ELSEVIER

Applied Numerical Mathematics 18 (1995) 367–386

APPLIED NUMERICAL MATHEMATICS

The effect of the stopping of the Newton iteration in implicit linear multistep methods

M.N. Spijker

Department of Mathematics and Computer Science, University of Leiden, Niels Bohrweg 1, 2333 CA Leiden, Netherlands

Received 19 September 1994; accepted 5 October 1994

Abstract

This paper concerns the application of implicit, linear multistep methods in the numerical solution of stiff, nonlinear initial value problems. Liniger (1971) studied the (modified) Newton iteration for coping with the nonlinear equations connected with the implicitness of the multistep methods. He gave formulas for the order of the corresponding Newton stopping errors.

Dorsselaer and Spijker (1994) showed, for a class of strongly nonlinear initial value problems, that estimates for the Newton stopping errors are in force with an order that is *lower* than that of Liniger.

In the present paper we find that the actual effect of the stopping of the Newton iteration differs remarkably from what one might expect in view of the paper by Dorsselaer and Spijker: This effect turns out to be of a *higher order* than suggested by a naive application of the orders of Dorsselaer and Spijker. We find that Liniger's order is reliable for stiff problems that are mildly nonlinear. Moreover, for stiff problems that are strongly nonlinear, the order of the accumulated effect of all Newton stopping errors turns out to be greater (by one) than the order which one may expect.

1. Introduction

This paper deals with the numerical solution of the initial value problem

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

Here u_0 is a given vector in the s-dimensional real vectorspace $\mathbb{R}^s$, and $U(t) \in \mathbb{R}^s$ is unknown. Further, f is a given mapping from $[0, T] \times \mathscr{D}$ to $\mathbb{R}^s$, where $\mathscr{D}$ is an open and convex subset of $\mathbb{R}^s$. We assume that the partial derivatives of f, up to the second order, exist and are continuous on $[0, T] \times \mathscr{D}$.

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 see e.g. [7,11,15].

SSDI 0168-9274(95)00064-X

Let $h > 0$ denote a *stepsize*, and consider gridpoints $t_n = nh$ in $[0, T]$ for integer values of n. Numerical approximations v_n to $U(t_n)$ can be defined by the *linear multistep method*

$$\alpha_0 v_n + \alpha_1 v_{n-1} + \cdots + \alpha_k v_{n-k} = h[\beta_0 f_n + \beta_1 f_{n-1} + \cdots + \beta_k f_{n-k}]. \tag{1.2}$$

Here the integer $k \geqslant 1$ and the real coefficients α_i and β_i are constants specifying the method. We assume

$$\alpha_0 = 1, \qquad \beta_0 > 0.$$

These assumptions are satisfied (after a simple normalization) by all methods that are of practical interest in the solution of stiff problems—see the above references. The method is said to have an *order of accuracy p* if, for any solution $U(t)$ to (1.1) with a continuous derivative of order $(p+1)$, the vectors $v_{n-i} = U(t_{n-i})$ and $f_{n-i} = U'(t_{n-i})$ $(i = 0, 1, \ldots, k)$ satisfy (1.2) up to a term which is $O(h^{p+1})$ (as $h \to 0$). We assume $p \geqslant 1$.

In order to apply the linear multistep formula one first calculates starting vectors $v_j \approx U(t_j)$ (for $j = 0, 1, \ldots, k-1$). Next, for $n = k, k+1, k+2, \ldots$, approximations v_n are defined recursively via (1.2) by putting $f_{n-i} = f(t_{n-i}, v_{n-i})$ (for $i = 0, 1, \ldots, k$).

Due to the fact that $\beta_0 \neq 0$ and $f_n = f(t_n, v_n)$ one has to solve a (nonlinear) equation in order to obtain v_n from (1.2). In general this equation cannot be solved exactly, and accordingly in practice the linear multistep method will produce only *approximations* to $v_k, v_{k+1}, v_{k+2}, \ldots$, which we denote by $u_k, u_{k+1}, u_{k+2}, \ldots$.

These vectors u_n will typically be obtained as numerical approximations to the solution x^* of the equation

$$F(x) = 0, \tag{1.3}$$

where

$$F(x) = -x + \gamma \cdot f(t_n, x) + c \quad \text{for } x \in \mathscr{D}, \tag{1.4a}$$

$$\gamma = h\beta_0, \qquad c = -(\alpha_1 u_{n-1} + \cdots + \alpha_k u_{n-k}) + h(\beta_1 f_{n-1} + \cdots + \beta_k f_{n-k}). \tag{1.4b}$$

In the actual numerical solution of nonlinear stiff problems the solution x^* to (1.3) is almost invariably approximated by Newton's method or a variant thereof. We consider the so-called *modified Newton process*

$$F'(x_0)(x_j - x_{j-1}) = -F(x_{j-1}) \quad \text{for } j = 1, 2, 3, \ldots. \tag{1.5}$$

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

In the following we study, for a given $j \geqslant 1$, the order of the errors due to the stopping of the iteration (1.5) after j steps and replacing x^* simply by x_j. We shall analyse these errors in terms of the stepsize h.

The Sections 2.1–2.3 deal with stiff initial value problems that are *mildly nonlinear*. In Section 2.1 we first review a classical estimate for the *Newton stopping error* $x^* - x_j$, which was derived by Liniger [16] *without* taking stiffness into account. Next we present a numerical experiment which appears to confirm Liniger's estimate in the situation of an initial value problem that is stiff. In Section 2.2 we present a theoretical framework explaining the values obtained in our numerical experiment. This framework reveals the fact that Liniger's estimate applies in full to the stiff, mildly nonlinear problems under consideration. Section 2.3 contains the proofs of the theorems formulated in Section 2.2.

In order to highlight the scope and limitations of our theoretical framework we consider in Section 2.4 a numerical experiment with a stiff problem that is *strongly nonlinear*. In Section 2.5 we derive a theoretical estimate of $x^* - x_j$ valid for a general class of strongly nonlinear problems. This estimate is fully in agreement with the experiment of Section 2.4. Its order is lower than the one of Liniger's estimate—and is equal to an order given by Dorsselaer and Spijker [5].

In Section 2.6 we deal with closely related questions among which the order of the stopping errors $x^* - x_j$ for variants to the iterative process (1.5).

In Section 3 we turn our attention to the accumulated effect that the Newton stopping errors $x^* - x_j$ may have on the approximations u_n actually obtained. In Section 3.1 we deal first with the *local stopping error*, which is still nothing but the error $x^* - x_j$ under the localizing assumptions $u_{n-i} = U(t_{n-i})$ and $f_{n-i} = U'(t_{n-i})$ $(1 \leqslant i \leqslant k)$. The estimates for $x^* - x_j$, presented in Section 2, are used to analyse these local stopping errors. Next, in the rest of Section 3, the results of this analysis are used to explore the order of the so-called *global stopping error* $v_n - u_n$, where v_n and u_n are the approximations to $U(t_n)$ defined above. In the situation where $t_n \in (0, T]$ is fixed and $h \to 0$, $n \to \infty$, the order of the global stopping error for strongly nonlinear problems appears to be higher than might be expected at first sight—it is equal to the order of the local stopping errors.

