Some Properties of a Classic Numerical Integration Formula

By I. W. SANDBERG

(Manuscript received May 19, 1967)

The numerical integration formula

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

can be used to obtain a numerical solution of the system of nonlinear differential equations

$$\dot{x} + f(x, t) = 0, \qquad t \geqq 0 \; [x(0) = x_0]. \tag{2}$$

In many instances, it is known beforehand that the solution of (2) possesses a particular property such as boundedness or asymptotic periodicity with a given period, and it is then of interest to analytically determine the range of values of the step size h such that the sequence $\{y_n\}$ *defined by (1) exhibits (at least) that property. In this paper, we consider problems of this type [but do not actually use assumptions concerning the character of the solution of (2)], and we study also the overall effect of solving instead of (1) the equation*

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

which takes into account the effect of local roundoff errors and errors in the starting values. We consider explicitly only the case in which x(t) is scalar valued.

I. INTRODUCTION

In this paper, we present some theorems concerning properties of the classic numerical integration formula[1]

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

a formula which can be used to obtain a numerical solution of the set of first-order nonlinear differential equations

$$\dot{x} + f(x, t) = 0, \qquad t \geqq 0 \; [x(0) = x_0]. \tag{2}$$

In (1) 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

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

Specializations of (1) include, for example, Euler's method:

$$y_{n+1} = y_n + hy'_n , \tag{3}$$

and the more useful formula

$$y_{n+1} = y_n + \tfrac{1}{2}h(y'_n + y'_{n+1}). \tag{4}$$

In many instances it is known beforehand that the solution of (2) possesses a particular property such as boundedness or asymptotic periodicity with a given period, and it is then of interest to analytically determine the range (or ranges) of step sizes that will lead to a sequence $\{y_n\}$ which exhibits (at least) that property. This is one type of problem that we consider. For related material concerned with the overall effect of local truncation errors, see Ref. 2. Our results dealing with questions of asymptotic periodicity of the y_n are restricted to cases in which the basic period is a multiple of the step size h. However, it is often reasonable to choose h in this way to reduce programming complexity.

In addition to the fact that the solution of (1) differs from the samples of the solution of (2) due to truncation effects,[1, 3] the problem of solving (2) is further complicated by the fact that the numbers obtained from the computer differ from the y_n of (1) as a result of roundoff errors. The local roundoff error R_n introduced in calculating y_{n+1} can be taken into account[1] by replacing (1) by

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h \sum_{k=-1}^{p} b_k y'_{n-k} + R_n , \qquad n \geqq p. \tag{5}$$

If $b_{-1} \neq 0$, the error in solving (1) for y_{n+1} , caused typically by truncating an iteration procedure[1,3] after a finite number of steps, can be accounted for by redefining R_n . The second type of problem that we treat is to bound (from below as well as from above) a measure of the overall error in solving (5) instead of (1). The problem of estimating the R_n before the calculations are performed is by no means trivial, and is not considered here. On the other hand, since there exist methods for bounding R_n *given* y_k for $(n - p) \leqq k \leqq n$ (see, for example, Wilkinson[4], for bounds on the effect of roundoff in forming sums, products, etc.), our results suggest the feasibility of programming the computer to evaluate overall error bounds as the calculation of the successive y_{n+1} proceeds.

We shall explicitly consider only the case in which $x(t)$ and the y_n are scalars. Without much difficulty, each of the theorems can be extended to cover the vector case. In this extension, requirements on, for example, the derivative $\partial f(x, t)/\partial x$ are replaced by conditions on the Jacobian matrix of $f(x, t)$ (see Ref. 2).

For reasons that will become clear to the reader, our theorems are quite naturally characterized as "frequency-domain" results. Some of these theorems are close relatives of earlier results concerned with the input-output stability of nonlinear feedback systems* (see Ref. 5 and the difference-equation theorems stated without proof of Ref. 6). To the writer's knowledge, the only even remotely related material concerning (1) in the numerical-analysis literature, with the exception of Ref. 2, is Hamming's transfer-function approach.[3]

II. RESULTS†

We begin by introducing some definitions and assumptions. We assume throughout this section that y_n and $f(y_n, nh)$ are real-valued scalars.

