Lecture Notes in Mathematics

Edited by A. Dold, Heidelberg and B. Eckmann, Zürich

395

Numerische Behandlung nichtlinearer Integrodifferential- und Differentialgleichungen

Vorträge einer Tagung im Mathematischen Forschungsinstitut Oberwolfach, 2. 12. – 7. 12. 1973

Springer-Verlag
Berlin · Heidelberg · New York

TWO-SIDED ERROR ESTIMATES IN THE NUMERICAL SOLUTION OF INITIAL VALUE PROBLEMS

by

M.N. Spijker

1. INTRODUCTION

Let u denote the solution of a finite-difference equation approximating a given differential equation. If the finite-difference equation is perturbed by a quantity w, e.g. due to round-off error, then instead of u we obtain a solution, say $\tilde{u}$. It is an interesting problem to find an upperbound for $||\tilde{u}-u||$ in terms of w (with $||-||$ we denote some seminorm). Such an error estimate, which is of the form say $||\tilde{u}-u|| \leq R(w)$, is of interest only if it doesn't overestimate $||\tilde{u}-u||$ to much. In fact, we have to require that also $||\tilde{u}-u|| \geq \beta \cdot R(w)$ for some constant β which is essentially greater than zero - and if possible is approximately equal to one. Thus we arrive at the task to derive estimates of the form

$$L(w) \leq ||\tilde{u}-u|| \leq R(w)$$

where the left-hand member $L(w)$ and the right-hand member $R(w)$ are identical, except for a finite factor $\beta > 0$. An estimate of this form will be called a *two-sided error estimate.*

In this paper there will be presented a condition on the seminorm $||-||$ which is necessary and sufficient for the existence of such a two-sided error estimate (see chapter 3). Further we shall show (in chapter 4) that in case a two-sided error estimate exists, it is possible to determine the order of magnitude of both $\sup_w ||\tilde{u}-u||$ and $\inf_w ||\tilde{u}-u||$ where w ranges over all perturbations whose seminorm $||w||$ equals an arbitrary given value $\varepsilon > 0$.

We shall restrict our considerations to finite-difference methods for approximating solutions of initial value problems for ordinary and partial semi-linear differential equations. The theorems of this article will be illustrated by an application to the numerical solution of a nonlinear parabolic initial-boundary

value problem (see chapter 5). For further applications and generalizations the reader is referred to the publications [2] - [6] listed at the end of this paper.

2. NOTATIONS AND ASSUMPTIONS

2.1. The finite-difference scheme we shall deal with, is of the form

$$(1) \qquad u_n = K(u_{n-1}) + h\cdot F(u_{n-1}) \ (n=1,2,\ldots,N), \qquad u_o = c \ ,$$

and the perturbed finite-difference scheme is assumed to be of the form

$$(2) \qquad \begin{aligned} \tilde{u}_n &= K(\tilde{u}_{n-1}) + h\cdot[F(\tilde{u}_{n-1}) + w_n] \ (n=1,2,\ldots,N) \ , \\ \tilde{u}_o &= c + w_o \ . \end{aligned}$$

In (1), (2) we denote by h the so-called stepsize, which belongs to some set H. It is assumed that H is contained in some interval $(0,T]$ and that $\inf H = 0$. Further u_n, $\tilde{u}_n$, w_n and c are vectors belonging to the normed real vectorspace $\mathfrak{R}_h$. The norm in $\mathfrak{R}_h$ is denoted by $|a| = |a|_h$ for $a \in \mathfrak{R}_h$. For $h \in H$ we use the notation $t_n = nh$ and N is the unique integer with $t_N \le T < t_{N+1}$. Clearly $N \to \infty$ if $h \to 0$. u_n is an approximation to the true solution of some given differential equation at $t = t_n$ and c is obtained from the initial condition of the original infinitesimal problem.

An example of (1) is provided by Euler's method for solving an initial value problem for an autonomous system of ordinary differential equations which, by using vector notation, can be written in the form $\frac{d}{dt} U(t) = f(U(t)) \ \ (0 \le t \le T)$, $U(o) = c$. In this example we have $F = f$ and K equals the identity. Further $\mathfrak{R}_h$ is a finite-dimensional vectorspace the dimension of which is equal to the dimension of the given system of differential equations. In this case H can be chosen equal to $H = (0,T]$ and u_n is an approximation of $U(t)$ at $t = t_n$. In this example the perturbation w_o in (2) may represent the error occuring if the initial value c cannot be represented exactly in the computer and the error w_n $(n \ge 1)$ in (2) may be due to the fact that, as a consequence of the possibly complicated structure of the function f, the value $F(\tilde{u}_{n-1}) = f(\tilde{u}_{n-1})$ is not computed exactly. For another example of (1), (2), in which K is different from the

identity operator, we refer to chapter 5.

