Numer. Math. 18, 73–100 (1971)

On the Structure of Error Estimates for Finite-Difference Methods

M. N. Spijker

Received January 1, 1971

Abstract. In this paper we study in an abstract setting the structure of estimates for the global (accumulated) error in semilinear finite-difference methods. We derive error estimates, which are the most refined ones (in a sense specified precisely in this paper) that are possible for the difference methods considered. Applications and (numerical) examples are presented in the following fields: 1. Numerical solution of ordinary as well as partial differential equations with prescribed initial or boundary values. 2. Accumulation of local round-off error as well as of local discretization error. 3. The problem of fixing which methods out of a given class of finite-difference methods are "most stable". 4. The construction of finite-difference methods which are convergent but not consistent with respect to a given differential equation.

1. Introduction

In this paper we investigate the stability of finite-difference methods for the solution of differential (and other functional) equations. In most current definitions of the concept of stability for such difference methods it is required that the error resulting from a local perturbation in the finite-difference equation admits a bound of a prescribed structure in terms of the perturbation (cf. e.g. [2, 3, 6, 13]). Depending on the structure prescribed we thus have different concepts of stability. In this paper we study the following general problem:

What is the structure of the most refined error bound that is valid for the error in a given difference equation?

This problem thus amounts to determining the most stringent definition of stability with regard to which the difference equation is still stable.

In order to illustrate this general problem we consider the numerical solution of the initial value problem

$$U'(x) = f(x, U(x)) \quad (a \leqq x \leqq b), \qquad U(a) = c$$

by Euler's method

$$u_0 - c = 0, \qquad h^{-1}(u_n - u_{n-1}) - f(x_{n-1}, u_{n-1}) = 0 \qquad (n = 1, 2, \ldots, N),$$

where u_n is an approximation of $U(x)$ at $x = x_n = a + nh$ and N is the greatest integer with $Nh \leqq b - a$. Let

$$\tilde{u}_0 - c = w_0, \qquad h^{-1}(\tilde{u}_n - \tilde{u}_{n-1}) - f(x_{n-1}, \tilde{u}_{n-1}) = w_n \qquad (n = 1, 2, \ldots, N)$$

where w_n are local perturbations (e.g. caused by round-off in the actual application of Euler's method). It may be proved that the accumulated error $\tilde{u}_n - u_n$ resulting from these local perturbations satisfies each of the following inequalities

$$|\tilde{u}_n - u_n| \leqq \gamma_1 \cdot \max_{0 \leqq i \leqq N} |w_i|, \tag{1.1}$$

$$|\tilde{u}_n - u_n| \leqq \gamma_2 \cdot \left\{ |w_0| + h \sum_{i=1}^{N} |w_i| \right\}, \tag{1.2}$$

$$|\tilde{u}_n - u_n| \leqq \gamma_3 \cdot \max_{1 \leqq j \leqq N} \left\{ |w_0| + \left| h \sum_{i=1}^{j} w_i \right| \right\} \tag{1.3}$$

for some constants $\gamma_1, \gamma_2, \gamma_3$ independent of n, h, w_i (provided f satisfies the continuity conditions of [5] p. 15; cf. [5, 6, 11], respectively). Error bounds of type (1.1), (1.2) appear in the stability definitions of [7, 6], respectively. Using the triangle inequality and the inequality $Nh \leqq b - a$ it is easily verified that (1.3) implies (1.2) and that (1.2) implies (1.1). But it is not possible to derive (1.3) directly from (1.2) or (1.2) from (1.1). Hence (1.3) is essentially a more refined error bound than (1.1) or (1.2). Since it may be proved (cf. Chapter 3) that the structure of (1.3) is even more refined than of any other bound for the error $\tilde{u}_n - u_n$, the general problem formulated above has thus been solved for Euler's method.

In Section 2.1 we introduce in an abstract setting the notion of a *stability functional*, which is the appropriate tool in comparing different concepts of stability. In Section 2.2 this notion is used to derive theorems which enable us to solve the general problem formulated above for the case of semilinear difference equations.

A second abstract notion introduced in Section 2.1 is the concept of *optimal stability*, which allows us, a certain set of difference equations being given, to single out the equations which are "most stable" (in a sense specified precisely in this section). Using the results of Sections 2.2 we present in Section 2.3 a general criterion for optimal stability of semilinear difference equations.

In the rest of this paper we present some (numerical) examples to demonstrate the application of the abstract (and general) considerations of Chapter 2 to initial and boundary value problems for ordinary and partial differential equations and to the accumulation of local discretization error and local round-off error (cf. [5] for definition of these concepts).

In Chapter 3 the results of Section 2.3 are applied to a set of finite-difference equations (including Runge-Kutta as well as general linear multistep methods) for solving initial value problems for systems of ordinary differential equations. Using the criterion of Section 2.3 it is fixed which of these difference equations are optimally stable in the sense of Section 2.1.

In Chapter 4 we consider a class of finite-difference equations for solving a nonlinear two-point boundary value problem. The refined error bounds of Section 2.2 are applied here to prove the second order accuracy of these difference methods. The results of Chapter 4 prove the interesting fact that there exist finite-difference methods for solving boundary value problems which are convergent (even of second order accuracy) but which fail to be consistent (cf. [14] for definition of the concept of consistency).

In Chapter 5 we derive a procedure (based on splitting of difference equations, cf. [12]) for reducing to linear growth the quadratic round-off error accumulation in the numerical solution of a partial hyperbolic problem. The theorem in Chapter 5 exhibiting this linear growth is proved using the theory of Chapter 2.

2. Stability Functionals and Optimal Stability

2.1. Notations and Definitions

Let $h_0>0$ and let H denote a nonempty subset of the interval $(0, h_0]$ with $\inf H=0$. Let Λ^h denote a real vectorspace (for each $h\in H$). Let $\|\ldots\|^h$ denote a seminorm and $\psi^h[\ldots]$ a real functional on Λ^h. Assume that C^h is a (nonlinear) operator mapping Λ^h onto itself.

Definition. *The operator C^h is stable with respect to the functional ψ^h if there are fixed numbers γ, $h_1>0$ such that for all vectors u^h, $\tilde{u}^h$, $w^h\in\Lambda^h$ satisfying*

$$C^h u^h=0, \tag{2.1}$$

$$C^h \tilde{u}^h=w^h \tag{2.2}$$

(with $h\in H$, $h\leqq h_1$) we have

$$\|\tilde{u}^h-u^h\|\leqq\gamma\cdot\psi^h[w^h]. \tag{2.3}$$

Since there is no danger of confusion, we shall frequently suppress the superscripts h in the following discussion.

In the applications formula (2.1) will stand for a finite-difference equation approximating a differential (or other functional) equation with given initial (or boundary) conditions (cf. the next chapters for examples). The solution u to (2.1) will denote an approximation to the solution of the original infinitesimal problem. The vector $\tilde{u}$ will denote the solution of the difference equation obtained in presence of a perturbation w and the estimate (2.3) for the resultant error $\tilde{u}-u$ is a generalization of error estimates like (1.1), (1.2), (1.3).

Let φ and ψ denote arbitrary real functionals on Λ (for each $h\in H$). In the following we use the notation

$$\varphi\prec\psi$$

if there are positive constants β and h_1 such that

$$\varphi[w]\leqq\beta\cdot\psi[w]$$

for all $w\in\Lambda$ and $h\in H$ with $h\leqq h_1$. The functionals φ and ψ are said to be *equivalent* if $\varphi\prec\psi$ and $\psi\prec\varphi$.

A functional ψ on Λ is called a *stability functional* for the operator C if C is stable with respect to ψ. A stability functional φ for C is called a *minimal stability functional* for C if all stability functionals ψ for C satisfy $\varphi\prec\psi$.

Assume C and D are operators mapping Λ onto itself. C is said to be *more stable* than D if C and D have minimal stability functionals φ and ψ, respectively which are not equivalent and satisfy $\varphi\prec\psi$. The operators C and D are said to be *equivalent* if they have minimal stability functionals which are equivalent.

Let K be a set of operators from Λ onto itself. An operator $C \in K$ is said to be *optimally stable in* K if C is more stable than each $D \in K$ to which it is not equivalent.

In the subsequent (Sections 2.2, 2.3) we are concerned with the following questions:

1. Assume C is a given operator mapping Λ onto itself. What is the form of (possible) minimal stability functionals for C?

2. Assume K is a given set of operators. What condition is necessary and sufficient for an operator $C \in K$ to be optimally stable in K?

These two questions will be answered (in Sections 2.2, 2.3, respectively) for the case of semilinear operators C.

2.2. Minimal Stability Functionals for Semilinear Operators

In the following P^h denotes (for each $h \in H$) a fixed linear, bijective operator from Λ^h onto itself.

We define

$$\psi_0[w] = \|P^{-1}w\| \tag{2.4}$$

for $w \in \Lambda = \Lambda^h$, P^{-1} denoting the inverse of $P = P^h$.

Let A, B, C and Q be arbitrary operators from Λ into itself (for each $h \in H$) satisfying the following conditions a), b), c).

a) $C = A + B$, $A = PQ$,

b) Q is linear and bounded (uniformly for $h \in H$), and B satisfies a Lipschitz condition

$$\|B\tilde{v} - Bv\| \leqq \lambda \cdot \|\tilde{v} - v\|$$

(λ being independent of $h \in H$ and $\tilde{v}$, $v \in \Lambda$),

c) A and C are bijective (for h sufficiently small), and stable with respect to the functional $\psi[w] = \|w\|$.

We define the functional $\varphi = \varphi(A)$ by

$$\varphi(A)[w] = \|A^{-1}w\|. \tag{2.5}$$

The following theorem shows that the operator C has a minimal stability functional which is independent of the (nonlinear) term B.

Theorem 1. *$\varphi(A)$ is a minimal stability functional for $C = A + B$.*

Proof. 1. We shall show that $\varphi(A)$ is a stability functional for C.

Let

$$Cu = 0, \quad C\tilde{u} = w.$$

For h sufficiently small A is invertible and we define

$$v = A^{-1}w, \quad z = \tilde{u} - v.$$

Hence

$$\begin{aligned} Cz &= Az + Bz = A\tilde{u} - Av + B\tilde{u} + Bz - B\tilde{u} \\ &= C\tilde{u} - w + B(\tilde{u} - v) - B\tilde{u} = B(\tilde{u} - v) - B\tilde{u}. \end{aligned}$$

In view of the stability of C (cf. c)) the equality

$$Cz = B(\tilde{u} - v) - B\tilde{u}$$

immediately leads to

$$\|z - u\| \leqq \gamma \cdot \|B(\tilde{u} - v) - B\tilde{u}\|,$$

γ being some constant (independent of $h \leqq$ some h_1).