Let α and β be two real constants, let $a_{-1} \triangleq 0$, and let

$$F(z) \triangleq 1 - \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta)hb_k]z^{-(k+1)} \tag{6}$$

for all complex $z \neq 0$.

Assumption 1: It is assumed throughout that $1 + \frac{1}{2}(\alpha + \beta)hb_{-1} \neq 0$, and that $F(z) \neq 0$ for all $|z| \geqq 1$.

This assumption implies that the sequence of approximations defined by (1) is bounded and approaches zero as $n \to \infty$ for all sets of starting values when $f(x, t) = \frac{1}{2}(\alpha + \beta)x$.

* The usual frequency-domain nonlinear system stability results such as Popov's criterion[7] are not directly related because they do not deal with systems subjected to external inputs.

† The proofs of the theorems stated here are given in Section III.

Definitions

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

(*ii*)*
$$l_2 \triangleq \left\{ \{s_n\} \;\middle|\; \sum_{n=0}^{\infty} | s_n |^2 < \infty \right\}$$
$$l_\infty \triangleq \{ \{s_n\} \mid \sup_{n \geqq 0} | s_n | < \infty \}$$

(*iii*)* Let K be a positive integer, and let
$$\mathcal{K} \triangleq \{ \{s_n\} \mid s_n = s_{n+K+1} \text{ for } n = 0, \pm 1, \pm 2, \cdots \}$$

(*iv*)
$$\rho_K \triangleq \tfrac{1}{2}(\beta - \alpha)h \max_{q \,\varepsilon\, \mathcal{R}} \left| \frac{\sum_{k=-1}^{p} b_k \exp\left[-\frac{i(k+1)2\pi q}{K+1} \right]}{F\left[\exp\left(\frac{i2\pi q}{K+1} \right) \right]} \right|$$

in which $\mathcal{R} \triangleq \{0, 1, 2, \cdots, K\}$.

2.1 *Properties of (1)*

Theorem 1: If

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

if $\rho < 1$, and if

$$\alpha \leqq \frac{f(u, nh) - f(0, nh)}{u} \leqq \beta, \qquad n \geqq 0$$

for all real $u \neq 0$, then

(*i*) $\{f(0, nh)\} \,\varepsilon\, l_2$ implies that $\{y_n\} \,\varepsilon\, l_2$
(*ii*) $\{f(0, nh)\} \,\varepsilon\, l_\infty$ *implies that* $\{y_n\} \,\varepsilon\, l_\infty$.

Remarks:

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

$$\Theta(\omega) \triangleq \frac{\sum_{k=0}^{p} a_k \exp\,(ik\omega) - \exp\,(-i\omega)}{\sum_{k=-1}^{p} b_k \exp\,(ik\omega)} \tag{7}$$

[*] We consider only real sequences.

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).

For Euler's formula (3), we have $F(z) = 1 - [1 - \frac{1}{2}(\alpha + \beta)h]z^{-1}$, so that $F(z) \neq 0$ for $| z | \geqq 1$ if and only if $0 < \frac{1}{2}(\alpha + \beta)h < 2$. For this formula 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, and $\rho < 1$. 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 $y_0, y_1, y_2, \cdots$ unbounded (assuming merely that $y_0 \neq 0$).

For the formula (4):

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

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

We have $1 + \frac{1}{4}(\alpha + \beta)h \neq 0$ and $F(z) \neq 0$ for $| z | \geqq 1$ if and only if $(\alpha + \beta)h > 0$. 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 $\alpha > 0$. On the other hand, if $\alpha < 0$, then for even the special case $f(x, t) = \alpha x : y_0, y_1, \cdots$ is unbounded provided that $y_0 \neq 0$.

The following theorem is concerned with conditions under which

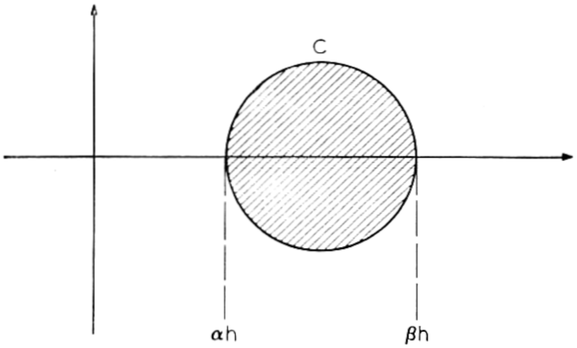

Fig. 1 — Location of the critical circle C.

