Topics in Numerical Analysis III
Academic Press (1977) (London, etc.)
pp. 383 – 400

The Behaviour of Error Bounds in the Numerical Solution of Initial Value Problems when the Stepsize h Tends to Zero

M. N. Spijker

1. Introduction

1.1. Let us consider an initial value problem for a system of s ordinary differential equations which, by using vector notation, can be written in the form

$$\frac{d}{dt}U(t) = f(t, U(t)) \quad (0 \leqslant t \leqslant T), \quad U(0) = c. \tag{1.1}$$

In (1.1) c denotes a given vector in the s-dimensional real vector space $\mathbb{R}^s$ and f is a given, continuous mapping form $[0, T] \times \mathbb{R}^s$ into $\mathbb{R}^s$. We assume that the Jacobian matrix of the function $f(t, x)$, denoted by

$$J(t, x) = \frac{\partial}{\partial x} f(t, x),$$

exists and is continuous on $[0, T] \times \mathbb{R}^s$. Let $|J(t, x)|$ denote the norm of the matrix $J(t, x)$ subordinate to some given vector norm $|x|$ in $\mathbb{R}^s$ and assume

$$|J(t, x)| \leqslant L \quad (0 \leqslant t \leqslant T, x \in \mathbb{R}^s), \tag{1.2}$$

where L denotes a given positive constant.

1.2. In this paper we are concerned with the approximation of the solution $U(t)$ to (1.1) at the end point $t = T$ of the interval $[0, T]$.

We consider finite-difference schemes that can be written in the form

$$h^{-1}(u_n - u_{n-1}) = F(t_n; u_0, u_1, \ldots, u_n; h) \quad (n = k, k+1, \ldots, N), \tag{1.3a}$$

$$u_n = F(t_n; h) \quad (n = 0, 1, \ldots, k-1). \tag{1.3b}$$

The so-called stepsize h, appearing in (1.3), is assumed to belong to the set H defined by

$$H = \{h \mid h = T/N; N = N_0, N_0 + 1, N_0 + 2, \ldots\}.$$

With N_0 and k we denote fixed integers with $N_0 \geqslant k \geqslant 1$. We use the notation $t_n = nh$ and u_n denotes an approximation of $U(t_n)$ (for $n = 0, 1, \ldots, N$ where $N = T/h$). The vector u_N thus stands for an approximation of $U(t)$ at $t = T$.

The function F—which depends on the given function f and the initial value c occurring in (1.1)—is assumed to satisfy the following conditions (1.4a), (1.4b).

For $h = T/N \in H$, $k \leqslant n \leqslant N$, $x_j \in \mathbb{R}^s$ $(0 \leqslant j \leqslant n)$ the element $y = F(t_n; x_0, x_1, \ldots, x_n; h)$ belongs to $\mathbb{R}^s$. Further, for all x_j and $\tilde{x}_j$ in $\mathbb{R}^s$ we have the inequality

$$|F(t_n; \tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_n; h) - F(t_n; x_0, x_1, \ldots, x_n; h)| \leqslant \sum_{i=0}^{n} \lambda_i |\tilde{x}_{n-i} - x_{n-i}|$$

where $\lambda_0, \lambda_1, \lambda_2, \ldots$ are arbitrary constants independent of $h, n, x_j, \tilde{x}_j$. It is assumed that there exist integers q, r with $0 \leqslant q \leqslant r$ such that the constants λ_i vanish for all $i < q$ and all $i > r$. (1.4a)

For $h = T/N \in H$, $0 \leqslant n \leqslant k - 1$ the element $y = F(t_n; h)$ belongs to $\mathbb{R}^s$. (1.4b)

It is easily verified that many well known numerical methods, when applied to the initial value problem (1.1), generate finite-difference equations of type (1.3a) satisfying condition (1.4a). Such numerical methods are to be found e.g. within the class of (explicit or implicit) Runge-Kutta methods or the class of (generalized) linear multistep methods (cf. [4], [8]).

The elements $F(t_n; h)$ occurring in (1.3b) denote starting vectors in $\mathbb{R}^s$ found e.g. by a Taylor expansion of $U(t)$ at $t = 0$.

We refer to Section 4.3 for a detailed example of a finite-difference scheme of the form (1.3).

1.3. We consider the following perturbed version (1.5) of the given finite-difference scheme (1.3).

$$h^{-1}(\tilde{u}_n - \tilde{u}_{n-1}) = F(t_n; \tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_n; h) + w_n \quad (n = k, k+1, \ldots, N), \tag{1.5a}$$

$$\tilde{u}_n = F(t_n; h) + w_n \quad (n = 0, 1, \ldots, k-1). \tag{1.5b}$$

$\tilde{u}_n$ denotes the approximation of $U(t_n)$ obtained in the presence of some local perturbations $w_0, w_1, \ldots, w_N$. For instance w_n may be caused by

rounding-off. Likewise w_n may be understood to be the local discretization error (see [8] for this terminology), in which case we have $\tilde{u}_n = U(t_n)$. Clearly, in both of these cases it is desirable to have error bounds by means of which the effect on the difference $\tilde{u}_N - u_N$ of the perturbations w_n $(n = 0, 1, \ldots, N)$ can be estimated.

Such error bounds are particularly interesting for large values of N. Since $N = T/h$, the cases where $T \to \infty$ and where $h \downarrow 0$ thus deserve special attention. In this paper we want to investigate in what manner $\tilde{u}_N - u_N$ behaves when $h \downarrow 0$ while $T > 0$ remains fixed. Consequently we shall only be interested in *the behaviour of error bounds as the stepsize h tends to zero and the interval* $[0, T]$ *is fixed.* For a discussion of the dependence of $\tilde{u}_N - u_N$ on T (for large T) see e.g. [1], [2], [8] p. 87 ff.