Consequently (cf. b))

$$\|\tilde{u} - u\| \leqq \|\tilde{u} - z\| + \|z - u\| \leqq \|v\| + \gamma \cdot \lambda \cdot \|v\| = (1 + \gamma\lambda) \cdot \|v\|.$$

Hence

$$\|\tilde{u} - u\| \leqq (1 + \gamma\lambda) \cdot \|A^{-1}w\| \quad (\text{for } h \leqq h_1)$$

and C is thus stable with respect to $\varphi(A)$ (cf. (2.5)).

2. We shall show that $\varphi(A)$ is minimal for C.

Let

$$Cu = 0, \quad C\tilde{u} = w.$$

Then $C\tilde{u} - Cu = w$. Since $C = A + B$ and A is linear we get

$$A(\tilde{u} - u) + B\tilde{u} - Bu = w.$$

"Multiplying" both sides of this equality by A^{-1} (which exists for h sufficiently small, cf. c)) we have

$$(\tilde{u} - u) + A^{-1}(B\tilde{u} - Bu) = A^{-1}w.$$

Hence

$$\|A^{-1}w\| \leqq \|\tilde{u} - u\| + \|A^{-1}(B\tilde{u} - Bu)\|.$$

Since A is stable (cf. c)) there is a constant $\alpha > 0$ such that

$$\|A^{-1}v\| \leqq \alpha \cdot \|v\|$$

for all $v \in \Lambda$ and h sufficiently small. Consequently

$$\|A^{-1}(B\tilde{u} - Bu)\| \leqq \alpha \|B\tilde{u} - Bu\| \leqq \alpha\lambda \cdot \|\tilde{u} - u\| \quad (\text{cf. b})).$$

It follows that

$$\|A^{-1}w\| \leqq \|\tilde{u} - u\| + \alpha\lambda \cdot \|\tilde{u} - u\| = (1 + \alpha\lambda) \cdot \|\tilde{u} - u\|$$

(provided $h \leqq$ some constant h_1).

Assume $\psi[w]$ is any stability functional for C. Then

$$\|\tilde{u} - u\| \leqq \gamma \cdot \psi[w]$$

for some constant γ (provided $h \leqq$ some constant h_2).

It follows that

$$\|A^{-1}w\| \leqq (1 + \alpha\lambda) \cdot \gamma \cdot \psi[w] \quad (\text{for } h \leqq h_3 = \min(h_1, h_2)).$$

Hence (cf. (2.5))

$$\varphi(A)[w] \leqq \beta \cdot \psi[w] \quad (\text{for } h \leqq h_3),$$

$\beta = (1+\alpha\lambda)\cdot\gamma$ being independent of h and w. Thus $\varphi(A) \prec \psi$ and the theorem is proved.

The following theorem shows that the functional ψ_0 (defined by (2.4)) is "smaller" than any stability functional for C.

Theorem 2. $\psi_0 \prec \varphi(A)$.

Proof. Since $A = PQ$ we have $P^{-1}A = Q$.

Hence for h sufficiently small $P^{-1} = QA^{-1}$ (cf. c)) and

$$\|P^{-1}w\| \leqq \|Q\| \cdot \|A^{-1}w\|$$

(for any operator D from Λ into Λ we put

$$\|D\| = \sup\{\|Dv\|/\|v\| : \|v\| \neq 0\}).$$

Consequently

$$\psi_0[w] \leqq \|Q\| \cdot \varphi(A)[w].$$

In view of b) this inequality implies $\psi_0 \prec \varphi(A)$.

The next theorem gives a necessary and sufficient condition for ψ_0 (cf. (2.4)) to be a stability functional for C. In the applications this theorem is particularly usefull when ψ_0 has a simpler structure than $\varphi(A)$.

Theorem 3. *$\varphi(A) \prec \psi_0$ if and only if* $\overline{\lim_{h\to 0}} \|Q^{-1}\| < \infty$.

Proof. We note that $Q = P^{-1}A$ and that P and A are bijective (for h sufficiently small). Hence Q is also bijective and $Q^{-1} = A^{-1}P$, $A^{-1} = Q^{-1}P^{-1}$.

1. Let $\overline{\lim_{h\to 0}} \|Q^{-1}\| < \infty$. Then $\|Q^{-1}\| \leqq \beta < \infty$ for $h \leqq$ some h_1 and $\varphi(A)[w] = \|A^{-1}w\| = \|Q^{-1}P^{-1}w\| \leqq \beta \cdot \psi_0[w]$ for h sufficiently small. Hence $\varphi(A) \prec \psi_0$.

2. Let $\varphi(A) \prec \psi_0$. Then $\|Q^{-1}P^{-1}w\| \leqq \beta \cdot \|P^{-1}w\|$ for some β and all $w \in \Lambda$, $h \in H$, $h \leqq$ some h_1.

Since P is bijective this leads to

$$\|Q^{-1}y\| \leqq \beta \cdot \|y\|$$

for all $y \in \Lambda$ and $h \leqq h_1$. Hence $\|Q^{-1}\| \leqq \beta$ (for $h \leqq h_1$) and $\overline{\lim_{h\to 0}} \|Q^{-1}\| < \infty$.

2.3 Optimal Stability of Semilinear Operators

Let P denote the operator of Section 2.2 and let K denote a set of operators C from Λ into itself with the following two properties:

(i) for each C in K the conditions a), b), c) of Section 2.2 are fulfilled,

(ii) there is a C_0 in K satisfying a) with $Q = I =$ the identity.

It is clear that all C in K are of the form

$$C = PQ + B,$$

where P is fixed and Q, B are variable.

The following theorem gives necessary and sufficient conditions for optimal stability in K.

Theorem 4. *Let $C = PQ + B \in K$. Then the following four propositions are equivalent.*

1. *C is optimally stable in K,*
2. *ψ_0 is a minimal stability functional for C,*
3. *ψ_0 is a stability functional for C,*
4. $\overline{\lim_{h \to 0}} \|Q^{-1}\| < \infty$.

Proof. Let $A = PQ$ and $\varphi = \varphi(A)$.

1) Assume 1. holds. Then C is more stable than or equivalent to C_0 (cf. (ii)) and consequently (cf. Theorem 1) $\varphi \prec \psi_0$. Hence (cf. Theorem 2) φ and ψ_0 are equivalent. Since φ is a minimal stability functional for C it follows that ψ_0 is also a minimal stability functional for C. Thus 2. is true.

2) Assume 2. holds. Then also 3. is true.

3) Assume 3. to be true. Then (cf. Theorem 1) $\varphi \prec \psi_0$, and in view of Theorem 3 this implies 4.

4) Finally, assume 4. to be satisfied. Then (cf. Theorem 3):

$$\varphi(A) \prec \psi_0.$$

Let $C' = PQ' + B' = A' + B'$ denote any operator in K.

Then (cf. Theorem 2)

$$\psi_0 \prec \varphi(A').$$

It follows that $\varphi(A) \prec \varphi(A')$. Consequently C is more stable than (or equivalent to) C' (cf. Theorem 1). Hence 1. holds.

This completes the proof of the theorem.

2.4. Notes

1. The definitions in Section 2.1 may easily be modified so as to include a number of stability definitions (cf. eg. [10, 14]) which, strictly speaking, are not covered by the formulations of Section 2.1. Since the definitions as given in Section 2.1 suffice for the semilinear equations discussed in the Chapters 3, 4, 5 we still have preferred here the simple definitions of Section 2.1.

2. Assume the operator P (cf. Section 2.3) is strongly stable (in the sense of [13, 16]) with respect to a given infinitesimal problem. In a number of examples (cf. also Chapter 3) we found that optimal stability of an operator C in K (cf. Section 2.3) always implies strong stability of C (but that the converse does not always hold). The question thus arises whether this is a general phenomenon and what are the general connections between optimal and strong stability.

3. The concept of stability as dealt with in Section 2.1 has essentially a qualitative character. By taking into account the value of the constant γ (cf. (2.3)) it would be possible to compare the stability of operators on a quantitative basis and one may wonder whether there are simple criteria for an operator C to be optimally stable in this quantitative sense.

4. Finally we note the limitedness of the theorems in Sections 2.2, 2.3—they only apply to semilinear operators—and the question arises whether similar theorems hold for (classes of) operators $C=A+B$ where B fails to satisfy the Lipschitz condition of Section 2.2.

3. Initial Value Problems for Systems of Ordinary Differential Equations

3.1. Step-by-step Methods

Let a and b be real numbers with $a<b$ and let $f(x, v)$ be a function defined for $a\leqq x\leqq b$, $v\in R_m$ (m-dimensional real vector space) and with range in R_m. Let $c\in R_m$ and let the vector-valued function $U(x)$ be a solution of the initial value problem

$$U'(x)=f\big(x, U(x)\big) \quad (a\leqq x\leqq b), \qquad U(a)=c. \tag{3.1}$$

Let k be a fixed integer $\geqq 1$. Let $h_0>0$, $H=(0, h_0]$ and let $h\in H$. We consider difference equations for the approximation of $U(x)$ of the type

$$\begin{aligned} &u_n-s_n=0 \qquad (0\leqq n\leqq k-1)\\ &h^{-1}\sum_{i=0}^{k} a_i u_{n-i}-F_n(u_{n-k}, \ldots, u_{n-1}, u_n; h)=0 \qquad (k\leqq n\leqq N) \end{aligned} \tag{3.2}$$

where N is the greatest integer with $Nh\leqq b-a$ and where the vectors u_n denote approximations of $U(x)$ at $x=x_n=a+nh$ $(n=0, 1, 2, \ldots, N)$. The vectors s_n $(0\leqq n\leqq k-1)$ are starting values found e.g. by a Taylor expansion (cf. [5]). The coefficients a_i in (3.2) are real constants with $a_0=1$. The vector-valued function $F_n(v_0, v_1, \ldots, v_k; h)$ (which depends on the given function f) is defined for $v_i\in R_m$ $(0\leqq i\leqq k)$, $h\in H$, $nh\leqq b-a$ and is assumed to satisfy a Lipschitz condition

$$\big|F_n(\tilde{v}_0, \ldots, \tilde{v}_k; h)-F_n(v_0, \ldots, v_k; h)\big|\leqq \lambda\cdot\max_{0\leqq i\leqq k}|\tilde{v}_i-v_i| \tag{3.3}$$

where $|\ldots|$ denotes the maximum norm in R_m. λ is some constant independent of $\tilde{v}_i$, $v_i\in R_m$, $h\in H$ and n (with $k\leqq n\leqq N$).