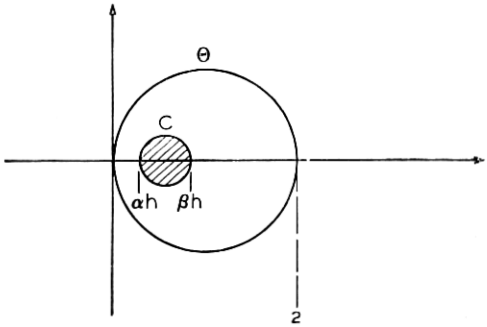

Fig. 2 — The locus of $\Theta(\omega)$ For Euler's method, and the critical circle C.

asymptotically periodic $f(0, nh)$ in (1) implies that $\{y_n\}$ is asymptotically periodic with the same period as that of $f(0, nh)$.

Theorem 2: If

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

if $\rho < 1$, *if* $[f(u, nh) - f(0, nh)] = [f(u, (n + K + 1)h) - f(0, (n + K + 1)h)]$ *for all real* u *and* $n \geqq 0$, *if*

$$\alpha \leqq \frac{\partial f(u, nh)}{\partial u} \leqq \beta, \qquad n \geqq 0$$

for all real u, *and if there exists a* $y_a^* \varepsilon \mathcal{K}$ *such that* $[f(0, nh) - y_a^*] \varepsilon l_2$, *then there exists a* $y_b^* \varepsilon \mathcal{K}$ *such that*

(*i*) $(y - y_b^*) \varepsilon l_2$
(*ii*) y_b^* *is independent of* $[f(0, nh) - y_a^*]$.

Remarks:

In many cases of interest $[f(u, nh) - f(0, nh)]$ is independent of n, and hence certainly satisfies the periodicity requirement.

Theorem 3, below, provides a condition under which the sequence $\{y_n\}$ of (1) *cannot* approach a "self sustained" limit cycle with period $(K + 1)$.

Theorem 3: If

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} - h \sum_{k=-1}^{p} f[y_{n-k}, (n-k)h], \qquad n \geqq p$$

if $[f(u, nh) - f(0, nh] = [f(u, (n + K + 1)h) - f(0, (n + K + 1)h)]$ *for all real* u *and* $n \geqq 0$, *if*

$$\alpha \leqq \frac{\partial f(u, nh)}{\partial u} \leqq \beta, \qquad n \geqq 0$$

for all real u, *if* $f(0, nh) \to 0$ *as* $n \to \infty$, *and if there exists a* $y^* \varepsilon \mathcal{K}$ *different from the zero element of* $\mathcal{K}$ *such that* $(y_n - y_n^*) \to 0$ *as* $n \to \infty$, *then* $\rho_K \geqq 1$.

Remark:

For $\rho_K \geqq 1$, at least one of the complex numbers

$$\Theta\left(\frac{2\pi q}{K + 1}\right) \qquad q = 0, 1, 2, \cdots, K$$

must lie on or within the circle C of Fig. 1.

2.2 *Results Concerning the Effect of* R_n *and Errors in the Starting Values*

Theorem 4, below, is essentially the same as a result concerning the effect of local roundoff and truncation errors proved in Ref. 2. The proof of Theorem 4 given in Section III is considerably more direct than the corresponding argument of Ref. 2.

Definition:

$$\langle s \rangle_N \triangleq \left(\frac{1}{N + 1} \sum_{n=0}^{N} | s_n |^2\right)^{\frac{1}{2}}$$

for all $N \geqq 0$ and every sequence $\{s_n\}$.