In conclusion, the paper shows that, both for mildly and strongly nonlinear problems, the actual effect of stopping Newton-type iterations is of a higher order than suggested by a naive application of the orders given by Dorsselaer and Spijker [5].

2. The Newton stopping error

2.1. Liniger's error estimate

In all of the following $|x|$ denotes, unless specified otherwise, an arbitrary norm for the vectors $x \in \mathbb{R}^s$.

Let $t_n \in [t_k, T]$, and vectors u_{n-i} and f_{n-i} $(i = 1, 2, \ldots, k)$ be given. Consider, for the corresponding function F defined by (1.4), the process (1.5). In assessing the norm $|x^* - x_j|$ of the Newton stopping error $x^* - x_j$ (for $j \geqslant 1$) we assume that the initial guess x_0 satisfies

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

(with an O-constant of moderate size, and $q > 0$).

By using Taylor series expansions in a straightforward way, the corresponding errors $x^* - x_j$ can be estimated. In this manner Liniger [16] arrived at

$$|x^* - x_j| = O(h^{R(j)}) \qquad \text{with } R(j) = (j+1)q + j \quad \text{for } j \geqslant 1. \tag{2.2}$$

However, there is a weak point in this derivation of (2.2). Stiff problems (1.1) may be characterized by very large magnitudes in the $s \times s$ Jacobian matrix (with respect to x) of the function $h \cdot f(t, x)$. Here h is a "natural" stepsize for the problem and $t \simeq t_n$ and $x \simeq U(t_n)$. In the above derivation the size of these magnitudes was not taken into account, and the (first- and second-order) derivatives with respect to x of $h \cdot f(t, x)$ were replaced simply by $O(h)$. For

stiff problems such quantities $O(h)$ cannot be interpreted, in the standard fashion, as the product of a moderate O-constant and a small stepsize h. Accordingly, one may expect that the O-constant in (2.2) is excessively large (or that (2.2) holds only for excessively small h).

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

Problem 1.

$$U_1'(t) = -10^8\left[U_1(t) - (t-2)^3\right] + \left[U_2(t) - 2\right]^2 + 2(t-2)^2, \qquad U_1(0) = -8,$$

$$U_2'(t) = -10^8\left[U_1(t) - (t-2)^3\right] + \tfrac{2}{3}\left[U_1(t) - (U_2(t) - 2)^3\right] + 1, \quad U_2(0) = 0,$$

with $0 \leqslant t \leqslant \frac{1}{2}$.

The true solution equals $U_1(t) = (t-2)^3$, $U_2(t) = t$, so that any "natural" stepsize h in the numerical solution of Problem 1 need not be very small. The large factor 10^8 causes the problem to be stiff. This stiffness also reflects itself in the eigenvalues λ_1 and λ_2 of the corresponding Jacobian matrix $J(t, x)$ obtained by differentiation, with respect to x, of $f(t, x)$ at $t = 0$, $x = u_0 = (-8, 0)^T$. We have $\lambda_1 \simeq -4$, $\lambda_2 \simeq -10^8$.

We consider the backward Euler method, i.e. (1.2) with

$$k = 1, \qquad \alpha_0 = 1, \qquad \alpha_1 = -1, \qquad \beta_0 = 1, \qquad \beta_1 = 0.$$

Further, we consider the corresponding function F given by (1.4), with

$$t_n = \tfrac{1}{10}, \qquad u_{n-1} = U(t_{n-1}),$$

and we choose the natural initial guess

$$x_0 = u_{n-1}.$$

We deal with the l_1-norm $|x| = |x|_1$ for $x \in \mathbb{R}^2$. It can be proved that the estimate (2.1) holds with an O-constant of moderate size and

$$q = 1.$$

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

$$|x^* - x_1| = O(h^3).$$

In order to check this estimate we have listed the (rounded) actual ratios $|x^* - x_1|/h^3$, for various choices of h, in Table 1. From Table 1 it is evident that the estimate (2.2) is reliable in the present example. In fact, this is surprising since its derivation was defective.

2.2. *A class of stiff, mildly nonlinear problems*

The question arises of whether the reliability of (2.2) in the above example is an exception just due to some coincidence. In the following we shall see that the agreement between the

Table 1
Ratios for the Newton stopping error in Problem 1

h	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
$\lvert x^* - x_1\rvert/h^3$	4.3	3.0	2.8	2.8	2.8	2.8

ratios in Table 1 and (2.2) is no exception but an illustration of a more general phenomenon. We shall specify a class of mildly nonlinear problems for which Liniger's estimate will be proved to be reliable.

We suppose throughout that f satisfies the assumptions already made in Section 1, and we consider the following two additional conditions on f:

C1. $|W(\tau_1) - V(\tau_1)| \leqslant |W(\tau_0) - V(\tau_0)|$ whenever V and W are any two solutions of the differential equation on $[\tau_0, \tau_1] \subset [0, T]$.

C2. $\|J(t, y) - J(t, x)\| \leqslant L_0 |y - x|$ for all $t \in [0, T]$ and $x, y \in \mathscr{D}$.

In these conditions $\|\cdot\|$ stands for the matrix (operator) norm for real $s \times s$ matrices that is induced by the given norm $|\cdot|$ on $\mathbb{R}^s$. Further, $J(t, x)$ stands for the $s \times s$ Jacobian matrix (with respect to x) of $f(t, x)$. The Lipschitz constant L_0 in C2 is assumed to be of *moderate size*.

There exist arbitrarily stiff, nonlinear problems satisfying C1 and C2 with L_0 of moderate size. As an example consider Problem 1. Here, the function given by

$$f(t, x) = \begin{pmatrix} -10^8\left[\xi_1 - (t-2)^3\right] + (\xi_2 - 2)^2 + 2(t-2)^2 \\ -10^8\left[\xi_1 - (t-2)^3\right] + \frac{2}{3}\left[\xi_1 - (\xi_2 - 2)^3\right] + 1 \end{pmatrix} \quad \text{for } x = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$$

can be proved to satisfy C1 and C2 with $L_0 = 11$ and

$$|x| = |\xi_1| + |\xi_2|, \qquad \mathscr{D} = \left\{x : -\infty < \xi_1 < \infty,\ -\tfrac{1}{4} < \xi_2 < 1\right\},$$
$$\text{where } x = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \in \mathbb{R}^2. \tag{2.3}$$

In the general situation we assume that U satisfies (1.1), and we denote the distance of the solution set $\{U(t): 0 \leqslant t \leqslant T\}$ to the boundary of $\mathscr{D}$ by ω, i.e.

$$\omega = \inf\{|U(t) - x| : 0 \leqslant t \leqslant T,\ x \in \mathbb{R}^s \setminus \mathscr{D}\}.$$

In case $\mathscr{D} = \mathbb{R}^s$ we put $\omega = \infty$. In the following theorems we refer to the assumption

$$|u_{n-i} - U(t_{n-i})| \leqslant \alpha \cdot \omega, \quad h|f_{n-i} - U'(t_{n-i})| \leqslant \alpha \cdot \omega \qquad \text{for } 1 \leqslant i \leqslant k,$$
$$h|U'(t)| \leqslant \alpha \cdot \omega \quad \text{for } 0 \leqslant t \leqslant T. \tag{2.4}$$

Theorem 2.1. *Assume C1. Then there is a factor $\alpha > 0$, depending only on the coefficients α_i and β_i $(0 \leqslant i \leqslant k)$, such that (1.3) has a unique solution $x^* \in \mathscr{D}$ whenever F is defined by (1.4) with any u_{n-i}, f_{n-i} and h satisfying (2.4).*

Note that, if $\mathscr{D} = \mathbb{R}^s$, condition (2.4) is always fulfilled so that, according to Theorem 2.1, Eq. (1.3) has a unique solution for any given u_{n-i}, f_{n-i} and h.

