are satisfied, we can similarly conclude that (10) holds for u.

Analogously, we split (33) in two parts:

$$y'(t) - u'(t) = (\mathrm{III}) + (\mathrm{IV}),$$

where

$$(\mathrm{III}) := \int_{t_0}^{t} \frac{\partial}{\partial t}K(t,x)[f(x,Pu(x)) - u'(x)]\,dx + f(t,Pu(t)) - u'(t)$$

and

$$\begin{aligned}(\mathrm{IV}) &:= \int_{t_0}^{t} \frac{\partial}{\partial t}K(t,x)[f(x,u(x)) - f(x,Pu(x))]\,dx + f(t,u(t)) - f(t,Pu(t)) \\ &= -\int_{t_0}^{t} \frac{\partial}{\partial t}K(t,x)\circ\frac{\partial}{\partial y}f(x,u(x))N_\nu\left(\frac{x-t_0}{h}\right)u^{(\nu)}(t_0)h^\nu\,dx \\ &\quad - \frac{\partial}{\partial y}f(t,u(t))N_\nu\left(\frac{t-t_0}{h}\right)u^{(\nu)}(t_0)h^\nu + O(h^{2\nu}).\end{aligned}$$

By Proposition 9 in [10], the derivatives of u remain uniformly bounded as $h \to 0$. Thus (IV) $= O(h^\nu)$ and, since $f(t, Pu(t)) - u'(t)$ vanishes at ν points (see (31)), also (III) $= O(h^\nu)$. Therefore (9), too, is satisfied by u. We can conclude that the perturbed collocation solution u is a NCE of degree ν.

Only if) If a NCE u of degree ν exists, then (24) becomes

$$\sum_{i=1}^{\nu} b_i'(\theta)c_i^{r-1} \equiv \theta^{r-1}, \qquad r = 1,\dots,\nu.$$

Therefore, all the three matrices in (25) are square of dimension ν, and C is invertible, since $\nu^* = \nu$. In particular, by choosing $\theta_j := c_j$, $j = 1,\dots,\nu$, we have $\Theta = C$ and hence $B = I$. This implies that the polynomials $b_i'(\theta)$ are uniquely determined and that they must coincide with the Lagrange polynomials relative to the nodes c_i. Hence, (7) yields $u'(t_0 + c_ih) = g^{(i)}$, $i = 1,\dots,\nu$. Moreover, the R-K process (2)–(3) turns out to be interpolatory and thus, by Theorem 6 in [10], it is equivalent to a PECO-method (31), the solution $\tilde{u}$ of which satisfies $\tilde{u}'(t_0 + c_ih) = g^{(i)}$, $i = 1,\dots,\nu$. This yields $\tilde{u} = u$.

Since (9) holds for u and since, by (31), also (III) $= O(h^\nu)$, it is immediately seen that one must have $N_j(\theta) \equiv 0$, $j = 1,\dots,\nu - 1$, in order that (IV) $= O(h^\nu)$, too.

The R-K process (2)–(3) being interpolatory of order p, the quadrature formula with the nodes c_i and weights b_i is exact for polynomials of degree $\leqslant p - 1$. This implies $\int_0^1 \theta^s M(\theta)\,d\theta = 0$, $s = 0,\dots,p-\nu-1$, and (I) $= O(h^{p+1})$. It follows that (II) $= O(h^{p+1})$, from which also $\int_0^1 \theta^s N_\nu(\theta)\,d\theta = 0$, $s = 0,\dots,p-\nu-1$. □

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

This theorem, together with Theorem 7, shows that every R-K process (2)–(3) of optimal order, that is such that $\nu = q$ (e.g., those in Butcher [5] and, in part, [6]), has a unique NCE, which is just the perturbed collocation solution of the equivalent PECO-method. Recall that, if $p = 2\nu$, then the equivalent PECO-method is the collocation method at ν Gaussian points.

By Theorem 6, if a R-K process (2)–(3) is not of optimal order (that is $\nu > q$) and satisfies the hypotheses of Theorem 9 and the respective sufficient and necessary condition, then it has more than one NCE: it has a NCE of degree d for every $d = q, \dots, \nu$. The NCE of degree ν is unique; it is the perturbed collocation solution of the equivalent PECO-method (which, in particular, may reduce to a collocation method).

It is interesting to point out that, whenever a R-K process (2)–(3) is equivalent to a PECO-method such that $N_{j^*}(\theta) \not\equiv 0$ for a $j^* < \nu - 1$, the perturbed collocation solution is not a NCE.

