Two Theorems on the Accuracy of Numerical Solutions of Systems of Ordinary Differential Equations

By I. W. SANDBERG

(Manuscript received March 13, 1967)

We consider the accuracy with which a numerical solution of the system of ordinary differential equations

$$\dot{x} + f(x, t) = 0, \qquad t \geqq 0$$

can be obtained by the use of a numerical integration formula of the well-known type

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h \sum_{k=-1}^{p} b_k y'_{n-k} .$$

For the scalar case, under some natural assumptions, and assuming that α *and* β *are real constants such that*

$$\alpha \leqq \frac{\partial f(x, t)}{\partial x} \leqq \beta, \qquad t \geqq 0$$

at every point x, *it is proved that if*

$$F(z) \triangleq 1 - \sum_{k=0}^{p} a_k z^{-(k+1)} + \tfrac{1}{2}(\alpha + \beta)h \sum_{k=-1}^{p} b_k z^{-(k+1)} \neq 0$$

for all $| z | \geqq 1$, *then* $\langle e \rangle$, *the root-mean-squared error over a given interval, between the true samples of* $x(t)$ *and the* y_n, *satisfies*

$$\langle e \rangle \geqq (1 + \rho)^{-1} \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \varphi \rangle$$

in which ρ *depends on* α, β, *the* a_k, *and the* b_k, *and* $\langle \varphi \rangle$ *takes into account the local roundoff and truncation errors as well as errors in the starting values for computing the* y_n.

If the condition on $F(z)$ *stated above holds, and if* $\rho < 1$, *then*

$$\langle e \rangle \leqq (1 - \rho)^{-1} \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \varphi \rangle.$$

The significance of the key assumptions is discussed and two examples are given.

I. INTRODUCTION

In this paper we present some results concerning the accuracy with which a numerical solution of the system of ordinary differential equations

$$\dot{x} + f(x, t) = 0, \qquad t \geqq 0 \quad [x(0) = x_0] \tag{1}$$

can be obtained by the use of a numerical integration formula of the well-known type[1]

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h \sum_{k=-1}^{p} b_k y'_{n-k}, \qquad n \geqq p. \tag{2}$$

In (2) the y_n are approximations to the $x_n \triangleq x(nh)$, where h, a positive number, is the step size parameter; $y_0, y_1, \cdots, y_p$ are starting vectors, the last p of which are obtained by an independent method; and

$$y'_n \triangleq -f(y_n, nh).$$

If $b_{-1} \neq 0$, then y_{n+1} is defined implicitly, and (2) is said to be of closed type. It is assumed throughout that (2) can be solved* for y_{n+1} for all $n \geqq p$. Specializations of (2) include, for example, Euler's method:

$$y_{n+1} = y_n + hy'_n,$$

and the more useful formula

$$y_{n+1} = y_n + \tfrac{1}{2}h(y'_n + y'_{n+1}).$$

It is assumed throughout that for $t \geqq 0$, $f(x, t)$ is a well-defined real N-vector-valued function defined in the set of all real N-vectors x, that $f(x, t)$ satisfies (the usual weak) conditions which guarantee the existence and uniqueness of a solution of (1), and that the Jacobian matrix $\partial f(x, t)/\partial x$ exists for all x and all $t \geqq 0$.

Equation (2) ignores the roundoff error R_n introduced in calculating y_{n+1}, and, in order to take R_n into account, we shall consider instead

* It is well known that if f satisfies a uniform Lipshitz condition, and if h is sufficiently small, then (2) possesses a unique solution y_{n+1} which can be obtained by a simple iterative process.[1,2] However, this smallness condition is by no means always necessary. For example, if $b_{-1} > 0$ and, with α as defined in Section 2.3, if $\alpha > 0$, then for *any* $h > 0$ a unique solution y_{n+1} exists and can be computed by an iterative process which is only slightly more complicated than the usual one (see Section IV).

of (2):

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h \sum_{k=-1}^{p} b_k y'_{n-k} + R_n , \qquad n \geqq p. \tag{3}$$

We may also assume that R_n takes into account the error in solving (2) for y_{n+1}, caused typically by truncating an iteration procedure after a finite number of steps. Predictor-corrector techniques can of course be viewed as giving rise to a degenerate (often one-step) iteration technique in which the "starting point" is generated by the predictor.

The truncation error T_n, a basic entity associated with the integration formula (2) and the differential equation (1), is defined for $n \geqq p$ by the relation

$$x_{n+1} = \sum_{k=0}^{p} a_k x_{n-k} + h \sum_{k=-1}^{p} b_k x'_{n-k} + T_n , \qquad n \geqq p$$

in which $x'_n = -f(x_n, nh)$. Clearly, T_n is a measure of how well the samples $x_{n-p}, x_{n-p+1}, \cdots, x_{n+1}$ of the solution of (1) satisfy the integration formula. The problem of estimating T_n is a classical one, and there are standard methods which lead to precise bounds.[1,2]

We now define a set of vectors $\{\varphi_n\}$ which plays a central role in the subsequent discussion:

$$\varphi_n = T_n - R_n , \qquad n \geqq p \tag{5}$$

$$\varphi_n = (x_{n+1} - y_{n+1}) - \sum_{k=0}^{p} a_k(x_{n-k} - y_{n-k}) + h \sum_{k=-1}^{p} b_k\{f[x_{n-k}, (n-k)h] - f[y_{n-k}, (n-k)h]\}, \qquad n = -1, 0, \cdots, (p-1) \tag{6}$$

(with the understanding that $x_n = y_n = f(x_n, nh) = f(y_n, nh) = 0$ for $n < 0$). Note that the φ_n for $n = -1, 0, \cdots, (p-1)$ are measures of the departures of the starting vectors from the exact values, and that $\varphi_n = 0$ for $n = -1, 0, \cdots, (p-1)$ if the starting vectors are exact.

We shall describe our results first for the scalar case (i.e., for N = 1).

II. RESULTS

Let e_n denote $(x_n - y_n)$, the difference between $x(nh)$ and its computed approximation. Suppose that $N = 1$, and that α and β are real constants such that

$$\alpha \leqq \frac{\partial f(x,\, t)}{\partial x} \leqq \beta \tag{7}$$

for all t $\geqq 0$ (at every point x), and that

$$F(z) \triangleq 1 - \sum_{k=0}^{p} a_k z^{-(k+1)} + \tfrac{1}{2}(\alpha + \beta)h \sum_{k=-1}^{p} b_k z^{-(k+1)} \neq 0 \tag{8}$$

for all $| z | \geqq 1$ (including "$z = \infty$"). We prove that then

$$\langle e \rangle \triangleq \left((M + 1)^{-1} \sum_{m=0}^{M} | e_m |^2 \right)^{\frac{1}{2}}, \tag{9}$$

the root-mean-squared value of the first $(M + 1)$ error terms [M is an *arbitrary* positive integer greater than or equal to $(p + 1)$] is bounded from *below* in terms of

$$\langle \varphi \rangle \triangleq \left((M + 1)^{-1} \sum_{m=0}^{M} | \varphi_{m-1} |^2 \right)^{\frac{1}{2}}, \tag{10}$$

the corresponding quantity for the φ's, in accordance with the inequality

$$\langle e \rangle \geqq (1 + \rho)^{-1} \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \varphi \rangle. \tag{11}$$

[$F(z)$ is defined in (8)], in which

$$\rho \triangleq \tfrac{1}{2}(\beta - \alpha)h \max_{0 \leqq \omega \leqq 2\pi} \left| \frac{\sum_{k=-1}^{p} b_k e^{-i(k+1)\omega}}{F(e^{i\omega})} \right| . \tag{12}$$

We also prove that if in addition to the assumptions stated above, we have $\rho < 1$, then the sequence $\{e_n\}$ is bounded (i.e., there exists a positive constant c such that $| e_n | \leqq c$ for all n $\geqq 0$) whenever the sequence $\{\varphi_{n-1}\}$ is bounded, and

$$\langle e \rangle \leqq (1 - \rho)^{-1} \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \varphi \rangle \tag{13}$$

(whether or not $\{\varphi_{n-1}\}$ is bounded).