Let positive K and q be given. The next theorem relates the assumption

$$x_0 \in \mathscr{D}, \qquad |x^* - x_0| \leqslant Kh^q \leqslant \omega/2, \qquad \beta_0 L_0 \cdot Kh^{q+1} < 2(\sqrt{2} - 1) \tag{2.5}$$

to the following property of the corresponding modified Newton process:

There exist unique vectors $x_j \in \mathscr{D}$ generated by process (1.5) with initial guess x_0, and

$$|x^* - x_j| \leqslant \left[\frac{1+\sqrt{2}}{2}\beta_0 L_0 K\right]^j \cdot K \cdot h^{(j+1)q+j} \quad \text{for } j \geqslant 1. \tag{2.6}$$

Theorem 2.2. *Assume C1 and C2. Then there is a factor $\alpha > 0$ with the properties already stated in Theorem 2.1 and with the additional property that (2.6) holds whenever u_{n-i}, f_{n-i}, h and x_0 satisfy (2.4) and (2.5).*

The last theorem shows that, under the assumptions C1 and C2, the relation (2.1) (with O-constant K of moderate size) implies that (2.2) is valid with an O-constant

$$K_j = \left[\frac{1+\sqrt{2}}{2}\beta_0 L_0 K\right]^j K.$$

This constant K_j is not affected by stiffness, and it is of moderate size when j is limited in a realistic way. Theorem 2.2 thus nicely explains the moderate values for the ratios $|x^* - x_1|/h^3$ displayed in Table 1.

2.3. Proof of the theorems on the Newton stopping error in stiff, mildly nonlinear problems

The above two theorems will be proved below by making use of the following lemma.

Lemma 2.3. *Let F be defined by (1.4) with arbitrary vectors u_{n-i}, $f_{n-i} \in \mathbb{R}^s$ $(1 \leqslant i \leqslant k)$, and assume C1. Let $z_0 \in \mathscr{D}$ be such that*

$$\{x: x \in \mathbb{R}^s \text{ with } |x - z_0| \leqslant |F(z_0)|\} \subset \mathscr{D}. \tag{2.7a}$$

Then there is a unique $x^* \in \mathscr{D}$ with $F(x^*) = 0$, and

$$|x^* - z_0| \leqslant |F(z_0)|. \tag{2.7b}$$

Proof. We denote the logarithmic norm, for real $s \times s$ matrices, induced by the given norm on $\mathbb{R}^s$ by $\mu(\cdot)$ (cf. e.g. [3,4]). It is well known that the inequality

$$\mu(J(t, x)) \leqslant 0 \qquad \text{for all } t \in [0, T], \quad x \in \mathscr{D},$$

is a sufficient condition in order that C1 holds. It is also a necessary condition (see [5]). Therefore, we can make use of that inequality.

Since

$$F'(x) = -I + \gamma \cdot J(t_n, x) \quad \text{for } x \in \mathscr{D},$$

where I denotes the $s \times s$ identity matrix, we obtain

$$\mu(F'(x)) = \mu(-I + \gamma \cdot J(t_n, x)) \leqslant \mu(-I) + \gamma \cdot \mu(J(t_n, x)).$$

Consequently,

$$\mu(F'(x)) \leqslant -1 \quad \text{for } x \in \mathscr{D}. \tag{2.7c}$$

Theorem 3.6 of [5] states that the last inequality in combination with (2.7a) implies the existence of a unique $x^* \in \mathscr{D}$ with $F(x^*) = 0$. That theorem also states that (2.7b) is implied. □

Proof of Theorem 2.1. For u_{n-i}, f_{n-i} and h satisfying (2.4), with any $\alpha > 0$, we have

$$F(U(t_n)) = S_1 + S_2,$$

with

$$S_1 = -\sum_{i=0}^{k} \alpha_i U(t_{n-i}) + h \sum_{i=0}^{k} \beta_i U'(t_{n-i}),$$

$$S_2 = \sum_{i=1}^{k} \{\alpha_i [U(t_{n-i}) - u_{n-i}] - h\beta_i [U'(t_{n-i}) - f_{n-i}]\}.$$

Since the order p of (1.2) satisfies $p \geqslant 1$, we have

$$\sum_{i=0}^{k} \alpha_i U(t_{n-i}) = \sum_{i=1}^{k} \alpha_i [U(t_{n-i}) - U(t_n)],$$

and therefore

$$|S_1| \leqslant \alpha\omega \sum_{i=0}^{k} (i|\alpha_i| + |\beta_i|).$$

Further,

$$|S_2| \leqslant \alpha\omega \sum_{i=1}^{k} (|\alpha_i| + |\beta_i|),$$

so that we can write

$$|F(U(t_n))| \leqslant \beta \cdot \alpha\omega,$$

where $\beta > 0$ depends only on the coefficients α_i and β_i.

We choose $\alpha = (4\beta)^{-1}$. Clearly, for any u_{n-i}, f_{n-i} and h satisfying (2.4) with this α, we have

$$|F(U(t_n))| \leqslant \omega/4.$$

In view of the definition of ω we see that, for the vector

$$z_0 = U(t_n),$$

the requirement (2.7a) is fulfilled. The proof is completed by applying Lemma 2.3. □

Proof of Theorem 2.2. We assume C1 and C2 and use the notations of the proof of Theorem 2.1. Choose $\alpha > 0$ as indicated in that proof, and let u_{n-i}, f_{n-i}, h and x_0 satisfy (2.4) and (2.5).

In view of Lemma 2.3, the inequality (2.7b) is fulfilled. Consequently,

$$|x^* - U(t_n)| \leqslant \omega/4.$$

By noting that $Kh^q \leqslant \omega/2$ (see (2.5)) and by using the definition of ω we conclude

$$\{x\colon x \in \mathbb{R}^s \text{ with } |x - x^*| \leqslant Kh^q\} \subset \mathscr{D}. \tag{2.8a}$$

For all $x,y \in \mathscr{D}$ we have $\| F'(y) - F'(x) \| = \gamma \| J(t_n, y) - J(t_n, x) \|$, and therefore

$$\| F'(y) - F'(x) \| \leqslant \Lambda | y - x |, \quad \text{with } \Lambda = \beta_0 L_0 h.$$