Returning to the general scheme (1) we now shall list what conditions K and F are supposed to fulfill in the rest of this paper.

a. Assumptions on K :

1. For each $h \in H$ we denote by $K = K_h$ a linear mapping from $\mathfrak{R}_h$ into $\mathfrak{R}_h$,
2. There is a constant $\alpha > 0$ such that the norm of the n-th power of K_h satisfies $|(K_h)^n| \leq \alpha$ for all $h \in H$ and integers n with $0 \leq nh \leq T$,
3. For each $h \in H$ there is an eigenvector e_h in $\mathfrak{R}_h$ with $|e_h| = 1$, $K(e_h) = \kappa e_h$ such that the corresponding eigenvalue $\kappa = \kappa_h$ satisfies the inequality $\kappa_h \geq 1 - \mu_1 h \geq \mu_o$ for some fixed positive constants μ_o and μ_1 .

b. Assumption on F :

For each $h \in H$ we denote by $F = F_h$ a mapping from $\mathfrak{R}_h$ into $\mathfrak{R}_h$ satisfying a Lipschitz condition

$$|F(\tilde{a}) - F(a)| \leq \lambda \cdot |\tilde{a}-a|$$

where λ is an arbitrary positive constant independent of $h \in H$ and $\tilde{a}, a \in \mathfrak{R}_h$.

2.2. In order to investigate estimates for the errors $\tilde{u}_n - u_n$, resulting from the local perturbations w_n occuring in (2), it is appropriate to introduce the vectors

$$u = (u_o, u_1, \ldots, u_N) \ , \quad \tilde{u} = (\tilde{u}_o, \tilde{u}_1, \ldots, \tilde{u}_N) \ , \quad w = (w_o, w_1, \ldots, w_N) \ .$$

These vectors belong to the vectorspace X_h given by

$$X_h = \{x | x = (x_o, x_1, \ldots, x_N) \text{ and each } x_n \in \mathfrak{R}_h\} \ ,$$

in which addition and multiplication with real numbers are defined coordinate-wise. Since we want to measure the difference between $\tilde{u}$ and u by means of a seminorm, we assume some seminorm $||x||_h$ to be given in X_h . It is assumed that this seminorm is *absolute*, i.e. it is required that $||x||_h = ||y||_h$ for any pair of vectors $x = (x_o, x_1, \ldots, x_N)$, $y = (y_o, y_1, \ldots, y_N)$ the coordinates of which satisfy $|x_n| = |y_n|$ $(n = 0, 1, \ldots, N)$.

Example 1. $||x||_h = |x_N|$,

Example 2. $||x||_h = \left(h \sum_{o}^{N} |x_n|^2 \right)^{\frac{1}{2}}$,

Example 3. $||x||_h = max \{ |x_n| : 0 \leq n \leq N \}$.

The following definition formalizes the concept of a two-sided error estimate discussed in the introduction.

DEFINITION. Let γ_o and γ_1 be positive constants and assume ϕ_h is a mapping from X_h into the set of real numbers, R. If for all $h \in H$ and all $w_n \in \mathring{R}_h$ the relations (1), (2) imply that

$$\gamma_o \cdot \phi_h[w] \le ||\tilde{u}-u||_h \le \gamma_1 \cdot \phi_h[w] , \tag{3}$$

then (3) is called a *two-sided error estimate* for the finite-difference scheme (1).

Note that (3) trivially holds with $\gamma_o = \gamma_1 = 1$, $\phi_h[w] \equiv ||\tilde{u}-u||_h$. It is clear that with this choice of the functional ϕ_h the error estimate is useless since the values $\phi_h[w]$ depend on w in a way which, due to the possibly complicated structure of the (nonlinear) F, is untransparent. It is in view of the existence of such trivial and simultaneously useless error estimates that we shall look for estimates of type (3) with a functional ϕ_h that is *independent of* F.

3. THE EXISTENCE OF TWO-SIDED ERROR ESTIMATES

In order to formulate our criterion for the existence of two-sided error estimates in a concise way, we introduce the *summation operator* S from X_h into X_h. For $x = (x_o, x_1, \ldots, x_N) \in X_h$ we define $y = Sx$ by $y = (y_o, y_1, \ldots, y_N)$ with $y_o = 0$, $y_n = h \cdot (x_o + x_1 + \ldots + x_{n-1})$ (for $1 \le n \le N$).

THEOREM 1 (*The existence of two-sided error estimates*). The following propositions (i) and (ii) are equivalent:

(i) There exists a two-sided error estimate for (1) with a functional ϕ_h independent of F,

(ii) There is a constant $\delta < \infty$ such that for all $h \in H$, $x \in X_h$ we have the inequality $||Sx||_h \le \delta \cdot ||x||_h$.