1.4. Let us assume that the maximal stepsize h_0, given by

$$h_0 = T/N_0,$$

is such that

$$\lambda_0 h_0 < 1. \tag{1.6}$$

Then it can be proved (cf. e.g. [4], [6]-[9] and Theorem 2 of the present paper) that the equations (1.3a), (1.5a) have unique solutions u_n and $\tilde{u}_n$ $(n = k, k+1, \ldots, N)$, respectively and that the difference $\tilde{u}_N - u_N$ admits the following bounds

$$|\tilde{u}_N - u_N| \leqslant \Gamma_1 \cdot \max_{0 \leqslant n \leqslant N} |w_n| \tag{1.7}$$

and

$$|\tilde{u}_N - u_N| \leqslant \Gamma_2 \cdot \left\{ |w_0| + |w_1| + \cdots + |w_{k-1}| + \max_{k \leqslant n \leqslant N} \left| h \sum_{j=k}^{n} w_j \right| \right\}. \tag{1.8}$$

In these error bounds Γ_1 and Γ_2 denote parameters (which depend on F and T, but) which do not depend on the stepsize $h \in H$ or on the perturbations $w_n \in \mathbb{R}^s$.

Using the triangle inequality and the inequality $(N - k + 1)\, h \leqslant T$, we see that (1.8) implies (1.7) (with $\Gamma_1 = (k + T) \cdot \Gamma_2$). Moreover, in case $w_n = 0$ $(0 \leqslant n \leqslant N, n \neq k)$ and $|w_k| = \epsilon$ ($\epsilon > 0$ denoting some constant independent of h) the error bound (1.8) establishes the fact that $|\tilde{u}_N - u_N|$ tends to zero if $h \downarrow 0$, while (1.7) only shows that $|\tilde{u}_N - u_N|$ remains bounded. Consequently, with regard to the behaviour as $h = T/N \downarrow 0$ and $T > 0$ is fixed, the error bound (1.8) is *more refined* than (1.7).

1.5. The question arises whether there exists an error bound being still more refined than (1.8). A further natural question is whether there exists an error bound that is "maximally refined".

In this paper an attempt is made to formulate these questions rigorously and to find answers to them.

We note that a similar investigation was performed by the author in [6], a major difference with the present paper being that in the setting of [6] the left-hand members $|\tilde{u}_N - u_N|$ of (1.7), (1.8) would read

$$\max_{0 \leqslant n \leqslant N} |\tilde{u}_n - u_n|.$$

In the subsequent chapters it will turn out that new and unexpected results arise as a consequence of the seemingly immaterial replacement of

$$\max_{0 \leqslant n \leqslant N} |\tilde{u}_n - u_n| \text{ by } |\tilde{u}_N - u_N|.$$

1.6. In Section 2 we present our basic definitions and notations.

In Section 3 we prove our first theorem, containing a (negative) result about the existence of an error bound that is maximally refined.

In Section 4 we relax the requirements to be fulfilled by error bounds in order that they are called maximally refined. We obtain a second theorem according to which there exists an error bound that is maximally refined. This error bound turns out to be more refined than (1.8). Section 4 is concluded with an application of this maximally refined error bound to an Adams-type finite-difference scheme for solving the initial value problem (1.1).

In the Sections 5 and 6 we have collected a series of lemmata, two of which are used in the proof of Theorem 2 (in Section 4). We note that the lemmata of Section 6 have applications to finite-difference schemes which are of a more general type than (1.3).

Let $\alpha_0, \alpha_1, \ldots, \alpha_k$ denote real constants with $\alpha_k = 1$, such that the root condition (1.9) is fulfilled.

All roots ζ of the equation

$$\alpha_k \zeta^k + \cdots + \alpha_1 \zeta + \alpha_0 = 0 \tag{1.9}$$

have a modulus $|\zeta| \leqslant 1$ and roots with modulus $|\zeta| = 1$ are simple.

The lemmata of Section 6 provide bounds for the error $\tilde{u}_N - u_N$ where u_N is computed from

$$h^{-1}(\alpha_k u_n + \alpha_{k-1} u_{n-1} + \cdots + \alpha_0 u_{n-k}) = F(t_n; u_0, u_1, \ldots, u_n; h) \quad (k \leqslant n \leqslant N), \tag{1.10a}$$

$$u_n = F(t_n; h) \quad (0 \leqslant n \leqslant k-1), \tag{1.10b}$$

and $\tilde{u}_N$ is computed from

$$h^{-1}(\alpha_k \tilde{u}_n + \alpha_{k-1} \tilde{u}_{n-1} + \cdots + \alpha_0 \tilde{u}_{n-k}) = F(t_n; \tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_n; h) + w_n \quad (k \leqslant n \leqslant N), \tag{1.11a}$$

$$\tilde{u}_n = F(t_n; h) + w_n \quad (0 \leqslant n \leqslant k-1). \tag{1.11b}$$

The final lemma of Section 6 gives an error bound which is a straightforward generalization of the maximally refined error bound stated in Theorem 2 for the case of the scheme (1.3).

2. Notations and Definitions

2.1. Let σ, σ_0 be given fixed constants with

$$\sigma > 0, \quad \sigma_0 \geqslant 0, \quad \sigma_0 h_0 < 1. \tag{2.1}$$

We shall investigate error bounds which are valid for finite-difference schemes (1.3) in which F fulfils condition (1.4) with arbitrary constants λ_i satisfying

$$\sum_{i=q}^{r} \lambda_i \leqslant \sigma, \quad \lambda_0 \leqslant \sigma_0. \tag{2.2}$$

We denote the set consisting of all these functions F by $S_0(\sigma, \sigma_0; N_0, k; T, s)$ or simply by $S_0(\sigma, \sigma_0)$ or by S_0. We thus have

$$S_0 = \{F \mid F \text{ fulfils (1.4) with constants } \lambda_i \text{ satisfying (2.2)}\}.$$

In order to keep the framework of this paper sufficiently general, so as to be able to cope with diverse applications, we give our subsequent definitions relative to an arbitrary set S of functions F which is only required to be such that

(i) S is a subset of S_0,

(ii) For each function f satisfying the conditions listed in Section 1.1 with $L = \sigma$, there can be found an $F \in S$ such that

$$F(t_n; x_0, x_1, \ldots, x_n; h) = \frac{1}{k} \cdot \sum_{i=1}^{k} f(t_{n-i}, x_{n-i})$$

(for $h = T/N \in H$, $k \leqslant n \leqslant N$, $x_j \in \mathbb{R}^s$).