Theorem 4: If

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

$$z_{n+1} = \sum_{k=0}^{p} a_k z_{n-k} - h \sum_{k=-1}^{p} b_k f[z_{n-k}, (n - k)h] + R_n, \qquad n \geqq p$$

if

$$\alpha \leqq \frac{\partial f(u, nh)}{\partial u} \leqq \beta, \qquad n \geqq 0$$

for all real u, *then for all* $N \geqq 0$:

(*i*) $$\langle y - z \rangle_N \geqq (1 + \rho)^{-1} \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \psi \rangle_N ,$$

and

(ii) if $\rho < 1$,

$$\langle y - z \rangle_N \leqq (1 - \rho)^{-1} \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \psi \rangle_N$$

in which

$$\begin{aligned} \psi_n &= -R_{n-1}, \qquad n \geqq (p + 1) \\ &= (y_n - z_n) - \sum_{k=0}^{p} a_k (y_{n-k-1} - z_{n-k-1}) \\ &\quad + h \sum_{k=-1}^{p} b_k \{ f[y_{n-k-1}, (n - k - 1)h] - f[z_{n-k-1}, (n - k - 1)h] \}, \\ &\qquad\qquad n = 0, 1, 2, \cdots, p \end{aligned}$$

with $y_n = f(y_n, nh) = z_n = f(z_n, nh) = 0$ *for* $n < 0$.

Remarks:

Ref. 2 considers two simple examples concerning the evaluation of the numbers

$$(1 + \rho)^{-1} \min_{\omega} | F(e^{i\omega}) |^{-1} \quad \text{and} \quad (1 - \rho)^{-1} \max_{\omega} | F(e^{i\omega}) |^{-1}.$$

Since

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

we see that ρ is the ratio of the radius of the circle C of Fig. 1 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 $\Theta(\omega)$.

The following corollary provides asymptotic bounds on the difference between the solutions of (1) and (5) when the solution $\{y_n\}$ of (1) is, for example, asymptotically periodic.

Corollary to Theorem 4: If

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

with

$$\alpha \leqq \frac{\partial f(u, nh)}{\partial u} \leqq \beta$$

for all real u and $n \geqq 0$, *if there exists a sequence* $\tilde{y}$ *such that* $(y_n - \tilde{y}_n) \to 0$ *as* $n \to \infty$, *and if*

$$z_{n+1} = \sum_{k=0}^{p} a_k z_{n-k} - h \sum_{k=-1}^{p} b_k f[z_{n-k}, (n-k)h] + R_n, \qquad n \geqq p.$$

Then

(*i*)

$$\langle z - \tilde{y} \rangle_N \geqq (1 + \rho)^{-1} \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \psi \rangle_N - | q_N |$$

with $q_N \to 0$ *as* $N \to \infty$,

and

(*ii*) *if* $\rho < 1$,

$$\langle z - \tilde{y} \rangle_N \leqq (1 - \rho)^{-1} \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \psi \rangle_N + | r_N |$$

with $r_N \to 0$ *as* $N \to \infty$

in which

$$\begin{aligned} \psi_n &= R_{n-1}, \qquad & n &\geqq (p + 1) \\ &= 0, & n &= 0, 1, 2, \cdots, p. \end{aligned}$$

Remark:

Note that the lower bound is valid under quite weak assumptions.

III. PROOFS

We first prove the following lemma which plays a role in the proofs of all of the theorems

Lemma 1: If

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} - h \sum_{k=-1}^{p} b_k f[y_{n-k}, (n-k)h] + R_n, \qquad n \geqq p$$

then

$$y_n = \sum_{k=0}^{n} w_{n-k} g(y_k, kh) + \sum_{k=0}^{n} w_{n-k} f(0, kh) + \sum_{k=0}^{n} v_{n-k} \varphi_k, \qquad n \geqq 0$$

in which $\{w_n\}$ *and* $\{v_n\}$ *are the inverse z-transforms of*

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

and

$$V(z) \triangleq \frac{1}{1 - \sum_{k=-1}^{p} [a_k - \frac{1}{2}(\alpha + \beta)hb_k]z^{-(k+1)}},$$

respectively;

$$\sum_{n=0}^{\infty} | w_n | < \infty, \qquad \sum_{n=0}^{\infty} | v_n | < \infty,$$

$$g(y_k, kh) \triangleq f(y_k, kh) - f(0, kh) - \tfrac{1}{2}(\alpha + \beta)y_k,$$

and

$$\begin{aligned} \varphi_n &= R_{n-1}, \qquad n \geqq (p+1) \\ &= y_n - \sum_{k=0}^{p} a_k y_{n-k-1} + h \sum_{k=-1}^{p} b_k f[y_{n-k-1}, (n-k-1)h], \\ &\qquad\qquad n = 0, 1, 2, \cdots, p \end{aligned}$$

with $y_n = f(y_n, nh) \triangleq 0$ *for* $n < 0$.