It should be understood that in statement (i) we deal with a *fixed* family $\{K_h\}_{h \in H}$ and a *variable* family $\{F_h\}_{h \in H}$, both families satisfying the as-

sumptions stated in chapter 2. Thus statement (i) could have been formulated alternatively as follows: (i*) There exists a family of functionals $\{\phi_h\}_{h\in H}$ such that for each family $\{F_h\}_{h\in H}$ satisfying a Lipschitz condition of the type described in chapter 2, there exist positive constants γ_o and γ_1 such that (1), (2) always imply (3) (γ_o and γ_1 may depend on $\{F_h\}_{h\in H}$ but not on h and w).

It is easily verified that with the seminorm $||x||_h$ of example 1 (chapter 2) condition (ii) of the above theorem is violated. On the other hand the seminorms of the examples 2, 3 satisfy condition (ii) with $\delta = T$.

THEOREM 2 (*The form of two-sided error estimates*). Let condition (ii) of theorem 1 be fulfilled. Then the two-sided error estimate (3) holds with

$$\gamma_o = (1+\alpha\delta\lambda)^{-1}\ ,\quad \gamma_1 = 1 + \alpha\delta\lambda\cdot exp(\alpha\lambda T)\ ,$$

$$\phi_h[w] = ||v||_h \quad \text{where} \quad v = (v_o, v_1, \ldots, v_N) \quad \text{and}$$

$$v_n = K^n w_o + h\cdot(w_n + Kw_{n-1} + \ldots + K^{n-1}w_1)\ .$$

For the proof of the theorems 1,2 we refer to chapter 3 of [5]. Note that already in the simple case where (1) stands for Euler's method for solving an ordinary differential equation, the theorems 1,2 are nontrivial.

4. APPLICATIONS OF TWO-SIDED ERROR ESTIMATES

4.1. Two sided error estimates have applications in several different fields such as the propagation of rounding errors, the propagation of local discretization errors and in the proof of so-called equivalence theorems stating necessary and sufficient conditions for convergence (cf. [2]-[6]).

In this chapter we shall confine ourselves to showing that two-sided error estimates may be used with succes in determining the orders of magnitude of the interesting functions $g_o(\varepsilon,h)$, $g_1(\varepsilon,h)$ defined by

$$g_o(\varepsilon,h) = \inf_{||w||_h=\varepsilon} ||\tilde{u}-u||_h\ ,\qquad g_1(\varepsilon,h) = \sup_{||w||_h=\varepsilon} ||\tilde{u}-u||_h$$

where $\varepsilon>0$, $h\in H$. Note that, in view of (2), $\tilde{u}$ is uniquely determined by w, and the seminorm $||\tilde{u}-u||_h$ appearing in the above definitions thus indeed is a

function of the variable vector $w \in X_h$.

For the sake of simplicity we only deal with the seminorm of example 2 (section 2), i.e.

$$(4) \qquad ||x||_h = \{h \sum_0^N |x_n|^2\}^{\frac{1}{2}}$$

and we only consider perturbations $w = (w_0, w_1, \ldots, w_N)$ the first component of which vanishes.

$$(5) \qquad w_0 = 0 .$$

In the subsequent sections we use the above definitions of g_0, g_1 with the restrictions expressed in (4), (5). Further we consider a fixed family $\{K_h\}_{h\in H}$ and a fixed family $\{F_h\}_{h\in H}$, both families satisfying the conditions stated in section 2.

4.2. We first give a lower bound for g_0 and an upperbound for g_1 :

$$(6) \qquad (1+\alpha+\lambda h)^{-1}\cdot h\cdot\varepsilon \le g_0(\varepsilon,h) ,$$

$$(7) \qquad g_1(\varepsilon,h) \le (2\lambda)^{-1}[exp(2\alpha\lambda T)-1-2\alpha\lambda T\cdot exp(-\alpha\lambda h)]^{\frac{1}{2}}\cdot\varepsilon .$$

Proof of (6). Using (4) with $x=w$ and substituting for w_n the expression which can be obtained by subtracting (1) from (2) we have, in case $w_0=0$,

$$||w||_h = \{h \sum_1^N h^{-2}|(\tilde{u}_n-u_n)-K(\tilde{u}_{n-1}-u_{n-1})-h[F(\tilde{u}_{n-1})-F(u_{n-1})]|^2\}^{\frac{1}{2}} .$$

In view of the assumptions on K and F we thus obtain

$$\begin{aligned} h||w||_h &\le \{h \sum_1^N [|\tilde{u}_n-u_n|+(\alpha+\lambda h)|\tilde{u}_{n-1}-u_{n-1}|]^2\}^{\frac{1}{2}} \\ &\le \{h \sum_1^N |\tilde{u}_n-u_n|^2\}^{\frac{1}{2}} + (\alpha+\lambda h)\cdot\{h \sum_1^N |\tilde{u}_{n-1}-u_{n-1}|^2\}^{\frac{1}{2}} \\ &\le (1+\alpha+\lambda h)\ ||\tilde{u}-u||_h , \end{aligned}$$

which proves (6).