Condition (i) implies that the set S should not be chosen too large, while (ii) prohibits S from being chosen too small.

Note that for instance

$$S = S_0 \tag{2.3}$$

is a set satisfying the conditions (i), (ii). Choosing $N_0 = k = 1$ and

$$S = \{F \mid F(t_0; h) \equiv c, F(t_n; x_0, x_1, \ldots, x_n; h) \equiv f(t_{n-1}, x_{n-1})$$

$(1 \leqslant n \leqslant N)$; $c \in \mathbb{R}^s$ and f fulfils the conditions of Section 1.1 with $L = \sigma\}$ (2.4)

we have another example of a set S satisfying (i), (ii). Clearly, the finite-difference schemes (1.3) with $F \in S$ where S is given by (2.4), are gener-

ated by applications of Euler's method to arbitrary initial value problems (1.1).

2.2. For $h \in H$ we define the space X_h, consisting of all possible perturbations, by

$$X_h = \{w \mid w = (w_0, w_1, \ldots, w_N) \text{ where } N = T/h,\ w_n \in \mathbb{R}^s \quad (0 \leqslant n \leqslant N)\}.$$

We shall deal with bounds for the errors $\tilde{u}_N - u_N$, resulting from local perturbations w_n in (1.5), that can be written in the form

$$|\tilde{u}_N - u_N| \leqslant \psi(w, h, F) \quad (\text{for all } h \in H,\ w \in X_h,\ F \in S). \tag{A}$$

In (A) w stands for $w = (w_0, w_1, \ldots, w_N)$ and w_n, h, F, $\tilde{u}_N$, u_N denote the quantities occurring in (1.3), (1.5). Further ψ denotes an arbitrary real function with domain

$$D = \{(w, h, F) \mid h \in H,\ w \in X_h,\ F \in S\}.$$

It can be shown (see e.g. [7] or the arguments used in Section 6) that (1.7), (1.8) hold with $\Gamma_1 = \gamma_1(F)$, $\Gamma_2 = \gamma_2(F)$ where

$$\gamma_1(F) = (1 + T) \cdot \exp(\theta\lambda T), \tag{2.5a}$$

$$\gamma_2(F) = \exp(\theta\lambda T) \tag{2.6a}$$

and

$$\lambda = \sum_{i=q}^{r} \lambda_i, \qquad \theta = (1 - \lambda_0 h_0)^{-1},$$

λ_i denoting the smallest constants for which F satisfies (1.4a). Consequently (1.7), (1.8) are examples of error bounds of type (A) (e.g. with S given by (2.3)) with $\psi = \psi_1$ and $\psi = \psi_2$, respectively where

$$\psi_1(w, h, F) = \gamma_1(F) \cdot \max_{0 \leqslant n \leqslant N} |w_n|, \tag{2.5b}$$

$$\psi_2(w, h, F) = \gamma_2(F) \cdot \left\{ |w_0| + |w_1| + \cdots + |w_{k-1}| + \max_{k \leqslant n \leqslant N} \left| h \sum_{j=k}^{n} w_j \right| \right\}. \tag{2.6b}$$

2.3. The following definitions will enable us to compare the structures of error bounds like (1.7) and (1.8) in a rigorous fashion.

Definition 1

Let ψ and ψ' be real functions with domain D. Assume there exists a constant $\beta > 0$ such that $\psi(w, h, F) \leqslant \beta \cdot \psi'(w, h, F)$ (for all $(w, h, F) \in D$). Then we use the notation

$$\psi \lhd \psi'.$$

Along with (A) we consider another error bound

$$|\tilde{u}_N - u_N| \leqslant \psi'(w, h, F) \quad \text{(for all } h \in H, w \in X_h, F \in S). \tag{A'}$$

Definition 2
The error bound (A') is *more refined* than (A) if $\psi' \preceq \psi$ and not $\psi \preceq \psi'$.

As an illustration of the above definitions we again consider the functions ψ_1, ψ_2 defined by (2.5), (2.6). By arguments already mentioned in Section 1.4 it is easily proved that $\psi_2 \preceq \psi_1$. By choosing $w_k \neq 0$, $w_n = 0$ $(n \neq k)$ and letting $h \downarrow 0$ the assumption $\psi_1 \preceq \psi_2$ is seen to lead to a contradiction. Hence we do not have $\psi_1 \preceq \psi_2$. According to Definition 2 the error bound (1.8) is thus more refined than (1.7).

2.4. In the following Sections we have adopted the convention that

$$\sum_{i=m}^{n} \ldots = 0 \tag{2.7}$$

whenever $m > n$.

3. A First Definition of "Maximally Refined"

3.1. The following definition is a natural sequel to Definition 2.

Definition 3
The error bound (A) is *maximally refined* if there exists no bound of type (A') which is more refined than (A).

We define the function ψ_0 by

$$\psi_0(w, h, F) = |\tilde{u}_N - u_N| \tag{3.1}$$

where $u_N, \tilde{u}_N$ are computed from (1.3), (1.5) and $|\tilde{u}_N - u_N|$ thus can be regarded as a function of $w = (w_0, w_1, \ldots, w_N)$, h and F. Clearly, with the choice $\psi = \psi_0$ the error bound (A) is maximally refined. However, at the same time the error bound (A) now has become useless, since the values $\psi_0(w, h, F)$ depend on $w = (w_0, w_1, \ldots, w_N)$ in a manner which, in general, is not transparent. This is due to the fact that $y = F(t_n; x_0, x_1, \ldots, x_n; h)$ may depend on $x_0, x_1, \ldots, x_n$ in a complicated (nonlinear) fashion.

