Annals of Numerical Mathematics 1 (1994) 199-212

On the error committed by stopping the Newton iteration in implicit Runge-Kutta methods

M.N. Spijker

Department of Mathematics and Computer Science, University of Leiden, The Netherlands

This paper concerns the numerical solution of initial value problems for nonlinear ordinary differential equations by implicit Runge–Kutta methods. For non-stiff problems, fixed point iteration, also called Picard iteration, is a classical approach to the solution of the system of (algebraic) equations occurring in each time step of these methods. The order of the error due to the stopping of this process after a fixed number of iterations is well understood. For stiff problems, Picard iteration is not appropriate and some variant of the Newton method is usually used instead. This paper addresses the problem of estimating the stopping error of Newton-like iterations. We aim for an understanding of this error comparable to what is known about Picard iterations. Because of stiffness, the theory is more delicate than for Picard iterations.

The first part of the paper reviews various estimates of the Newton stopping errors. The second part explores the effect of these errors on the Runge–Kutta approximations to the solution of the given initial value problem. Most of the error estimates established in this paper reflect order reduction phenomena in the presence of stiffness. This order reduction is confirmed by numerical experiments.

Keywords: ordinary differential equations, stiff initial value problems, implicit Runge–Kutta methods, Newton's method, order reduction due to stiffness.

Subject classification: AMS(MOS): primary 65L05, 65L06; secondary 65H10.

1. Introduction

In this paper we deal with the numerical solution of initial value problems of the form

$$U'(t) = f(t, U(t)) \qquad \text{for} \quad 0 \leq t \leq T, \quad U(0) = u_0. \tag{1.1}$$

Here $u_0 \in \mathbb{R}^s$ is given, and $U(t) \in \mathbb{R}^s$ is unknown, whereas $s \geq 1$. Further f is a given mapping from $[0, T] \times \mathcal{D}$ to $\mathbb{R}^s$, where $\mathcal{D} \subset \mathbb{R}^s$ is open and convex. We assume the function f to possess continuous partial derivatives up to the second order on its domain of definition.

In the following the initial value problem (1.1) will be assumed to be *stiff*. For the general concept of stiffness in initial value problems we refer to Dekker & Verwer (1984), Frank et al. (1985), Butcher (1987), Hairer & Wanner (1991).

Let $h > 0$ denote a *stepsize*, and consider *gridpoints* $t_n = nh$ in $[0, T]$ for integer values of n. The general Runge–Kutta method for solving (1.1) provides approximations $u_n \simeq U(t_n)$ computed recursively by

$$u_{n+1} = u_n + h \sum_{i=1}^{m} b_i f(t_n + c_i h, \xi_i) \qquad \text{for} \quad n = 0, 1, 2, \ldots . \tag{1.2}$$

Here ξ_i are vectors in $\mathcal{D}$ which depend on n and have to satisfy the system of (nonlinear) equations

$$-\xi_i + u_n + h \sum_{k=1}^{m} a_{ik} f(t_n + c_k h, \xi_k) = 0 \qquad \text{for} \quad 1 \le i \le m. \tag{1.3}$$

m denotes the *number of stages*, and a_{ik}, b_i, c_i are coefficients defining the Runge–Kutta method. We assume

$$\sum_{k=1}^{m} b_k = 1, \quad \text{and} \quad \sum_{k=1}^{m} a_{ik} = c_i \quad \text{with} \quad 0 \le c_i \le 1 \quad \text{for} \quad i = 1, 2, \ldots, m.$$

We introduce the $m \times m$ matrix $A = (a_{ik})$, and suppose that an $m \times m$ diagonal matrix D exists such that

$$\text{both } D \text{ and } (DA + A^T D) \text{ are positive definite.} \tag{1.4}$$

Many Runge–Kutta methods that are of interest in the solution of stiff problems satisfy (1.4), see Dekker & Verwer (1984), Frank et al. (1985), Butcher (1987), Hairer & Wanner (1991).

It is convenient to introduce vectors

$$x = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_m \end{pmatrix} \in \mathcal{D}^m \quad \text{and} \quad F(x) = \begin{pmatrix} F_1(x) \\ \vdots \\ F_m(x) \end{pmatrix} \in \mathbb{R}^{sm}$$

so that the left-hand members of (1.3) are equal to $F_i(x) (1 \le i \le m)$. The system (1.3) thus becomes equivalent to

$$F(x) = 0. \tag{1.5}$$

Newton's method for the numerical solution of (1.5) reads

$$F'(x_j)(x_{j+1} - x_j) = -F(x_j) \quad (j = 0, 1, 2, \ldots). \tag{1.6}$$

Here $F'(x)$ denotes the Jacobian matrix of F at x, and x_j are approximations to the solution x^* of (1.5).