Proof of (7). The upperbound (7) can be proved in a standard fashion by subtracting (1) from (2), which yields for the difference

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

the recurrence relation

$$d_n = K\, d_{n-1} + h[F(\tilde{u}_{n-1}) - F(u_{n-1})] + h\, w_n \ .$$

Solving for d_n we obtain (provided $\tilde{u}_o - u_o = w_o = 0$) the representation

$$d_n = z_n + v_n$$

where v_n is defined in theorem 2 and $z_n = h \sum_{i=1}^{n-1} K^{n-1-i}[F(\tilde{u}_i) - F(u_i)]$

Using the assumptions on K and F stated in chapter 2 we get

$$|z_n| \le \alpha\lambda h \sum_{i=1}^{n-1} |d_i| \ , \qquad |v_n| \le \alpha h \sum_{i=1}^{n} |w_i| \ .$$

Consequently

$$|d_n| \le \alpha\lambda h \sum_{i=1}^{n-1} |d_i| + \alpha h \sum_{i=1}^{n} |w_i| \ ,$$

from which it follows by induction with respect to n that

$$|d_n| \le \alpha h \sum_{i=1}^{n} (1+\alpha\lambda h)^{n-i} |w_i| \quad (n=1,2,\ldots,N) \ .$$

Using the Schwarz inequality we thus obtain

$$\begin{aligned}(||d||_h)^2 &\le h \sum_{n=1}^{N} \alpha^2 [h \sum_{i=1}^{n} (1+\alpha\lambda h)^{n-i} |w_i|]^2 \\ &\le \alpha^2 h \sum_{n=1}^{N} [h \sum_{i=1}^{n} (1+\alpha\lambda h)^{2(n-i)}] \cdot [h \sum_{i=1}^{n} |w_i|^2] \\ &\le (\alpha ||w||_h)^2 \cdot h^2 \sum_{n=1}^{N} \sum_{i=o}^{n-1} (1+\alpha\lambda h)^{2i} \ .\end{aligned}$$

Since $\sum_{n=1}^{N} \sum_{i=o}^{n-1} (1+\alpha\lambda h)^{2i} \le \sum_{n=1}^{N} \sum_{i=o}^{n-1} e^{2\alpha\lambda h i}$

$$= \sum_{n=1}^{N} (e^{2\alpha\lambda h n} - 1)/(e^{2\alpha\lambda h} - 1)$$

$$= (e^{2\alpha\lambda h} - 1)^{-2} \cdot [e^{2\alpha\lambda h}(e^{2\alpha\lambda h N} - 1) - (e^{2\alpha\lambda h} - 1)N]$$

and $e^{2t} - 1 \ge 2t\, e^{t}$ (for $t = \alpha\lambda h$) we get

$$(||d||_h)^2 \le (||w||_h)^2 \cdot (2\lambda)^{-2} \cdot [e^{2\alpha\lambda N h} - 1 - 2\alpha\lambda N h e^{-\alpha\lambda h}] \ .$$

In view of $Nh = t_N \le T$ and the definition of d_n we thus have

$$||\tilde{u} - u||_h \le ||w||_h \cdot (2\lambda)^{-1} \cdot [e^{2\alpha\lambda T} - 1 - 2\alpha\lambda T \cdot e^{-\alpha\lambda h}]^{\frac{1}{2}} \ ,$$

which completes the proof of (7).

4.3. Next we present a theorem showing that the lower and upperbound (6), (7) are realistic, as $h \to 0$, at least in the sense that they contain the correct powers of h.