For $\zeta\neq 0$ we define

$$r(\zeta)=\sum_{i=0}^{k} a_i\zeta^{-i}$$

and the characteristic polynomial $\varrho(\zeta)$ is defined by

$$\varrho(\zeta)=\zeta^k\cdot r(\zeta).$$

$\varrho(\zeta)$ is assumed to satisfy the following conditions (3.4), (3.5):

$$\varrho(1)=0, \tag{3.4}$$

(3.5) all complex roots of the equation $\varrho(\zeta)=0$ have a modulus $\leqq 1$ and roots with modulus 1 are simple.

It is easily verified that e.g. (generalized) linear multistep methods (cf. e.g. [4]) and (explicit or implicit) Runge-Kutta methods (cf. e.g. [1]) are step-by-step methods of the form (3.2) satisfying the conditions (3.3), (3.4), (3.5)—provided

$f(x, v)$ satisfies a Lipschitz condition with respect to the variable v. In the subsequent we investigate the stability of the methods of type (3.2) using the ideas of Chapter 2. In Section 3.2 we describe the results of this investigation. The proofs and exact formulations of theorems have been postponed to Section 3.3.

3.2. *Formulation of Problems and Results*

In order to use the concepts of Chapter 2 we define for $h \in H = (0, h_0]$ the vector space $\Lambda = \Lambda^h$ by

$$\Lambda = \{u: u = (u_0, u_1, \ldots, u_N), \quad u_n \in R_m \, (0 \leqq n \leqq N)\}$$

where N is the greatest integer with $Nh \leqq b - a$. For $u \in \Lambda$ we define the norm

$$\|u\| = \max_{0 \leqq n \leqq N} |u_n|,$$

$|\ldots|$ denoting the maximum norm in R_m. The operator P is defined by

$$(3.6) \qquad (Pu)_n = u_n \quad (0 \leqq n \leqq k-1), \qquad (Pu)_n = h^{-1} \cdot (u_n - u_{n-1}) \quad (k \leqq n \leqq N)$$

where $u = (u_0, u_1, \ldots, u_N) \in \Lambda$.

With any method of type (3.2) satisfying (3.3) (for some value of λ), (3.4), (3.5) we associate an operator C defined by

$$(3.7a) \qquad C = A + B,$$

where

$$(3.7b) \quad (Au)_n = u_n \quad (0 \leqq n \leqq k-1), \qquad (Au)_n = h^{-1} \sum_{i=0}^{k} a_i u_{n-i} \quad (k \leqq n \leqq N),$$

$$(3.7c) \quad (Bu)_n = -s_n \quad (0 \leqq n \leqq k-1), \qquad (Bu)_n = -F_n (u_{n-k}, \ldots, u_n; h) \quad (k \leqq n \leqq N)$$

(for $u = (u_0, u_1, \ldots, u_N) \in \Lambda$). The operator Q is defined by

$$(3.8) \qquad Q = P^{-1}A$$

(note that P (cf. (3.6)) is bijective and the inverse P^{-1} thus exists).

Using (3.3), (3.4), (3.5) it may be verified (cf. e.g. [7, 10]) that the conditions a), b), c) of Chapter 2 are satisfied here. Further it is clear that the equation $Cu = 0$ (cf. (2.1)) is equivalent to (3.2) with $u = (u_0, u_1, \ldots, u_N)$ and that $C\tilde{u} = w$ (cf. (2.2)) is equivalent to

$$(3.9) \qquad \begin{aligned} \tilde{u}_n - s_n &= w_n \qquad (0 \leqq n \leqq k-1) \\ h^{-1} \sum_{i=0}^{k} a_i \tilde{u}_{n-i} - F_n (\tilde{u}_{n-k}, \ldots, \tilde{u}_n; h) &= w_n \qquad (k \leqq n \leqq N) \end{aligned}$$

(with $\tilde{u} = (\tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_N)$, $w = (w_0, w_1, \ldots, w_N)$).

K will denote the set of all operators C (cf. (3.7)) associated with any of the methods of type (3.2) satisfying (3.3), (3.4), (3.5). It is easily verified that K satisfies the general conditions of Section 2.3.

We deal with the following two questions:

1. What conditions on the method (3.2) are necessary and sufficient in order that it be associated with an operator C which is optimally stable in K?

2. What functional is a minimal stability functional for an operator C which is optimally stable in K?

In the next section (cf. Theorem 5) it will be shown that (3.2) is associated with an optimally stable C if and only if the following condition (3.10) (which is stronger than (3.5)) is satisfied:

(3.10) apart from $\zeta=1$ all complex roots of the equation $\varrho(\zeta)=0$ have a modulus <1.

It is striking that the same condition (3.10) also appears in the study of so-called "strong stability" of general linear multistep methods (cf. [5, 13]).

The second question will be answered by Theorem 6 of the next section in which it is proved that the functional φ_0 defined by

$$\varphi_0[w]=\sum_{i=0}^{k-1}|w_i|+\max_{k\leqq n\leqq N}\left|h\sum_{i=k}^{n}w_i\right| \tag{3.11}$$

is a minimal stability functional for any optimally stable operator C.

As a result of the Theorems 5, 6 it will be shown (cf. Theorem 7 of Section 3.3) that the conditions (3.3), (3.4), (3.10) imply the inequality

$$\max_{0\leqq n\leqq N}|\tilde{u}_n-u_n|\leqq\gamma\cdot\left\{\sum_{i=0}^{k-1}|w_i|+\max_{k\leqq n\leqq N}\left|h\sum_{i=k}^{n}w_i\right|\right\} \tag{3.12}$$

which holds for all vectors u_n, $\tilde{u}_n$ satisfying (3.2), (3.9), respectively with $0<h\leqq$ some h_1, γ denoting a constant independent of h and w_n. Theorem 7 also implies that (3.12) does not hold if condition (3.10) is violated.

The inequality (3.12) may be used to obtain bounds for the so-called global (accumulated) discretization error by applying (3.12) with $\tilde{u}_n=U(x_n)$ where $U(x)$ solves (3.1). For an example of such an application (where the usual inequality $\max_{0\leqq n\leqq N}|\tilde{u}_n-u_n|\leqq\gamma'\cdot\max_{0\leqq n\leqq N}|w_n|$ leads to an essentially too pessimistic error bound) we refer to [11].

3.3. Proof of the Theorems

In this section we prove the theorems already discussed in Section 3.2. Throughout this section K denotes the set defined above (Section 3.2).

Theorem 5. *Let $C=A+B\in K$. Then C is optimally stable in K if and only if the polynomial $\varrho(\zeta)$ satisfies condition* (3.10).

Proof. In view of Theorem 4 we only have to show that (3.10) is equivalent to $\overline{\lim_{h\to 0}}\|Q^{-1}\|<\infty$ where Q is defined by (3.8).

1. Assume (3.10) to be satisfied. We shall show that there are constants $h_1>0$, $\gamma>0$ such that

$$\|Q^{-1}w\|\leqq\gamma\cdot\|w\| \quad (0<h<h_1,\ w\in\Lambda). \tag{3.13}$$

Let $w\in\Lambda$ and let $z\in\Lambda$ satisfy $Qz=w$, i.e. (cf. (3.8)): $Az=Pw$. Hence (cf. (3.6), (3.7b))

$$z_n=w_n \qquad (0\leqq n\leqq k-1), \tag{3.14}$$

$$r(E)z_n=(I-E^{-1})w_n \qquad (k\leqq n\leqq N) \tag{3.15}$$

where E is the shifting operator ($E^p z_n=z_{n+p}$), I is the identity operator and r is the function introduced in Section 3.1. In view of (3.4) and $a_0=1$ we have

$$r(\zeta)=(1-\zeta^{-1})\cdot t(\zeta)$$

where $t(\zeta)$ is a polynomial of degree $\leqq k-1$ in the variable ζ^{-1} with lowest coefficient $=1$. (3.15) thus leads to

$$(I-E^{-1})\ \big(t(E)z_n-w_n\big)=0 \qquad (k\leqq n\leqq N),$$

from which we get (by performing a summation from k up to n):

$$t(E)z_n-w_n=t(E)z_{k-1}-w_{k-1} \qquad (k-1\leqq n\leqq N).$$

In view of (3.14) we have

$$t(E)z_n=w_n+t(E)w_{k-1}-w_{k-1} \qquad (k-1\leqq n\leqq N). \tag{3.16}$$

We define the characteristic polynomial $\tau(\zeta)$ by

$$\tau(\zeta)=\zeta^{k-1}\cdot t(\zeta).$$

It is easily verified that $\tau(\zeta)=(\zeta-1)^{-1}\cdot\varrho(\zeta)$. In view of (3.10) it follows that all complex roots of the equation $\tau(\zeta)=0$ have a modulus <1. Hence, by applying the lemma of [10] (with $p=0$), it follows that the vectors z_n (cf. (3.16)) satisfy the following inequality:

$$\max_{0\leqq n\leqq N}|z_n|\leqq\alpha\cdot\max_{0\leqq n\leqq N}|v_n| \qquad (0<h<\text{some } h_1)$$

where α is some constant and v_n is defined by

$$v_n=z_n \qquad (0\leqq n\leqq k-2),$$

$$v_n=w_n+t(E)w_{k-1}-w_{k-1} \qquad (k-1\leqq n\leqq N).$$

It follows (cf. (3.14)) that

$$\|z\|=\max_{0\leqq n\leqq N}|z_n|\leqq\gamma\cdot\max_{0\leqq n\leqq N}|w_n|=\gamma\cdot\|w\| \qquad (0<h<h_1)$$

γ being a constant independent of h and w. Since $z=Q^{-1}w$ this proves inequality (3.13), which is equivalent to $\overline{\lim_{h\to 0}}\,\|Q^{-1}\|<\infty$.