In the actual numerical solution of nonlinear stiff problems (1.1) the vector x^* is almost invariably approximated by an iterate x_j obtained from Newton's method (1.6) or a simplified version thereof (in which all matrices $F'(x_j)$ in (1.6) are replaced by some approximation, e.g. $F'(x_0)$). In the following we study, for a given $j \geq 1$, the order of the errors due to the stopping of the iteration after j steps and replacing x^* simply by x_j. We shall measure all errors by their Euclidean norms and analyse their orders in terms of the stepsize h.

In the Sections 2.1 – 2.4 we review various estimates for the Euclidean norm $|x_j - x^*|$ of the Newton stopping error $x_j - x^*$. Some of these estimates reflect order reduction phenomena in the presence of stiffness. This order reduction is confirmed by a numerical experiment presented in Section 2.2. In Section 2.5 closely related questions are discussed among which the order of the stopping error $x_j - x^*$ for a simplified Newton process.

In Chapter 3 we deal with the effect that the stopping errors $x_j - x^*$, as discussed in Chapter 2, may have on the approximations u_n actually obtained. In Section 3.1 we analyse the local stopping error, which is nothing but the above effect after one time step of the Runge–Kutta method. In the rest of the paper the result of this analysis is used to explore the order of the global stopping error. The latter error amounts to the accumulated effect of n stopping errors of the form $x_j - x^*$ after n time steps of the Runge–Kutta method. We focus on the situation where $t_n \in (0, T]$ is fixed and $h \to 0$, $n \to \infty$.

2. The Newton stopping error

2.1. AN ERROR ANALYSIS NEGLECTING STIFFNES

Let $t_n \in [0, T)$, $u_n \in \mathcal{D}$ be given. In assessing the Euclidean norm $|x_j - x^*|$ of the error $x_j - x^*$ we assume that the initial guess x_0 satisfies

$$|x_0 - x^*| = \mathcal{O}(h^q) \tag{2.1}$$

(with $\mathcal{O}$-constant of moderate size, and $q > 0$). The left hand member of (2.1) stands for the Euclidean norm of the starting error $x_0 - x^*$.

By using Taylor series expansions in a straightforward way the corresponding errors $x_j - x^*$ can be estimated. For the exact Newton process (1.6) one arrives in this manner at

$$|x_j - x^*| = \mathcal{O}(h^{R(j)}) \quad \text{with} \quad R(j) = 2^j(q+1) - 1 \quad \text{for} \quad j = 1, 2, 3, \ldots . \tag{2.2}$$

Estimates corresponding to (2.2) can be found e.g. in Sugiura & Torii (1991) and Jackson, Kvaernø & Nørsett (1991).

However, stiff problems (1.1) may be characterized by very large magnitudes in the $s \times s$ Jacobian matrix (with respect to ξ) of the function $h \cdot f(t, \xi)$. Here

h is a "natural" stepsize for the problem and $t \simeq t_n$, $\xi \simeq U(t_n)$. In the above derivation of (2.2) the 1^{st} and 2^{nd} order derivatives with respect to ξ of $h \cdot f(t, \xi)$ are replaced simply by $\mathcal{O}(h)$. Accordingly, the conclusion (2.2) based on these replacements may be relevant only to nonstiff problems. For stiff problems it is thus questionable whether the $\mathcal{O}$-constant in (2.2) is still of moderate size.

2.2. A NUMERICAL EXPERIMENT

In order to check the relevance of (2.2) to arbitrary stiff problems (1.1) we consider the following example:

$$\begin{array}{lcl} U'(t) & = & -10^{11}[U(t)]^3 + 1 + 10^{11}(1+t)^3 \quad \text{for} \quad 0 \le t \le 1/4, \\ U(0) & = & 1. \end{array} \tag{2.3}$$

The true solution to this problem equals $U(t) = 1 + t$, so that any "natural" stepsize in the numerical solution of (2.3) need not be small. The large factor 10^{11} causes (2.3) to be stiff.

We consider the implicit midpoint rule, i.e. (1.2), (1.3) with

$$m = 1, \quad a_{11} = 1/2. \tag{2.4}$$

We choose $n = 0$ and consider the corresponding equation (1.5) in the situation (2.3), (2.4). Further, we choose the natural starting value $x_0 = u_0 = 1$ for Newton's method (1.6). Since (1.5) has a unique solution $x^* = 1 + h/2$, the starting error equals $x_0 - x^* = -h/2$. Hence (2.1) holds with an $\mathcal{O}$-constant of moderate size and $q = 1$.

For $j = 1$ the error estimate (2.2) with $q = 1$ reduces to

$$|x_1 - x^*| = \mathcal{O}(h^3). \tag{2.5}$$

In order to check (2.5) we have listed the (rounded) actual ratios $|x_1 - x^*|/h^3$ for various choices of h.

h	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
$\lvert x_1 - x^*\rvert/h^3$	0.25×10^{1}	0.25×10^{2}	0.25×10^{3}	0.25×10^{4}	0.25×10^{5}	0.25×10^{6}

We see from the table that the estimate (2.5) is misleading and provides no insight regarding the actual order of the errors $x_1 - x^*$.