Inequality (11) provides a limitation on obtainable accuracy under essentially only the weak assumption that the sequence of approximations $\{y_n\}$ defined by (2) approaches zero as $n \to \infty$ for all sets of starting values when $f(x, t) = \frac{1}{2}(\alpha + \beta)x$. Since, by assumption, $F(e^{i\omega}) \neq 0$ for $0 \leqq \omega \leqq 2\pi$, it is clear that $\rho < \infty$.

The condition that $\rho < 1$ is satisfied if and only if the locus of

$$\Theta(\omega) \triangleq \frac{\sum_{k=0}^{p} a_k e^{ik\omega} - e^{-i\omega}}{\sum_{k=-1}^{p} b_k e^{ik\omega}} \tag{14}$$

for $0 \leqq \omega \leqq 2\pi$ lies outside the "critical circle" C of radius $\frac{1}{2}(\beta - \alpha)h$ centered in the complex plane at $[\frac{1}{2}(\alpha+\beta)h, 0]$ (see Fig. 1).*

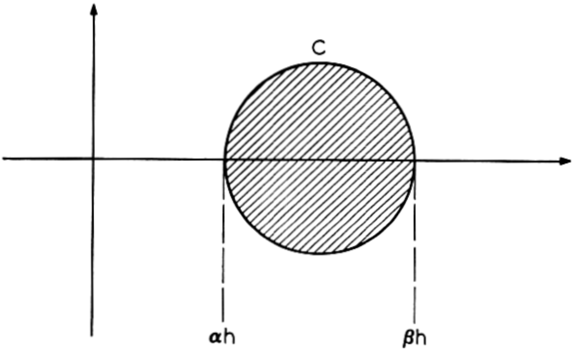

Fig. 1 — Location of the critical circle C (for $N = 1$).

Since

$$\rho = \tfrac{1}{2}(\beta - \alpha)h\{\min_{\omega} \mid \Theta(\omega) - \tfrac{1}{2}(\alpha + \beta)h \mid\}^{-1}, \tag{15}$$

we see that ρ is the ratio of the radius of C to the distance between c and θ, where c is the center of C and θ is a point nearest c on the locus of $\Theta(\omega)$.

2.1 *Discussion*

The quantity $\langle e \rangle$ of course of interest in problems in which we are concerned with a measure of the accuracy of a numerical solution

* If $\alpha > 0$, we can express both the conditions that $\rho < 1$ and $F(z) \neq 0$ for $|z| \geqq 1$ entirely in terms of a condition on the locus of $\Theta(\omega)^{-1}$ for $\omega \epsilon [0, 2\pi]$. The requirements on $F(z)$ and ρ are met if the disk of radius $\frac{1}{2}[(\alpha h)^{-1} - (\beta h)^{-1}]$ centered at $\{\frac{1}{2}[(\alpha h)^{-1} + (\beta h)^{-1}], 0\}$ is not "encircled" or intersected by the locus of $\Theta(\omega)^{-1}$. There is a complication that arises as a result of the fact that $\Theta(\omega)^{-1}$ is typically not bounded. This complication often stems from a "consistency requirement" which implies that $1 - \sum^{p}{}_{k=0} a_k z^{-(k+1)}$ has at least one zero on the unit circle. However, due to also typical "convergence requirements," $1 - \sum^{p}{}_{k=0} a_k z^{-(k+1)}$ normally has only *simple* zeros on the unit circle, a fact that can be used to suitably define what is meant by the locus of $\Theta(\omega)^{-1}$ not encircling the disk. We leave the details of the necessary "principle of the argument" argument to the sufficiently interested reader.

over a large number of steps, as opposed to the accuracy of some final value obtained at the end of a large number of steps.

Although there is a vast and interesting literature concerned with various aspects of the problem of error estimation in digital computation (see, for example, Refs. 3, 4, and 5), the results presented above, and their proofs, appear to be most closely related to earlier results concerning the input-output stability of continuous-time nonlinear feedback systems.[6,*] Indeed, the writer is not aware of any lower-bound results in the numerical analysis literature of the type described above. However, some upper bounds concerning (2) of (for example) the form $|e_n| \leqq K$ with K independent of n (which imply $\langle e \rangle \leqq K$) have been obtained in certain cases. In this connection, our condition that guarantees the boundedness of $\{e_n\}$ is often weaker, and our upper bounds on $\langle e \rangle$ are often much stronger, because, for example, the φ_{m-1} can become very small as m becomes large.

Our approach can be applied to several other problems in numerical analysis. In particular, with reasonably direct modifications of our proofs, analogous theorems can be proved concerning the numerical integration of systems of second-order ordinary differential equations.

2.2 *Examples*

Euler's Method: $y_{n+1} = y_n + hy'_n$

Here $F(z) = 1 - [1 - \frac{1}{2}(\alpha + \beta)h]z^{-1}$, so that $F(z) \neq 0$ for $|z| \geqq 1$ if and only if $0 < \frac{1}{2}(\alpha + \beta)h < 2$, and $|F(e^{i\omega})| = |1 - [1 - \frac{1}{2}(\alpha + \beta)h]e^{-i\omega}|$. Thus (with $0 < \frac{1}{2}(\alpha + \beta)h < 2$),

$$\min_{\omega} |F(e^{i\omega})|^{-1} = [1 + |1 - \tfrac{1}{2}(\alpha + \beta)h|]^{-1}$$

$$\max_{\omega} |F(e^{i\omega})|^{-1} = [1 - |1 - \tfrac{1}{2}(\alpha + \beta)h|]^{-1}.$$

The locus of Θ is the circle shown in Fig. 2, since $\Theta(\omega) = 1 - e^{-i\omega}$. If $\alpha h > 0$ and $\beta h < 2$, then the critical disk (Fig. 2) is not intersected by the locus of Θ, the condition that $0 < \frac{1}{2}(\alpha + \beta)h < 2$ is satisfied, $\rho < 1$, and in accordance with the last paragraph of the section preceding Section 2.1:

$$\rho = \tfrac{1}{2}(\beta - \alpha)h \max\left([\tfrac{1}{2}(\beta + \alpha)h]^{-1}, [2 - \tfrac{1}{2}(\beta + \alpha)h]^{-1}\right).$$

[*] It is interesting to note that the possibility of exploiting feedback-theoretic ideas has been emphasized by Hamming.[2]

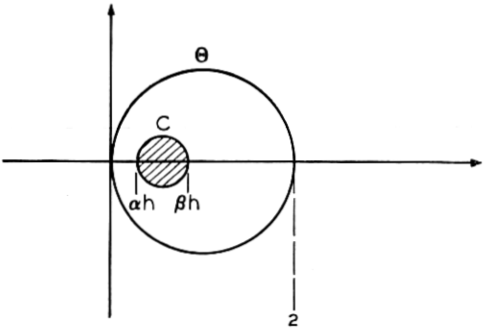

Fig. 2 — The locus of $\Theta(\omega)$ for Euler's method, and the critical circle C.

If $0 < (\alpha + \beta)h < 2$, then

$$\min_{\omega} | F(e^{i\omega}) |^{-1} = [2 - \tfrac{1}{2}(\alpha + \beta)h]^{-1},$$

$$\max_{\omega} | F(e^{i\omega}) |^{-1} = [\tfrac{1}{2}(\alpha + \beta)h]^{-1},$$

$$\rho = \frac{\beta - \alpha}{\beta + \alpha},$$

and

$$\langle e \rangle \geqq [1 + (\beta - \alpha)(\beta + \alpha)^{-1}]^{-1}[2 - \tfrac{1}{2}(\alpha + \beta)h]^{-1}\langle \varphi \rangle$$

$$\langle e \rangle \leqq [1 - (\beta - \alpha)(\beta + \alpha)^{-1}]^{-1}[\tfrac{1}{2}(\alpha + \beta)h]^{-1}\langle \varphi \rangle.$$

For estimates of T_n, see Ref. 1 or 2.