It is in view of the existence of such trivial and simultaneously untransparent error bounds that we shall focus on error bounds in which, apart from factors independent of w and h, the right-hand members are independent of the (complicated) function F. Thus we are lead to consider bounds of type (A) in which the function ψ is of the form

$$\psi(w, h, F) = \gamma(F) \cdot \Phi(w, h) \quad \text{(for } (w, h, F) \in D), \tag{B}$$

γ denoting an arbitrary positive function with domain S and Φ denoting an arbitrary real function with domain $\{(w, h) \mid h \in H, w \in X_h\}$.

We note that (2.5b) and (2.6b) provide examples of functions ψ that are of type (B).

3.2. In contrast with what was found in the analogous situation considered in [6], we have the following negative result.

Theorem 1

There exists no error bound of type (A) with the following two properties.

(1) (A) is maximally refined (in the sense of Definition 3),

(2) The right-hand member $\psi(w, h, F)$ of (A) is of type (B).

Proof

(a) Suppose (A) stands for an error bound with the properties 1, 2.

From (3.1) we then have

$$\psi_0(w, h, F) \leqslant \psi(w, h, F) = \gamma(F) \cdot \Phi(w, h).$$

Consequently, $\psi_0 \dashv \psi$, and since (A) is maximally refined there follows

$$\psi \dashv \psi_0. \tag{3.2}$$

For any two functions F_1, F_2 belonging to the set S we have

$$\psi_0(w, h, F_1) \leqslant \psi(w, h, F_1) = \frac{\gamma(F_1)}{\gamma(F_2)} \cdot \psi(w, h, F_2).$$

In view of (3.2) we thus obtain the inequality

$$\psi_0(w, h, F_1) \leqslant \beta \cdot [\gamma(F_1)/\gamma(F_2)] \cdot \psi_0(w, h, F_2) \quad (\text{for all } h \in H, w \in X_h), \tag{3.3}$$

β denoting a positive constant independent of h and w.

(b) In view of condition (ii), imposed on S in Section 2.1, we may choose F_1 and F_2 in such a way that

$$F_1(t_n; x_0, x_1, \ldots, x_n; h) \equiv \frac{\sigma}{k} \cdot \sum_{i=1}^{k} x_{n-i}, \quad F_2(t_n; x_0, x_1, \ldots, x_n; h) \equiv 0.$$

Let $w = (w_0, w_1, \ldots, w_N)$ be such that

$$w_n = 0 \quad (0 \leqslant n \leqslant N-2), \quad w_N = -w_{N-1} \neq 0. \tag{3.4}$$

A little calculation shows that for $N \geqslant k+1$ we have

$$\psi_0(w, h, F_2) = 0, \quad \psi_0(w, h, F_1) = h^2 \cdot \frac{\sigma}{k} \cdot |w_N| \neq 0.$$

In view of (3.3) there follows

$$\psi_0(w, h, F_1) = 0.$$

We have obtained a contradiction and the theorem has thus been proved.

We note that the above theorem is closely related to Lemma 6 in [7].

4. A Second Definition of "Maximally Refined"

4.1. In view of the negative result formulated in Theorem 1 we now turn to a weaker version of Definition 3. We shall impose less restrictive requirements on an error bound (A) if it is to be called maximally refined. In fact, we shall compare a given error bound (A) not with arbitrary bounds (A′), as was done in Definition 3, but only with certain bounds which have the same structure as the bounds (A) considered in Section 3 (see (B)). Thus we are going to restrict our considerations to error bounds (A′) whose right-hand members are of type

$$\psi'(w, h, F) = \gamma'(F) \,.\, \Phi'(w, h) \quad (\text{for } (w, h, F) \in D). \tag{B′}$$

Definition 4
The error bound (A) is *maximally refined* if there exists no bound of type (A′) with the following two properties.

(1) (A′) is more refined than (A),

(2) The right-hand member $\psi'(w, h, F)$ of (A′) is of the form (B′) with $0 < \inf_{F \in S} \gamma'(F), \sup_{F \in S} \gamma'(F) < \infty$.

We note that (1.7), (1.8) are examples of bounds of type (A′) with the second property stated in the above definition (see (2.5), (2.6)).

In the Section 4.2 it will be seen that Definition 4 carries us farther than Definition 3.

4.2. In the following theorem we shall refer to the functions γ and Φ defined by

$$\Phi(w, h) = \left| w_{k-1} + h \sum_{j=k}^{N} w_j \right| + h \sum_{j=0}^{k-1} |w_j| + h \sum_{n=k}^{N-1} \left| w_{k-1} + h \sum_{j=k}^{n} w_j \right|, \tag{4.1}$$

$$\gamma(F) = \max \{1 + \theta\lambda h_0, \theta\lambda \,.\, \exp[\theta\lambda T]\} \tag{4.2a}$$

where

$$\lambda = \sum_{i=q}^{r} \lambda_i, \quad \theta = (1 - \lambda_0 h_0)^{-1}, \tag{4.2b}$$

λ_i denoting the smallest constants for which F satisfies (1.4a).

Theorem 2
Let ψ be defined by $\psi(w, h, F) = \gamma(F) \,.\, \Phi(w, h)$ where γ and Φ are given by (4.2), (4.1).

Then the error bound (A) holds. Further, with this choice of ψ, the bound (A) has the following three properties.

(1) (A) is maximally refined (in the sense of Definition 4),

(2) The right-hand member $\psi(w, h, F)$ of (A) is of type (B).
(3) (A) is more refined than (1.8).