2.3. A CLASS OF NONLINEAR STIFF PROBLEMS

In the following we present a framework in which reliable estimates for $|x_j - x^*|$ can be derived.

We consider the following two additional assumptions concerning f:

$$J(t,\eta) = J(t,\xi)\{I + e[t,\xi,\eta]\} \quad \text{with} \quad \|e[t,\xi,\eta]\| \leq \lambda(t)|\xi - \eta| \quad \text{for all} \quad t \in [0,T] \quad \text{and} \quad \xi,\eta \in \mathcal{D}; \tag{2.6}$$

$$|\widetilde{V}(\tau_1) - V(\tau_1)| \leq |\widetilde{V}(\tau_0) - V(\tau_0)| \quad \text{whenever} \quad V,\ \widetilde{V} \quad \text{are any two solutions of the differential equation on} \quad [\tau_0,\tau_1] \subset [0,T]. \tag{2.7}$$

In (2.6) we denote the $s \times s$ Jacobian matrix (with respect to ξ) of $f(t,\xi)$ by $J(t,\xi)$. Further, $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^s$, and $\|\cdot\|$ the corresponding induced matrix norm. The parameter $\lambda(t)$ in (2.6) is assumed to be of moderate size. It provides an upper bound for the relative variation of $J(t,\eta)$ when η varies through $\mathcal{D}$.

There exist arbitrarily stiff, nonlinear problems satisfying (2.6), (2.7) with $\lambda(t)$ of moderate size. As an example consider problem (2.3). The function

$$f(t,\xi) = -10^{11}\xi^3 + 1 + 10^{11}(1+t)^3$$

satisfies (2.6), (2.7) e.g. with $T = 1/4$, $\mathcal{D} = (1/2, 3/2)$, $\lambda(t) = 8$.

2.4. A RIGOROUS ERROR ANALYSIS RELEVANT TO NONLINEAR STIFF PROBLEMS

We make the same assumptions concerning f as in Chapter 1, and assume U satisfies (1.1). We denote the distance of the set $\{U(t) : 0 \leq t \leq T\}$ to the boundary of $\mathcal{D}$ by δ; if $\mathcal{D} = \mathbb{R}^s$ we put $\delta = \infty$. We assume (1.4), and use the notations

$$\|U'\| = \sup_{t\in[0,T]} |U'(t)|, \quad \|\lambda\| = \sup_{t\in[0,T]} |\lambda(t)|.$$

Let $t_n \in [0,T)$, $u_n \in \mathcal{D}$ be given and consider h with $0 < h \leq T - t_n$.

In the following theorems we use the assumption

$$|u_n - U(t_n)| \leq \alpha \cdot \delta \quad \text{and} \quad 0 < h \cdot \|U'\| \leq \alpha \cdot \delta. \tag{2.8}$$

THEOREM 2.1

Assume (2.7). Then there is a factor $\alpha > 0$ such that (1.5) has a unique solution $x^* \in \mathcal{D}^m$ whenever u_n, h satisfy (2.8). Here $\alpha = \alpha(A)$ only depends on the matrix A.

Let $K > 0$, $q > 0$ be given. The next theorem relates the assumption

$$x_0 \in \mathcal{D}^m \quad \text{with} \quad |x_0 - x^*| \leq Kh^q \leq \frac{1}{2}\min[\delta, \alpha/\|\lambda\|] \tag{2.9}$$

to the following property:

$$\text{There are unique } x_j \in \mathcal{D}^m \text{ generated by Newton's method with initial guess } x_0, \text{ and } |x_j - x^*| \leq [\alpha^{-1}\|\lambda\|Kh^q]^{2^j-1} \cdot Kh^q \quad \text{for} \quad j = 1,2,3,\ldots. \tag{2.10}$$

THEOREM 2.2

Assume (2.6), (2.7). Then there is a factor $\alpha > 0$ with the properties already stated in Theorem 2.1 and with the additional property that (2.10) holds whenever u_n, h, x_0, K and q satisfy (2.8), (2.9).

The Theorems 2.1 and 2.2 easily follow from the material in Dorsselaer & Spijker (1992). Theorem 2.2 shows that under the assumptions (2.6), (2.7), the relation (2.1) (with $\mathcal{O}$-constant K of moderate size) implies

$$|x_j - x^*| = \mathcal{O}(h^{Q(j)}) \quad \text{with} \quad Q(j) = 2^j \cdot q \quad \text{for} \quad j = 1, 2, 3, \ldots. \tag{2.11}$$

In (2.11) we have an $\mathcal{O}$-constant $K_j = [\alpha^{-1}\|\lambda\|K]^{2^j-1}K$, which is unaffected by the stiffness of (1.1). K_j is of moderate size when j is limited in a realistic way.