THEOREM 3. (a) For each $\varepsilon > 0$ and $h \in H$ there exists a vector $w = (w_0, w_1, \ldots, w_N)$ with $w_0 = 0$, $||w||_h = \varepsilon$ such that $\tilde{u}_n$ and u_n, computed from (2), (1), respectively satisfy

$$||\tilde{u}-u||_h \leq \frac{1+\alpha}{1+\mu_0} \cdot [1+\alpha\lambda T exp(\alpha\lambda T)] \cdot h \cdot ||w||_h .$$

(b) For each $\varepsilon > 0$ and $h \in H$ there also exists a vector $w = (w_0, w_1, \ldots, w_N)$ with $w_0 = 0$, $||w||_h = \varepsilon$ such that $\tilde{u}_n$ and u_n computed from (2), (1) satisfy

$$||\tilde{u}-u||_h \geq \frac{1}{\mu_1(1+\alpha\lambda T)} \cdot \{\frac{1}{T} \int_0^T [1-exp(-\mu_1 t)]^2 dt\}^{\frac{1}{2}} \cdot ||w||_h .$$

It follows from part (a) of this theorem that

$$(6^*) \qquad g_0(\varepsilon,h) \leq \frac{1+\alpha}{1+\mu_0} \cdot [1+\alpha\lambda T exp(\alpha\lambda T)] \cdot h \cdot \varepsilon$$

and from part (b) there follows

$$(7^*) \qquad g_1(\varepsilon,h) \geq \frac{1}{\mu_1(1+\alpha\lambda T)} \cdot \{\frac{1}{T} \int_0^T [1-exp(-\mu_1 t)]^2 dt\}^{\frac{1}{2}} \cdot \varepsilon .$$

Formula (7^*) shows that (7) cannot essentially be improved since apparently it is not possible to replace (7) by an estimate of the form $g_1(\varepsilon,h) \leq \gamma \cdot h^p \cdot \varepsilon$ with any $p > 0$. Consequently, the order of magnitude of $g_1(\varepsilon,h)$ is exactly equal to ε. Similarly, (6), (6^*) prove that the order of magnitude of $g_0(\varepsilon,h)$ is exactly equal to $h \cdot \varepsilon$. In fact, according to a definition of a two-sided error estimate differing only slightly from the one given in chapter 2, one might say that (6), (6^*) and (7), (7^*) provide two-sided error estimates of $g_0(\varepsilon,h)$ and $g_1(\varepsilon,h)$, respectively.

In the above proofs of (6), (7) we have not made use of the two-sided error estimate for $||\tilde{u}-u||_h$ which exists in view of theorem 1. On the other hand, the following proof of theorem 3 is entirely based on this two-sided error estimate.

Proof of theorem 3. (a) Choosing $w_0 = 0$ and $w_n = [(-1)^n \cdot T^{-\frac{1}{2}} \varepsilon] \cdot e_h$ (for $n=1,2,\ldots,N$) where e_h is the eigenvector whose existence was postulated in chapter 2, we have for the vectors v_n defined in theorem 2

$$v_n = (-1)^n T^{-\frac{1}{2}} \varepsilon \cdot h \cdot [1-\kappa+\kappa^2-\ldots+(-1)^{n-1}\kappa^{n-1}] \cdot e_h .$$

Consequently

$$|v_n| = T^{-\frac{1}{2}}\,\varepsilon h\cdot[1-(-\kappa)^n]/[1+\kappa] .$$

Since $(\kappa)^n \le |(K_h)^n| \le \alpha$ and $\kappa \ge \mu_o$ (cf. chapter 2) we thus obtain

$$|v_n| \le T^{-\frac{1}{2}}\,\varepsilon h\cdot(1+\alpha)/(1+\mu_o) .$$

Hence $||v||_h \le \varepsilon h\cdot(1+\alpha)/(1+\mu_o)$. By virtue of theorem 2 we have $||\tilde{u}-u||_h \le [1+\alpha\lambda T\cdot exp(\alpha\lambda T)] \cdot ||v||_h$, from which it follows that

$$||\tilde{u}-u||_h \le [1+\alpha\lambda T\cdot exp(\alpha\lambda T)]\cdot(1+\alpha)(1+\mu_o)^{-1}\cdot h\varepsilon .$$

Since $||w||_h = \varepsilon$ part (a) of the theorem has thus been proved.

(b) In order to prove part (b) we choose the perturbations $w_o = 0$, $w_n = T^{-\frac{1}{2}}\,\varepsilon\cdot e_h$ $(n=1,2,\ldots,N)$ the seminorm $||w||_h$ of which again equals ε . We have

$$\begin{aligned} |v_n| &= T^{-\frac{1}{2}}\,\varepsilon h\cdot(1+\kappa+\ldots+\kappa^{n-1}) \\ &\ge T^{-\frac{1}{2}}\,\varepsilon h\cdot[1+(1-\mu_1 h)+\ldots+(1-\mu_1 h)^{n-1}] \\ &= T^{-\frac{1}{2}}\,\varepsilon(\mu_1)^{-1}\cdot[1-(1-\mu_1 h)^n] \\ &\ge T^{-\frac{1}{2}}\,\varepsilon(\mu_1)^{-1}\cdot[1-exp(-\mu_1 nh)] . \end{aligned}$$

Consequently $(||v||_h)^2 \ge (\frac{\varepsilon}{\mu_1})^2\cdot T^{-1}\cdot h \sum_{n=1}^{N} [1-exp(-\mu_1 nh)]^2$ $\ge (\frac{\varepsilon}{\mu_1})^2\cdot\frac{1}{T}\cdot\int_o^T [1-exp(-\mu_1 t)]^2 dt$. In view of theorem 2 we have $||\tilde{u}-u||_h \ge (1+\alpha\lambda T)^{-1}\cdot||v||_h$ from which it thus follows that part (b) of theorem 3 holds. This completes the proof of the theorem.