We conclude this section by listing all the NCEs for some of the most popular explicit R-K processes (2)–(3), that is, methods for which $a_{ij} = 0$ for $i \leqslant j$.

1-*stage R-K process of order* 1 (*Euler method*).

$$d = q = \nu = p = 1 \qquad b_1(\theta) \equiv \theta.$$

2-*stage R-K processes of order* 2.

$$d = q = 1 \qquad b_i(\theta) \equiv b_i\theta, \qquad i = 1, 2.$$

$$d = \nu = p = 2 \qquad \begin{cases} b_1(\theta) \equiv (b_1 - 1)\theta^2 + \theta \\ b_2(\theta) \equiv b_2\theta^2. \end{cases}$$

3-*stage R-K processes of order* 3 (*with* $c_2, c_3 \neq 0$).

$$d = q = 2 \qquad b_i(\theta) \equiv w_i\theta^2 + (b_i - w_i)\theta, \qquad i = 1, 2, 3,$$

where $w_1 := -[\lambda(c_3 - c_2) + c_2]/2c_2c_3$, $w_2 := \lambda/2c_2$, $w_3 := (1 - \lambda)/2c_3$ and $\lambda \in \mathbf{R}$.

Here we have a one-parameter family of NCEs, that is, an example of nonuniqueness of NCEs of minimal degree q (observe that $\nu > q$). The NCE furnished by Theorem 7 is

$$b_i(\theta) \equiv 3(2c_i - 1)b_i\theta^2 + 2(2 - 3c_i)b_i\theta, \qquad i = 1, 2, 3,$$

which corresponds to the particular value $\lambda = 6c_2(2c_2 - 1)b_2$.

In this case no NCE of degree $\nu = p$ exists.

4-*stage R-K processes of order* 4.

$$d = q = 2 \qquad b_i(\theta) \equiv 3(2c_i - 1)b_i\theta^2 + 2(2 - 3c_i)b_i\theta, \qquad i = 1, 2, 3, 4.$$

For $d = q$ the NCE is only the one furnished by Theorem 7.

$$d = 3 \qquad \begin{cases} b_1(\theta) \equiv 2(1 - 4b_1)\theta^3 + 3(3b_1 - 1)\theta^2 + \theta, \\ b_i(\theta) \equiv 4(3c_i - 2)b_i\theta^3 + 3(3 - 4c_i)b_i\theta^2, \qquad i = 2, 3, 4. \end{cases}$$

Also in this case no NCE of degree $\nu = p$ exists.

4. Applications. The main application of the NCEs can be found, in our opinion, in the next theorem.

Before stating it, we consider the following IVP with driving equation:

$$(35) \qquad \begin{cases} z'(\tau) = F_{(y)}(\tau, z(\tau)), \\ z(\tau_0) = z_0 \end{cases}$$

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

where $F_{(y)}(\tau, x) :\equiv F(\tau, x, y(\phi(\tau)), y'(\phi(\tau)))$, z, z_0 and F are m-vectors, τ is a real variable and y is the solution of (1). Also in this case F is supposed to be sufficiently smooth. Moreover, the function ϕ is one-to-one between $[\tau_0, \tau_0 + \tilde{\rho}]$ and $[t_0 = \phi(\tau_0), t_0 + \rho = \phi(\tau_0 + \tilde{\rho})]$ and is sufficiently smooth together with its inverse ψ.

THEOREM 10. *Consider a R-K process* (2)–(3) *of order p to approximate the solution of* (1) *at the point* $t_0 + h$, $h \leqslant \rho$, *and let u be a NCE* (*of degree d*). *Moreover, consider the IVP*

$$\text{(36)} \qquad \begin{cases} w'(\tau) = F_{(u)}(\tau, w(\tau)), \\ w(\tau_0) = z_0, \end{cases}$$

in the interval $[\tau_0, \tau_0 + \tilde{h} = \psi(t_0 + h)]$, *where*

$$F_{(u)}(\tau, x) :\equiv F\big(\tau, x, u(\phi(\tau)), u'(\phi(\tau))\big).$$

If z is the solution of (35), *then*

$$\text{(37)} \qquad \max_{\tau_0 \leqslant \tau \leqslant \tau_0 + \tilde{h}} |z^{(k)}(\tau) - w^{(k)}(\tau)| = O(h^{d+1-k}), \qquad k = 0, \dots, d,$$

and all the higher derivatives of w remain uniformly bounded as $h \to 0$;

$$\text{(38)} \qquad |z(\tau_0 + \tilde{h}) - w(\tau_0 + \tilde{h})| = O(h^{p+1});$$

$$\text{(39)} \qquad \int_{\tau_0}^{\tau_0 + \tilde{h}} G(\tau)\big[z^{(k)}(\tau) - w^{(k)}(\tau)\big]\, d\tau = O(h^{p+1}), \qquad k = 0, 1,$$

for every sufficiently smooth matrix-valued function G.