In the example of Section 2.2, (2.11) reduces to the error estimate

$$|x_1 - x^*| = \mathcal{O}(h^2).$$

This estimate agrees perfectly with the values in the table of Section 2.2.

2.5. VARIOUS RELATED QUESTIONS

Orders $R(j)$ and $Q(j)$ analogous to those in (2.2), (2.11) can be derived for the *simplified Newton process*

$$F'(y_0)(x_{j+1} - x_j) = -F(x_j) \quad (j = 0, 1, 2, \ldots). \tag{2.12}$$

Here y_0 is an approximation to x^*, possibly different from x_0, satisfying

$$|y_0 - x^*| = \mathcal{O}(h^r) \tag{2.13}$$

(with $\mathcal{O}$-constant of moderate size, and $0 < r \leq q$). Estimating the errors $x_j - x^*$ for this process by Taylor series expansions while neglecting stiffness, gives

$$|x_j - x^*| = \mathcal{O}(h^{R(j)}) \quad \text{with} \quad R(j) = (r+1)j + q \quad \text{for} \quad j = 1, 2, 3, \ldots. \tag{2.14}$$

Taking stiffness into account, and assuming (2.6), (2.7) as before, it can be seen from Dorsselaer & Spijker (1992) that

$$|x_j - x^*| = \mathcal{O}(h^{Q(j)}) \quad \text{with} \quad Q(j) = rj + q \quad \text{for} \quad j = 1, 2, 3, \ldots \tag{2.15}$$

with an $\mathcal{O}$-constant that is not affected by stiffness.

We note that the discussion in the Sections 2.1, 2.2 about the relevance in stiff problems of estimates involving terms $\mathcal{O}(h^R)$ is analogous to considerations regarding B-consistency in Frank et al. (1985), Dekker & Verwer (1984), Butcher

(1987), Hairer & Wanner (1991). Further, the above error estimates (2.11), (2.15) are related to interesting investigations by Alexander (1991), who studied the convergence behaviour of the iterates generated by modified Newton processes in implicit Runge–Kutta methods. The class of nonlinear stiff problems considered by Alexander has a nonempty intersection with our class specified by (2.6), (2.7) (but neither class is contained in the other one). We also note that for $\delta = \infty$ the above Theorem 2.1 is an immediate consequence of the material presented e.g. in Crouzeix et al. (1983), Butcher (1987), Hairer & Wanner (1991), and that for $\delta < \infty$ it is related to material in Alexander (1991) and Frank et al. (1985) (Section 5).

In the above we have refrained from discussing under what conditions (2.1) is fulfilled. For this question we refer to Dorsselaer & Spijker (1992), Sand (1992) and the subsequent section.

3. The local and global stopping errors

3.1. RELATING THE LOCAL STOPPING ERROR TO THE NEWTON STOPPING ERROR

Let v_n denote the approximations to $U(t_n)$ generated by the theoretical Runge–Kutta method (1.2), (1.3) starting with $v_0 = u_0$. By u_n we denote the approximations obtained when, in each time step of the Runge–Kutta method, the solution to the equation (1.3) is approximated by performing j iteration steps with Newton's method (or a variant thereof). We want to explore the order of the *global stopping errors* $D_{n,j}$ defined by

$$D_{n,j} = v_n - u_n.$$

We assume that

$$u_{n+1} = u_n + h\sum_{k=1}^{m} b_k f_k, \tag{3.1}$$

with vectors $f_k \in \mathbb{R}^s$ computed from the linear system

$$h\sum_{k=1}^{m} a_{ik} f_k = \xi_{i,j} - u_n \qquad (i = 1, 2, \ldots, m). \tag{3.2}$$

Here $\xi_{1,j}, \xi_{2,j}, \ldots, \xi_{m,j}$ belong to $\mathcal{D}$ and are equal to the subvectors of the j-th iterate x_j in $\mathcal{D}^m$, computed during the $(n+1)$-st time step, i.e.

$$\begin{pmatrix} \xi_{1,j} \\ \xi_{2,j} \\ \vdots \\ \xi_{m,j} \end{pmatrix} = x_j.$$

Note that, in view of (1.4), the matrix A is regular so that unique f_k exist satisfying (3.2).

In the following we confine ourselves to the (important) situation where the simplified Newton process (2.12) is used, and where

$$y_0 = x_0 \in \mathcal{D}^m \quad \text{with subvectors} \quad \xi_{i,0} = u_n \in \mathcal{D} \quad (i = 1, 2, \ldots, m). \tag{3.3}$$

In deriving estimates for the errors $D_{n,j}$ it is useful to define *local stopping errors* $d_{n,j}$ by

$$d_{n,j} = \widetilde{v}_{n+1} - \widetilde{u}_{n+1},$$

and to analyse the latter errors first. Here $\widetilde{v}_{n+1}$ stands for the approximation to $U(t_{n+1})$ defined by the Runge–Kutta formulae (1.2), (1.3) in which u_n is replaced by $U(t_n)$. Similarly $\widetilde{u}_{n+1}$ denotes the approximation to $U(t_{n+1})$ generated by the formulae (3.1), (3.2), (3.3) in which u_n is replaced throughout by $U(t_n)$, and where j steps are performed of the iteration (2.12).