5. A PARABOLIC INITIAL-BOUNDARY VALUE PROBLEM

5.1. In order to illustrate the theory of the preceding chapters we consider the numerical solution by a very simple explicit finite-difference scheme of the initial-boundary value problem

$$\begin{aligned} &\frac{\partial}{\partial t}U(s,t) = \frac{\partial^2}{\partial s^2}U(s,t) + f(s,U(s,t)) , \\ &U(0,t) = U(1,t) = 0 , \quad U(s,0) = c(s), \\ &0 \le s \le 1, \quad 0 \le t \le T . \end{aligned} \tag{8}$$

In (8) f and c denote given functions and f is assumed to satisfy the Lipschitz condition

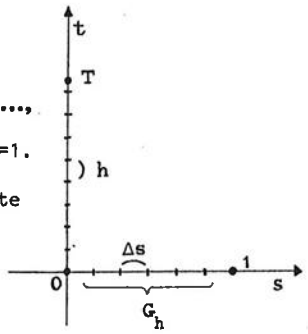
(9) $$|f(s,\tilde{r}) - f(s,r)| \le \lambda\cdot|\tilde{r}-r| ,$$

where λ is an arbitrary positive constant, uniformly for $0 \le s \le 1$ and $-\infty < r, \tilde{r} < \infty$. The finite-difference scheme we deal with, is of the form

(10) $$\frac{1}{h}\cdot[u_n(s)-u_{n-1}(s)] = \frac{1}{(\Delta s)^2}\cdot[u_{n-1}(s-\Delta s)-2u_{n-1}(s)+u_{n-1}(s+\Delta s)] + f(s,u_{n-1}(s)) ,$$

$$u_o(s) = c(s).$$

In (10) the variable s takes on the values $s = \Delta s, 2\Delta s, \ldots, M\Delta s$ where M is an arbitrary integer ≥ 2 with $(M+1)\Delta s=1$. Applying (10) successively with $n=1,2,\ldots,N$ we can compute approximations $u_n(s)$ of $U(s,t)$ at the points $(s,t) = (m\Delta s, nh)$ with $m=1,2,\ldots,M$ and $n=1,2,\ldots,N$.



5.2. We give some definitions which will enable us to write the finite-difference scheme (10) in the general form (1). We assume throughout that the ratio $h/(\Delta s)^2$ equals some constant σ with

(11) $$0 < \sigma \le 1/2 .$$

The set H of stepsizes h is defined by

$$H = \{h \mid h=\sigma\cdot(\Delta s)^2 , \quad \Delta s = (M+1)^{-1} \text{ with } M=2,3,4,\ldots\}$$

and for $h \in H$ the grid G_h on the s-axis is defined by

$$G_h = \{s \mid s=m\cdot\Delta s \text{ with } m=1,2,\ldots,M\} .$$

The vectorspace $\mathcal{R}_h$ consists of all real functions defined on the grid G_h and the norm $|a|_h$ in $\mathcal{R}_h$ is defined by

(12) $$|a|_h = \{\Delta s \sum_{m=1}^{M} [a(m\cdot\Delta s)]^2\}^{\frac{1}{2}}$$

for $a \in \mathcal{R}_h$. The operators K and F mapping $\mathcal{R}_h$ into itself are defined by

$$(Ka)(s) = \begin{cases} (1-2\sigma)\cdot a(s) + \sigma\cdot a(s+\Delta s) & (\text{for } s=\Delta s) , \\ \sigma\cdot a(s-\Delta s) + (1-2\sigma)\cdot a(s) + \sigma\cdot a(s+\Delta s) & \\ \qquad\qquad (\text{for } s=m\Delta s,\ 1<m<M) , & \\ \sigma\cdot a(s-\Delta s) + (1-2\sigma)\cdot a(s) & (\text{for } s=M\Delta s) , \end{cases}$$

$(Fa)(s) = f(s,a(s))$ (for $s=m\Delta s$, $1\le m\le M$)

where a denotes an arbitrary element belonging to $\mathcal{R}_h$. With these definitions (1) is equivalent tot (10) and (2) is equivalent to

$$\frac{1}{h}[\tilde{u}_n(s)-\tilde{u}_{n-1}(s)] = \frac{1}{(\Delta s)^2}\cdot[\tilde{u}_{n-1}(s-\Delta s)-2\tilde{u}_{n-1}(s)+\tilde{u}_{n-1}(s+\Delta s)]$$

$$+ f(s,\tilde{u}_{n-1}(s)) + w_n(s) \,, \tag{13}$$

$$\tilde{u}_o(s) = c(s) + w_o(s)$$

where $w_n(s)$ thus represents a perturbation occuring when the approximation $\tilde{u}_n(s)$ of $U(s,nh)$ is calculated.