Proof. By (9) and (12) we have

$$\text{(40)} \qquad \begin{aligned} &\max_{\tau_0 \leqslant \tau \leqslant \tau_0 + \tilde{h}} |y^{(k)}(\phi(\tau)) - u^{(k)}(\phi(\tau))| \\ &\quad = \max_{t_0 \leqslant t \leqslant t_0 + h} |y^{(k)}(t) - u^{(k)}(t)| = O(h^{d+1-k}), \qquad k = 0, 1. \end{aligned}$$

This yields

$$\max_{\tau_0 \leqslant \tau \leqslant \tau_0 + \tilde{h}} |F_{(y)}(\tau, x) - F_{(u)}(\tau, x)| = O(h^d)$$

uniformly with respect to x in any bounded set of **R**.

By standard arguments on IVPs for ODEs, we get

$$\max_{\tau_0 \leqslant \tau \leqslant \tau_0 + \tilde{h}} |z'(\tau) - w'(\tau)| = O(h^d).$$

This easily implies (37) for every $k = 0, \dots, d$. The uniform boundedness of the higher derivatives of w (as $h \to 0$) follows by (11) and by the smoothness of ϕ and F.

Before proving (39), observe that (38) is a particular case for $G(\tau) \equiv I$ (identity matrix) and $k = 1$.

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

If G is a sufficiently smooth matrix-valued function, then, by (35) and (40), the analogue of (34) for the IVP (36) becomes

$$\begin{aligned}\int_{\tau_0}^{\tau_0+\tilde{h}} & G(\tau)[z'(\tau) - w'(\tau)]\,d\tau \\ &= \int_{\tau_0}^{\tau_0+\tilde{h}} H(\tau_0+\tilde{h}, x)\left[\frac{\partial}{\partial y}F_{(y)}(x, z(x))(y(\phi(x)) - u(\phi(x)))\right. \\ &\qquad\qquad \left.\frac{\partial}{\partial y'}F_{(y)}(x, z(x))(y'(\phi(x)) - u'(\phi(x)))\right] dx \\ &\quad + O(h^{2d+1}),\end{aligned}$$

where $H(\tau, x)$ is a suitable matrix depending on w, the derivatives of which remain uniformly bounded as $h \to 0$.

Since $2d \geqslant 2q \geqslant p$, by (10) and (13) and by substituting $x := \psi(\xi)$, we obtain

$$\begin{aligned}\int_{\tau_0}^{\tau_0+\tilde{h}} & G(\tau)[z'(\tau) - w'(\tau)]\,d\tau \\ &= \int_{t_0}^{t_0+h} H(\tau_0+\tilde{h}, \psi(\xi))\left[\frac{\partial}{\partial y}F_{(y)}(\psi(\xi), z(\psi(\xi)))(y(\xi) - u(\xi))\right. \\ &\qquad\qquad \left.\frac{\partial}{\partial y'}F_{(y)}(\psi(\xi), z(\psi(\xi)))(y'(\xi) - u'(\xi))\right]\psi'(\xi)\,d\xi \\ &\quad + O(h^{2d+1}) = O(h^{p+1}).\end{aligned}$$

Hence (39) is proved for $k = 1$. Simple integration by parts, together with (38) and (39) for $k = 1$, finally yields (39) for $k = 0$. □

Roughly speaking, the properties (9), (10), (11), (12) and (13) of the perturbation $y - u$ are reproduced for the error $z - w$. In particular, the order $p + 1$ of the local error at the nodes is preserved, although the uniform rate of convergence to zero of the perturbation $y - u$ can be lower.

As an application of Theorem 10 we consider the following IVP for DDEs of neutral type:

$$(41) \qquad \begin{cases} y'(t) = f(t, y(t), y(t-\alpha(t)), y'(t-\alpha(t))), \\ y(t) \equiv \sigma(t) \quad \text{for } t \leqslant t_0, \\ y'(t) \equiv \omega(t) \quad \text{for } t < t_0, \end{cases}$$

where the delay α satisfies $\alpha(t) \geqslant \bar{\alpha} > 0$.