2. Assume (3.13) to be satisfied. We shall show that (3.10) holds.

Let v_n $(0 \leqq n \leqq N)$ be given vectors $\in R_m$ and assume the vectors z_n are determined by the Eqs. (3.17):

$$z_n = v_n \quad (0 \leqq n \leqq k-2) \tag{3.17a}$$

$$t(E) z_n = v_n \quad (k-1 \leqq n \leqq N). \tag{3.17b}$$

By "multiplying" both members of (3.17b) by $(I - E^{-1})$ we get

$$r(E) z_n = (I - E^{-1}) v_n \quad (k \leqq n \leqq N).$$

Defining $s = (I - t(E)) v_{k-1}$, we thus have

$$r(E) z_n = (I - E^{-1}) \{v_n + s\} \quad (k \leqq n \leqq N).$$

In view of (3.17) it follows that

$$\begin{aligned} z_n &= w_n && (0 \leqq n \leqq k-1) \\ r(E) z_n &= (I - E^{-1}) w_n && (k \leqq n \leqq N) \end{aligned}$$

where the vectors w_n are defined by

$$w_n = v_n \quad (0 \leqq n \leqq k-2), \qquad w_n = v_n + s \quad (k-1 \leqq n \leqq N).$$

Writing $z = (z_0, z_1, \ldots, z_N)$, $w = (w_0, w_1, \ldots, w_N)$ we thus have (cf. (3.6), (3.7b))

$$A z = P w.$$

Consequently $Q z = w$ and from (3.13) we obtain the inequality

$$\|z\| \leqq \gamma \cdot \|w\| \quad (\text{for } 0 < h < h_1).$$

In view of the definition of the vectors w_n we have

$$\|z\| \leqq \alpha \cdot \|v\| \quad (\text{for } 0 < h < h_1)$$

α being a constant independent of h and v. Since z_n is independent of v_i with $i > n$ (cf. (3.17)) we thus have

$$|z_n| \leqq \alpha \cdot \max_{0 \leqq i \leqq n} |v_i| \quad (\text{for } 0 < h < h_1). \tag{3.18}$$

Since (3.18) holds for all vectors v_n, z_n satisfying (3.17) we may apply the lemma of [10] (with $p = 0$). This yields the result that all zeros of the polynomial $\tau(\zeta) = \zeta^{k-1} \cdot t(\zeta) = (\zeta - 1)^{-1} \cdot \varrho(\zeta)$ have a modulus < 1. Hence (3.10) is fulfilled. This completes the proof of the theorem.

Theorem 6. *Let $C \in K$. Then the functional φ_0 defined by* (3.11) *is a minimal stability functional for C if and only if C is optimally stable in K.*

Proof. 1. The functional ψ_0 is defined by

$$\psi_0[w] = \|P^{-1} w\| \quad (\text{for } w \in \Lambda) \tag{3.19}$$

where P is the operator defined by (3.6).

Let $C \in K$. In view of Theorem 4 ψ_0 is a minimal stability functional for C if and only if C is optimally stable in K. We shall prove below that φ_0 (cf. (3.11)) and ψ_0 are equivalent. Hence φ_0 is a minimal stability functional for C if and only if ψ_0 is a minimal stability functional for C. It thus follows that φ_0 is a minimal stability functional for C if and only if C is optimally stable in K and Theorem 6 is thus proved.

2. It remains to be shown that φ_0 and ψ_0 are equivalent.

Let $w \in \Lambda$ and let $z \in \Lambda$ be defined by $Pz = w$. It follows (cf. (3.19)) that

$$\psi_0[w] = \|z\|.$$

In view of (3.6) we have

$$\begin{aligned} z_n &= w_n && (0 \leqq n \leqq k-1), \\ z_n &= w_{k-1} + h \sum_{i=k}^{n} w_i && (k \leqq n \leqq N). \end{aligned} \tag{3.20}$$

Combining (3.11) and (3.20) we get

$$\|z\| = \max_{0 \leqq n \leqq N} |z_n| \leqq \varphi_0[w]$$

and it follows that

$$\psi_0 \prec \varphi_0. \tag{3.21}$$

In view of (3.11), (3.20) we also have

$$\varphi_0[w] = \sum_{i=0}^{k-1} |z_i| + \max_{k \leqq n \leqq N} |z_n - z_{k-1}| \leqq (k+2) \cdot \max_{0 \leqq n \leqq N} |z_n| = (k+2) \cdot \|z\|$$

and it follows that

$$\varphi_0 \prec \psi_0. \tag{3.22}$$

The equivalence of φ_0 and ψ_0 has thus been established (cf. (3.21), (3.22)), which completes the proof of the theorem.

Theorem 7. *Let* (3.2) *be a step-by-step method satisfying the conditions* (3.3), (3.4), (3.5). *Then the following statement* (3.23) *is true if and only if condition* (3.10) *is fulfilled.*

(3.23) There are fixed numbers $\gamma, h_1 > 0$ such that whenever vectors u_n, $\tilde{u}_n$ satisfy (3.2), (3.9), respectively with $0 < h < h_1$, then $|\tilde{u}_n - u_n|$ satisfies the inequality (3.12).

Proof. Let (3.2) be a given method (satisfying (3.3), (3.4), (3.5)) and let C denote the associated operator (cf. (3.7)). Since $C \in K$ we may apply the Theorems 5, 6.

1. Assume (3.10) to be fulfilled. In view of the Theorems 5, 6 φ_0 is a minimal stability functional for C. Since (3.23) is equivalent to stability of C with respect to φ_0 (cf. (3.11)), it follows that (3.23) is true.

2. Assume (3.23) to hold, i.e. assume C to be stable with respect to φ_0. In view of $\varphi_0 \prec \psi_0$ (cf. (3.22)), it follows that C is also stable with respect to ψ_0. By virtue of Theorem 4 C is optimally stable in K. Hence (cf. Theorem 5) condition (3.10) is satisfied.

4. Boundary Value Problems for Ordinary Differential Equations

4.1. Finite-Difference Methods

Let $f(x, v)$ denote a real-valued function defined and continuous on the set

$$S=\{(x, v): 0 \leqq x \leqq 1, -\infty < v < \infty\}$$

with a continuous partial derivative $f_v(x, v)$ satisfying

$$\inf_S f_v(x, v) = -\eta > -\pi^2. \tag{4.1}$$

These conditions on f imply that the boundary value problem

$$U''(x) = f(x, U(x)) \quad (0 \leqq x \leqq 1), \qquad U(0) = U(1) = 0 \tag{4.2}$$

has a unique solution $U(x)$ (cf. [8]). In the following it is assumed that the solution $U(x)$ has a continuous 4th order derivative on $[0,1]$.

Let

$$H=\{h: h=(N+1)^{-1} \text{ where } N=1, 2, 3, \ldots\}$$

and let $h=(N+1)^{-1} \in H$. The difference methods for approximating $U(x)$ we shall consider are of the form

$$\begin{aligned} h^{-2} \cdot (u_{n+1}-2u_n+u_{n-1}) - \beta_n \cdot f(x_n, u_n) = 0 \qquad (n=1, 2, \ldots, N) \\ u_0 = u_{N+1} = 0, \end{aligned} \tag{4.3}$$

where the numbers u_n denote approximations of $U(x)$ at $x=x_n=nh$ $(n=0, 1, \ldots, N+1)$. The coefficients β_n in (4.3) are real numbers which may depend on n and h but which are independent of f and U. It is assumed that the β_n are chosen in such a way that

$$\inf \{\beta_n f_v(x_n, v): h \in H, 1 \leqq n \leqq N, -\infty < v < \infty\} = -\vartheta > -\pi^2, \tag{4.4}$$

$$\sup \{|\beta_n|: h \in H, 1 \leqq n \leqq N\} = \sigma < \infty. \tag{4.5}$$

By applying the techniques used by Lees in [8] it may be verified that, for h sufficiently small, (4.3) has a unique solution $(u_0, u_1, \ldots, u_{N+1})$.

For $\beta_n \equiv 1$ formula (4.3) reduces to a well known difference method (cf. e.g. [5, 8]) and (4.4), (4.5) are fulfilled with $\vartheta=\eta$, $\sigma=1$.

For $\beta_n \not\equiv 1$ formula (4.3) doesn't represent a direct approximation of the boundary value problem (4.2) and it is in general not consistent (in the maximum norm) with (4.2) (cf. e.g. [14] for definition of consistency). Nevertheless, in Section 4.4 (cf. Table 2) we shall present an example where the approximations u_n obtained from (4.3) with $\beta_n \not\equiv 1$ are more accurate than those obtained from (4.3) with $\beta_n \equiv 1$. This phenomenon is explained by Theorem 9 of Section 4.3.

In Section 4.2 we shall apply the concepts of Chapter 2 to (4.3) and determine a minimal stability functional for (4.3). The results of Section 4.2 are essential in our proof of Theorem 9.

4.2. Stability of Method (4.3)

In order to use the concepts of Chapter 2 we define for $h=(N+1)^{-1}\in H$ the vectorspace Λ by

$$\Lambda=\{u\colon u=(u_0, u_1, \ldots, u_{N+1}),\ -\infty<u_n<\infty,\ u_0=u_{N+1}=0\}$$

and for $u=(u_0, u_1, \ldots, u_{N+1})\in\Lambda$ we define the norm

$$\|u\|=\max_{1\leqq n\leqq N}|u_n|.$$

With method (4.3) we associate an operator C defined by

$$C=A+B \tag{4.6a}$$

where A and B are operators from Λ into Λ satisfying

$$(Au)_n=h^{-2}(u_{n+1}-2u_n+u_{n-1}), \tag{4.6b}$$

$$(Bu)_n=-\beta_n\cdot f(x_n, u_n) \tag{4.6c}$$

for $u\in\Lambda$, $1\leqq n\leqq N$. Operators P and Q from Λ into Λ are defined by $P=A$ and $Q=I$, the identity.

With these definitions the equation $Cu=0$ (cf. (2.1)) is equivalent to (4.3) and $C\tilde{u}=w$ (cf. (2.2)) is equivalent to

$$\begin{aligned} &h^{-2}\cdot(\tilde{u}_{n+1}-2\tilde{u}_n+\tilde{u}_{n-1})-\beta_n f(x_n, \tilde{u}_n)=w_n \qquad (1\leqq n\leqq N)\\ &\tilde{u}_0=\tilde{u}_{N+1}=0. \end{aligned} \tag{4.7}$$

Applying the techniques used by Lees in his study of (4.3) with $\beta_n\equiv 1$ (cf. [8]) it may be verified that the conditions imposed on the operators A, B, C, P, Q in Section 2.2 are satisfied here with the exception of the Lipschitz condition on B (cf. Section 2.2, b)).