Proof of Lemma 1:

From

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} - h \sum_{k=-1}^{p} b_k f[y_{n-k}, (n-k)h] + R_n, \qquad n \geqq p$$

we have

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

and, with the φ_n as defined in the lemma,

$$y_n = \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta)hb_k]y_{n-k-1} - h \sum_{k=-1}^{p} b_k \delta_{n-k-1} + \varphi_n, \qquad n \geqq 0$$

where

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

Let $M > 0$. Then $y_n = \hat{y}_n$ for $n = 0, 1, \cdots, M$, in which

$$\hat{y}_n = \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta)hb_k]\hat{y}_{n-k-1} - h \sum_{k=-1}^{p} b_k \hat{\delta}_{n-k-1} + \hat{\varphi}_n, \qquad n \geqq 0,$$

where

$$\begin{aligned}\hat{\delta}_n &= \delta_n \quad \text{for} \quad n \leqq M\\ &= 0 \quad \text{for} \quad n > M,\\ \hat{\varphi}_n &= \varphi_n \quad \text{for} \quad n \leqq M\\ &= 0 \quad \text{for} \quad n > M,\end{aligned}$$

and

$$\hat{y}_n = f(\hat{y}_n\,, nh) = 0 \quad \text{for} \quad n < 0.$$

It is clear that $\{\hat{\varphi}_n\}$, $\{\hat{\varphi}_n\}$, and $\{\hat{y}_n\}$ are z-transformable. Let

$$\psi(z) \triangleq \sum_{n=0}^{\infty} \hat{\varphi}_n z^{-n}, \qquad \Delta(z) \triangleq \sum_{n=0}^{\infty} \hat{\delta}_n z^{-n},$$

and

$$Y(z) \triangleq \sum_{n=0}^{\infty} \hat{y}_n z^{-n}.$$

Then

$$\left[1 - \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta)hb_k]z^{-(k+1)}\right]Y(z) = -h \sum_{k=-1}^{p} b_k z^{-(k+1)}\Delta(z) + \psi(z).$$

Therefore,

$$Y(z) = \frac{-h \sum_{k=-1}^{p} b_k z^{-(k+1)}}{1 - \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta)hb_k]z^{-(k+1)}}\,\Delta(z) + \frac{\psi(z)}{1 - \sum_{k=-1}^{p} [a_k - \tfrac{1}{2}(\alpha + \beta)hb_k]z^{-(k+1)}}$$

and, with $\{w_n\}$ and $\{v_n\}$ the inverse z-transform of $W(z)$ and $V(z)$, respectively,* we have

$$\hat{y}_n = \sum_{k=0}^{n} w_{n-k}\hat{\delta}_k + \sum_{k=0}^{n} v_{n-k}\hat{\varphi}_k\,, \qquad n \geqq 0$$

* Recall that $W(z)$ and $V(z)$ are defined in Lemma 1.

with (in view of Assumption 1)

$$\sum_{n=0}^{\infty} | w_n | < \infty, \quad \text{and} \quad \sum_{n=0}^{\infty} | v_n | < \infty . \tag{8}$$

Thus,

$$y_n = \sum_{k=0}^{n} w_{n-k}\delta_k + \sum_{k=0}^{n} v_{n-k}\varphi_k \tag{9}$$

for $n = 0, 1, 2, \cdots , M$. Since M is arbitrary, (9) is satisfied for all $n \geqq 0$. Finally, with

$$g(y_k , kh) \triangleq f(y_k , kh) - f(0, kh) - \tfrac{1}{2}(\alpha + \beta)y_k$$

$$y_n = \sum_{k=0}^{n} w_{n-k}g(y_k , kh) + \sum_{k=0}^{n} w_{n-k}f(0, kh) + \sum_{k=0}^{n} v_{n-k}\varphi_k , \qquad n \geqq 0.$$

We now prove a lemma which is used in the proofs of most of the theorems. We repeat the

Definition:

$$\langle s \rangle_N \triangleq \left(\frac{1}{N+1} \sum_{n=0}^{N} | s_n |^2 \right)^{\frac{1}{2}}$$

for all $N \geqq 0$ and every sequence $\{s_n\}$.