Recently, many papers have appeared in the literature on this subject, even in much more general contexts. Numerical methods for (41) have been investigated, for example, by Zverkina [15] and [16]. In the particular case that the equation is not of neutral type (i.e., f is independent of y'), we quote Oberle and Pesh [11], Arndt [1], Bellen and Zennaro [3] and the references therein. However, we do not treat the problem in detail here, and refer the reader to the cited papers.

We want only to remark that, in solving (41), one generally has to approximate the solution y at the retarded argument with the same (or greater) order of accuracy of the method he is using. Nevertheless, for nonneutral equations, Bellen [2] has proved that, if one performs the one-step collocation method at ν Gaussian points (of order

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

2ν), then he can use the collocation solution u to approximate the solution y at the retarded argument (which, by (5), is uniformly accurate of order $\nu + 1$ only) without losing the order 2ν at the nodes, provided a particular choice of the mesh Δ is made imposed by the delay α. The same result has been obtained by Vermiglio [12] for a one-step subregion method of order $2\nu - 2$.

They both assume the following hypothesis:

CONDITION 11. *The delay α is such that the functional argument $\phi(t)$: $\equiv t - \alpha(t)$ is strictly increasing and sufficiently smooth together with its inverse ψ.*

We recall the cited restriction on the mesh Δ.

CONDITION 12. *The mesh Δ includes the breaking points (i.e., the points at which the solution y has discontinuities in its derivatives, caused by the delay α) and, moreover, each node of the mesh is mapped exactly into another one by the functional argument ϕ.*

By virtue of Theorem 10, similar results are valid for all R-K processes. More precisely, by induction on the intervals defined by the breaking points, one easily gets the following very general result.

THEOREM 13. *If Conditions* 11 *and* 12 *hold, then any R-K process* (2)–(3) *maintains its order p for the DDE* (41), *provided the solution y is approximated by a NCE at the retarded argument.*

Vermiglio [12] has observed that the method she proposes is equivalent to that ν-stage R-K process of order $2\nu - 2$ which is equivalent to the collocation method based on the Lobatto quadrature formula at ν points. One can verify by some calculations that the continuous solution given by the subregion method is a NCE of minimal degree $\nu - 1$. Moreover, one need only observe that the collocation solution is, by Theorem 9, a NCE of degree ν.

Hence, we can conclude that the order results in [2] and in [12] are included in Theorem 13.

Going back to ODEs such as (1), another application of the NCEs could be the tabulation of the solution y at nonnodal points. However, since in general the degree d of the NCEs is lower than the order p of the R-K method, the accuracy at nonnodal points is often worse. On the other hand, by (4) and (12), for the uniform order p to be attained, it is necessary and sufficient to have a NCE of degree $d \geqslant p - 1$. If d^* is the maximum degree possible for a NCE of the R-K process (2)–(3), Theorem 5 states that $d^* \leqslant \min\{\nu^*, p\}$. More generally, only NCEs of minimal degree q are assured to exist, and this is the case for the one-step collocation method at ν Gaussian points.

For this method, the author [14] has proposed a sort of Iterated Defect Correction method in order to obtain uniform order 2ν approximation. The procedure gives rise to a sequence of $\nu - 1$ uniform corrections w_k, $k = 1, \ldots, \nu - 1$, of the collocation solution u. These improved uniform approximations w_k are polynomials of degree $\leqslant \nu + k$ which do not change the value of u at the endpoints t_0 and $t_0 + h$. Moreover, they satisfy (9) and (10) with $d = \nu + k$ and $p = 2\nu$. They are found in an explicit way by making further evaluations of the function f. An algorithm describing the procedure for $\nu = 2, 3$ is given by Bellen and Zennaro [3]. In the general case, it consists in formally embedding the equivalent ν-stage R-K process

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

(2)–(3) into another "larger" equivalent one, which has further stages $g^{(i)}$, $i = \nu + 1, \ldots, \nu'$, with corresponding weights $b_i = 0$. These further stages are utilized only to compute the improved uniform approximations w_k, which turn out to be NCEs of the new "larger" R-K process.

Furthermore, a last uniform correction w_ν is considered. It is a NCE of a still "larger" equivalent R-K process and is utilized to give an accurate estimation of the local discretization error of the original collocation method.

One can easily see that these results can be generalized to all R-K methods. By starting from a NCE of degree d^*, one performs $p - d^* - 1$ uniform corrections, by applying the same procedure as in [14], in order to find a NCE of degree $p - 1$ of a "larger" equivalent R-K process. The cost is, obviously, some extra evaluations of the function f. In this way one reaches uniform order p approximation.