If

$$\sup_S f_v(x, v)<\infty \tag{4.8}$$

we have for $v=(v_0, \ldots, v_{N+1})$, $\tilde{v}=(\tilde{v}_0, \ldots, \tilde{v}_{N+1})\in\Lambda$:

$$\|B\tilde{v}-Bv\|=\max_{1\leqq n\leqq N}|\beta_n f(x_n, \tilde{v}_n)-\beta_n f(x_n, v_n)|\leqq\max_{1\leqq n\leqq N}\sigma L\cdot|\tilde{v}_n-v_n|=\lambda\cdot\|\tilde{v}-v\|$$

where $\lambda=\sigma L$ and $L=\sup_S|f_v(x, v)|<\infty$ (cf. (4.6c), (4.5), (4.1), (4.8)). Consequently if (4.8) holds, the operator B satisfies the Lipschitz condition required in Section 2.2 and all conditions of Section 2.2 are thus satisfied here. By applying Theorem 1 we arrive at the following theorem.

Theorem 8. *Let f satisfy the conditions of Section* 4.1 *and condition* (4.8). *Then there are fixed numbers γ, $h_1>0$ such that for all numbers u_n, $\tilde{u}_n$, w_n satisfying* (4.3), (4.7), *respectively with $h\in H$, $h\leqq h_1$ we have the inequality*

$$|\tilde{u}_n-u_n|\leqq\gamma\cdot\max_{1\leqq n\leqq N}\left|-x_n\cdot h\sum_{i=1}^{N}(1-x_i)w_i+h\sum_{i=1}^{n}(x_n-x_i)w_i\right| \quad (n=1,2,\ldots,N). \tag{4.9}$$

Proof. In view of Theorem 1 $\|A^{-1}w\|$ is a stability functional for C (cf. (4.6)). Hence there are γ, $h_1>0$ such that whenever u_n, $\tilde{u}_n$, w_n satisfy (4.3) (4.7), respectively with $h\in H$, $h\leqq h_1$ then

$$\max_{1\leqq n\leqq N}|\tilde{u}_n-u_n|\leqq\gamma\cdot\max_{1\leqq n\leqq N}|v_n| \tag{4.10}$$

where the v_n are defined by

$$h^{-2}(v_{n+1}-2v_n+v_{n-1})=w_n \quad (1\leqq n\leqq N), \qquad v_0=v_{N+1}=0.$$

It is readily verified (cf. e.g. [5], p. 363) that

$$v_n=-x_n\cdot h\sum_{i=1}^{N}(1-x_i)w_i+h\sum_{i=1}^{n}(x_n-x_i)w_i. \tag{4.11}$$

In view of (4.10) this proves (4.9).

4.3. Accuracy of Method (4.3)

In this section we consider methods of type (4.3) where the coefficients β_n satisfy (in addition to (4.4), (4.5)):

$$\beta_n=b \quad (\text{for } n\equiv 0,1 \pmod 3), \qquad \beta_n=3-2b \quad (\text{for } n\equiv 2 \pmod 3), \tag{4.12}$$

b denoting an arbitrary real constant which is independent of n and h. For $b=1$ we get the conventional coefficients $\beta_n\equiv 1$ while for $b\neq 1$ (4.3) reduces to a non-consistent approximation of (4.2) (cf. Section 4.1). Similarly as stability has been defined with respect to a functional ψ, consistency may also be defined with respect to a functional ψ. For $b\neq 1$ the operator C defined by (4.6), (4.12) is in general not consistent with respect to $\psi[w]\equiv\|w\|$. However, if the right hand side of (4.9) would be used as a definition of $\psi[w]$, then C would become consistent with respect to ψ (see the proof of Theorem 9). The following theorem shows that the methods generated by any value $b\neq 1$ are still of the same order of accuracy as (4.3) with $b=1$—cf. [8] for a detailed treatment of the case $b=1$. Note that in this theorem we have got rid of condition (4.8).

Theorem 9. *Let f satisfy the conditions of Section* 4.1 *and let the coefficients β_n satisfy* (4.4), (4.12).

Let u_n satisfy (4.3). *Then*

$$\max_{1\leqq n\leqq N}|u_n-U(x_n)|=\mathcal{O}(h^2). \tag{4.13}$$

Proof. 1. Suppose the theorem has already been proved for the case of functions f satisfying (4.8). We shall show that the theorem then also holds for f violating (4.8).

Let f be an arbitrary function satisfying the conditions of Section 4.1 violating (4.8) and let

$$\max\{|U(x)|: 0\leqq x\leqq 1\}<M<\infty.$$

The function f^* on S is defined by

$$f^*(x,v)=\begin{cases} f(x,v) & (|v|\leqq M)\\ f(x,M)+(v-M)\cdot f_v(x,M) & (v>M)\\ f(x,-M)+(v+M)\cdot f_v(x,-M) & (v<-M).\end{cases}$$

$U^*(x)$ and u_n^* are defined by requiring that they satisfy (4.2), (4.3), respectively with f replaced by f^*. Since $\sup_S f_v^*(x,v)<\infty$ we have

$$\max_{1\leqq n\leqq N}|u_n^*-U^*(x_n)|=\mathcal{O}(h^2) \tag{4.14}$$

and as $U''(x)=f^*(x,U(x))$ $(0\leqq x\leqq 1)$ we have $U(x)=U^*(x)$.

Consequently

$$\max_{1\leqq n\leqq N}|u_n^*-U(x_n)|\to 0 \quad (\text{for } h\to 0).$$

Therefore, for $h\leqq$ some h_2 we have $|u_n^*|\leqq M$. Thus u_n^* and u_n satisfy the same set of Eqs. (4.3) for $h\leqq h_2$. It follows that $u_n=u_n^*$ and consequently (cf. (4.14))

$$\max_{1\leqq n\leqq N}|u_n-U(x_n)|=\mathcal{O}(h^2),$$

which proves that the theorem holds for the function f.

2. In view of the above we may assume with no loss of generality that (4.8) is fulfilled. Consequently we may apply Theorem 8. Taking $\tilde{u}_n=U(x_n)$ formula (4.9) reduces to

$$\max_{1\leqq n\leqq N}|u_n-U(x_n)|\leqq\gamma\cdot\max_{1\leqq n\leqq N}|v_n| \tag{4.15}$$

where v_n is given by (4.11) with

$$w_n=h^{-2}\big(U(x_{n+1})-2U(x_n)+U(x_{n-1})\big)-\beta_n f\big(x_n,U(x_n)\big).$$

Using Taylor's formula we have

$$w_n=(1-\beta_n)U''(x_n)+\mathcal{O}(h^2) \qquad (\text{uniformly for } 1\leqq n\leqq N;\ h\in H).$$

Substituting this expression for w_n in (4.11) we obtain the formula

$$v_n=-x_n\cdot y_{N+1}+y_n+\mathcal{O}(h^2) \qquad (\text{uniformly for } 1\leqq n\leqq N;\ h\in H), \tag{4.16}$$

where y_m is defined by

$$y_m=h\sum_{i=1}^{m}(x_m-x_i)\,(1-\beta_i)\,U''(x_i) \qquad (1\leqq m\leqq N+1). \tag{4.17}$$

It will be shown (cf. 3. below) that $y_m=\mathcal{O}(h^2)$ (uniformly for $1\leqq m\leqq N+1$; $h\in H$). Consequently (cf. (4.16))

$$\max_{1\leqq n\leqq N}|v_n|=\mathcal{O}(h^2)$$

and in view of (4.15) this yields (4.13).

3. It remains to be shown that y_m defined by (4.17) is $\mathcal{O}(h^2)$ uniformly for $1\leqq m\leqq N+1$.

Let $1\leqq m\leqq N+1$ and $k=[m/3]$. Hence $m=3k+r$ with $0\leqq r<3$. In view of (4.12) we have for $r>0$:

$$\left|h\sum_{i=3k+1}^{m}(x_m-x_i)\,(1-\beta_i)\,U''(x_i)\right|\leqq h^2\cdot|1-b|\max_x|U''(x)|.$$

Consequently (cf. (4.17))

$$y_m=h\sum_{i=1}^{3k}(x_m-x_i)\,(1-\beta_i)\,U''(x_i)+\mathcal{O}(h^2) \tag{4.18}$$

uniformly in m (for $h\in H$). Using (4.12) and writing $U_i''=U''(x_i)$, $x=x_m$ we have

$$\begin{aligned}
&h\sum_{i=1}^{3k}(x_m-x_i)\,(1-\beta_i)\,U''(x_i)\\
&=h\sum_{j=1}^{k}\{(x-x_{3j-2})\,(1-b)\,U''_{3j-2}+(x-x_{3j-1})\,(-2+2b)\,U''_{3j-1}+(x-x_{3j})\,(1-b)\,U''_{3j}\}\\
&=h(1-b)\sum_{j=1}^{k}z_j,
\end{aligned}$$

where z_j is defined by

$$z_j=(x-x_{3j-2})\,U''_{3j-2}-2(x-x_{3j-1})\,U''_{3j-1}+(x-x_{3j})\,U''_{3j}.$$

Writing $g(\xi)=(x-\xi)\,U''(\xi)$ and using Taylor's formula we readily obtain

$$|z_j|\leqq h^2\cdot\max_\xi|g''(\xi)|. \tag{4.19}$$

It follows (cf. (4.18)) that

$$\begin{aligned}
|y_m|&\leqq h|1-b|\sum_{j=1}^{k}|z_j|+\mathcal{O}(h^2)\\
&\leqq h|1-b|\cdot k\cdot\max_j|z_j|+\mathcal{O}(h^2)\\
&\leqq \tfrac{1}{3}\cdot|1-b|\cdot\max_j|z_j|+\mathcal{O}(h^2).
\end{aligned}$$

Hence (cf. (4.19))

$$y_m=\mathcal{O}(h^2)\qquad(\text{uniformly for } 1\leqq m\leqq N+1;\ h\in H).$$

This completes the proof of Theorem 9.