Proof

(a) Applying lemma 4 (see Section 6) with

$$\alpha_k = 1, \quad \alpha_{k-1} = -1, \quad \alpha_i = 0 \quad (0 \leqslant i \leqslant k-2)$$

we obtain the error bound

$$|\tilde{u}_N - u_N| \leqslant |v_N| + \theta\lambda \,.\, h \sum_{n=0}^{N} e^{\theta\lambda t_{N-n}} \,.\, |v_n|$$

where θ, λ and v_n are defined by (6.6) and (6.4), respectively. From (6.4) we easily obtain the expressions (5.4) for the vectors v_n. The error bound (A) thus holds with $\psi(w, h, F) = \gamma(F) \,.\, \Phi(w, h)$ where γ and Φ are defined in (4.2) and (4.1). Further it is easily seen that our bound (A) has property 2 stated in the theorem.

(b) Let ψ be as defined in the theorem. We shall prove that the error bound (A) now has property 1.

By applying Lemma 1 (of Section 5) with $L = \sigma$ and by using condition (ii), imposed on S in Section 2.1, it follows that for all $h \in H$, $w \in X_h$ there can be found an $F \in S$ such that

$$\psi_0(w, h, F) \geqslant \beta \,.\, \Phi(w, h) \tag{4.3}$$

where ψ_0 is defined by (3.1), Φ is defined by (4.1) and $\beta = \min\{1, [(1 + \sigma T) \,.\, k]^{-1} \,.\, \sigma\}$.

Suppose (A′) is an error bound with the properties 1, 2 appearing in Definition 4. We put

$$\epsilon_0 = \inf_{F \in S} \gamma'(F), \qquad \epsilon_1 = \sup_{F \in S} \gamma'(F).$$

It follows that

$$\psi_0(w, h, F) \leqslant \epsilon_1 \,.\, \Phi'(w, h) \quad \text{(for all } (w, h, F) \in D\text{)}. \tag{4.4}$$

A combination of (4.3) and (4.4) yields

$$\beta \,.\, \Phi(w, h) \leqslant \epsilon_1 \,.\, \Phi'(w, h) \quad \text{(for all } h \in H, w \in X_h\text{)}.$$

Consequently, for all $(w, h, F) \in S$ we have

$$\psi(w, h, F) = \gamma(F) \,.\, \Phi(w, h) \leqslant \beta^{-1}\gamma(F)\epsilon_1 \,.\, \Phi'(w, h) \leqslant [\beta\epsilon_0]^{-1}\gamma(F)\epsilon_1 \,.\, \psi'(w, h, F).$$

Since

$$\sup_{F \in S} \gamma(F) < \infty$$

(see (4.2) and condition (i) in Section 2.1) we thus obtain

$$\psi \not\prec \psi'.$$

In view of property 1 (see Definition 4) we have obtained a contradiction. It follows that (A), with ψ as defined in the Theorem, is maximally refined.

(c) It remains to be shown that our error bound (A) has property 3.

From (2.6), (4.1) there follows

$$\Phi(w, h) \leqslant (1 + T) \cdot [\gamma_2(F)]^{-1} \cdot \psi_2(w, h, F).$$

Hence

$$\psi(w, h, F) = \gamma(F) \cdot \Phi(w, h) \leqslant (1 + T) \cdot [\gamma(F)/\gamma_2(F)] \cdot \psi_2(w, h, F).$$

Since $\sup_{F \in S} [\gamma(F)/\gamma_2(F)] < \infty$ we thus have

$$\psi \prec \psi_2.$$

By choosing $w = (w_0, w_1, \ldots, w_N)$ in such a way that (3.4) holds we obtain, for $N \geqslant k + 1$,

$$\psi(w, h, F) = \gamma(F) \cdot h^2 \cdot |w_N|, \qquad \psi_2(w, h, F) = \gamma_2(F) \cdot h \cdot |w_N|.$$

By letting $h \downarrow 0$ it follows that we do not have $\psi_2 \prec \psi$. In view of Definition 2 this completes the proof of the theorem.

4.3. In order to illustrate Theorem 2 we consider the following Adams-type finite-difference scheme for solving the initial value problem (1.1).

$$h^{-1}(u_1 - u_0) = f(t_0, u_0)$$

$$h^{-1}(u_2 - u_1) = 2f(t_1, u_1) - f(t_0, u_0)$$

$$h^{-1}(u_n - u_{n-1}) = \tfrac{1}{12} \cdot [23f(t_{n-1}, u_{n-1}) - 16f(t_{n-2}, u_{n-2}) + 5f(t_{n-3}, u_{n-3})] \quad (3 \leqslant n \leqslant N), \tag{4.5a}$$

$$u_0 = c. \tag{4.5b}$$

It is easily verified that (4.5) is of type (1.3), where F satisfies (1.4) with $k = 1, N_0 = 3, q = 1, r = 3$.

Let $\tilde{u}_n$ denote the solution of the finite-difference scheme (4.5) obtained in the presence of local perturbations w_n added to the right-hand members of (4.5a) (we assume $\tilde{u}_0 - u_0 = w_0 = 0$). Then an application of the error bound (A) established in Theorem 2 yields the estimate

$$|\tilde{u}_N - u_N| \leqslant \Gamma \cdot \left\{ \left| h \sum_{j=1}^{N} w_j \right| + h \sum_{n=1}^{N-1} \left| h \sum_{j=1}^{n} w_j \right| \right\}, \tag{4.6}$$

Γ denoting a parameter independent of $h = T/N$ and w_n.

In case the w_n stand for rounding errors, the bound (4.6) clearly shows that there are (hypothetical) cases where the effect (as $h \downarrow 0$) of these errors on $\tilde{u}_N - u_N$ is less serious than could be expected from the corresponding bounds (1.7) or (1.8). For instance let $w_m = -w_{m-1}$, $|w_m| = \epsilon$, $w_j = 0$ (for $j \neq m, m-1$) where m is an arbitrary integer with $1 < m \leqslant N$ and $\epsilon > 0$ is a constant independent of h. Then (1.7) shows that $|\tilde{u}_N - u_N|$ remains bounded as $h \downarrow 0$ and (1.8) proves that $|\tilde{u}_N - u_N| = O(h)$. But from (4.6) we have the still stronger estimate $|\tilde{u}_N - u_N| = O(h^2)$.