From the relations

$$\begin{aligned} \widetilde{v}_{n+1} &= U(t_n) + h\textstyle\sum_k b_k f_k^*, \quad h\sum_k a_{ik} f_k^* = -U(t_n) + \xi_i^*, \ f_i^* = f(t_n + c_i h, \xi_i^*), \\ \widetilde{u}_{n+1} &= U(t_n) + h\textstyle\sum_k b_k f_k, \quad h\sum_k a_{ik} f_k = -U(t_n) + \xi_{i,j} \end{aligned}$$

we obtain

$$d_{n,j} = h\sum_k b_k (f_k^* - f_k), \quad h\sum_k a_{ik}(f_k^* - f_k) = \xi_i^* - \xi_{i,j}.$$

Denoting the inverse of the matrix A by (c_{ki}) there follows

$$d_{n,j} = \sum_k \sum_i b_k c_{ki} (\xi_i^* - \xi_{i,j}).$$

Consequently

$$|d_{n,j}| \leq \beta \cdot |x^* - x_j|, \tag{3.4}$$

where the constant β only depends on the coefficients b_k, a_{ik} of the Runge–Kutta method.

The right-hand member of (3.4) can be estimated by using (2.14) or (2.15) if values for q, r are available. In order to determine these values we denote by $F_i(x)$ the left-hand member of (1.3) in which u_n is replaced by $U(t_n)$. Further, $F(x) \in \mathbb{R}^{sm}$ is made up from the subvectors

$$F_1(x), F_2(x), \ldots, F_m(x) \in \mathbb{R}^s,$$

and $z_0 \in \mathcal{D}^m$ is made up from the subvectors

$$U(t_n + c_1 h), U(t_n + c_2 h), \ldots, U(t_n + c_m h) \in \mathcal{D}.$$

Clearly $|F(z_0)| \leq \gamma_0 \cdot \|U'\| \cdot h$, with a constant γ_0 only depending on the matrix A. From the material in Dorsselaer & Spijker (1992) (which is closely related to the considerations regarding BSI-stability in Dekker & Verwer (1984) and Hairer & Wanner (1991)) our last inequality can be seen to imply

$$|x^* - z_0| \leq \gamma_1 \cdot \|U'\| \cdot h.$$

Here γ_1 only depends on A, and x^* denotes the solution to the equation $F(x) = 0$. Applying (3.3) (with u_n replaced by $U(t_n)$) there follows

$$|x^* - y_0| = |x^* - x_0| \leq \gamma \cdot \|U'\| \cdot h,$$

with $\gamma = (\gamma_1 + \sqrt{m})$ only depending on A. Clearly (2.1), (2.13) hold with

$$q = r = 1 \tag{3.5}$$

and an $\mathcal{O}$-constant $K = \gamma \cdot \|U'\|$ which is not affected by stiffness.

In view of (3.4), (3.5) we can apply (2.14), (2.15) so as to obtain

$$|d_{n,j}| = \mathcal{O}(h^{R(j)}) \quad \text{with} \quad R(j) = 2j + 1 \quad \text{if stiffness is neglected,} \tag{3.6}$$

and

$$|d_{n,j}| = \mathcal{O}(h^{Q(j)}) \text{ with } Q(j) = j + 1 \text{ and an } \mathcal{O}\text{-constant not affected by stiffness.} \tag{3.7}$$

3.2. NUMERICAL EXPERIMENTS ON THE GLOBAL STOPPING ERROR

Let $Nh = T$ with integer $N \geq 1$. Then for stable Runge–Kutta methods, the global error $D_{N,j}$ may be expected to be related to the N local errors $d_{n,j}$ (with $n < N$) in such a way that

$$|D_{N,j}| = \mathcal{O}(|d_{0,j}| + |d_{1,j}| + \ldots + |d_{N-1,j}|),$$

with an $\mathcal{O}$-constant that is independent of $N = T/h$ and not affected by stiffness. In view of (3.6), (3.7) we may thus expect

$$|D_{N,j}| = \mathcal{O}(h^{2j}) \quad \text{if stiffness is neglected,} \tag{3.8}$$

and

$$|D_{N,j}| = \mathcal{O}(h^j) \quad \text{with an } \mathcal{O}\text{-constant not affected by stiffness.} \tag{3.9}$$

We shall check (3.8), (3.9) for the following 3 Runge–Kutta methods:

Backward Euler, i.e. $m = 1,\ a_{11} = 1,\ b_1 = 1,$ (3.10)

Implicit Midpoint Rule, i.e. $m = 1,\ a_{11} = \frac{1}{2},\ b_1 = 1,$ (3.11)

2-stage Gauss, i.e. $m = 2,\ a_{11} = a_{22} = 1/4,\ a_{12} = 1/4 - \sqrt{3}/6,$
$a_{21} = 1/4 + \sqrt{3}/6,\ b_1 = b_2 = 1/2.$ (3.12)

All of these methods are A-stable, B-stable and satisfy (1.4) for some $m \times m$ diagonal matrix D (see Dekker & Verwer (1984), Butcher (1987) or Hairer & Wanner (1991)).

In order to check whether the orders $2j$ or j, corresponding to (3.8), (3.9), manifest themselves in reality we introduce

$$S_j(h) = \log_2 [D_{N,j}/D_{2N,j}], \quad \text{where} \quad N = T/h.$$

Clearly $S_j(h)$ equals the order of the global stopping error observed in actual calculations.

We consider the stiff problem

$$\begin{aligned} U'(t) &= -10^{11}\{[U(t)]^3 - [1 + \sin t]^3\} + \cos t \quad \text{for} \quad 0 \le t \le T, \\ U(0) &= 1, \end{aligned} \tag{3.13}$$

the exact solution of which equals $U(t) = 1 + \sin t$.

In the table we have displayed values for $S_j(h)$.

j	$2j$	$S_j(h)$ for (3.10)	$S_j(h)$ for (3.11)	$S_j(h)$ for (3.12)
1	2	2.0008	1.9943	1.0063
2	4	3.0205	3.0044	2.0196
3	6	4.0286	4.0191	3.0180
4	8	5.0371	5.0239	4.0240

Values for $S_j(h)$ with $h = 0.01$ in problem (3.13) with $T = 1/2$.

We see that the values $S_j(h)$ for method (3.12) are in excellent agreement with (3.9), and thus confirm our estimates (3.7), (2.15). The values in the table for (3.10), (3.11) are neither in agreement with (3.8) nor (3.9). Below we go further into this anomaly; at the end of Section 3.3 the values $S_j(h)$ for (3.10), (3.11) will be explained.

We note that for various other nonlinear stiff problems, among which (2.3), values for $S_j(h)$ were found that are nearly equal to those in the above table.

3.3. RELATING THE GLOBAL STOPPING ERROR TO THE LOCAL STOPPING ERRORS

We relate the theoretical Runge–Kutta process to a function Ψ_n, and the Runge–Kutta process carried out with j simplified Newton iterations (according

to (3.1), (3.2), (3.3)) to a function Φ_n, so that we can write

$$\begin{aligned} v_{n+1} &= \Psi_n(v_n, h), \qquad \widetilde{v}_{n+1} = \Psi_n(U(t_n), h), \\ u_{n+1} &= \Phi_n(u_n, h), \qquad \widetilde{u}_{n+1} = \Phi_n(U(t_n), h). \end{aligned}$$

Since

$$d_{n,j} = \Psi_n(U(t_n), h) - \Phi_n(U(t_n), h),$$

we have the following basic relation for the global stopping error:

$$D_{n+1,j} = [\Psi_n(v_n, h) - \Psi_n(U(t_n), h)] + [\Phi_n(U(t_n), h) - \Phi_n(u_n, h)] + d_{n,j}. \quad (3.14)$$

In the following we use the notation $J_n = J(t_n, U(t_n))$, and we denote the stability function of the Runge–Kutta method by ϕ, i.e.

$$\phi(\zeta) = 1 + \zeta b^T (1 - \zeta A)^{-1} \mathbb{1}, \quad b^T = (b_1, b_2, \ldots, b_m), \quad \mathbb{1} = (1, 1, \ldots, 1)^T$$

(see Dekker & Verwer (1984), Butcher (1987), Hairer & Wanner (1991)). If the Jacobian $J(t, \xi)$ would be constant for $t_n \le t \le t_{n+1}$, $\xi \in \mathcal{D}$, we would have

$$\begin{aligned} \Psi_n(v_n, h) - \Psi_n(U(t_n), h) &= \phi(hJ_n)[v_n - U(t_n)], \\ \Phi_n(U(t_n), h) - \Phi_n(u_n, h) &= \phi(hJ_n)[U(t_n) - u_n], \end{aligned}$$

and therefore (3.14) would reduce to $D_{n+1,j} = \phi(hJ_n) D_{n,j} + d_{n,j}$. Therefore, the vectors $\overline{D}_{n,j}$ defined by

$$\begin{aligned} \overline{D}_{n+1,j} &= \phi(hJ_n)\overline{D}_{n,j} + d_{n,j} \quad \text{for} \quad n = 0, 1, \ldots, N-1, \\ \overline{D}_{0,j} &= 0 \end{aligned} \quad (3.15)$$

may be expected to be useful approximations to $D_{n,j}$.