From (2.7c) it follows (see e.g. [4,5]) that

$$F'(x) \text{ is invertible} \quad \text{for } x \in \mathscr{D}, \tag{2.8b}$$

with

$$\| [F'(x)]^{-1} \| \leqslant 1. \tag{2.8c}$$

Consequently,

$$F'(y) = F'(x)[I + E(x, y)] \quad \text{for all } x,y \in \mathscr{D}, \tag{2.8d}$$

with an $s \times s$ matrix $E(x, y)$ satisfying

$$\| E(x, y) \| = \| F'(x)^{-1}(F'(y) - F'(x)) \| \leqslant \| F'(y) - F'(x) \| \leqslant \Lambda | y - x |,$$

so that

$$\| E(x, y) \| \leqslant \Lambda | y - x | \quad \text{for all } x,y \in \mathscr{D}. \tag{2.8e}$$

Theorem 3.2 of [5] states that, in the situation where (2.8b), (2.8d) and (2.8e) hold, the modified Newton process generates unique x_j in $\mathscr{D}$, with

$$| x^* - x_j | \leqslant [\Lambda\delta(1 + \Lambda\delta/4)]^{j-1} \Lambda\delta^2/2 \quad \text{for } j \geqslant 1. \tag{2.8f}$$

Here $\delta > 0$ is only required to be so small that

$$\Lambda\delta < 2(\sqrt{2} - 1), \qquad \{x: x \in \mathbb{R}^s \text{ with } | x - x^* | \leqslant \delta\} \subset \mathscr{D},$$

whereas the initial guess x_0 is required to satisfy

$$| x^* - x_0 | \leqslant \delta.$$

In the present situation all of these requirements are fulfilled with $\delta = Kh^q$—this follows immediately from (2.5) and (2.8a). By substituting in (2.8f) the values $\Lambda = \beta_0 L_0 h$ and $\delta = Kh^q$ we arrive at (2.6). □

2.4. An example of a stiff, highly nonlinear problem

We turn our attention to stiff problems where C2 is *not* fulfilled with L_0 of moderate size. We consider:

Problem 2.

$$U_1'(t) = -10^8\left[U_1(t) - (U_2(t) - 2)^3\right] + 3(U_2(t) - 2)^2, \quad U_1(0) = -8,$$

$$U_2'(t) = 10^8\left[U_1(t) - (U_2(t) - 2)^3\right] + 1, \quad U_2(0) = 0,$$

with $0 \leqslant t \leqslant \frac{1}{2}$.

Table 2
Ratios for the Newton stopping error in Problem 2

h	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
$\lvert x^* - x_1\rvert/h^3$	10	97	96×10	96×10^2	96×10^3	96×10^4

The true solution is the same as for Problem 1 so that also here any "natural" stepsize h need not be very small. The large factors 10^8 make the problem stiff. This stiffness reflects itself in the eigenvalues of the Jacobian matrix $J(t, x)$ at $t = 0$, $x = u_0 = (-8, 0)^T$. They are approximately equal to -1 and -10^9 (cf. [9]).

Similarly as in the above analysis of Problem 1 we assume (2.3). For Problem 2 it can still be proved that C1 is fulfilled. But, for all $t \in [0, \frac{1}{2}]$, $x = (\xi_1, \xi_2)^T \in \mathscr{D}$, $y = (\eta_1, \eta_2)^T \in \mathscr{D}$, with $\xi_1 = \eta_1$, one can show that the Jacobian matrix corresponding to Problem 2 satisfies

$$\| J(t, y) - J(t, x) \| \geqslant 10^9 \, | y - x |.$$

Therefore, C2 is not fulfilled here with a moderate constant L_0, and Theorem 2.2 cannot be applied.

The fact that Theorem 2.2 cannot be applied to Problem 2 leaves the question about the reliability of Liniger's estimate for that problem unsettled. Therefore, we consider a numerical experiment analogous to the one of Section 2.1: We choose $k = 1$, $\alpha_0 = 1$, $\alpha_1 = -1$, $\beta_0 = 1$, $\beta_1 = 0$ and $t_n = \frac{1}{10}$, $x_0 = u_{n-1} = U(t_{n-1})$. We define F by (1.4), where f is the function corresponding to Problem 2. Again Eq. (1.3) can be proved to have a solution x^* for which (2.1) holds with $q = 1$ (and an O-constant of moderate size). Clearly, Liniger's estimate would predict

$$| x^* - x_1 | = O(h^3).$$

In order to check this estimate for Problem 2 we have listed in Table 2 values for the actual ratios $| x^* - x_1 | / h^3$ (cf. [9]). The entries in the table are quite different from those in Table 1. It is clear that the estimate (2.2) is *not* reliable in the present situation; it provides no insight regarding the actual error $x^* - x_1$ for Problem 2.

The above numerical experiment shows that our Theorem 2.2 is sharp in that condition C2, with moderate L_0, cannot be omitted if (2.2) is to be valid (with moderate O-constant).

2.5. *A class of stiff, highly nonlinear problems*

In this section we shortly review a framework that provides reliable estimates of $x^* - x_j$ for a class of highly nonlinear problems including Problem 2.

We assume, in addition to the assumptions concerning f made in Section 1, that:

C3. $J(t, y) - J(t, x) = J(t, x)e(t, x, y)$ with an $s \times s$ matrix $e(t, x, y)$ satisfying $\| e(t, x, y) \| \leqslant L_1 | x - y |$ for all $t \in [0, T]$ and $x, y \in \mathscr{D}$.

Here $| \cdot |$, $\| \cdot \|$ and J have the same meaning as in C2. The Lipschitz constant L_1 in C3 is again assumed to be of *moderate size*. It provides an upperbound for the *relative* variation of

$J(t, y)$ when y varies through $\mathscr{D}$—whereas the Lipschitz constant in C2 provides an upper-bound for the *absolute* variation of $J(t, y)$.

There exist arbitrarily stiff, nonlinear problems satisfying C1 and C3 with L_1 of moderate size. In particular, C3 can be fulfilled in cases where C2 is not fulfilled with moderate L_0.

As an example consider Problem 2, and assume (2.3). In this situation the corresponding Jacobian matrix can be proved to satisfy C3 with $L_1 = 12$ (cf. [9]).

Below we formulate a theorem that is valid for the general situation where C1 and C3 hold. The theorem relates the assumption

$$x_0 \in \mathscr{D}, \qquad |x^* - x_0| \leqslant Kh^q \leqslant \omega/2, \qquad L_1 Kh^q < (\sqrt{2} - 1) \tag{2.9}$$

to the following property of the corresponding iterative process:

> There are unique $x_j \in \mathscr{D}$ generated by process (1.5)
> with initial guess x_0, and
>
> $$|x^* - x_j| \leqslant \left[(1 + \sqrt{2})L_1 K\right]^j \cdot K \cdot h^{(j+1)q} \quad \text{for } j \geqslant 1. \tag{2.10}$$

Theorem 2.4. *Assume C1 and C3. 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-i}, f_{n-i}, h and x_0 satisfy (2.4) and (2.9).*

This theorem shows that, under the assumptions C1 and C3, the relation (2.1) (with a moderate O-constant K) implies

$$|x^* - x_j| = O(h^{Q(j)}) \qquad \text{with } Q(j) = (j+1)q \quad \text{for } j \geqslant 1. \tag{2.11}$$

In (2.11) we have an O-constant

$$K_j = \left[(1 + \sqrt{2})L_1 K\right]^j K,$$

which is not affected by the stiffness of (1.1); K_j is of moderate size for realistic values of j. The order $Q(j)$ occurring in (2.11) was given first by Dorsselaer and Spijker [5].