4.4. Numerical Illustration

As an illustration of the above we consider the numerical solution of the following boundary value problems A and B:

A. $U''(x)=U(x)+\sin(\pi x)\ \ (0\leqq x\leqq 1),\ U(0)=U(1)=0,$

B. $U''(x)=4^{-x}\cdot e^{U(x)}+(1+x)^{-2}\ \ (0\leqq x\leqq 1),\ U(0)=U(1)=0,$

the true solutions of which are

$$U(x) = -(1+\pi^2)^{-1} \cdot \sin(\pi x) \quad \text{and} \quad U(x) = x \log 4 - 2 \log(1+x),$$

respectively. We have solved the boundary value problem A by method (4.3) with $b=1$ and $b=9/8$, respectively and problem B by (4.3) with $b=1$ and $b=3/4$, respectively (cf. (4.12)). In Table 1 we list the error $u_n - U(x)$ (multiplied by 10^7) which is present in the approximation of $U(x)$ satisfying boundary value problem A. In view of the symmetry of problem A we only list the errors at $x = x_n \leqq 1/2$. In Table 2 referring to the boundary value problem B we list 10^7 times the error $u_n - U(x)$ at the points $x = x_n = 0.2, 0.4, 0.6, 0.8$.

Table 1. *Errors in solving problem A*

h	x	Error $u_n - U(x)$ using method (4.3), (4.12) with	
		$b=1$	$b=9/8$
1/64	0.125	−64	+12
	0.250	−119	−175
	0.375	−155	−236
	0.500	−168	+23
1/256	0.125	−4	+1
	0.250	−7	−11
	0.375	−10	−15
	0.500	−10	+1

Table 2. *Errors in solving problem B*

h	x	Error $u_n - U(x)$ using method (4.3), (4.12) with	
		$b=1$	$b=3/4$
1/75	0.2	44	13
	0.4	54	15
	0.6	45	13
	0.8	26	7
1/150	0.2	11	3
	0.4	13	4
	0.6	11	3
	0.8	6	2

The computations have been performed on the IBM 360-50 at the Centraal Reken-Instituut of Leiden University (56 binary digits in the mantissa). The numerical results clearly confirm the theory of Section 4.3.

We note that the interesting question for which differential equations (4.2) and at which points x_n a given method (4.3), (4.12) with $b \neq 1$ yields more accurate approximations (as $h \to 0$) than (4.3), (4.12) with $b = 1$, can be answered by deriving asymptotic estimates for the errors $u_n - U(x)$ along the lines described in [5, 15].

We shall not go further into this matter here.

5. Partial Hyperbolic Differential Equations

5.1. Finite-Difference Methods

Let $U(x, y)$ denote the solution of the Goursat problem

$$\begin{aligned} &U_{xy}(x, y) = f(x, y, U(x, y)) \quad (x \geqq 0, y \geqq 0) \\ &U(x, 0) = f_1(x) \quad (x \geqq 0), \qquad U(0, y) = f_2(y) \quad (y \geqq 0) \end{aligned} \tag{5.1}$$

where f, f_1, f_2 are given real-valued functions and f satisfies the Lipschitz condition

$$|f(x, y, \tilde{v}) - f(x, y, v)| \leqq L \cdot |\tilde{v} - v|, \tag{5.2}$$

L being a constant independent of the variables $x \geqq 0$, $y \geqq 0$, $-\infty < \tilde{v}, v < \infty$.

Let k be a fixed integer $\geqq 1$ and let $h_0 > 0$, $H = (0, h_0]$, $h \in H$. There are many methods (cf. e.g. [9, 17, 18]) for the approximation of $U(x, y)$ of the type

$$u_{m,n} - s_{m,n} = 0 \quad (\text{for } m < k \text{ or } n < k), \tag{5.3a}$$

$$h^{-2} \sum_{i=0}^{k} \sum_{j=0}^{k} a_i a_j u_{m-i, n-j} - F_{m,n}(u; h) = 0 \quad (\text{for } m \geqq k, n \geqq k) \tag{5.3b}$$

where m and n are integers $\geqq 0$ and $u_{m,n}$ is an approximation of $U(x, y)$ for $x = mh$, $y = nh$. The numbers $s_{m,n}$ are starting values (found from the characteristic values f_1, f_2 and, if $k > 1$, e.g. by applying a method of type (5.3) with $k = 1$). The coefficients a_i are real numbers with $a_0 = 1$ with which we associate the same polynomials $r(\zeta)$ and $\varrho(\zeta) = \zeta^k \cdot r(\zeta)$ as in Section 3.1. It is assumed that the conditions (3.4), (3.5) on $\varrho(\zeta)$ are satisfied here. In (5.3b) u denotes the (infinite dimensional) vector with components $u_{m,n}$ $(m \geqq 0, n \geqq 0)$ and F is a function (which depends on f) satisfying the Lipschitz condition

$$|F_{m,n}(\tilde{v}; h) - F_{m,n}(v; h)| \leqq \lambda \cdot \max\{|\tilde{v}_{i,j} - v_{i,j}| : i \geqq 0, j \geqq 0, i + j \leqq m + n\} \tag{5.4}$$

(for $m \geqq k$, $n \geqq k$) where λ is a constant independent of m, n, h and of the vectors $\tilde{v} = (\tilde{v}_{i,j})$, $v = (v_{i,j})$. Using Banach's theorem on contraction operators it may be proved (in a similar way as e.g. in [5], p. 299) that for $h < \lambda^{-1}$ (5.3) has a unique solution $u = (u_{m,n})$. In the subsequent we therefore assume that

$$h_0 < \lambda^{-1}. \tag{5.5}$$

Let a and b be given real numbers with $a > 0$, $b > 0$. Let M and N denote the greatest integers with $Mh \leqq a$, $Nh \leqq b$, respectively. In the subsequent sections we shall study the stability and the propagation of round-off error if (5.3) is used to approximate $U(x, y)$ for $0 < x \leqq a$, $0 < y \leqq b$.

We make the obvious assumption that $F_{m,n}(v;h)$ with $k \leqq m \leqq M$, $k \leqq n \leqq N$ is independent of the components $v_{i,j}$ of v with $i > M$ or $j > N$.

In Section 5.2 we apply the concepts of Chapter 2 to method (5.3) and determine a minimal stability functional for (5.3) (cf. Theorem 10). In Section 5.3 we apply Theorem 10 to the propagation of round-off error in method (5.3). In the remainder of this chapter we focus on a so-called split form (cf. [12]) of (5.3) for reducing the propagation of round-off error in (5.3). The main theorem on this split form (Theorem 12 of Section 5.5) is a generalization of Theorem 6 of [12]. Our proof of Theorem 12 is essentially based on Theorem 10. In Section 5.6 numerical examples are presented illustrating the better error propagation of the split form.

5.2. *Minimal Stability Functional*

In order to use the formulations of Chapter 2 we define $A = \{u: u = (u_{m,n})\}$ where $u_{m,n}$ are real numbers $(m \geqq 0, n \geqq 0)\}$ and for $u = (u_{m,n}) \in A$ we define the seminorm

$$\|u\| = \max\{|u_{m,n}|: 0 \leqq m \leqq M, 0 \leqq n \leqq N\}.$$

With method (5.3) we associate an operator C mapping A into itself. For $u \in A$ the numbers $(Cu)_{m,n}$ are defined by the left-hand members of (5.3) and the equation $Cu = 0$ (cf. (2.1)) is thus equivalent to (5.3).

The operator A is defined by

$$(5.6) \qquad \begin{aligned} (Au)_{m,n} &= u_{m,n} && (m < k \text{ or } n < k) \\ (Au)_{m,n} &= h^{-2} \cdot \sum_{i=0}^{k} \sum_{j=0}^{k} a_i a_j u_{m-i,n-j} && (m \geqq k,\ n \geqq k). \end{aligned}$$

Operators B, P and Q are defined by $B = C - A$, $P = A$, $Q = I$, the identity. Using techniques discussed e.g. in [10, 11, 17] it may be verified that the conditions imposed on the operators A, B, C, P and Q in Section 2.2 are satisfied here. Since the equation $C\,\tilde{u} = w$ (cf. (2.2)) is equivalent to

$$(5.7) \qquad \begin{aligned} \tilde{u}_{m,n} - s_{m,n} &= w_{m,n} && (m < k \text{ or } n < k) \\ h^{-2} \sum_i \sum_j a_i a_j \tilde{u}_{m-i,n-j} - F_{m,n}(\tilde{u}; h) &= w_{m,n} && (m \geqq k,\ n \geqq k) \end{aligned}$$

it follows that there are fixed numbers γ_1, $h_1 > 0$ such that for all $u = (u_{m,n})$, $\tilde{u} = (\tilde{u}_{m,n})$ satisfying (5.3), (5.7), respectively with $0 < h < h_1$ we have the inequality

$$(5.8) \qquad \|\tilde{u} - u\| \leqq \gamma_1 \cdot \|w\|$$

(cf. condition c) of Section 2.2).

Applying Theorem 1 it follows that $\varphi[w] = \|A^{-1}w\|$ is a minimal stability functional for C and—in view of (5.6)—we thus have the following theorem, which yields a refinement of the error bound (5.8):

Theorem 10. *There are fixed numbers γ_2, $h_2 > 0$ such that whenever $u = (u_{m,n})$ and $\tilde{u} = (\tilde{u}_{m,n})$ satisfy (5.3), (5.7), respectively with $0 < h < h_2$, then*

$$(5.9) \qquad \|\tilde{u} - u\| \leqq \gamma_2 \cdot \|z\|$$

where $z=(z_{m,n})$ *is defined by*

$$z_{m,n}=w_{m,n} \qquad (m<k \text{ or } n<k)$$
$$h^{-2}\sum_i\sum_j a_i a_j z_{m-i,n-j}=w_{m,n} \qquad (m\geqq k,\ n\geqq k). \tag{5.10}$$

5.3. *Round-off Error in Method* (5.3)

In actual computation the relations (5.3) defining the numerical method are not satisfied exactly because of round-off error (and because of the incomplete solution of (nonlinear) equations if (5.3 b) is an implicit procedure, cf. e.g. [18]). We assume that the values $\tilde{u}_{m,n}$ actually calculated satisfy (5.3 a) exactly and instead of (5.3 b) satisfy

$$\sum_i\sum_j a_i a_j \tilde{u}_{m-i,n-j}=h^2 F_{m,n}(\tilde{u};h)+\delta_{m-k,n-k} \qquad (m\geqq k,\ n\geqq k) \tag{5.11}$$

(note that (5.11) is normalized so that $\tilde{u}_{m,n}$ enters in the left member of (5.11) with a coefficient $=1$). The quantities $\delta_{m,n}$ $(m\geqq 0,\ n\geqq 0)$ are called local round-off errors and $\tilde{u}_{m,n}-u_{m,n}$ accumulated round-off errors (cf. [5]).