We next choose $\tilde{u}_n = U(t_n)$ where $U(t)$ solves (1.1). Assuming that the function $U(t)$ has derivatives of sufficiently high order we obtain, by using Taylor's theorem, for the corresponding perturbations w_n the expressions

$$w_1 = \frac{h}{2} \cdot \frac{d^2}{dt^2} U(t_0) + O(h^2), \qquad w_2 = -\frac{h}{2} \cdot \frac{d^2}{dt^2} U(t_0) + O(h^2),$$

$$w_n = O(h^3) \qquad \text{(uniformly for } 3 \leqslant n \leqslant N\text{)}.$$

On substituting these expressions into (4.6) there follows

$$|U(T) - u_N| = O(h^3).$$

Note that this estimate could not have been obtained in the same straightforward fashion from (1.7) or (1.8).

5. A Lower Bound for $|\tilde{u}_N - u_N|$

The purpose of the present section is to prove the following lemma, which has already been used in the proof of Theorem 2. The lemma deals with the finite-difference scheme

$$h^{-1}(u_n - u_{n-1}) = \frac{1}{k} \cdot \sum_{i=1}^{k} f(t_{n-i}, u_{n-i}) \qquad (n = k, k+1, \ldots, N), \tag{5.1a}$$

$$u_n = c_n \qquad (n = 0, 1, \ldots, k-1) \tag{5.1b}$$

and the perturbed scheme

$$h^{-1}(\tilde{u}_n - \tilde{u}_{n-1}) = \frac{1}{k} \cdot \sum_{i=1}^{k} f(t_{n-i}, \tilde{u}_{n-i}) + w_n \qquad (n = k, k+1, \ldots, N), \tag{5.2a}$$

$$\tilde{u}_n = c_n + w_n \qquad (n = 0, 1, \ldots, k-1). \tag{5.2b}$$

Lemma 1
Let $h \in H$, $w = (w_0, w_1, \ldots, w_N) \in X_h$ be given and assume L is an arbitrary positive constant. Then there exists a function f, with the properties stated

in Section 1.1, such that whenever $c_0, c_1, \ldots, c_{k-1}$ are vectors $\in \mathbb{R}^s$ and the vectors $u_N, \tilde{u}_N$ are computed from (5.1), (5.2), then

$$|\tilde{u}_N - u_N| \geqslant |v_N| + \frac{L}{k \,.\, (1 + LT)} \cdot h \sum_{n=0}^{N-1} |v_n| \tag{5.3}$$

where

$$v_n = w_n \quad (0 \leqslant n \leqslant k-1), \quad v_n = w_{k-1} + h \sum_{j=k}^{n} w_j \quad (k \leqslant n \leqslant N). \tag{5.4}$$

Proof

(a) Let $h \in H$, $w = (w_0, w_1, \ldots, w_N) \in X_h$, $L > 0$ be given. Let v_n be defined by (5.4) and put

$$a = \frac{1}{|v_N|} \cdot v_N$$

if $v_N \neq 0$. Let a be an arbitrary vector in $\mathbb{R}^s$ with norm $|a| = 1$ if $v_N = 0$. We define

$$f(t_n, x) = L \,.\, g_n(x) \,.\, a \quad (n = 0, 1, \ldots, N)$$

where g_n denotes a linear mapping from $\mathbb{R}^s$ into $\mathbb{R}$. g_n will be defined below. We define $f(t, x)$ by linear interpolation for $t_{n-1} < t < t_n$ $(n = 1, 2, \ldots, N)$.

Let n be an integer with $0 \leqslant n \leqslant k-1$. We define g_n to be a linear functional with $g_n(w_n) = |w_n|$ and with norm

$$|g_n| = \sup\{|g_n(x)|/|x| : x \in \mathbb{R}^s, x \neq 0\} \leqslant 1$$

(Note that such a functional exists by [5], Section 3).

Next let n be an integer with $k \leqslant n \leqslant N$ and assume g_j has already been defined for all $j \leqslant n-1$. Then $f(t, x)$ is defined for $0 \leqslant t \leqslant t_{n-1}$ as indicated above and $u_j, \tilde{u}_j$ are defined by (5.1), (5.2) for $0 \leqslant j \leqslant n$. Now the functional g_n is chosen in such a way that $g_n(\tilde{u}_n - u_n) = |\tilde{u}_n - u_n|$ and again $|g_n| \leqslant 1$. Note that $\tilde{u}_n - u_n$ and consequently g_n, too, are independent of $c_0, c_1, \ldots, c_{k-1}$.

In this way we have constructed a function f satisfying the conditions stated in Section 1.1.

(b) We define

$$d_n = \tilde{u}_n - u_n, \quad z_n = d_n - v_n \quad (n = 0, 1, \ldots, N)$$

where $u_n, \tilde{u}_n, v_n$ denote the vectors defined by (5.1), (5.2), (5.4), respectively.