In the example of Section 2.4 our relation (2.11) reduces to

$$|x^* - x_1| = O(h^2).$$

This error estimate is nicely in agreement with the numerical experiment displayed in Table 2.

Proof of Theorem 2.4. Assume C1 and C3. We use the notations of the proof of Theorem 2.1 and choose $\alpha > 0$ as indicated there. Let u_{n-i}, f_{n-i}, h and x_0 satisfy (2.4) and (2.9).

Similarly as in the proof of Theorem 2.2 we conclude that (2.8a)–(2.8c) hold.

In view of (1.4), C3 and (2.8b) we can write

$$F'(y) - F'(x) = \gamma\left[J(t_n, y) - J(t_n, x)\right] = F'(x)\left\{I + \left[F'(x)\right]^{-1}\right\} e(t_n, x, y).$$

We define $E(x, y) = \{I + [F'(x)]^{-1}\}e(t_n, x, y)$ and apply (2.8c) and C3 so as to arrive at (2.8d) and (2.8e) with $\Lambda = 2L_1$.

Similarly as in the proof of Theorem 2.2 we can apply, in the present situation where (2.9) is fulfilled, Theorem 3.2 of [5]. By substituting in the right-hand member of (2.8f) the values $\Lambda = 2L_1$ and $\delta = Kh^q$ we obtain (2.10). □

2.6. Remarks

Theorem 2.1 is related to existence and uniqueness results obtained by Desoer and Haneda [4], Söderlind [18], Alexander [1], Hairer and Wanner [11]. It has some similarity to Theorem 4.3 of [5], and it is a generalized version of a theorem announced, without proof, by Groeneweg and Spijker [9].

Theorem 2.4 is related to Theorem 4.2 of [5], and can be viewed as an improved version of a theorem announced, without proof, in [9].

Under the conditions of Theorem 2.2 and 2.4 the iterates x_j are easily seen to converge linearly towards x^* (see the estimates for $|x^* - x_j|$ in (2.6) and (2.10), respectively). This conclusion is related to interesting work by Alexander [1], who studied the rate of convergence of similar iterates. But, the class of nonlinear problems dealt with by Alexander is different from our classes specified by C1, C2 or C1, C3.

The Lipschitz condition in C2, with moderate L_0, is satisfied by a class of stiff, nonlinear problems related to the so-called D2-case specified by Frank et al. [6]; cf. also [2]. Here the stiffness is not allowed to manifest itself in the second-order derivatives, with respect to x, of $f(t, x)$. Note that the value L_0 is *not* invariant under a rescaling of the independent variable t. Therefore, from a formal point of view, it might have been more correct to require in Section 2.2 that the product hL_0, instead of L_0, is of moderate size. But, in all of the above we assumed (tacidly) that the stepsizes h under consideration are not larger than 1. Under this assumption a moderate value of L_0, as postulated in Section 2.2, already implies a moderate value of hL_0.

"Relative" Lipschitz conditions different from, but related to, C3 were used in the numerical analysis of stiff, nonlinear problems by Alexander [1] (cf. also [12,17,19]). A crucial point in C3 is the fact that it can be satisfied, with moderate L_1, in cases where the stiffness manifests itself also in the second-order derivatives of $f(t, x)$. Note that the value L_1 is invariant under a rescaling of t.

We finally note that orders $R(j)$ and $Q(j)$, analogous to those in (2.2) and (2.11), can be derived for the *exact Newton process*

$$F'(x_{j-1})(x_j - x_{j-1}) = -F(x_{j-1}), \quad j \geqslant 1, \tag{1.5'}$$

and for the *simplified Newton process*

$$F'(y_0)(x_j - x_{j-1}) = -F(x_{j-1}), \quad j \geqslant 1, \tag{1.5''}$$

where y_0 is an approximation, possibly different from x_0, satisfying

$$|x^* - y_0| = O(h^r)$$

(with an O-constant of moderate size, and $0 < r \leqslant q$).

In the general situation where C1 and C2 hold, the iterates x_j specified by (1.5') and (1.5'') satisfy

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

and

$$|x^* - x_j| = O(h^{R(j)}) \qquad \text{with } R(j) = jr + q + j \quad \text{for } j \geqslant 1, \tag{2.2''}$$

respectively (with O-constants of moderate size).

In the situation where C1 and C3 hold, the corresponding iterates satisfy

$$|x^* - x_j| = O(h^{Q(j)}) \qquad \text{with } Q(j) = 2^j q \quad \text{for } j \geqslant 1, \tag{2.11'}$$

and

$$|x^* - x_j| = O(h^{Q(j)}) \qquad \text{with } Q(j) = jr + q \quad \text{for } j \geqslant 1, \tag{2.11''}$$

respectively (with O-constants of moderate size).

The order $R(j)$ in (2.2') was given first by Liniger [16], and the orders in (2.2''), (2.11') and (2.11'') by Dorsselaer and Spijker [5].

The above two expressions for $R(j)$ can be proved to be valid, in the situation (C1, C2), by using material from [5]. Since we want to keep the paper reasonably concise, and the arguments leading to (2.2') and (2.2'') are analogous to those used to establish (2.6), we omit the details of the proof. A similar remark applies to the proof, in the general situation (C1, C3), of (2.11') and (2.11'').

3. The global stopping error

3.1. Estimating the local stopping errors

Let v_n denote the approximations to $U(t_n)$ generated by the theoretical linear multistep method (1.2), with $f_{n-i} = f(t_{n-i}, v_{n-i})$ $(0 \leqslant i \leqslant k)$. We assume $u_i = v_i$ $(0 \leqslant i \leqslant k-1)$; and by u_n (for $n \geqslant k$) we denote the actual approximations obtained when, in each application of the linear multistep formula, the solution to Eq. (1.3) is approximated by performing j iteration steps with a Newton-like method. In the following we explore the order of the *global stopping error* $D_{n,j}$ defined by

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

We shall confine ourselves mainly to the (important) situation where the modified Newton process (1.5) is used, and where the initial guess equals

$$x_0 = u_{n-1}. \tag{3.1}$$

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

$$d_{n,j} = \tilde{v}_n - \tilde{u}_n, \tag{3.2a}$$

and to analyse the latter errors first. In order to specify, and analyse, the vectors $\tilde{v}_n$ and $\tilde{u}_n$ we define, throughout this subsection, the function F by the relations (1.4) in which u_{n-i} and f_{n-i} are replaced by $U(t_{n-i})$ and $U'(t_{n-i})$, respectively $(1 \leqslant i \leqslant k)$. We put

$$\tilde{v}_n = x^*, \qquad \tilde{u}_n = x_j, \tag{3.2b}$$

where, with the above definition of F in force, $F(x^*) = 0$ and x_j is generated by the modified Newton process (1.5) with initial guess $x_0 = U(t_{n-1})$.