Lemma 2: If

$$y_n = \sum_{k=0}^{n} w_{n-k}a(k)y_k + b_n , \qquad n \geqq 0$$

and if $-\frac{1}{2}(\beta - \alpha) \leqq a(k) \leqq \frac{1}{2}(\beta - \alpha)$ *for all* $k \geqq 0$, *then*

(*i*) $\langle y \rangle_N \geqq (1 + \rho)^{-1}\langle b \rangle_N$ for $N \geqq 0$,

and

(*ii*) *if* $\rho < 1$, *then* $\langle y \rangle_N \geqq (1 - \rho)^{-1}\langle b \rangle_N$ for $N \geqq 0$.

Proof of Lemma 2:

Let

$$q_n \triangleq \sum_{k=0}^{n} w_{n-k}a(k)y_k , \qquad n \geqq 0.$$

By Minkowski's inequality,

$$\langle y \rangle_N \leqq \langle q \rangle_N + \langle b \rangle_N \tag{10}$$

and

$$\langle b\rangle_N \leqq \langle y\rangle_N + \langle q\rangle_N . \tag{11}$$

Lemma 2 follows from (10), (11), and the inequality[2]

$$\langle q\rangle_N \leqq \rho\langle y\rangle_N .$$

3.1 *Proof of Theorem 1:*

By Lemma 1, we have

$$y_n = \sum_{k=0}^{n} w_{n-k}g(y_k , kh) + \sum_{k=0}^{n} w_{n-k}f(0, kh) + \sum_{k=0}^{n} v_{n-k}\varphi_k , \qquad n \geqq 0$$

with (because $R_n = 0$ for all $n \geqq p$) $\varphi_n = 0$ for all $n \geqq (p + 1)$.

Let

$$b_n = \sum_{k=0}^{n} w_{n-k}f(0, kh) + \sum_{k=0}^{n} v_{n-k}\varphi_k , \qquad n \geqq 0.$$

Since both $\{w_n\}$ and $\{v_n\}$ belong to l_1 [i.e., since (8) is satisfied], $b \,\varepsilon\, l_2$ if $\{f(0, kh)\} \,\varepsilon\, l_2$ and $b \,\varepsilon\, l_\infty$ if $\{f(0, kh)\} \,\varepsilon\, l_\infty$.

Suppose that $b \,\varepsilon\, l_2$, and let

$$a(k) = \frac{g(y_k , kh)}{y_k} , \quad \text{for} \quad y_k \neq 0$$

$$= 0, \qquad \text{for} \quad y_k = 0.$$

The function $a(k)$ satisfies the bounds of Lemma 2, and

$$y_n = \sum_{k=0}^{n} w_{n-k}a(k)y_k + b_n , \qquad n \geqq 0 \tag{12}$$

Therefore, by Lemma 2,

$$\sum_{n=0}^{N} | y_n |^2 \leqq (1 - \rho)^{-2} \sum_{n=0}^{N} | b_n |^2 \leqq (1 - \rho)^{-2} \sum_{n=0}^{\infty} | b_n |^2$$

for all $N \geqq 0$, from which it is clear that $y \,\varepsilon\, l_2$.

If $b \,\varepsilon\, l_\infty$, then $\{y_n\}$ satisfies (12) with $b \,\varepsilon\, l_\infty$. According to the first conclusion of the following lemma, this implies that $y \,\varepsilon\, l_\infty$.