As a remark concerning the necessity of the condition $\rho < 1$, we note that if $\alpha h > 0$, but $\beta h > 2$, then for even the special case in which $f(x, t) = \beta x$, we have $e_0, e_1, e_2, \cdots$ unbounded, since $y_0, y_1, y_2, \cdots$ is unbounded (assuming merely that $y_0 \neq 0$).

The Formula $y_{n+1} = y_n + \tfrac{1}{2}h(y'_n + y'_{n+1})$:

In this important case

$$F(z) = 1 + \tfrac{1}{4}(\alpha + \beta)h - [1 - \tfrac{1}{4}(\alpha + \beta)h]z^{-1}, \qquad \text{and}$$

$$\Theta(\omega) = \frac{1 - e^{-i\omega}}{\tfrac{1}{2}(1 + e^{-i\omega})} = 2i \tan\left(\frac{\omega}{2}\right).$$

We have $F(z) \neq 0$ for $|z| \geqq 1$ if and only if $(\alpha + \beta)h > 0$, and [assuming that $(\alpha + \beta)h > 0$]:

$$\min_{\omega} \mid F(e^{i\omega}) \mid^{-1} = [1 + \tfrac{1}{4}(\alpha + \beta)h + \mid 1 - \tfrac{1}{4}(\alpha + \beta)h \mid]^{-1}$$

$$\max_{\omega} \mid F(e^{i\omega}) \mid^{-1} = [1 + \tfrac{1}{4}(\alpha + \beta)h - \mid 1 - \tfrac{1}{4}(\alpha + \beta)h \mid]^{-1}.$$

The locus of Θ lies entirely on the imaginary axis of the complex plane,

$$\rho = \frac{\beta - \alpha}{\beta + \alpha},$$

and obviously $\rho < 1$ if $\alpha > 0$.

If $\alpha > 0$ and $(\alpha + \beta)h < 4$, then

$$\min_{\omega} \mid F(e^{i\omega}) \mid^{-1} = \tfrac{1}{2}$$

$$\max_{\omega} \mid F(e^{i\omega}) \mid^{-1} = [\tfrac{1}{2}(\alpha + \beta)h]^{-1}$$

and

$$\langle e \rangle \geqq [1 + (\beta - \alpha)(\beta + \alpha)^{-1}]^{-1}\tfrac{1}{2}\langle \varphi \rangle$$

$$\langle e \rangle \leqq [1 - (\beta - \alpha)(\beta + \alpha)^{-1}]^{-1}[\tfrac{1}{2}(\alpha + \beta)h]^{-1}\langle \varphi \rangle.$$

The last inequality can be written as simply

$$\langle e \rangle \leqq (\alpha h)^{-1}\langle \varphi \rangle.$$

For the integration formula under consideration,

$$T_n = \frac{h^3 x'''(\eta_n)}{12}$$

where η_n lies in the interval $[(n - p)h, (n + 1)h]$.

Here $p = 0$, and

$$\varphi_{-1} = (x_0 - y_0) + \tfrac{1}{2}h[f(x_0, 0) - f(y_0, 0)].$$

Thus, assuming for the purpose of illustration that roundoff errors can be neglected:

$$\langle e \rangle \leqq (\alpha h)^{-1}\left((M + 1)^{-1} \mid \varphi_{-1} \mid^2 + (M + 1)^{-1} \sum_{m=1}^{M} \left| \frac{h^3 x'''(\eta_m)}{12} \right|\right)^{\frac{1}{2}}$$

provided that $\alpha > 0$ and $(\alpha + \beta)h < 4$. If, for example, $(\beta - \alpha) \cdot (\beta + \alpha)^{-1} = \frac{1}{2}$ and $\frac{1}{2}(\alpha + \beta)h = \frac{1}{3}$, then the ratio of our upper bound on $\langle e \rangle$ to our lower bound is 18. If $(\beta - \alpha)(\beta + \alpha)^{-1} = \frac{3}{4}$ and $\frac{1}{2}(\alpha + \beta)h = \frac{1}{3}$, then the ratio is 42.

2.3 *Results for the Vector Case* ($N \geqq 1$)

We shall state our results for $N \geqq 1$ in a slightly more formal fashion.

Definitions:

(*i*) Let $\| q \|$ denote $(\sum_{k=1}^{N} q_k^2)^{\frac{1}{2}}$ for every real N-vector $q = (q_1, q_2, \cdots, q_N)$.

(*ii*) Let $\{\partial f(x, t)/\partial x\}_S$ and $\{\partial f(x, t)/\partial x\}_A$ denote, respectively, the symmetric and antisymmetric (i.e., skew symmetric) part of $\partial f(x, t)/\partial x$, the Jacobian matrix of $f(x, t)$.

(*iii*) Let $F(z) \triangleq 1 - \sum_{k=0}^{p} a_k z^{-(k+1)} + \frac{1}{2}(\alpha + \beta)h \sum_{k=-1}^{p} b_k z^{-(k+1)}$

(*iv*) $e_n \triangleq x(nh) - y_n$, $n \geqq 0$ with the y_n for $n \geqq (p + 1)$ defined by (3).

(*v*) With M an arbitrary positive integer such that $M \geqq (p + 1)$, let

$$\langle e \rangle \triangleq \left((M + 1)^{-1} \sum_{m=0}^{M} \| e_m \|^2\right)^{\frac{1}{2}}$$

$$\langle \varphi \rangle \triangleq \left((M + 1)^{-1} \sum_{m=0}^{M} \| \varphi_{m-1} \|^2\right)^{\frac{1}{2}},$$

where the φ_{m-1} are defined in (5) and (6).

Assumptions:

Let the smallest eigenvalue of $\{\partial f(x, t)/\partial x\}_S$ be bounded from below by the real constant $\alpha(\alpha > -\infty)$ for all $t \geqq 0$, and let the largest eigenvalue of $\{\partial f(x, t)/\partial x\}_S$ be bounded from above by the real constant $\beta(\beta < \infty)$ for all $t \geqq 0$. Let the modulus of the largest eigenvalue of $\{\partial f(x, t)/\partial x\}_A$ be bounded from above by the real constant $\gamma(\gamma < \infty)$ for all $t \geqq 0$.

Definition:

Let

$$\rho \triangleq [\tfrac{1}{2}(\beta - \alpha)h + \gamma h] \cdot \max_{0 \leqq \omega \leqq 2\pi} \left| \frac{\sum_{k=-1}^{p} b_k e^{-i(k+1)\omega}}{1 - \sum_{k=0}^{p} a_k e^{-i(k+1)\omega} + \frac{1}{2}(\alpha + \beta)h \sum_{k=-1}^{p} b_k e^{-i(k+1)\omega}} \right| .$$

Theorem 1: If

(*i*) *the assumptions of Section I concerning* $f(x, t)$ *and* (2) *are satisfied,*

(*ii*) $1 + \frac{1}{2}(\alpha + \beta)hb_{-1} \neq 0$,

(*iii*) $F(z) \neq 0$ for all $| z | \geqq 1$, *then*

$$\langle e \rangle \geqq (1 + \rho)^{-1} \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \varphi \rangle.$$

Theorem 2: If (i), (ii), and (iii) of Theorem 1 are satisfied, and if $\rho < 1$, then

$$(i) \qquad \langle e \rangle \leqq (1 - \rho)^{-1} \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \varphi \rangle,$$

and

$$(ii) \qquad \sup_{m \geqq 0} \| \varphi_{m-1} \| < \infty \quad \text{implies} \quad \sup_{m \geqq 0} \| e_m \| < \infty .$$

Corollary to Theorem 1: If (i) of Theorem 1 is satisfied and there exists at least one real constant k_1 such that

$$1 - \sum_{k=0}^{p} a_k z^{-(k+1)} + k_1 h \sum_{k=-1}^{p} b_k z^{-(k+1)} \neq 0$$

for all $| z | \geqq 1$, then there exists a positive constant k_2, which depends only on $a_0, a_1, \cdots, a_p, b_{-1}, b_0, \cdots, b_p, \alpha, \beta$, and γ such that

$$\langle e \rangle \geqq k_2 \langle \varphi \rangle.$$

Theorems 1 and 2 are proved* in the following section. The proof of the corollary is very simple.