The local stopping errors can be estimated by using (2.2) or (2.11) provided a value for q is available. In order to determine q we consider the residual $F(z_0)$, of the function F just defined, at the point

$$z_0 = U(t_n).$$

A straightforward calculation (cf. the proof of Theorem 2.1) yields

$$|F(z_0)| \leqslant h\lambda \|U'\|,$$

with

$$\lambda = \sum_{i=0}^{k} (i|\alpha_i| + |\beta_i|), \qquad \|U'\| = \max\{|U'(t)|: 0 \leqslant t \leqslant T\}.$$

We see that our function F satisfies condition (2.7a) (provided $h > 0$ is so small that $h\lambda\|U'\| < \omega$, where ω is defined in Section 2.2). Assuming C1 we conclude from Lemma 2.3 that $|x^* - z_0| \leqslant |F(z_0)|$. With $x_0 = U(t_{n-1})$, we thus obtain

$$|x^* - x_0| \leqslant |F(z_0)| + |U(t_n) - U(t_{n-1})| \leqslant (\lambda + 1)\|U'\| \cdot h.$$

Clearly, (2.1) is fulfilled with

$$q = 1$$

and an O-constant $K = (\lambda + 1)\|U'\|$ which is not affected by stiffness.

Applying (2.2) and (2.11), with $q = 1$, to the local stopping error $d_{n,j} = x^* - x_j$ we obtain

$$|d_{n,j}| = O(h^{R(j)}) \quad \text{with } R(j) = 2j + 1 \text{ in the situation (C1, C2),} \tag{3.3a}$$

$$|d_{n,j}| = O(h^{Q(j)}) \quad \text{with } Q(j) = j + 1 \text{ in the situation (C1, C3),} \tag{3.3b}$$

with O-constants of moderate size.

3.2. *Numerical experiments regarding the global stopping error*

Let $Nh = T$ with integer $N \geqslant k$. The global error $D_{N,j} = v_N - u_N$ amounts to the accumulated effect of Newton stopping errors at the points $t_k, t_{k+1}, \ldots, t_N$. For stable linear multistep methods the error $D_{N,j}$ may thus be expected to satisfy

$$|D_{N,j}| = O\big(|d_{k,j}| + |d_{k+1,j}| + \cdots + |d_{N,j}|\big),$$

with an O-constant that is independent of $N = T/h$ and not affected by stiffness. In view of (3.3a) and (3.3b) we arrive at

$$|D_{N,j}| = O(h^{2j}) \quad \text{in the situation (C1, C2),} \tag{3.4a}$$

and

$$|D_{N,j}| = O(h^{j}) \quad \text{in the situation (C1, C3).} \tag{3.4b}$$

Table 3
$S_j(h)$ for Problem 1 (Section 2.1); $h = 0.01$, $T = 1/2$

j	$2j$	$S_j(h)$ for (3.5)	$S_j(h)$ for (3.6)
1	2	2.0425	1.9911
2	4	4.0418	3.9787
3	6	6.0533	5.9945

We shall check (3.4a) and (3.4b) for the following two linear multistep methods:

Backward Euler, i.e. $k = 1$, $\alpha_0 = 1$, $\alpha_1 = -1$, $\beta_0 = 1$, $\beta_1 = 0$. (3.5)

Backward Differentiation Formula of order 2, i.e. $k = 2$, $\alpha_0 = 1$,

$\alpha_1 = -\frac{4}{3}$, $\alpha_2 = \frac{1}{3}$, $\beta_0 = \frac{2}{3}$, $\beta_1 = \beta_2 = 0$. (3.6)

Both methods are A-stable (and have a positive damping order at infinity).

In order to check whether the orders $2j$ and j, corresponding to (3.4a) and (3.4b) manifest themselves in reality we introduce

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

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

In Table 3 we have displayed values of $S_j(h)$ (where the ℓ_1-norm is used to measure $D_{N,j}$ and $D_{2N,j}$) for the mildly nonlinear Problem 1 of Section 2.1. We see that the values in Table 3 are in excellent agreement with (3.4a), and thus confirm the estimates (3.3a) and (2.2).

In Table 4 we have displayed values of $S_j(h)$ (still using the ℓ_1-norm) for the strongly nonlinear Problem 2 (Section 2.4). These values were obtained by Groeneweg [8]. The values in Table 4 are neither in agreement with (3.4a) nor (3.4b). Since the conditions C1 and C3 (with moderate L_1) are fulfilled in Problem 2 we would have expected agreement with (3.4b). Below we go further into this anomaly; at the end of Section 3.3 the values $S_j(h)$ in Table 4 will be explained.

We note that, for various other nonlinear stiff problems, values for $S_j(h)$ were found by Groeneweg [8] that are nearly equal to those in Tables 3 and 4.

3.3. Relating the global stopping error to the local stopping errors

We confine ourselves in the following to the (important) situation where

$$\beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

Table 4
$S_j(h)$ for Problem 2 (Section 2.4); $h = 0.01$, $T = 1/2$

j	$2j$	$S_j(h)$ for (3.5)	$S_j(h)$ for (3.6)
1	2	1.9946	1.9976
2	4	2.9713	2.9806
3	6	3.9575	3.9732

This assumption facilitates our presentation and is fulfilled for the case of (3.5), (3.6).

We introduce the notations

$$\boldsymbol{v}_{n-1} = (v_{n-1}, \ldots, v_{n-k}), \qquad \boldsymbol{U}_{n-1} = (U(t_{n-1}), \ldots, U(t_{n-k})),$$

and we relate the theoretical linear multistep process to a function Ψ_n, so that the vectors v_n and $\tilde{v}_n$ (defined in Section 3.1) satisfy

$$v_n = \Psi_n(\boldsymbol{v}_{n-1}), \qquad \tilde{v}_n = \Psi_n(\boldsymbol{U}_{n-1}).$$

Similarly, we introduce

$$\boldsymbol{u}_{n-1} = (u_{n-1}, \ldots, u_{n-k}),$$

and we relate the linear multistep process carried out with j modified Newton iterations (according to (1.5) and (3.1)) to a function Φ_n, so that

$$u_n = \Phi_n(\boldsymbol{u}_{n-1}), \qquad \tilde{u}_n = \Phi_n(\boldsymbol{U}_{n-1}).$$

In view of the definitions of global and local stopping errors (Section 3.1) we have

$$D_{n,j} = [\Psi_n(\boldsymbol{v}_{n-1}) - \Psi_n(\boldsymbol{U}_{n-1})] + [\Phi_n(\boldsymbol{U}_{n-1}) - \Phi_n(\boldsymbol{u}_{n-1})] + d_{n,j}. \tag{3.7}$$