From (2.7) it follows that the logarithmic norm $\mu(hJ_n)$, induced by the Euclidean norm in $\mathbb{R}^s$, is nonpositive (see Dorsselaer & Spijker (1992)). For any real constant ξ with $\mu(hJ_n) \le \xi$ and any rational function $\psi(\zeta)$ which is holomorphic on the complex half-plane $\mathrm{Re}\,\zeta \le \xi$ we have

$$\|\psi(hJ_n)\| \le \sup_{\mathrm{Re}\,\zeta \le \xi} |\psi(\zeta)|. \quad (3.16)$$

This inequality is a consequence of a theorem of Von Neumann (see Dekker & Verwer (1984), Hairer & Wanner (1991)). Under the assumption of A-stability we can apply (3.16) with $\xi = 0$ and $\psi = \phi$ so as to get

$$\|\phi(hJ_n)\| \le 1.$$

In view of (3.15) we thus conclude that

$$|D_{N,j}| \simeq |\overline{D}_{N,j}| \le |d_{0,j}| + |d_{1,j}| + \cdots + |d_{N-1,j}|,$$

which agrees perfectly with our considerations at the beginning of Section 3.2.

In order to explain the orders $S_j(h)$ observed for (3.10), (3.11) in Section 3.2, a more refined analysis of (3.15) is required. In the following we use a general device for obtaining refined error estimates due to Hundsdorfer (1992) and Hundsdorfer & Steiniger (1991).

LEMMA 3.1.

Let the Runge–Kutta method be A-stable, and $\phi(\infty) \neq 1$. Assume that, for a given integer $j \geq 1$, we can write the local stopping errors as

$$\begin{aligned} &d_{n,j} = h^k d(t_n) + h^{k+1}\epsilon_n, \\ &\text{with vectors } d(t),\ \epsilon_n \in \mathbb{R}^s \text{ satisfying} \\ &|d(t)| \leq K_0, \quad |d(t+h) - d(t)| \leq K_1 h, \quad |\epsilon_n| \leq K_2. \end{aligned} \tag{3.17}$$

Then the vector $\overline{D}_{N,j}$ defined by (3.15) with $Nh = T$ satisfies

$$|\overline{D}_{N,j}| \leq \frac{2K_0 + (K_1 + 2K_2)T}{|1 - \phi(\infty)|} \cdot h^k + \frac{K_0 T}{|1 - \phi(\infty)|} \cdot \omega(h\nu) \cdot h^{k-1}, \tag{3.18}$$

where

$$\omega(\xi) = \max_{\mathrm{Re}\,\zeta \leq \xi} |\phi(\zeta) - \phi(\infty)|, \quad \nu = \max_{0 \leq t \leq T} \mu[J(t, U(t))].$$

Proof

Following the device in the two papers mentioned above we write $d_{n,j}$ as

$$d_{n,j} = [I - \phi(hJ_n)]y_n + z_n,$$

with

$$y_n = \frac{h^k}{1 - \phi(\infty)} d(t_n), \quad z_n = \frac{h^k}{1 - \phi(\infty)} [\phi(hJ_n) - \phi(\infty)I] d(t_n) + h^{k+1}\epsilon_n.$$

Using (3.15) there follows

$$(\overline{D}_{n+1,j} - y_{n+1}) = \phi(hJ_n)(\overline{D}_{n,j} - y_n) + [y_n - y_{n+1} + z_n] \quad (0 \leq n \leq N-1).$$

By expressing $\overline{D}_{N,j}$ in y_{N-1}, y_0, z_{N-1} and $[y_n - y_{n+1} + z_n]$, and by using the inequality $\|\phi(hJ_n)\| \leq 1$ we obtain

$$|\overline{D}_{N,j}| \leq |y_{N-1}| + |y_0| + |z_{N-1}| + \sum_{n=0}^{N-2} [|y_n - y_{n+1}| + |z_n|].$$

From this inequality we easily arrive at (3.18) by using the definitions of y_n, z_n and by applying (3.16) with $\xi = h\nu$ and $\psi(\zeta) = \phi(\zeta) - \phi(\infty)$. □

The following observations explain the values $S_j(h)$ for (3.10), (3.11) given in Section 3.2.