5.3. It is easily verified, by using (9), that the assumption on F postulated in chapter 2 is satisfied here. Further, the first assumption on K stated in chapter 2 is satisfied here as well. The second assumption on K of chapter 2 may be proved by a standard argument (cf. Rjabenki, Filippow [1]) using the inner product

$$(a,b) = \Delta s \sum_{m=1}^{M} a(m\Delta s)\cdot b(m\Delta s)$$

in $\mathcal{R}_h$. It can be verified by a straightforward calculation that the elements $e^{(1)},e^{(2)},\ldots,e^{(M)}$ in $\mathcal{R}_h$ given by

$$e^{(j)}(s) = \sqrt{2}\cdot sin(sj\pi) \quad (\text{for } s \in G_h)$$

are orthonormal with respect to this inner product and that they are eigenvectors of K with eigenvalues

$$\kappa^{(j)} = 1 - 4\sigma\cdot[sin(j\Delta s\cdot\pi/2)]^2 \,.$$

In view of (11) we have $-1 < \kappa^{(j)} < 1$ and consequently $|(K_h)^n| \le |K_h|^n = [\max_j|\kappa^{(j)}|]^n < 1$, which yields the value

$$\alpha = 1 \,. \tag{14}$$

The third assumption on K of chapter 2 is fulfilled here with $e_h = e^{(1)}$, $\kappa_h = \kappa^{(1)}$ and

$$\mu_o = 1 - (\pi/3)^2\cdot\sigma \,, \qquad \mu_1 = \pi^2 \,. \tag{15}$$

5.4. All general assumptions on K and F of chapter 2 being fulfilled

we may apply the results of chapter 4 to the finite-difference scheme (10) and the perturbed scheme (13). The seminorm $||x||_h$ (see (4)) used throughout chapter 4 yields in combination with the norm $|a|_h$ (see (12)) the formula $||x||_h = \{h\Delta s \cdot \sum_{n=0}^{N} \sum_{m=1}^{M} [x_n(m\Delta s)]^2\}^{\frac{1}{2}}$ and if $x_o = 0$ we thus have

$$||x||_h = \{h\Delta s \cdot \sum_{n=1}^{N} \sum_{m=1}^{M} [x_n(m\Delta s)]^2\}^{\frac{1}{2}} . \tag{16}$$

We combine the results (6), (6*), (7), (7*) of chapter 4 when applied to the situation at hand (see (14), (15)) into the following theorem.

THEOREM 4. Let u be computed from (10) and consider $\tilde{u}$ computed from (13) with arbitrary perturbation-vector w for which $w_o(s) \equiv 0$. Then, using the notation (16), we have for any $\varepsilon > 0$, $h \in H$ the inequalities

$$\inf_{||w||_h=\varepsilon} ||\tilde{u}-u||_h \geq \frac{h\varepsilon}{2+\lambda h} , \tag{17}$$

$$\inf_{||w||_h=\varepsilon} ||\tilde{u}-u||_h \leq \frac{18}{18-\pi^2\sigma} \cdot [1+\lambda T \cdot exp(\lambda T)] \cdot h\varepsilon , \tag{17*}$$

$$\sup_{||w||_h=\varepsilon} ||\tilde{u}-u||_h \leq \frac{1}{2\lambda}[exp(2\lambda T)-1-2\lambda T \cdot exp(-\lambda h)]^{\frac{1}{2}} \cdot \varepsilon , \tag{18}$$

$$\sup_{||w||_h=\varepsilon} ||\tilde{u}-u||_h \geq \frac{1}{\pi^2(1+\lambda T)} \{\frac{1}{T} \cdot \int_o^T [1-exp(-\pi^2 t)]^2 dt\}^{\frac{1}{2}} \cdot \varepsilon . \tag{18*}$$