Since

$$1 - \sum_{k=0}^{p} a_k z^{-(k+1)} + k_1 h \sum_{k=-1}^{p} b_k z^{-(k+1)} \neq 0$$

for all $| z | \geqq 1$, there exists a k_1' such that

$$1 - \sum_{k=0}^{p} a_k z^{-(k+1)} + k_1' h \sum_{k=-1}^{p} b_k z^{-(k+1)} \neq 0$$

for all $| z | \geqq 1$, and

$$1 + k_1' h b_{-1} \neq 0.$$

Choose α' and β' such that $\alpha' \leqq \alpha$, $\beta' \geqq \beta$, and $\frac{1}{2}(\alpha' + \beta') = k_1'$. If we replace α and β with α' and β', respectively, we see that Theorem 1 applies.

* Our proofs actually yield sharper, but less explicit, bounds on $\langle e \rangle$ than those of Theorems 1 and 2. See (31) and (37).

III. PROOFS

Proof of Theorem 1:

We have

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h \sum_{k=-1}^{p} b_k y'_{n-k} + R_n , \qquad n \geqq p$$

and

$$x_{n+1} = \sum_{k=0}^{p} a_k x_{n-k} + h \sum_{k=-1}^{p} b_k x'_{n-k} + T_n , \qquad n \geqq p.$$

Thus,

$$\begin{aligned} e_{n+1} = {} & \sum_{k=0}^{p} a_k e_{n-k} \\ & - h \sum_{k=-1}^{p} b_k \{ f[x_{n-k}(n - k)h] - f[y_{n-k} , (n - k)h] \} + \varphi_n , \quad n \geqq p \end{aligned}$$

and, with φ_n defined for $n = -1, 0, \cdots, (p - 1)$ by (6),

$$\begin{aligned} e_n = \sum_{k=0}^{p} a_k e_{n-1-k} - h \sum_{k=-1}^{p} b_k \{ & f[x_{n-1-k} , (n - 1 - k)h] \\ & - f[y_{n-1-k} , (n - 1 - k)h] \} + \varphi_{n-1} \end{aligned} \tag{16}$$

for $n \geqq 0$.

As a matter of convenience we define $a_{-1} \triangleq 0$.

Lemma 1: There exist real sequences $\{w_k\}_{k=0}^{\infty}$ and $\{v_k\}_{k=0}^{\infty}$ both belonging to l_1 (i.e., with the property that $\sum_{k=0}^{\infty} | w_k | < \infty$ and $\sum_{k=0}^{\infty} | v_k | < \infty$) such that

$$W(z) \triangleq \sum_{k=0}^{\infty} w_k z^{-k} = \frac{-h \sum_{k=-1}^{p} b_k z^{-(k+1)}}{1 - \sum_{k=-1}^{p} [a_k - \frac{1}{2}(\alpha + \beta) h b_k] z^{-(k+1)}} \tag{17}$$

$$V(z) \triangleq \sum_{k=0}^{\infty} v_k z^{-k} = \frac{1}{1 - \sum_{k=-1}^{p} [a_k - \frac{1}{2}(\alpha + \beta) h b_k] z^{-(k+1)}} \tag{18}$$

for all $| z | \geqq 1$.

The proof follows at once from the standard theory of z-transforms, in view of assumptions (*ii*) and (*iii*) of Theorem 1. The details are omitted.

Lemma 2: *Let* $\delta_n \triangleq f(x_n, nh) - f(y_n, nh) - \frac{1}{2}(\alpha + \beta)(x_n - y_n)$ *for all* $n \geqq 0$, *and let* $\{w_k\}$ *and* $\{v_k\}$ *be as described in Lemma 1. Then*

$$e_n = \sum_{k=0}^{n} w_{n-k}\,\delta_k + \sum_{k=0}^{n} v_{n-k}\varphi_{k-1} \tag{19}$$

for $n = 0, 1, \cdots, M$.

Proof of Lemma 2:

From (16) and our definition of δ_n:

$$\begin{aligned} e_n = \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta)hb_k]e_{n-k-1} \\ - h \sum_{k=-1}^{p} b_k\,\delta_{n-k-1} + \varphi_{n-1}, \qquad n \geqq 0. \end{aligned} \tag{20}$$

We multiply both sides of (20) by $e^{-in\omega}$ and then sum from $n = 0$ to $n = M$ to obtain

$$\begin{aligned} \sum_{n=0}^{M} e^{-in\omega}e_n = \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta)hb_k] \sum_{n=0}^{M} e^{-in\omega}e_{n-k-1} \\ - h \sum_{k=-1}^{p} b_k \sum_{n=0}^{M} e^{-in\omega}\,\delta_{n-k-1} + \sum_{n=0}^{M} e^{-in\omega}\varphi_{n-1} \end{aligned} \tag{21}$$

for all $\omega \in [0, 2\pi]$. Using $e_n = \delta_n = 0$ for $n < 0$, we have

$$\sum_{n=0}^{M} e^{-in\omega}e_{n-k-1} = e^{-i(1+k)\omega} \sum_{n=0}^{M} e^{-in\omega}e_n - e^{-i(1+k)\omega} \sum_{n=M-k}^{M} e^{-in\omega}e_n \tag{22}$$

and

$$\sum_{n=0}^{M} e^{-in\omega}\,\delta_{n-k-1} = e^{-i(1+k)\omega} \sum_{n=0}^{M} e^{-in\omega}\,\delta_n - e^{-i(1+k)\omega} \sum_{n=M-k}^{M} e^{-in\omega}\,\delta_n\,. \tag{23}$$

Thus, from (21), (22), and (23)

$$\begin{aligned} \sum_{n=0}^{M} e^{-in\omega}e_n = W(e^{i\omega}) \sum_{n=0}^{M} e^{-in\omega}\,\delta_n + V(e^{i\omega}) \sum_{n=0}^{M} e^{-in\omega}\varphi_{n-1} \\ + V(e^{i\omega})\Big\{h \sum_{k=-1}^{p} b_k e^{-i(1+k)\omega} \sum_{n=M-k}^{M} e^{-in\omega}\,\delta_n \\ - \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta)hb_k]e^{-i(1+k)\omega} \sum_{n=M-k}^{M} e^{-in\omega}e_n\Big\} \end{aligned} \tag{24}$$

for all $\omega \in [0, 2\pi]$.