In the following we shall use the Jacobian matrix $J_n = J(t_n, U(t_n))$. We introduce—similarly as in [14]—the rational functions

$$\psi_i(z) = (\beta_0 z - 1)^{-1}\alpha_i \quad \text{for } i = 1, 2, \ldots, k,$$

where β_0 and α_i are the coefficients of the linear multistep formula. If the Jacobian matrix $J(t_n, x)$ would be constant, for $x \in \mathscr{D}$, we would have

$$\Psi_n(\boldsymbol{v}_{n-1}) - \Psi_n(\boldsymbol{U}_{n-1}) = \sum_{i=1}^{k} \psi_i(hJ_n)[v_{n-i} - U(t_{n-i})],$$

$$\Phi_n(\boldsymbol{U}_{n-1}) - \Phi_n(\boldsymbol{u}_{n-1}) = \sum_{i=1}^{k} \psi_i(hJ_n)[U(t_{n-i}) - u_{n-i}],$$

and therefore (3.7) would reduce to

$$D_{n,j} = \sum_{i=1}^{k} \psi_i(hJ_n) D_{n-i,j} + d_{n,j}.$$

Clearly, the vectors $\overline{D}_{n,j}$ defined by

$$\overline{D}_{n,j} = \sum_{i=1}^{k} \psi_i(hJ_n)\overline{D}_{n-i,j} + d_{n,j} \quad \text{for } n = k, k+1, \ldots, N,$$

$$\overline{D}_{i,j} = 0 \quad \text{for } i = 0, 1, \ldots, k-1, \tag{3.8}$$

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

In the following analysis it is convenient to consider, along with (3.8), the general relations

$$y_n = \sum_{i=1}^{k} \psi_i(hJ_n) y_{n-i} + x_n, \quad k \leqslant n \leqslant N,$$

$$y_i = 0, \quad 0 \leqslant i \leqslant k-1, \tag{3.9a}$$

for arbitrary vectors $x_n, y_n \in \mathbb{R}^s$. We assume that the linear multistep method is *stable* in that a fixed factor σ, of moderate size, exists such that

$$|y_n| \leqslant \sigma \cdot \{|x_k| + |x_{k+1}| + \cdots + |x_N|\} \tag{3.9b}$$

whenever (3.9a) holds. In the literature many conditions for (3.9b) can be found that are relevant to the situation of stiff problems (see e.g. [11], or the review of such conditions in [14]).

Applying (3.9) to the vectors $x_n = d_{n,j}$, $y_n = \bar{D}_{n,j}$ we thus conclude that

$$|D_{N,j}| \simeq |\bar{D}_{N,j}| \leqslant \sigma \cdot \{|d_{k,j}| + |d_{k+1,j}| + \cdots + |d_{N,j}|\},$$

which is in agreement with our considerations at the beginning of Section 3.2. In order to explain the orders $S_j(h)$ observed in Table 4, a more refined analysis of (3.8) is required.

In view of (3.2) we have $d_{n,j} = x^* - x_j$; and from (1.5) there follows

$$F'(x_0)(x^* - x_j) = \int_0^1 \left[F'\left(x_{j-1} + t(x^* - x_{j-1})\right) - F'(x_0)\right](x_{j-1} - x^*)\, \mathrm{d}t,$$

where $j \geqslant 1$, $x_0 = U(t_{n-1})$ and F is defined by (1.4) (with u_{n-i}, f_{n-i} replaced by $U(t_{n-i})$, $U'(t_{n-i})$ for $1 \leqslant i \leqslant k$). By applying condition C3 we obtain

$$[I - h\beta_0 J(t_n, x_0)] d_{n,j}$$

$$= h\beta_0 J(t_n, x_0) \int_0^1 \left[e(t_n, x_0, x_{j-1} + t d_{n,j-1}) d_{n,j-1}\right] \mathrm{d}t.$$

Therefore,

$$d_{n,j} = \left\{[I - h\beta_0 J(t_n, U(t_{n-1}))]^{-1} - I\right\} a_n, \tag{3.10}$$

with

$$a_n = \int_0^1 \left[e(t_n, x_0, x_{j-1} + t d_{n,j-1}) d_{n,j-1}\right] \mathrm{d}t. \tag{3.11a}$$

In the situation where C1 and C3 are fulfilled there follows, by using (3.3b),

$$|a_n| = O(h^{j+1}) \tag{3.11b}$$

(with an O-constant of moderate size).

In view of (3.8) and (3.10) the following lemma will be useful. The lemma is based on a general device for obtaining refined error estimates due to Hundsdorfer [13] and Hundsdorfer and Steiniger [14].

Lemma 3.1. *Assume C1 and C3, and let the linear multistep method be stable in the sense of* (3.9). *Assume that, for a given integer* $j \geqslant 1$, *we can write the local stopping error in the form* (3.10), *where*

$$\begin{aligned} &a_n = h^r a(t_n) + h^{r+1} b_n, \text{ with } a(t), b_n \in \mathbb{R}^s \text{ satisfying} \\ &|a(t)| \leqslant K_0, \qquad |a(t') - a(t)| \leqslant K_1 |t' - t|, \qquad |b_n| \leqslant K_2. \end{aligned} \tag{3.12}$$

Then the vector $\overline{D}_{N,j}$ *defined by* (3.8), *with* $Nh = T$, *satisfies* $|\overline{D}_{N,j}| \leqslant Kh^r$, *with*

$$\begin{aligned} &K = K_0 + 2\sigma T(K_2 + K_0 L_1 \|U'\|) + \sigma(K_0 + K_1 T) \sum_{i=1}^{k} i|\alpha_i|, \\ &\|U'\| = \max\{|U'(t)|: 0 \leqslant t \leqslant T\}. \end{aligned} \tag{3.13}$$

The following four observations explain the values $S_j(h)$ in Table 4.

(1) Both methods (3.5) and (3.6) are A-stable, and may therefore be expected to satisfy the stability requirement (3.9) in the situation of Problem 2 (Section 2.4). (Method (3.5) can even be proved to satisfy (3.9b) with $\sigma = 1$.)

(2) In view of (3.11a) (where x_0 stands for $U(t_{n-1})$) and (3.11b) it is to be expected that (3.12) holds with $r = j + 1$ and K_i of moderate size.

(3) The fact that the conditions C1 and C3 are fulfilled in the situation of Problem 2 suggests to apply Lemma 3.1 so as to obtain

$$|D_{N,j}| \simeq |\overline{D}_{N,j}| \leqslant Kh^{j+1}.$$

(4) The order $j + 1$ just established is in excellent agreement with the values of $S_j(h)$ in Table 4.

3.4. Proof of Lemma 3.1

(1) We introduce the notations

$$R_n = [I - \beta_0 h J_n]^{-1}, \qquad S_n = [I - \beta_0 h \cdot J(t_n, U(t_{n-1}))]^{-1},$$

and we note that

$$\sum_{i=1}^{k} \psi_i(hJ_n) = R_n.$$

According to Hundsdorfer's device (see the above references) we write $d_{n,j}$, for $k \leqslant n \leqslant N$, in the form

$$d_{n,j} = [R_n - I] p_n + q_n,$$

with

$$p_n = h^r a(t_n), \qquad q_n = (S_n - R_n) p_n + h^{r+1}(S_n - I) b_n.$$

Substituting this expression for $d_{n,j}$ in (3.8) there follows

$$\left(\overline{D}_{n,j}+p_n\right)=\sum_{i=1}^{k}\psi_i(hJ_n)\left(\overline{D}_{n-i,j}+p_{n-i}\right)+x_n,\quad k\leqslant n\leqslant N,$$

where

$$p_n=0,\quad 0\leqslant n\leqslant k-1,$$

$$x_n=q_n+\sum_{i=1}^{k}\psi_i(hJ_n)(p_n-p_{n-i}),\quad k\leqslant n\leqslant N.$$