Applying (5.8) with $w_{m,n}=0$ $(m<k$ or $n<k)$, $w_{m,n}=h^{-2}\cdot\delta_{m-k,n-k}$ $(m\geqq k,$ $n\geqq k)$ we get the following bound for the accumulated round-off error:

$$\|\tilde{u}-u\|\leqq\gamma_1\cdot h^{-2}\|\delta\| \qquad (\text{for } h<h_1), \tag{5.12}$$

exhibiting a quadratic round-off error accumulation in method (5.3) (cf. also [17]). We note that it is not possible to replace the exponent -2 in (5.12) by a number >-2.

Applying (5.9) in a similar fashion as (5.8) we arrive at the following more refined error bound for the accumulated round-off error

$$\|\tilde{u}-u\|\leqq\mathfrak{N}(\delta) \qquad (\text{for } h<h_2), \tag{5.13}$$

where the functional $\mathfrak{N}$ is defined by

$$\mathfrak{N}(\delta)=\gamma_2\cdot\|z\|, \tag{5.14a}$$

$$z_{m,n}=0 \qquad (m<k \text{ or } n<k)$$
$$\sum_i\sum_j a_i a_j z_{m-i,n-j}=\delta_{m-k,n-k} \qquad (m\geqq k,\ n\geqq k). \tag{5.14b}$$

The inequality (5.13) will be essential in our proof of Theorem 12 (Section 5.5).

Since the operator A (cf. (5.6)) is stable with respect to $\psi[w]=\|w\|$ (cf. condition c), Section 2.2) we have $\|A^{-1}w\|\leqq\beta\|w\|$ for some constant β and h sufficiently small. Taking $w_{m,n}=0$ $(m<k$ or $n<k)$, $w_{m,n}=h^{-2}\cdot\delta_{m-k,n-k}$ $(m\geqq k,\ n\geqq k)$ we thus have the following inequality for the functional $\mathfrak{N}$:

$$\mathfrak{N}(\delta)\leqq\gamma_3\cdot h^{-2}\|\delta\| \qquad (\text{for } h<h_3) \tag{5.15}$$

where the constants $\gamma_3=\gamma_2\cdot\beta$ and $h_3>0$ are independent of δ and h.

5.4. *Split Form of* (5.3)

We consider the system of difference equations

$$\lambda_1(X)\,\mu_1(Y)\,u_{m,n}=h^p v_{m,n} \qquad (m\geqq 0,\ n\geqq 0) \tag{5.16a}$$

$$\lambda_2(X)\,\mu_2(Y)\,v_{m,n}=h^q F_{m+k,n+k}(u;h) \qquad (m\geqq 0,\ n\geqq 0) \tag{5.16b}$$

where λ_1, λ_2, μ_1, μ_2 are polynomials of a degree s, $k-s$, t, $k-t$, respectively with real coefficients and highest coefficient $=1$. It is assumed that

$$\lambda_1(\zeta)\,\lambda_2(\zeta)\equiv\mu_1(\zeta)\,\mu_2(\zeta)\equiv\varrho(\zeta) \tag{5.17a}$$

(ζ denoting a complex variable). X and Y denote shifting operators in x and y directions, respectively ($Xz_{m,n}=z_{m+1,n}$, $Yz_{m,n}=z_{m,n+1}$) and p, q are real numbers with

$$p+q=2. \tag{5.17b}$$

Theorem 11. a) *Let $u_{m,n}$, $v_{m,n}$ satisfy* (5.16). *Then the numbers $u_{m,n}$ also satisfy* (5.3b).

b) Let $u_{m,n}$ satisfy (5.3b). *Then there are numbers $v_{m,n}$ such that $u_{m,n}$, $v_{m,n}$ satisfy* (5.16).

Proof. a) Part a of the theorem is easily proved by solving $v_{m,n}$ from (5.16a), substituting the result in the left member of (5.16b) and then applying (5.17) in order to get

$$h^{-2}\cdot\varrho(X)\,\varrho(Y)\,u_{m,n}=F_{m+k,n+k}(u;h) \qquad (m\geqq 0,\ n\geqq 0). \tag{5.18}$$

Since $\varrho(\zeta)=\zeta^k\cdot r(\zeta)$ formula (5.18) is equivalent to (5.3b).

Part a of the theorem may likewise be proved by a direct application of Theorem 1 of [12].

b) Let $u_{m,n}$ satisfy (5.3b). We define numbers $v_{m,n}$ by

$$v_{m,n}=h^{-p}\,\lambda_1(X)\,\mu_1(Y)\,u_{m,n} \qquad (m\geqq 0,\ n\geqq 0). \tag{5.19}$$

Hence (5.16a) is fulfilled. By performing the same substitution as in the proof of part a and using (5.18) it follows easily that (5.16b) is fulfilled. The theorem is thus proved.

In view of this theorem the system (5.16) and (5.3b) are equivalent and (5.16) is called a *split form* of (5.3b). If no round-off is present and starting values $u_{m,n}=s_{m,n}$ ($m<k$ or $n<k$) are prescribed, (5.3b) and its split form evidently produce the same approximations $u_{m,n}$. However, as will be shown in Section 5.5, they behave quite differently with respect to the propagation of round-off error.

We conclude this section by indicating in which order the numbers $u_{m,n}$, $v_{m,n}$ should be computed from (5.16), starting values $u_{m,n}=s_{m,n}$ being given:

1. Compute $v_{m,n}$ ($m<k-s$ or $n<k-t$) from (5.16a) (cf. (5.19)).

2. In view of (5.4) $F_{k,k}(u;h)$ only depends on components $u_{m,n}$ of u which are already known and possibly on $u_{k,k}$.

If $F_{k,k}(u;h)$ is independent of $u_{k,k}$ compute $v_{k-s,k-t}$ directly from (5.16b) (with $m=n=0$) and $u_{k,k}$ from (5.16a) (with $m=k-s$, $n=k-t$).

If $F_{k,k}(u; h)$ depends on $u_{k,k}$, compute $v_{k-s,k-t}$ and $u_{k,k}$ by the method of successive substitutions from (5.16b) (with $m=n=0$), (5.16a) (with $m=k-s$, $n=k-t$) (cf. [12] proof of Theorem 4 for a similar computation).

3. Compute $u_{m,n}$ and $v_{m-s,n-t}$ successively for $m+n=2k+1,\ 2k+2,\ldots$ by using (5.16b), (5.16a) cyclically in a similar way as indicated above for $m=n=k$.

5.5. Round-off Error in the Split Form (5.16)

In order to study the propagation of round-off error if the split form (5.16) is used to approximate $U(x, y)$ we assume that

$$\lambda_1(X)\,\mu_1(Y)\,\tilde{u}_{m,n}=h^p\,\tilde{v}_{m,n}+\xi_{m,n} \qquad (m\geqq 0,\ n\geqq 0) \tag{5.20a}$$

$$\lambda_2(X)\,\mu_2(Y)\,\tilde{v}_{m,n}=h^q F_{m+k,n+k}(\tilde{u}; h)+\eta_{m,n} \qquad (m\geqq 0,\ n\geqq 0), \tag{5.20b}$$

where $\xi_{m,n}, \eta_{m,n}$ are local round-off errors. Assume $u_{m,n}$ satisfies (5.3b) and assume $\tilde{u}_{m,n}=u_{m,n}=s_{m,n}$ ($m<k$ or $n<k$). $\tilde{u}-u=(\tilde{u}_{m,n}-u_{m,n})$ thus represents the accumulated round-off error in method (5.16) resulting from the local errors $\xi_{m,n}, \eta_{m,n}$.

Theorem 12. *Assume that at most one of the polynomials $\lambda_1(\zeta)$, $\mu_1(\zeta)$ has zeros ζ with modulus $|\zeta|=1$.*

Then there are constants γ, $\varepsilon>0$ such that

$$\|\tilde{u}-u\|\leqq\gamma\cdot h^{-1}\{\|\xi\|+h^{p-1}\cdot\|\eta\|\} \tag{5.21}$$

for all $\xi=(\xi_{m,n})$, $\eta=(\eta_{m,n})$ and h with $0<h<\varepsilon$.

Proof. 1. In view of the error bound (5.13) we obtain by performing a similar substitution as in the proof of Theorem 11a or by a direct application of Theorem 2 of [12] the following inequality for the error $\tilde{u}-u$ resulting from the local perturbations ξ, η:

$$\|\tilde{u}-u\|\leqq\mathfrak{N}(w)+\mathfrak{N}(h^p\cdot\eta) \qquad (\text{for } h<h_2),$$

where $w=(w_{m,n})$, $\eta=(\eta_{m,n})$ and

$$w_{m,n}=\lambda_2(X)\,\mu_2(Y)\,\xi_{m,n} \qquad (m\geqq 0,\ n\geqq 0). \tag{5.22}$$

In view of (5.14), (5.15) we have

$$\mathfrak{N}(h^p\eta)=h^p\cdot\mathfrak{N}(\eta)\leqq\gamma_3\cdot h^{p-2}\|\eta\| \qquad (\text{for } h<h_3).$$

Hence, for $h<\varepsilon=\min(h_2, h_3)$:

$$\|\tilde{u}-u\|\leqq\mathfrak{N}(w)+\gamma_3\,h^{p-2}\cdot\|\eta\|.$$

It will be proved (cf. 2. below) that

$$\mathfrak{N}(w)\leqq\gamma_4\cdot h^{-1}\|\xi\| \tag{5.23}$$

γ_4 being independent of $\xi\in\Lambda$, $h\in H$. Consequently (5.21) holds with $\gamma=\max(\gamma_3, \gamma_4)$ and the theorem is proved.

2. We shall prove (5.23). In view of (5.14), (5.22) we have

$$\mathfrak{N}(w) = \gamma_2 \cdot \|z\| \tag{5.24}$$

where, using operator notation, the equations defining z may be written

$$z_{m,n} = 0 \qquad (m < k \text{ or } n < k) \tag{5.25a}$$

$$\varrho(X)\,\varrho(Y)\,z_{m,n} = \lambda_2(X)\,\mu_2(Y)\,\xi_{m,n} \qquad (m \geqq 0,\ n \geqq 0). \tag{5.25b}$$

Since at most one of the polynomials $\lambda_1(\zeta)$, $\mu_1(\zeta)$ has zeros with modulus $= 1$ we may assume (in view of (3.5), (5.17a)) that e.g.

(5.26) all zeros of $\lambda_1(\zeta)$ have a modulus < 1

(the case where $\lambda_1(\zeta)$ violates (5.26) is treated similarly by interchanging the roles of λ_1 and μ_1).