The expression within the braces in (24) can be written as

$$\sum_{n=0}^{\infty} s_n e^{-in\omega}$$

with $s_n = 0$ for $n = 0, 1, \cdots, M$ and for $n > (1 + M + p)$. Since $\{v_k\} \varepsilon\, l_1$, we have

$$V(e^{i\omega}) \sum_{n=0}^{\infty} s_n e^{-in\omega} = \sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} v_{n-k} s_k .$$

Similarly,

$$V(e^{i\omega}) \sum_{n=0}^{M} e^{-in\omega} \varphi_{n-1} = \sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} v_{n-k} (\varphi_{k-1})_M$$

in which

$$\begin{aligned} (\varphi_{k-1})_M &= \varphi_{k-1} , && k \leqq M \\ &= 0 && k > M \end{aligned}$$

and finally, since $\{w_k\} \epsilon\, l_1$,

$$W(e^{i\omega}) \sum_{n=0}^{M} e^{-in\omega}\ \delta_n = \sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} w_{n-k} (\delta_k)_M ,$$

where

$$\begin{aligned} (\delta_k)_M &= \delta_k , && k \leqq M \\ &= 0, && k > M . \end{aligned}$$

Thus,

$$\begin{aligned} \sum_{n=0}^{M} e^{-in\omega} e_n = {} & \sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} w_{n-k} (\delta_k)_M \\ & + \sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} v_{n-k} (\varphi_{k-1})_M \\ & + \sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} v_{n-k} s_k \end{aligned}$$

for all $\omega \epsilon [0, 2\pi]$. Since

$$\sum_{k=0}^{n} v_{n-k} s_k = 0$$

for $n = 0, 1, \cdots, M$ we have*

$$e_n = \sum_{k=0}^{n} w_{n-k}\,\delta_k + \sum_{k=0}^{n} v_{n-k}\varphi_{k-1} \tag{25}$$

for $n = 0, 1, \cdots, M$. This completes the proof of Lemma 2.

Lemma 3: If (19) holds for $n = 0, 1, \cdots, M$, then

$$\left(\sum_{n=0}^{M} \left\| \sum_{k=0}^{n} v_{n-k}\varphi_{k-1} \right\|^2\right)^{\frac{1}{2}} \leqq \left(\sum_{n=0}^{M} \| e_n \|^2\right)^{\frac{1}{2}} + \left(\sum_{n=0}^{M} \left\| \sum_{k=0}^{n} w_{n-k}\,\delta_k \right\|^2\right)^{\frac{1}{2}} \tag{26}$$

and

$$\left(\sum_{n=0}^{M} \| e_n \|^2\right)^{\frac{1}{2}} \leqq \left(\sum_{n=0}^{M} \left\| \sum_{k=0}^{n} w_{n-k}\,\delta_k \right\|^2\right)^{\frac{1}{2}} + \left(\sum_{n=0}^{M} \left\| \sum_{k=0}^{n} v_{n-k}\varphi_{k-1} \right\|^2\right)^{\frac{1}{2}}. \tag{27}$$

Proof of Lemma 3:

Inequality (26) or (27) follows from (19) by two applications of Minkowski's inequality. We leave the details to the reader.

Inequality (27) is used only in the proof of Theorem 2.

Lemma 4:

$$\sum_{n=0}^{M} \left\| \sum_{k=0}^{n} w_{n-k}\,\delta_k \right\|^2 \leqq \max_{0 \leqq \omega \leqq 2\pi} | W(e^{i\omega}) |^2 \sum_{n=0}^{M} \| \delta_n \|^2.$$

Proof of Lemma 4:

By Parseval's identity,

$$\begin{aligned}\sum_{n=0}^{M} \left\| \sum_{k=0}^{n} w_{n-k}\,\delta_k \right\|^2 &= \frac{1}{2\pi}\int_0^{2\pi} \left\| \sum_{n=0}^{M} e^{-i\omega n} \sum_{k=0}^{n} w_{n-k}\,\delta_k \right\|^2 d\omega \\ &= \frac{1}{2\pi}\int_0^{2\pi} \left\| \sum_{n=0}^{M} e^{-i\omega n} \sum_{k=0}^{n} w_{n-k}(\delta_k)_M \right\|^2 d\omega\end{aligned}$$

in which

$$\begin{aligned}(\delta_k)_M &= \delta_k\,, \qquad k \leqq M \\ &= 0, \qquad k > M.\end{aligned}$$

Therefore,

$$\sum_{n=0}^{M} \left\| \sum_{k=0}^{n} w_{n-k}\,\delta_k \right\|^2 \leqq \frac{1}{2\pi}\int_0^{2\pi} \left\| \sum_{n=0}^{\infty} e^{-i\omega n} \sum_{k=0}^{n} w_{n-k}(\delta_k)_M \right\|^2 d\omega.$$

* We could have obtained (25) from (20) directly using standard z-transform theory, if we had introduced further assumptions which guarantee that the sequences $\{y_n\}$, $\{x(nh)\}$, and $\{\varphi_{k-1}\}$ are transformable. However, this would have complicated the statement of our results and would have weakened them in a nontrivial manner.

But since $\{w_k\} \epsilon\, l_1$,

$$\sum_{n=0}^{\infty} e^{-i\omega n} \sum_{k=0}^{n} w_{n-k}(\delta_k)_M = \sum_{n=0}^{\infty} e^{-i\omega n} w_n \sum_{k=0}^{M} e^{-i\omega k}\, \delta_k$$

$$= W(e^{i\omega}) \sum_{k=0}^{M} e^{-i\omega k}\, \delta_k \,.$$

Thus,

$$\sum_{n=0}^{M} \left\| \sum_{k=0}^{n} w_{n-k}\, \delta_k \right\|^2 \leqq \frac{1}{2\pi} \int_0^{2\pi} \left\| W(e^{i\omega}) \sum_{k=0}^{M} e^{-i\omega k}\, \delta_k \right\|^2 d\omega$$

$$\leqq \max_{0 \leqq \omega \leqq 2\pi} |\, W(e^{i\omega})\, |^2 \frac{1}{2\pi} \int_0^{2\pi} \left\| \sum_{k=0}^{M} e^{-i\omega k}\, \delta_k \right\|^2 d\omega$$

$$\leqq \max_{0 \leqq \omega \leqq 2\pi} |\, W(e^{i\omega})\, |^2 \sum_{k=0}^{M} \|\, \delta_k \,\|^2$$

which proves Lemma 4.

Lemma 5:

$$\left(\sum_{n=0}^{M} \|\, \delta_n \,\|^2 \right)^{\frac{1}{2}} \leqq [\tfrac{1}{2}(\beta - \alpha) + \gamma] \left(\sum_{n=0}^{M} \|\, e_n \,\|^2 \right)^{\frac{1}{2}}.$$

Proof of Lemma 5:

We shall prove that

$$\|\, \delta_n \,\| \leqq [\tfrac{1}{2}(\beta - \alpha) + \gamma]\, \|\, e_n \,\|$$

for $n = 0, 1, \cdots, M$.

By definition

$$\|\, \delta_n \,\| = \|\, f(x_n\,, nh) - f(y_n\,, nh) - \tfrac{1}{2}(\alpha + \beta)(x_n - y_n)\, \|.$$

Let $q(a) = f[ax_n + (1 - a)y_n, nh]$ for $a \,\epsilon\, [0, 1]$. Then

$$\frac{\partial q}{\partial a} = f'[ax_n + (1 - a)y_n\,, nh](x_n - y_n)$$

in which $f'[ax_n + (1 - a)y_n\,, nh]$ denotes the Jacobian matrix of $f(x, t)$, evaluated at $x = ax_n + (1 - a)y_n\,, t = nh$. Now, since $q(1) - q(0) = f(x_n\,, nh) - f(y_n\,, nh)$, we have

$$\int_0^1 \frac{\partial q}{\partial a}\, da = f(x_n\,, nh) - f(y_n\,, nh).$$

Therefore,

$$\| \delta_n \| = \left\| \int_0^1 \{f'[ax_n + (1-a)y_n , nh] - (\alpha + \beta)1_N\} \, da(x_n - y_n) \right\|$$

in which 1_N denotes the identity matrix of order N.