Subtracting (5.1) from (5.2) there follows

$$h^{-1}(z_n - z_{n-1}) = \frac{L}{k} \cdot \sum_{i=1}^{k} g_{n-i}(d_{n-i}) \,.\, a \quad (k \leqslant n \leqslant N), \quad z_n = 0$$
$$(0 \leqslant n \leqslant k-1).$$

In view of the definition of the functionals g_j we thus have

$$z_n - z_{n-1} = \left(\frac{L}{k} \cdot h \sum_{i=1}^{k} |d_{n-i}|\right) \,.\, a \quad (k \leqslant n \leqslant N).$$

Consequently

$$z_n = \left\{\frac{L}{k} \cdot h \sum_{j=k}^{n} (|d_{j-1}| + |d_{j-2}| + \cdots + |d_{j-k}|)\right\} \,.\, a \quad (k \leqslant n \leqslant N). \tag{5.5}$$

Applying (5.5) with $n = N$ we obtain

$$d_N = \left\{|v_N| + \frac{L}{k} \cdot h \sum_{j=k}^{N} (|d_{j-1}| + |d_{j-2}| + \cdots + |d_{j-k}|)\right\} \,.\, a.$$

It follows that

$$|d_N| \geqslant |v_N| + \frac{L}{k} \cdot h \sum_{n=0}^{N-1} |d_n|. \tag{5.6}$$

(c) Since $|v_n| \leqslant |d_n| + |z_n|$ we obtain from (5.5) the inequality

$$|v_n| \leqslant |d_n| + L \,.\, h \sum_{j=0}^{n-1} |d_j| \quad (0 \leqslant n \leqslant N).$$

Hence

$$h \sum_{n=0}^{N-1} |v_n| \leqslant h \sum_{n=0}^{N-1} |d_n| + h \sum_{n=0}^{N-1} \left(Lh \sum_{j=0}^{n-1} |d_j|\right).$$

It follows that

$$(1 + LT)^{-1} \,.\, h \sum_{n=0}^{N-1} |v_n| \leqslant h \sum_{n=0}^{N-1} |d_n|.$$

A combination of this inequality with (5.6) proves the inequality (5.3).

6. Error Bounds for Finite-Difference Schemes of Type (1.10)

6.1. Throughout this section F denotes a fixed function satisfying condition (1.4) and with λ_i, q, r we denote the constants appearing in (1.4a).

We first consider a finite-difference scheme of type (1.10) satisfying the following general condition (6.1).

Whenever $h \in H$ and $w_0, w_1, \ldots, w_N$ are vectors in $\mathbb{R}^s$ with $w_n = 0$ (for $n = 0, 1, \ldots, k-1$) and $u_n, \tilde{u}_n$ satisfy (1.10), (1.11), then

$$|\tilde{u}_N - u_N| \leqslant P(h) \cdot \sum_{j=k}^{N} [Q(h)]^{N-j} |w_j|$$

where $P(h)$, $Q(h)$ are real functions, defined for $h \in H$, independent of w_n. (6.1)

Lemma 2
Let the finite-difference scheme (1.10) satisfy condition (6.1). Let $h \in H$ and let $w_0, w_1, \ldots, w_N$ be arbitrary perturbations in $\mathbb{R}^s$. Assume $u_n, \tilde{u}_n$ satisfy (1.10), (1.11), respectively. Then we have the error bound

$$|\tilde{u}_N - u_N| \leqslant |v_N| + \sum_{n=0}^{N} R_n(h) |v_n|. \tag{6.2}$$

The factors $R_n(h)$ and the vectors v_n are defined by

$$R_n(h) = P(h) \cdot \sum_{i=\max(q,k-n)}^{\min(r,N-n)} \lambda_i [Q(h)]^{N-n-i}, \tag{6.3}$$

$$\alpha_k v_n + \alpha_{k-1} v_{n-1} + \cdots + \alpha_0 v_{n-k} = h w_n \quad (k \leqslant n \leqslant N),$$
$$v_n = w_n \quad (0 \leqslant n \leqslant k-1). \tag{6.4}$$

Proof
From (1.11) it follows that the vectors r_n defined by

$$r_n = \tilde{u}_n - v_n$$

satisfy

$$h^{-1}(\alpha_k r_n + \alpha_{k-1} r_{n-1} + \cdots + \alpha_0 r_{n-k}) = F(t_n; r_0, r_1, \ldots, r_n; h) + s_n$$
$$(k \leqslant n \leqslant N),$$

where

$$s_n = F(t_n; \tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_n; h) - F(t_n; r_0, r_1, \ldots, r_n; h).$$

Applying (6.1) (with $\tilde{u}_n, w_n$ replaced by r_n, s_n, respectively) we obtain

$$|r_N - u_N| \leqslant P(h) \sum_{j=k}^{N} [Q(h)]^{N-j} |s_j|.$$

From the definition of s_j and condition (1.4) there follows

$$|s_j| \leqslant \sum_{i=q}^{r} \lambda_i |v_{j-i}| \quad (\text{we put } v_n = 0 \text{ for } n < 0).$$

Consequently

$$|\tilde{u}_N - u_N| \leqslant |v_N| + P(h) \sum_{j=k}^{N} \sum_{i=q}^{r} \lambda_i [Q(h)]^{N-j} |v_{j-i}|.$$

By a rearrangement of the sum appearing in the right-hand member of this inequality and by using the convention (2.7) we arrive at the error bound (6.2).

6.2. The following lemma states a simple condition under which the general requirement (6.1) is fulfilled by the finite-difference scheme (1.10). The proof of the lemma has been included to keep the present paper reasonably self-contained. It is similar to proofs of related theorems to be found in the literature (see e.g. [4], [8]).

It will be assumed that λ_0 is so small that

$$\lambda_0 h_0 < 1 \tag{6.5}$$

and we put

$$\theta = (1 - \lambda_0 h_0)^{-1}, \quad \lambda = \sum_{i=q}^{r} \lambda_i. \tag{6.6}$$

Lemma 3

Let (6.5) hold. Assume the coefficients α_i occurring in (1.10) are such that $\alpha_k = 1$ and the root condition (1.9) is fulfilled. Then the finite-difference scheme (1.10) satisfies requirement (6.1) with

$$P(h) = \alpha\theta \,.\, h, \quad Q(h) = (1 + \alpha\theta\lambda \,.\, h). \tag{6.7}$$

In (6.7) α denotes a constant, which only depends on the coefficients α_i and which is given by

$$\alpha = 1 \tag{6.8}$$

in case $\alpha_k = 1$, $\alpha_{k-1} = -1$, $\alpha_i = 0$ $(0 \leqslant i \leqslant k - 2)$.