5.5. (17*) and (18*) show that the bounds (17), (18) cannot essentially be improved (e.g. by including a factor h^p with $p \neq 0$) and that the precise orders of magnitude of $\inf||\tilde{u}-u||_h$ and $\sup||\tilde{u}-u||_h$ are $h\varepsilon$ and ε, respectively. In order also to test how realistic the error bounds (17), (18) are from a numerical point of view[1] a number of experiments have been performed with the initial-boundary value problems

$$\frac{\partial}{\partial t} U(s,t) = \frac{\partial^2}{\partial s^2} U(s,t) + \frac{1}{1+[U(s,t)]^2} + \frac{1+2s^2(1-s)^2}{1+s^2(1-s)^2} ,$$
$$U(0,t) = U(1,t) = 0 , \quad U(s,0) = s(1-s) , \tag{A}$$
$$0 \leq s \leq 1 , \quad 0 \leq t \leq 1$$

and

---

[1] Here we follow a suggestion made by L. Collatz at the Conference in Oberwolfach.

$$\frac{\partial}{\partial t} U(s,t) = \frac{\partial^2}{\partial s^2} U(s,t) + \frac{U(s,t)}{100} ,$$

(B)

$$U(0,t) = U(1,t) = 0 , \quad U(s,0) = \sin(\pi s) ,$$

$$0 \le s \le 1 , \quad 0 \le t \le 1 .$$

In the following tables we compare the numerical values obtained for $||\tilde{u}-u||_h$ with the theoretical bounds (17), (18). In (13) we have used the perturbations $w_n(s)$ defined by $w_o(s) = 0$ and $w_n(s) = \frac{1}{10}(-1)^n(1-\Delta s)^{-\frac{1}{2}}$ (for problem (A)), $w_n(s) = \frac{1}{10}(1-\Delta s)^{-\frac{1}{2}}$ (for problem (B)) .

$N=h^{-1}$	Theoretical lower bound for $\|\|\tilde{u}-u\|\|_h$ (cf.(17)), $h\varepsilon/(2+h)$	$\|\|\tilde{u}-u\|\|_h$	Theoretical upper bound for $\|\|\tilde{u}-u\|\|_h$ (cf.(18)) $\frac{\varepsilon}{2}\cdot[\exp(2)-1-2\cdot\exp(-h)]^{\frac{1}{2}}$
48	0.00103	0.00124	0.10524
192	0.00026	0.00028	0.10487
768	0.00007	0.00007	0.10478
3072	0.00002	0.00002	0.10476
Problem (A), $\lambda=1$, $T=1$, $\sigma=1/3$, $\|\|w\|\|_h=\varepsilon=0.1$, $w_n(m\Delta s)=(-1)^n(1-\Delta s)^{-\frac{1}{2}}\varepsilon$			

$N=h^{-1}$	Theoretical lower bound for $\|\|\tilde{u}-u\|\|_h$ (cf.(17)), $h\varepsilon/(2+\lambda h)$	$\|\|\tilde{u}-u\|\|_h$	Theoretical upper bound for $\|\|\tilde{u}-u\|\|_h$ (cf.(18)), $\frac{\varepsilon}{2\lambda}\cdot[\exp(2\lambda)-1-2\lambda\cdot\exp(-\lambda h)]^{\frac{1}{2}}$
32	0.00156	0.00988	0.07204
128	0.00039	0.00904	0.07122
512	0.00010	0.00870	0.07102
2048	0.00002	0.00855	0.07096
Problem (B), $\lambda=0.01$, $T=1$, $\sigma=\frac{1}{2}$, $\|\|w\|\|_h=\varepsilon=0.1$, $w_n(m\Delta s)=(1-\Delta s)^{-\frac{1}{2}}\varepsilon$			

The numerical results show that there are perturbations $w_n(s)$ for which the bounds (17), (18) don't underestimate or overestimate the actual error $||\tilde{u}-u||_h$ very much. In particular the lower bound (17) fits in surprisingly well with the numerical results in the first table, referring to problem (A). Thus (17), (18), when applied to the problems (A), (B), respectively appear to be realistic bounds both from a theoretical and from a more practical point of view.

Acknowledgement. I wish to thank F. Bakker and C. den Heijer who have been most helpful in performing on the IBM 370/158 of Leiden University the computations reported in section 5.5.

REFERENCES

[1] RJABENKI, V.S., A.F. FILIPPOW, Über die Stabilität von Differenzengleichungen. VEB Deutscher Verlag der Wissenschaften, Berlin (1960).

[2] STETTER, H.J., Analysis of discretization methods for ordinary differential equations. Springer-Verlag, Berlin etc. (1973).

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

[4] SPIJKER, M.N., Optimum error estimates for finite-difference methods, to appear in Acta Universitatis Carolinae (1974).

[5] SPIJKER, M.N., The existence of optimum error estimates in the numerical solution of differential equations, in preparation.

[6] STUMMEL, F., Approximation methods in analysis, Lecture Notes Series of Aarhus Universitet, Matematisk Institut (1973).