For H an arbitrary $N \times N$ matrix, let $\| H \| \triangleq$ (largest eigenvalue of $H^*H)^{\frac{1}{2}}$, in which H^* denotes the complex-conjugate transpose of H (i.e., let $\| H \|$ denote the "spectral norm" of H). Then

$$\begin{aligned} \| \delta_n \| &\leqq \left\| \int_0^1 \{f'[ax_n + (1-a)y_n , nh] - \tfrac{1}{2}(\alpha + \beta)1_N\} \, da \right\| \\ &\qquad \cdot \| x_n - y_n \| \qquad\qquad (28) \\ &\leqq \int_0^1 \| f'[ax_n + (1-a)y_n , nh] - \tfrac{1}{2}(\alpha+\beta)1_N \| \, da \, \| e_n \|. \end{aligned}$$

With $\{f'\}s$ and $\{f'\}_A$, respectively, the symmetric and antisymmetric parts of f', we have

$$\| f' - \tfrac{1}{2}(\alpha + \beta)1_N \| \leqq \| \{f'\}_S - \tfrac{1}{2}(\alpha+\beta)1_N \| + \| \{f'\}_A \|.$$

For each $a \; \varepsilon \; [0, 1]$, there exists[7] an orthogonal matrix T_1 such that $T_1\{f'\}_S T_1^{-1} \triangleq D = \text{diag} \; (\zeta_1 , \zeta_2 , \cdots \zeta_N)$ in which, since ζ_j is an eigenvalue of $\{f'\}_S$,

$$\alpha \leqq \zeta_j \leqq \beta$$

for $j = 1, 2, \cdots , N$. Using $\| T_1 \| = \| T_1^{-1} \| = 1$,

$$\begin{aligned} \| \{f'\}_S - \tfrac{1}{2}(\alpha + \beta)1_N \| &= \| T_1^{-1} DT_1 - \tfrac{1}{2}(\alpha+\beta)T_1^{-1}T_1 \| \\ &\leqq \| T_1^{-1} D - \tfrac{1}{2}(\alpha+\beta)T_1^{-1} \| \cdot \| T_1 \| \\ &\leqq \| T_1^{-1} \| \cdot \| D - \tfrac{1}{2}(\alpha+\beta)1_N \| \cdot \| T_1 \| \\ &\leqq \max_j \mid \zeta_j - \tfrac{1}{2}(\alpha + \beta) \\ &\leqq \tfrac{1}{2}(\beta - \alpha). \qquad\qquad (29) \end{aligned}$$

Thus, $\| f' - \frac{1}{2}(\alpha + \beta)1_N \| \leqq \frac{1}{2}(\beta - \alpha) + \| \{f'\}_A \|$.

Consider $\| \{f'\}_A \|$. For each $a \; \varepsilon \; [0, 1]$ there exists[7] an orthogonal matrix T_2 such that $T_2\{f'\}_A T_2^{-1} \triangleq S$ is a direct sum of 2×2 block matrices of the form

$$B_j = \begin{bmatrix} 0 & b_j \\ -b_j & 0 \end{bmatrix}$$

and, if N is odd, and "1 × 1 matrix" containing the zero element. Clearly,

$$B_j^* B_j = \begin{bmatrix} b_j^2 & 0 \\ 0 & b_j^2 \end{bmatrix},$$

S^*S is a diagonal matrix, and its largest element is not greater than γ^2. That is,

$$\| \{f'\}_A \| = \| T_2^{-1} S T_2 \| \leqq \| S \| \leqq \gamma,$$

and consequently

$$\| f' - \tfrac{1}{2}(\alpha + \beta) 1_N \| \leqq \tfrac{1}{2}(\beta - \alpha) + \gamma \tag{30}$$

for all $a \, \varepsilon \, [0, 1]$. Finally, from (28) and (30)

$$\| \delta_n \| \leqq [\tfrac{1}{2}(\beta - \alpha) + \gamma] \, \| e_n \|.$$

At this point we have proved that with ρ as defined in Section 2.3

$$\left(\sum_{n=0}^{M} \left\| \sum_{k=0}^{n} v_{n-k} \varphi_{k-1} \right\|^2 \right)^{\frac{1}{2}} \leqq (1 + \rho) \left(\sum_{n=0}^{M} \| e_n \|^2 \right)^{\frac{1}{2}}. \tag{31}$$

We now need the following result.

Lemma 6:

$$\sum_{n=0}^{M} \left\| \sum_{k=0}^{n} v_{n-k} \varphi_{k-1} \right\|^2 \geqq \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-2} \sum_{k=0}^{M} \| \varphi_{k-1} \|^2.$$

Proof of Lemma 6:

Consider $S \triangleq \sum_{n=0}^{M} \| \sum_{k=0}^{n} v_{n-k} c_k \|^2$, with the c_k's N-vectors. Choose $c_{M+1}, c_{M+2}, \cdots$ so that

$$\sum_{k=0}^{n} v_{n-k} c_k = 0, \qquad n \geqq (M + 1).$$

This is possible since $v_0 \neq 0$ [$v_0 = \lim_{z \to \infty} V(z)$]. Then

$$\begin{aligned} S &= \sum_{n=0}^{\infty} \left\| \sum_{k=0}^{n} v_{n-k} c_k \right\|^2 \\ &= \frac{1}{2\pi} \int_0^{2\pi} \left\| \sum_{n=0}^{\infty} e^{-i\omega n} \sum_{k=0}^{n} v_{n-k} c_k \right\|^2 d\omega. \end{aligned} \tag{32}$$

Under the *assumption* that

$$\sum_{k=0}^{\infty} \| c_k \|^2 < \infty,$$

we can write

$$\sum_{n=0}^{\infty} e^{-i\omega n} \sum_{k=0}^{n} v_{n-k}c_k = \sum_{n=0}^{\infty} e^{-i\omega n} v_n \sum_{k=0}^{\infty} e^{-i\omega k} c_k \tag{33}$$

in which the last sum over k is interpreted as the usual limit in the mean. From (32) and (33)

$$\begin{aligned} S &= \frac{1}{2\pi}\int_0^{2\pi} \left\| \sum_{n=0}^{\infty} e^{-i\omega n} v_n \sum_{k=0}^{\infty} e^{-i\omega k} c_k \right\|^2 d\omega \\ &\geqq \min_{0\leqq\omega\leqq 2\pi} \mid F(e^{i\omega}) \mid^{-2} \frac{1}{2\pi}\int_0^{2\pi} \left\| \sum_{k=0}^{\infty} e^{-i\omega k} c_k \right\|^2 d\omega \\ &\geqq \min_{0\leqq\omega\leqq 2\pi} \mid F(e^{i\omega}) \mid^{-2} \sum_{k=0}^{\infty} \| c_k \|^2 \\ &\geqq \min_{0\leqq\omega\leqq 2\pi} \mid F(e^{i\omega}) \mid^{-2} \sum_{k=0}^{M} \| c_k \|^2 . \end{aligned} \tag{34}$$

We now *prove* that

$$\sum_{k=0}^{\infty} \| c_k \|^2 < \infty .$$

Let

$$q_n \triangleq \sum_{k=0}^{n} v_{n-k}c_k , \qquad n \geqq 0.$$

Of course: $q_n = 0$ for $n \geqq (M + 1)$.