Since the vectors $y_n=\overline{D}_{n,j}+p_n$ satisfy (3.9a), we can apply (3.9b) so as to obtain

$$|\overline{D}_{N,j}|\leqslant K_0h^r+\sigma\sum_{n=k}^{N}|x_n|, \tag{3.14a}$$

with

$$|x_n|\leqslant\|S_n-R_n\|K_0h^r+\|S_n-I\|K_2h^{r+1}+\sum_{i=1}^{k}\|\psi_i(hJ_n)\|\cdot|p_n-p_{n-i}|. \tag{3.14b}$$

(2) In view of C1 we have, similarly as in the proof of Theorem 2.2, the inequality (2.8c). Therefore,

$$\|R_n\|\leqslant 1,\qquad \|S_n\|\leqslant 1,\qquad \|\psi_i(hJ_n)\|\leqslant|\alpha_i|.$$

Further, by C3, we have

$$\begin{aligned}S_n-R_n&=R_n\left[\beta_0hJ_n\cdot e(t_n,U(t_n),U(t_{n-1}))\right]S_n\\&=(R_n-I)e(t_n,U(t_n),U(t_{n-1}))S_n,\end{aligned}$$

and therefore

$$\|S_n-R_n\|\leqslant 2L_1\|U'\|\cdot h.$$

A combination of the last three inequalities with (3.14b) shows that

$$|x_n|\leqslant(K_0L_1\|U'\|+K_2)2h^{r+1}+\sum_{i=1}^{k}|\alpha_i|\cdot|p_n-p_{n-i}|.$$

Substituting this upperbound for $|x_n|$ in (3.14a) we obtain

$$|\overline{D}_{N,j}|\leqslant K_0h^r+(K_0L_1\|U'\|+K_2)2\sigma T\cdot h^r+\sigma\sum_{n=k}^{N}\sum_{i=1}^{k}|\alpha_i|\cdot|p_n-p_{n-i}|.$$

By applying (3.12) we arrive at (3.13). □

3.5. Remarks

The above analysis of the global stopping error v_n-u_n may be modified by defining local stopping errors differently, viz.

$$d_{n,j}=\Psi_n(v_{n-1})-\Phi_n(v_{n-1})\quad\text{or}\quad d_{n,j}=\Psi_n(u_{n-1})-\Phi_n(u_{n-1}).$$

In this way one would arrive at a relation for the global stopping error that is simpler than (3.7). However, it would become more difficult to assess the order of such modified local stopping errors.

The theoretical analysis in the Sections 3.1–3.4 leads to the conclusion that, in the stiff situation where C1 and C3 hold, the global stopping error is of the *same order* as the local stopping error. For variable stepsizes, and variants to (3.1) and (1.5) (see (1.5′) and (1.5″) in Section 2.6), one can arrive at the same conclusion by carrying out an analysis similar to the one above. This theoretical conclusion is confirmed by numerical experiments of Groeneweg [8].

The above conclusions are related to an analysis in [20] of the global stopping error in implicit Runge–Kutta methods. In [20] it was shown that this error, measured by the Euclidean norm in $\mathbf{R}^s$, can be of the same order as the local stopping error in the stiff situation (C1, C3). However, it was assumed that $h \cdot \mu(J_n) \ll -1$. Note that this assumption is not used in the present paper nor satisfied in Problem 2.

We note that the order $S_j(h) = j + 1$ (instead of j), which manifests itself in Table 4, is essentially due to the *special structure* of the local stopping error as expressed in (3.10), (3.11); it is *not* due to a strong damping at the point $T = t_N$ of all preceeding local stopping errors $d_{n,j}$ (with $n < N$). In this context it is instructive to consider the *discretization* errors in the applications of (3.5) and (3.6) to Problem 2. Let $e_n = U(t_n) - \Psi_n(\boldsymbol{U}_{n-1})$ and $E_n = U(t_n) - v_n$ denote the *local* and *global* discretization errors, respectively. If a strong error damping would be present, we would expect the global error E_N also to be of the same order as the local errors e_n. But, numerical experiments of Groeneweg [8] show that, e.g. in the application of (3.5) to Problem 2, we have $|e_n| = O(h^2)$ and $|E_N| = O(h)$; the global discretization error is *not* of the same order as the local discretization errors.

The above estimates for the global stopping error are believed to be relevant to the question of how many Newton-type iterations should be carried out in order that the stopping error does not interfere with the intrinsic accuracy of the linear multistep method. For instance, in the application of (3.5) to Problem 2, just one iteration step of (1.5) (starting according to (3.1)) yields an error $|v_N - u_N| = |D_{N,1}| = O(h^2)$, whereas $|U(t_N) - v_N| = |E_N| = O(h)$. According to these estimates it certainly does not pay to perform more than one iteration step. This observation is confirmed by the numerical experiments of Groeneweg [8].

Acknowledgements

The author is indebted to J.L.M. van Dorsselaer, W.H. Hundsdorfer 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] G.J. Cooper, The convergence of some iterative schemes in the solution of stiff ordinary differential equations, Report (1993).

[3] K. Dekker and J.G. Verwer, *Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations* (North Holland, Amsterdam, 1984).
[4] C.A. Desoer, and H. Haneda, The measure of a matrix as a tool to analyse computer algorithms for circuit analysis, *IEEE Trans. Circuit Theory* 19 (1972) 480–486.
[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, *IMA J. Numer. Anal.* 14 (1994) 183–209.
[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] C.W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations* (Prentice-Hall, Englewood Cliffs, 1971).
[8] J. Groeneweg, On the error committed by stopping the Newton iteration in Runge–Kutta methods and linear multistep methods, Leiden University, Department of Mathematics and Computer Science, Report TW-93-13 (1993).
[9] J. Groeneweg, and M.N. Spijker, (1994): On the error due to the stopping of the Newton iteration in implicit linear multistep methods, in: D.F. Griffiths and G.A. Watson, eds., *Numerical analysis 1993, Proceedings of the 15th Dundee Conference* (Longman Scientific and Technical, Harlow, 1994) 157–166.
[10] E. Hairer, S.P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I* (Springer, Berlin, 1987).
[11] E. Hairer, and G. Wanner, *Solving Ordinary Differential Equations II* (Springer, Berlin, 1991).
[12] W.H. Hundsdorfer, A note on monotonicity of a Rosenbrock method, *J. Comput. Appl. Math.* 20 (1987) 267–274.
[13] W.H. Hundsdorfer, Unconditional convergence of some Crank–Nicolson LOD methods for initial-boundary value problems, *Math. Comp.* 58 (1992) 35–53.
[14] 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.
[15] J.D. Lambert, *Numerical Methods for Ordinary Differential Equations* (Wiley, Chichester, 1991).
[16] W. Liniger, A stopping criterion for the Newton–Raphson method in implicit multistep integration algorithms for nonlinear systems of ordinary differential equations, *Comm. ACM* 14 (1971) 600–601.
[17] C. Lubich, On the convergence of multistep methods for nonlinear stiff differential equations, *Numer. Math.* 58 (1991) 839–853.
[18] G. Söderlind, G. The Lipschitz algebra and its extensions, Report (1992).
[19] M.N. Spijker, A note on contractivity in the numerical solution of initial value problems *BIT* 27 (1987) 424–437.
[20] M.N. Spijker, On the error committed by stopping the Newton iteration in implicit Runge–Kutta methods, *Ann. Numer. Math.* 1 (1994) 199–212.