A last uniform correction yields a NCE of degree p of a still "larger" equivalent R-K process, which can be utilized to estimate the local discretization error of the original R-K method. In fact we have the following theorem.

THEOREM 14. *If u is a NCE of degree p of a R-K process* (2)–(3) *of order p, then*

$$y_0 + \int_{t_0}^{t_0+h} f(t, u(t))\, dt - \bar{y} = y(t_0 + h) - \bar{y} + O(h^{p+2}). \tag{42}$$

Proof. By (12) with $d = p$ and by the smoothness of f, we get

$$\begin{aligned} f(t, u(t)) &= y'(t) - f(t, y(t)) + f(t, u(t)) \\ &= y'(t) + O(h^{p+1}), \end{aligned}$$

and this implies (42). □

Therefore, by using a quadrature formula which is exact for polynomials of degree $\leqslant p$, the left-hand side of (42) gives the desired estimation of the local discretization error (it is convenient to use Gauss, or Radau, or Lobatto quadrature formulas).

This manner of estimating the local discretization error is not convenient for explicit R-K processes, since the well-known R-K-Fehlberg and R-K-Merson methods certainly require less evaluations of the function f. On the other hand, if one uses an implicit R-K method (e.g., when some stiffness is present), then this procedure can be quite practical, if compared to Richardson's extrapolation technique.

Istituto di Matematica
Università di Trieste
Piazzale Europa 1
I-34100 Trieste, Italy

1. H. ARNDT, "Numerical solution of retarded initial value problems: Local and global error and stepsize control," *Numer. Math.*, v. 43, 1984, pp. 343–360.

2. A. BELLEN, "One-step collocation for delay differential equations," *J. Comput. Appl. Math.*, v. 10, 1984, pp. 275–283.

3. A. BELLEN & M. ZENNARO, "Numerical solution of delay differential equations by uniform corrections to an implicit Runge-Kutta method," *Numer. Math.*, v. 47, 1985, pp. 301–316.

4. J. C. BUTCHER, "Coefficients for the study of Runge-Kutta integration processes," *J. Austral. Math. Soc.*, v. 3, 1963, pp. 185–201.

5. J. C. BUTCHER, "Implicit Runge-Kutta processes," *Math. Comp.*, v. 18, 1964, pp. 50–64.

6. J. C. BUTCHER, "Integration processes based on Radau quadrature formulas," *Math. Comp.*, v. 18, 1964, pp. 233–243.

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms

7. A. GUILLOU & F. L. SOULÉ, "La résolution numérique des problèmes differentielles aux conditions initials par des méthodes de collocation," *RAIRO*, v. 3, 1969, pp. 17–44.

8. B. L. HULME, "One-step piecewise polynomial Galerkin methods for initial value problems," *Math. Comp.*, v. 26, 1972, pp. 415–426.

9. S. P. NØRSETT & G. WANNER, "The real-pole sandwich for rational approximation and oscillation equations," *BIT*, v. 19, 1979, pp. 79–94.

10. S. P. NØRSETT & G. WANNER, "Perturbed collocation and Runge-Kutta methods," *Numer. Math.*, v. 38, 1981, pp. 193–208.

11. H. J. OBERLE & H. J. PESH, "Numerical treatment of delay differential equations by Hermite interpolation," *Numer. Math.*, v. 37, 1981, pp. 235–255.

12. R. VERMIGLIO, "A one-step subregion method for delay differential equations," *Calcolo*. (In press.)

13. K. WRIGHT, "Some relationships between implicit Runge-Kutta, collocation and Lanczos τ-method, and their stability properties," *BIT*, v. 10, 1970, pp. 217-227.

14. M. ZENNARO, "One-step collocation: Uniform superconvergence, predictor-corrector method, local error estimate," *SIAM J. Numer. Anal.*, v. 22, 1985.

15. T. S. ZVERKINA, "A modification of finite-difference methods for integrating ordinary differential equations with nonsmooth solutions," *Zh. Vychisl. Mat. i Mat. Fiz.*, v. 4 (suppl.), 1964, pp. 149–160. (In Russian.)

16. T. S. ZVERKINA, "A modified Adams' formula for the integration of equations with deviating argument," *Trudy Sem. Teor. Differencial. Uravnenĭi s Otklon. Argumentom*, Univ. Družby Narodov Patrisa Lumumby, vol. 3, 1965, pp. 221–232. (In Russian.)

This content downloaded from 155.198.30.43 on Wed, 13 Nov 2019 14:58:40 UTC
All use subject to https://about.jstor.org/terms