Let K be an arbitrary positive integer. Then

$$\sum_{n=0}^{K} e^{-in\omega} \sum_{k=0}^{n} v_{n-k}c_k = \sum_{k=0}^{K} e^{-in\omega} q_n .$$

With

$$\begin{aligned} (c_k)_K &= c_k , \qquad & k \leqq K \\ &= 0, \qquad & k > K \end{aligned}$$

we obtain

$$\sum_{n=0}^{\infty} e^{-in\omega} \sum_{k=0}^{n} v_{n-k}(c_k)_K - \sum_{K+1}^{\infty} e^{-i\omega n} \sum_{k=0}^{n} v_{n-k}(c_k)_K = \sum_{n=0}^{K} e^{-in\omega} q_n .$$

Therefore,

$$\begin{aligned} \sum_{k=0}^{K} e^{-ik\omega} c_k = {} & \left[1 - \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta)hb_k] e^{-i(k+1)\omega} \right] \\ & \cdot \sum_{K+1}^{\infty} e^{-i\omega n} \sum_{k=0}^{n} v_{n-k}(c_k)_K + F(e^{i\omega}) \sum_{n=0}^{K} e^{-in\omega} q_n . \end{aligned} \tag{35}$$

The first term on the right side of (35) can be written as

$$\sum_{K+1}^{\infty} e^{-i\omega k}\, d_k \ .$$

Thus,

$$\begin{aligned}\sum_{k=0}^{K} \| c_k \|^2 &\leqq \sum_{k=0}^{K} \| c_k \|^2 + \sum_{K+1}^{\infty} \| d_k \|^2 \\ &= \frac{1}{2\pi} \int_0^{2\pi} \left\| F(e^{i\omega}) \sum_{n=0}^{K} e^{-in\omega} q_n \right\|^2 d\omega,\end{aligned}$$

from which we obtain

$$\begin{aligned}\sum_{k=0}^{K} \| c_k \|^2 &\leqq \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^2 \sum_{n=0}^{K} \| q_n \|^2 \\ &\leqq \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^2 \sum_{n=0}^{M} \| q_n \|^2.\end{aligned} \tag{36}$$

Since (36) holds for all $K > 0$, we have completed the proof of Lemma 6.*

Inequality (31) and Lemma 6 prove Theorem 1.

Proof of Theorem 2:

By Lemma 3 [inequality (27)], Lemma 4, and Lemma 5 of the preceding proof, we have

$$\left(\sum_{n=0}^{M} \| e_n \|^2\right)^{\frac{1}{2}} \leqq \rho\left(\sum_{n=0}^{M} \| e_n \|^2\right)^{\frac{1}{2}} + \left(\sum_{n=0}^{M} \left\| \sum_{k=0}^{n} v_{n-k}\varphi_{k-1} \right\|^2\right)^{\frac{1}{2}}. \tag{37}$$

Furthermore, by essentially the same argument used to prove Lemma 4, we find that

$$\left(\sum_{n=0}^{M} \left\| \sum_{k=0}^{n} v_{n-k}\varphi_{k-1} \right\|^2\right)^{\frac{1}{2}} \leqq \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \left(\sum_{k=0}^{M} \| \varphi_{k-1} \|^2\right)^{\frac{1}{2}}.$$

Therefore, with $\rho < 1$,

$$\left(\sum_{n=0}^{M} \| e_n \|^2\right)^{\frac{1}{2}} \leqq (1 - \rho)^{-1} \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \left(\sum_{k=0}^{M} \| \varphi_{k-1} \|^2\right)^{\frac{1}{2}}.$$

We now prove that $\sup_{n \geqq 0} \| \varphi_{n-1} \| < \infty$ implies $\sup_{n \geqq 0} \| e_n \| < \infty$. Assume

* Alternatively, we can show using (35), that only a *finite* number of c_k are nonzero.

that $\sup_{n \geqq 0} \| \varphi_{n-1} \| < \infty$. Let

$$u_n = \sum_{k=0}^{n} v_{n-k} \varphi_{k-1} .$$

Then we have

$$e_n = \sum_{k=0}^{n} w_{n-k} \, \delta_k + u_n , \qquad n = 0, 1, 2, \cdots , M \tag{38}$$

$$\delta_k = f(x_k , kh) - f(y_k , kh) - \tfrac{1}{2}(\alpha + \beta)(x_k - y_k) \tag{39}$$

in which, since $\{v_k\} \, \varepsilon \, l_1$, $\sup_{n \geq 0} \| u_n \| < \infty$.

There exist positive constants c_1 and c_2 such that $| w_n | \leqq c_1 e^{-c_2 n}$ for all $n \geqq 0$. By continuity, since $\rho < 1$, there exists $a \; \sigma \; \varepsilon \; (0, c_2)$ such that

$$\rho_\sigma \triangleq \max_{0 \leqq \omega \leqq 2\pi} | W(e^{i\omega - \sigma}) | \, [\tfrac{1}{2}(\beta - \alpha) + \gamma] < 1$$

in which of course

$$W(e^{i\omega - \sigma}) = \sum_{n=0}^{\infty} w_n e^{-(i\omega - \sigma)n} .$$

From (38) and (39):

$$\tilde{e}_n = \sum_{k=0}^{n} \tilde{w}_{n-k} \, \tilde{\delta}_k + \tilde{u}_n , \qquad n = 0, 1, 2, \cdots , M$$

$$\tilde{\delta}_k = \tilde{f}(\tilde{x}_k , kh) - \tilde{f}(\tilde{y}_k , kh) - \tfrac{1}{2}(\alpha + \beta)(\tilde{x}_k - \tilde{y}_k)$$

where $\tilde{e}_n = e^{\sigma n} e_n$, $\tilde{w}_n = e^{\sigma n} w_n$, $\tilde{u}_n = e^{\sigma n} u_n$, $\tilde{\delta}_k = e^{\sigma k} \delta_k$, $\tilde{x}_k = e^{\sigma k} x_k$, $\tilde{y}_k = e^{\sigma k} y_k$, and $\tilde{f}(q, kh) = e^{\sigma k} f(e^{-\sigma k} q, kh)$ for all N-vectors q.

The Jacobian matrix $\tilde{f}'$ of $\tilde{f}$ is related to the Jacobian matrix of f by

$$\tilde{f}'(q, kh) = f'(e^{-\sigma k} q, kh)$$

from which we see that $\tilde{f}'$ satisfies the assumption concerning f' relative to the numbers α, β, and γ. Therefore, by the proof of Theorem 1,

$$\left(\sum_{n=0}^{M} \| \tilde{e}_n \|^2 \right)^{\frac{1}{2}} \leqq (1 - \rho_\sigma)^{-1} \left(\sum_{n=0}^{M} \| \tilde{u}_n \|^2 \right)^{\frac{1}{2}}$$

for all $M \geqq (p + 1)$.

Now,

$$\left(\sum_{n=0}^{M} \| \tilde{u}_n \|^2 \right)^{\frac{1}{2}} = \left(\sum_{n=0}^{M} \| e^{\sigma n} u_n \|^2 \right)^{\frac{1}{2}} \leqq \sup_{n \geqq 0} \| u_n \| \left(\sum_{n=0}^{M} e^{2\sigma n} \right)^{\frac{1}{2}}$$

$$\leqq \sup_{n \geqq 0} \| u_n \| \left(\frac{e^{2\sigma(M+1)} - 1}{e^{2\sigma} - 1} \right)^{\frac{1}{2}}$$

$$\leqq \sup_{n \geqq 0} \| u_n \| \frac{e^{\sigma(M+1)}}{(e^{2\sigma} - 1)^{\frac{1}{2}}}$$

and so,

$$\left(\sum_{n=0}^{M} \| \tilde{e}_n \|^2 \right)^{\frac{1}{2}} \leqq (1 - \rho_\sigma)^{-1} \frac{e^{\sigma(M+1)}}{(e^{2\sigma} - 1)^{\frac{1}{2}}} \sup_{n \geqq 0} \| u_n \| \tag{40}$$

for all $M \geqq (\mathrm{p} + 1)$.