1. For (3.10), (3.11) we have $\phi(\zeta) = (1-\zeta)^{-1}$ and $\phi(\zeta) = (1+\zeta/2)(1-\zeta/2)^{-1}$, respectively, so that $\phi(\infty) = 0$ or $\phi(\infty) = -1$. For both Runge–Kutta methods the assumption $\phi(\infty) \neq 1$ of Lemma 3.1 is fulfilled.
2. In view of the definition of $d_{n,j}$ and (3.7) it is to be expected that (3.17) holds with $k = j+1$ and K_i of moderate size.
3. Since $\lim_{\xi \to -\infty} \omega(\xi) = 0$, the second term in the right-hand member of (3.18) will be small for large values of $|h\nu|$. In the situation of Section 3.2 we have

$$\nu = \max\{-3 \times 10^{11}[1+\sin t]^2 : \quad 0 \leq t \leq 1/2\} = -3 \times 10^{11},$$

and $h = 0.01$ or $h = 0.005$, so that

$$\begin{aligned} \omega(h\nu) &\lesssim 6.6 \times 10^{-10} \qquad \text{for (3.10)}, \\ \omega(h\nu) &\lesssim 2.7 \times 10^{-9} \qquad \text{for (3.11)}. \end{aligned}$$

4. These small values of $\omega(h\nu)$ suggest to suppress the second term in the right-hand member of (3.18) so as to obtain

$$|D_{N,j}| \simeq |\overline{D}_{N,j}| \lesssim K \cdot h^{j+1}, \quad K = \frac{2K_0 + (K_1 + 2K_2)T}{|1-\phi(\infty)|},$$

which is in agreement with the values for $S_j(h)$ given in Section 3.2.
5. Relation (3.18) does not apply to (3.12) since for the corresponding stability function one has

$$\phi(\zeta) = \frac{12 + 6\zeta + \zeta^2}{12 - 6\zeta + \zeta^2} \quad \text{with} \quad \phi(\infty) = 1.$$

3.4. VARIANTS TO THE ABOVE ERROR ANALYSIS

The analysis of the global stopping error $v_n - u_n$ as presented above can be modified by defining local stopping errors by the formula

$$d_{n,j} = \Psi_n(v_n, h) - \Phi_n(v_n, h).$$

In this way one would arrive at a basic relation for the global stopping error that is simpler than (3.14). Consequently, it would be easier to analyse the relation between global and local stopping errors. However, it would become more difficult to determine the order of such modified local stopping errors.

We note that, using Hundsdorfer's device, modified versions of Lemma 3.1 can be derived, by means of which the numerical experiments of Section 3.2 can be

explained as well. A variant of (3.18) can be proved with a right-hand member that is $\mathcal{O}(h^k)$. However, such a variant of (3.18) is valid only for $\nu < 0$, and the $\mathcal{O}$-constant in the estimate tends to ∞ when $h\nu \to 0$.

Acknowledgment

The author is indebted to W.H. Hundsdorfer, J.L.M. van Dorsselaer and J. Groeneweg for stimulating discussions on the topic of this paper. The numerical results presented in the tables were obtained by J. Groeneweg.

References

[1] R. Alexander, The modified Newton method in the solution of stiff ordinary differential equations, Math. Comp. 57 (1991) 673–701.

[2] J.C. Butcher, *The numerical analysis of ordinary differential equations,* (John Wiley & Sons (Chichester), 1987).

[3] M. Crouzeix, W.H. Hundsdorfer and M.N. Spijker, On the existence of solutions to the algebraic equations in implicit Runge–Kutta methods, BIT 23 (1983) 84–91.

[4] K. Dekker and J.G. Verwer, *Stability of Runge–Kutta methods for stiff nonlinear differential equations,* (North Holland Publ. Comp. Amsterdam, 1984)

[5] J.L.M. van Dorsselaer and M.N. Spijker, The error committed by stopping the Newton iteration in the numerical solution of stiff initial value problems, to appear in IMA Journ. Numer. Anal.

[6] R. Frank, J. Schneid and C.W. Ueberhuber, Order results for implicit Runge–Kutta methods applied to stiff systems, SIAM J. Numer. Anal. 22 (1985) 515–534.

[7] E. Hairer and G. Wanner, *Solving ordinary differential equations,* Vol. II, (Springer, Berlin, 1991).

[8] W.H. Hundsdorfer, Unconditional convergence of some Crank-Nicolson LOD methods for initial-boundary value problems, Math. Comp. 58 (1992) 35–53.

[9] W.H. Hundsdorfer and B.I. Steiniger, Convergence of linear multistep and one-leg methods for stiff nonlinear initial value problems, BIT 31 (1991) 124–143.

[10] K.R. Jackson, A. Kvaernø and S.P. Nørsett, Order of Runge–Kutta methods when using Newton-type iteration. Report 1/91, Div. Math., Norw. Inst. Techn., Trondheim, (1991).

[11] J. Sand, Methods for starting iteration schemes for implicit Runge–Kutta formulae, in *Computational ordinary differential equations,* 115–126. eds. J.R. Cash and I. Gladwell, (Clarendon Press, Oxford, 1992).

[12] H. Sugiura and T. Torii, A method for constructing generalized Runge–Kutta methods, J. Comp. Appl. Math. 38 (1991) 399–410.