Proof

Let $h \in H$ and $w_0, w_1, \ldots, w_N \in \mathbb{R}^s$ with $w_n = 0$ $(0 \leqslant n \leqslant k - 1)$ be given. Subtracting (1.10) from (1.11) and writing

$$d_n = \tilde{u}_n - u_n$$

there follows

$$\alpha_k d_n + \alpha_{k-1} d_{n-1} + \cdots + \alpha_0 d_{n-k} = h s_n \quad (k \leqslant n \leqslant N)$$

where

$$s_n = w_n + \{F(t_n; \tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_n; h) - F(t_n; u_0, u_1, \ldots, u_n; h)\}.$$

By virtue of the root condition (1.9) we have

$$|d_n| \leqslant h|s_n| + \alpha \,.\, h \sum_{j=k}^{n-1} |s_j| \quad (k \leqslant n \leqslant N)$$

where $\alpha \geqslant 1$ is a constant with the properties stated in the lemma (cf. e.g. [8] p. 205).

In view of the Lipschitz condition appearing in (1.4a) we have

$$|s_j| \leqslant |w_j| + \lambda_0 |d_j| + \sum_{i=1}^{j} \lambda_i |d_{j-i}| \quad (k \leqslant j \leqslant N).$$

Consequently

$$(1 - \lambda_0 h)|d_n| \leqslant \alpha\lambda \,.\, h \sum_{j=k}^{n-1} |d_j| + \alpha \,.\, h \sum_{j=k}^{n} |w_j| \quad (k \leqslant n \leqslant N).$$

Since $h \leqslant h_0$ and (6.5) holds, we obtain, by induction with respect to n, the inequality

$$|d_n| \leqslant \alpha\theta \,.\, h \sum_{j=k}^{n} (1 + \alpha\theta\lambda \,.\, h)^{n-j} |w_j| \quad (k \leqslant n \leqslant N). \tag{6.9}$$

The proof of the lemma is completed by choosing in this inequality

$$n = N.$$

6.3. We next turn to our final lemma. It contains the main result of the present section. The proof of the lemma is obtained by a straightforward combination of the lemmata 2 and 3.

Lemma 4

Let λ_0 be so small that $\lambda_0 h_0 < 1$. Assume the coefficients α_i occurring in (1.10) are such that $\alpha_k = 1$ and the root condition (1.9) is fulfilled.

Let $h \in H$ and let $w_0, w_1, \ldots, w_N$ be arbitrary perturbations in $\mathbb{R}^s$. Assume $u_n, \tilde{u}_n$ satisfy (1.10), (1.11), respectively.

Then we have the error bound

$$|\tilde{u}_N - u_N| \leqslant |v_N| + \alpha\theta\lambda \,.\, h \sum_{n=\max(0,\, k-r)}^{N-q} e^{\alpha\theta\lambda \,.\, t_{N-q-n}} \,.\, |v_n| \tag{6.10}$$

where the constant α is as in Lemma 3. The constants θ, λ and the vectors v_n are defined by (6.6) and (6.4), respectively.

Proof

By Lemma 3 condition (6.1) is fulfilled with $P(h)$, $Q(h)$ given by (6.7).

Applying Lemma 2 we obtain the error bound (6.2) with

$$R_n(h) = \alpha\theta \cdot h \sum_{i=\max(q,\,k-n)}^{\min(r,N-n)} \lambda_i (1 + \alpha\theta\lambda \cdot h)^{N-n-i}.$$

Since $(1 + \alpha\theta\lambda \cdot h)^{N-n-i} \leqslant e^{\alpha\theta\lambda \cdot t_{N-n-q}}$ we have

$$R_n(h) \leqslant \alpha\theta\lambda \cdot \left(\sum_{i=\max(q,k-n)}^{\min(r,N-n)} \lambda_i \right) \cdot e^{\alpha\theta\lambda \cdot t_{N-n-q}}.$$

In view of the convention (2.7) the error bound (6.2) thus proves (6.10).

6.4. We note that Lemma 2 also has applications in situations different from those in Lemma 4. Particularly interesting are applications where (6.1) can be shown to hold with a function $Q(h)$ that is smaller than $(1 + \alpha\Theta\lambda \cdot h)$—as is the case in certain finite-difference schemes for solving stiff differential equations (cf. e.g. [3]). In such applications one obtains bounds similar to (6.10) but with factors $< e^{\alpha\theta\lambda \cdot t_{N-q-n}}$.

References

[1] Dahlquist, G. (1959). Stability and error bounds in the numerical integration of ordinary differential equations, *Trans. Roy. Inst. Technol., Stockholm,* Nr. 130.

[2] Dahlquist, G. (1976). Error analysis for a class of methods for stiff nonlinear initial value problems. In: *Numerical Analysis, Proceedings of the Dundee Conference on Numerical Analysis, 1975. Lecture Notes in Mathematics 506,* pp. 60–72. Springer-Verlag, Berlin.

[3] Desoer, C. A. and Haneda, H. (1972). The measure of a matrix as a tool to analyze computer algorithms for circuit analysis, *IEEE Trans. Circuit Theory* **19**, 480–486.

[4] Henrici, P. (1962). *Discrete Variable Methods in Ordinary Differential Equations.* J. Wiley & Sons, New York.

[5] Rudin, W. (1973). *Functional Analysis.* McGraw-Hill Book Company, New York.

[6] Spijker, M. N. (1971). On the structure of error estimates for finite-difference methods, *Numer. Math.* **18**, 73–100.

[7] Spijker, M. N. (1976). On the possibility of two-sided error bounds in the numerical solution of initial value problems, *Numer. Math.* **26**, 271–300.

[8] Stetter, H. J. (1973). *Analysis of Discretization Methods for Ordinary Differential Equations.* Springer-Verlag, Berlin.

[9] Stummel, F. (1973). *Approximation Methods in Analysis.* Lecture Notes Series of Aarhus Universitet.