From (38):

$$\| e_M \| \leqq \left\| \sum_{k=0}^{M} w_{M-k} \, \delta_k \right\| + \sup_{n \geqq 0} \| u_n \|. \tag{41}$$

We shall now use (40) to bound the first term on the right side of (41). Using the Schwarz inequality,

$$\begin{aligned} \left\| \sum_{n=0}^{M} w_{M-k} \, \delta_k \right\| &= e^{-\sigma M} \left\| \sum_{k=0}^{M} \tilde{w}_{M-k} \, \tilde{\delta}_k \right\| \\ &\leqq e^{-\sigma M} \left(\sum_{n=0}^{\infty} | \tilde{w}_n |^2 \right)^{\frac{1}{2}} \left(\sum_{k=0}^{M} \| \tilde{\delta}_k \|^2 \right)^{\frac{1}{2}}. \end{aligned} \tag{42}$$

By the proof of Lemma 5,

$$\| \tilde{\delta}_k \| \leqq [\tfrac{1}{2}(\beta - \alpha) + \gamma] \, \| \tilde{e}_k \|,$$

which leads to

$$\left\| \sum_{k=0}^{M} w_{M-k} \, \delta_k \right\|$$

$$\leqq \frac{e^{\sigma}}{(e^{2\sigma} - 1)^{\frac{1}{2}}} [\tfrac{1}{2}(\beta - \alpha) + \gamma](1 - \rho_\sigma)^{-1} \sup_{n \geqq 0} \| u_n \| \left(\sum_{n=0}^{\infty} | \tilde{w}_n |^2 \right)^{\frac{1}{2}}.$$

Since $\sigma \epsilon (0, c_2)$ [see the paragraph below (39)], $|\tilde{w}_n|^2 \varepsilon \, l_1$, and therefore,

$$\sup_{M > (p+1)} \left\| \sum_{k=0}^{M} w_{M-k} \, \delta_k \right\| < \infty. \tag{43}$$

Finally, from (41) and (43), we have $\sup_{n \geqq 0} \| e_n \| < \infty$, which completes the proof of Theorem 2.

IV. A CONDITION FOR THE SOLVABILITY OF (2) FOR y_{n+1}

The problem of solving (2) for y_{n+1} is that of solving the equation

$$y + hb_{-1}f(y, t) = g \tag{44}$$

for y, with g a given N-vector. We write (44) as

$$Qy = g \tag{45}$$

in which the operator Q is defined by the condition that $Qv = v + hb_{-1}f(v, t)$ for all real N-vectors v.

We prove below that (with $\langle \cdot , \cdot \rangle$ denoting the usual inner product of real N-vectors):

$$\langle Qy_a - Qy_b , y_a - y_b \rangle \geqq k_1 \| y_a - y_b \|^2 \tag{46}$$

$$\| Qy_a - Qy_b \| \leqq k_2 \| y_a - y_b \| \tag{47}$$

for every pair of real N-vectors y_a and y_b, in which $k_1 = (1 + hb_{-1}\alpha)$ if $b_{-1} \geqq 0$, $k_1 = 1 + hb_{-1}\beta$ if $b_{-1} \leqq 0$: and $k_2 = \{1 + h \mid b_{-1} \mid \cdot [\max (| \alpha |, | \beta |) + \gamma]\}$. Since $k_2 < \infty$, according to a special case of Theorem I of Ref. 8, if $k_1 > 0$ (e.g., if $b_{-1} \geqq 0$ and $1 + hb_{-1}\alpha > 0$), then (45) possesses exactly one solution which can be determined by an iteration procedure that is only slightly more complicated than the usual procedure[1,2] (which is valid only under much stronger conditions on h).

To derive (46), let

$$q(\eta) = \eta y_a + (1 - \eta)y_b + hb_{-1}f[\eta y_a + (1 - \eta)y_b , t]$$

for $\eta \epsilon [0, 1]$. Then

$$q'(\eta) = (y_a - y_b) + hb_{-1}f'[\eta y_a + (1 - \eta)y_b , t](y_a - y_b),$$

and so

$$\begin{aligned} Qy_a - Qy_b &= \int_0^1 q'(\eta)\, d\eta \\ &= \int_0^1 \{1_N + hb_{-1}f'[\eta y_a + (1 - \eta)y_b , t]\}(y_a - y_b)\, d\eta. \end{aligned} \tag{48}$$

Thus,

$$\begin{aligned} &\langle Qy_a - Qy_b , y_a - y_b \rangle \\ &\quad = \left\langle \int_0^1 \{1_N + hb_{-1}f'[\eta y_a + (1 - \eta)y_b , t]\}\, d\eta (y_a - y_b), (y_a - y_b) \right\rangle \end{aligned}$$

$$
\begin{aligned}
&= \int_0^1 \langle \{1_N + hb_{-1}f'[\eta y_a + (1-\eta)y_b\,,\,t]\}(y_a - y_b), (y_a - y_b)\rangle\, d\eta \\
&= \|\, y_a - y_b \,\|^2 \\
&\qquad + hb_{-1} \int_0^1 \langle f'_S[\eta y_a + (1-\eta)y_b\,,\,t](y_a - y_b), (y_a - y_b)\rangle\, d\eta,
\end{aligned}
$$

in which f'_S denotes the symmetric part of f'. Thus, since the eigenvalues of f'_S are bounded from below by α, and from above by β:

$$
\begin{aligned}
\langle Qy_a - Qy_b\,, y_a - y_b\rangle &\geqq (1 + hb_{-1}\alpha)\,\|\, y_a - y_b \,\|^2, &\quad b_{-1} \geqq 0 \\
&\geqq (1 + hb_{-1}\beta)\,\|\, y_a - y_b \,\|^2, &\quad b_{-1} \leqq 0.
\end{aligned}
$$

Consider now the derivation of (47). By (48),

$$
\begin{aligned}
&\|\, Qy_a - Qy_b \,\| \\
&\quad = \left\| \int_0^1 \{1_N + hb_{-1}f'[\eta y_a + (1-\eta)y_b\,,\,t]\}\, d\eta(y_a - y_b) \right\| \\
&\quad \leqq \left\| \int_0^1 \{1_N + hb_{-1}f'[\eta y_a + (1-\eta)y_b\,,\,t]\}\, d\eta \right\| \cdot \|\, y_a - y_b \,\| \\
&\quad \leqq \int_0^1 \|\, 1_N + hb_{-1}f'[\eta y_a + (1-\eta)y_b\,,\,t] \,\|\, d\eta\, \|\, y_a - y_b \,\|.
\end{aligned}
$$

But, with f'_A the antisymmetric part of f',

$$
\begin{aligned}
&\|\, 1_N + hb_{-1}f'[\eta y_a + (1-\eta)y_b\,,\,t] \,\| \\
&\qquad \leqq 1 + h \mid b_{-1} \mid \|\, f'[\eta y_a + (1-\eta)y_b\,,\,t] \,\| \\
&\qquad \leqq 1 + h \mid b_{-1} \mid \cdot \|\, f'_S \,\| + h \mid b_{-1} \mid \cdot \|\, f'_A \,\| \\
&\qquad \leqq 1 + h \mid b_{-1} \mid \max(\mid \alpha \mid, \mid \beta \mid) + h \mid b_{-1} \mid \gamma.
\end{aligned}
$$

Therefore,

$$
\|\, Qy_a - Qy_b \,\| \leqq \{1 + h \mid b_{-1} \mid [\max(\mid \alpha \mid, \mid \beta \mid) + \gamma]\}\, \|\, y_a - y_b \,\|
$$

which is equivalent to (47).

REFERENCES

1. Ralston, A., *First Course in Numerical Analysis,* McGraw-Hill Book Co., New York, 1965.
2. Hamming, R. W., *Numerical Methods for Scientists and Engineers,* McGraw-Hill Book Co., New York, 1962.
3. Henrici, P., *Error Propagation for Difference Methods,* Wiley & Sons, Inc., New York, 1963.

4. Hildebrand, F. B., *Introduction to Numerical Analysis,* McGraw-Hill Book Co., New York, 1956.
5. Rall, L. B. (editor), *Error in Digital Computation,* Volumes 1 and 2, Wiley & Sons, Inc., New York, 1965.
6. Sandberg, I. W., On the Theory of Physical Systems Governed by Nonlinear Functional Equations, B.S.T.J., *44,* May-June, 1965, p. 871.
7. Macduffee, C. C., *The Theory of Matrices,* Chelsea Publishing Co., New York, 1956.
8. Sandberg, I. W., On the Properties of Some Systems that Distort Signals—I. B.S.T.J., *42,* September, 1963, p. 2033.