Let n be a fixed integer $\geqq 0$. Defining

$$\hat{z}_m = \varrho(Y)\,z_{m,n}, \qquad \hat{\xi}_m = \mu_2(Y)\,\xi_{m,n} \qquad (m \geqq 0)$$

the formulas (5.25) reduce to

$$\hat{z}_m = 0 \quad (m < k), \qquad \varrho(X)\,\hat{z}_m = \lambda_2(X)\,\hat{\xi}_m \quad (m \geqq 0).$$

Since $\varrho(X) = \lambda_2(X)\,\lambda_1(X)$ (see (5.17a)) we have

$$\lambda_2(X)\{\lambda_1(X)\,\hat{z}_m - \hat{\xi}_m\} = 0 \qquad (m \geqq 0). \tag{5.27}$$

Since all zeros of the polynomial $\lambda_2(\zeta)$ have a modulus $\leqq 1$ and zeros with modulus 1 are simple (cf. (3.5), (5.17a)) the difference equation (5.27) implies that

$$|\lambda_1(X)\,\hat{z}_m - \hat{\xi}_m| \leqq \alpha \cdot \sum_{i=0}^{k-s-1} |\lambda_1(X)\,\hat{z}_i - \hat{\xi}_i|$$

α being a constant independent of $m \geqq 0$, $h > 0$ (see e.g. the lemma in [10] with $p = 1$). As $\hat{z}_i = 0$ $(i < k)$ this inequality leads to

$$|\lambda_1(X)\,\hat{z}_j| \leqq |\lambda_1(X)\,\hat{z}_j - \hat{\xi}_j| + |\hat{\xi}_j| \leqq \alpha \cdot \sum_{i=0}^{k-s-1} |\hat{\xi}_i| + |\hat{\xi}_j|.$$

Hence

$$|\lambda_1(X)\hat{z}_j| \leqq \beta_1 \cdot \max_{0 \leqq i \leqq j} |\hat{\xi}_i| \qquad (\text{for } j \geqq 0),$$

where $\beta_1 = 1 + \alpha \cdot (k - s)$. By virtue of (5.26) we have

$$|\hat{z}_m| \leqq \beta_2 \cdot \max_{0 \leqq j \leqq m} |\lambda_1(X)\,\hat{z}_j|$$

β_2 being a constant independent of m, h (see the lemma in [10] with $p = 0$). Consequently

$$|\hat{z}_m| \leqq \beta_2\,\beta_1 \cdot \max_{0 \leqq i \leqq m} |\hat{\xi}_i| \qquad (m \geqq 0).$$

Using the definitions of $\hat{z}_m$, $\hat{\xi}_i$ we obtain

$$|\varrho(Y)z_{m,n}| \leqq \beta_2 \beta_1 \cdot \max_{0 \leqq i \leqq m} |\mu_2(Y)\xi_{i,n}|.$$

Hence for $0 \leqq m \leqq M$, $0 \leqq n \leqq N-k$ we have

$$|\varrho(Y)z_{m,n}| \leqq \beta_3 \cdot \|\xi\| \tag{5.28}$$

β_3 being a constant independent of m, n, h. In view of the root condition (3.5) the inequality (5.28) implies that

$$|z_{m,n}| \leqq \beta_4 \cdot h^{-1}\|\xi\| \quad (0 \leqq m \leqq M,\ 0 \leqq n \leqq N)$$

for some constant β_4 (cf. the lemma loc. cit.). Combining this inequality with (5.24) we get (5.23) with $\gamma_4 = \gamma_2 \beta_4$.

This completes the proof of the theorem.

5.6. Numerical Illustration

As an illustration of the above we consider the numerical solution of (5.1) by the mid-point rule

$$u_{m+2,n+2} - u_{m,n+2} - u_{m+2,n} + u_{m,n} = 4h^2 \cdot f_{m+1,n+1} \quad (m \geqq 0,\ n \geqq 0) \tag{5.29}$$

where $f_{i,j} = f(ih, jh, u_{i,j})$. As is easily verified (5.29) is a method of type (5.3b) with $k=2$, $\varrho(\zeta) = \zeta^2 - 1$, satisfying all requirements of Section 5.1. The following procedure is a split form of type (5.16) with $p=0$, $q=2$:

$$\begin{aligned} u_{m+1,n} - u_{m,n} &= v_{m,n} \\ v_{m+1,n+2} + v_{m,n+2} - v_{m+1,n} - v_{m,n} &= 4h^2 f_{m+1,n+1} \end{aligned} \quad (m \geqq 0,\ n \geqq 0). \tag{5.30}$$

In (5.30) we have $\lambda_1(\zeta) = \zeta - 1$, $\lambda_2(\zeta) = \zeta + 1$, $\mu_1(\zeta) = 1$, $\mu_2(\zeta) = \zeta^2 - 1$. The conditions on $\lambda_1(\zeta)$, $\mu_1(\zeta)$ of Theorem 12 being satisfied here we may apply (5.21). This yields the following bound for the accumulated round-off error in method (5.30):

$$\|\tilde{u} - u\| \leqq \gamma \cdot h^{-1} \cdot \{\|\xi\| + h^{-1}\|\eta\|\}. \tag{5.31}$$

In view of (5.20a), (5.30) we may expect that $\tilde{v}_{m,n} = \mathcal{O}(h)$.

Consequently, if floating point arithmetic is used, $h^{-1}\|\eta\|$ (cf. (5.20b), (5.30)) may be expected to be of the same order of magnitude as $\|\xi\|$ (and as $\|\delta\|$ in (5.11)) and (5.31) thus exhibits the linear round-off error accumulation for the split form mentioned in Chapter 1.

We have solved the Goursat problem

$$\begin{aligned} &U_{xy}(x, y) = 2[U(x, y)]^3 \\ &U(x, 0) = (1+x)^{-1}, \quad U(0, y) = (1+y)^{-1} \end{aligned} \quad (x \geqq 0,\ y \geqq 0) \tag{5.32}$$

(the true solution of which is $U(x,y) = (1+x+y)^{-1}$) by each of the two procedures (5.29), (5.30), respectively. In the following table we list the accumulated round-

off error $\tilde{u}-u$ (multiplied by 10^6) which is present in the approximation of $U(x, y)$ at $(x, y)=(2, 2)$.

h^{-1}	Accumulated round-off error using procedure	
	(5.29)	(5.30)
32	−249	−5
64	−1073	−8
128	−4351	−16

The starting values $u_{m,n}=s_{m,n}$ ($m<2$ or $n<2$) used in the calculations agree with the exact solution $U(x, y)$ (to the number of digits used in the computations) and the $u_{m,n}$, $v_{m,n}$ for (5.30) have been calculated as indicated in Section 5.4. The computations have been performed on the IBM 360-50 at the Centraal Reken-Institut of Leiden University in short floating-point arithmetic (24 binary digits in the mantissa).

Since the function $f(x, y, v)\equiv 2\cdot v^3$ (cf. (5.1), (5.32)) violates condition (5.2), strictly speaking the theory of the preceding sections does not apply directly to (5.29), (5.30) as used in solving (5.32). However, by an argument analogous to part 1 of the proof of Theorem 9 it follows that the results derived above are still valid for (5.32) provided $\|\delta\|$, $\|\xi\|$ and $\|\eta\|$ are small enough.

The numerical results listed in the table (and further results not listed here) clearly confirm the theory described above.

Acknowledgement. I wish to thank Mr. L. H. Deckers and Mr. D. W. Kuilman who wrote the programs for the computations reported in Section 4.4 and 5.6.

References

1. Ceschino, F., Kuntzmann, J.: Problèmes différentiels de conditions initiales. Paris: Dunod 1963.
2. Forsythe, G. E., Wasow, W. R.: Finite-difference methods for partial differential equations. New York: J. Wiley & Sons 1960.
3. Godunov, S. K., Ryabenki, V. S.: Theory of difference schemes. Amsterdam: North-Holland Publishing Company 1964.
4. Gragg, W. B., Stetter, H. J.: Generalized multistep predictor-corrector methods. J. Assoc. Comput. Mach. **11**, 188–209 (1964).
5. Henrici, P.: Discrete variable methods in ordinary differential equations. New York: J. Wiley & Sons 1962.
6. Hull, T. E., Luxemburg, W. A. J.: Numerical methods and existence theorems for ordinary differential equations. Numer. Math. **2**, 30–41 (1960).
7. Isaacson, E., Keller, H. B.: Analysis of numerical methods. New York: J. Wiley & Sons 1966.
8. Lees, M.: Discrete methods for nonlinear two-point boundary value problems. In: Numerical solution of partial differential equations, e. d. J. H. Bramble. New York: Academic Press 1966.
9. Metté, A.: Essai de résolution du problème de Goursat par la methode de Runge-Kutta pour une equation aux dérivées partielles du type hyperbolique. Rev. Francaise Informat. Recherche Opérationnelle **1**, 67–90 (1967).

10. Spijker, M. N.: Convergence and stability of step-by-step methods for the numerical solution of initial-value problems. Numer. Math. **8**, 161–177 (1966).
11. — Stability and convergence of finite-difference methods (Thesis). Leiden University 1968.
12. — Reduction of round-off error by splitting of difference formulae. J. Soc. Indust. Appl. Math. Ser. B. Num. Anal., to appear.
13. Stetter, H. J.: A study of strong and weak stability in discretization algorithms. J. Soc. Indust. Appl. Math. Ser. B. Numer. Anal. **2**, 265–280 (1965).
14. — Stability of nonlinear discretization algorithms. In: Numerical solution of partial differential equations, ed. J. H. Bramble. New York: Academic Press 1966.
15. — Asymptotic expansions for the error of discretization algorithms for non-linear functional equations. Numer. Math. **7**, 18–31 (1965).
16. — Instability and non-monotonicity phenomena in discretizations to boundary-value problems. Numer. Math. **12**, 139–145 (1968).
17. — Törnig, W.: General multistep finite difference methods for the solution of $u_{xy}=f(x, y, u, u_x, u_y)$. Rend. Circ. Mat. Palermo **12**, 281–298 (1963).
18. Törnig, W.: Zur numerischen Behandlung von Anfangswertproblemen partieller hyperbolischer Differentialgleichungen zweiter Ordnung in zwei unabhängigen Veränderlichen. Arch. Rat. Mech. Anal. **4**, 428–466 (1960).

Dr. M. N. Spijker
Centraal Reken-Instituut
Rijksuniversiteit Leiden
Stationsplein 20
Leiden
The Netherlands