Lemma 3: If

$$y_n = \sum_{k=0}^{n} w_{n-k}a(k)y_k + b_n , \qquad n \geqq 0$$

with $b \,\varepsilon\, l_\infty$ *, if* $\rho < 1$*, and if* $-\tfrac{1}{2}(\beta - \alpha) \leqq a(k) \leqq \tfrac{1}{2}(\beta - \alpha)$ *for all* $k \geqq 0$*, then* *(i)* $y \,\varepsilon\, l_\infty$

(ii) there exists a constant c_∞, *which depends on only the* a_k, *the* b_k, α, *and* β *such that*

$$\sup_{n \geqq 0} | y_n | \leqq c_\infty \sup_{n \geqq 0} | b_n |.$$

Proof of Lemma 3:

The proof is essentially the same as that of the second part of Theorem 2 of Ref. 2. The details are omitted.*

3.2 *Proof of Theorem 2*

Definitions: Let $\mathcal{K}$ denote the set of all real sequences $\{s_n\}$ such that $s_n = s_{n+K+1}$ for all $n = 0, \pm 1, \pm 2, \cdots$, and let $\mathcal{R} \triangleq \{0, 1, 2, \cdots, K\}$.

Lemma 4: Let $g^*(x, nh)$ *be defined for all real* x *and all* $n = 0, \pm 1, \pm 2, \cdots$, *such that:* $g^*(x, nh) = g^*[x, (n + K + 1)h]$ *for all* x *and* n, *and*

$$-\tfrac{1}{2}(\beta - \alpha) \leqq \frac{\partial g^*(x, nh)}{\partial x} \leqq \tfrac{1}{2}(\beta - \alpha)$$

for all x *and* n. *If* $p \,\varepsilon\, \mathcal{K}$ *and if* $\rho_K < 1$, *then* $\mathcal{K}$ *contains exactly one element* y^* *such that*

$$y_n^* = \sum_{k=-\infty}^{n} w_{n-k} g^*(y_k, kh) + p_n$$

for $n = 0, \pm 1, \pm 2, \cdots$.

Proof of Lemma 4:

With the norm

$$|| s || \triangleq \left(\sum_{k=0}^{K} | s_n |^2 \right)^{\frac{1}{2}},$$

the set $\mathcal{K}$ is a Banach space. The operator WG defined on $\mathcal{K}$ by

$$(WGs)_n = \sum_{k=-\infty}^{n} w_{n-k} g^*(s_k, kh), \qquad s \,\varepsilon\, \mathcal{K}$$

maps $\mathcal{K}$ into itself. By the contraction-mapping fixed-point theorem, it suffices to show that WG is a contraction when $\rho_K < 1$. It is clear that

$$\begin{aligned} || WGs_a - WGs_b || &\leqq || W || \cdot || Gs_a - Gs_b || \\ &\leqq \tfrac{1}{2}(\beta - \alpha) \, || W || \cdot || s_a - s_b || \end{aligned}$$

for all $s_a \,\varepsilon\, \mathcal{K}$ and all $s_b \,\varepsilon\, \mathcal{K}$.

* See also Ref. 6.

If $s \in \mathcal{K}$, then

$$s_k = \sum_{l=0}^{K} \hat{s}_l \exp\left(\frac{i2\pi lk}{K+1}\right) \quad \text{for} \quad k = 0, \pm 1, \pm 2, \cdots$$

in which

$$\hat{s}_l = (K+1)^{-1} \sum_{n=0}^{K} s_n \exp\left(-\frac{i2\pi ln}{K+1}\right)$$

and

$$\sum_{n=0}^{K} | s_n |^2 = (K+1) \sum_{n=0}^{K} | \hat{s}_n |^2.$$

Thus, if

$$u_n = \sum_{k=-\infty}^{n} w_{n-k} s_k \quad \text{for} \quad n = 0, \pm 1, \pm 2, \cdots$$

with $s \in \mathcal{K}$, we find that

$$\begin{aligned} u_n &= \sum_{k=-\infty}^{n} w_{n-k} \sum_{l=0}^{K} \hat{s}_l \exp\left(\frac{i2\pi lk}{K+1}\right) \\ &= \sum_{l=0}^{K} \hat{s}_l \sum_{k=-\infty}^{\infty} w_{n-k} \exp\left(\frac{i2\pi lk}{K+1}\right) \qquad (w_n = 0, n < 0) \\ &= \sum_{l=0}^{K} \hat{s}_l \exp\left(\frac{i2\pi ln}{K+1}\right) \sum_{n=0}^{\infty} w_n \exp\left(-\frac{i2\pi ln}{K+1}\right) \\ &= \sum_{l=0}^{K} W\left[\exp\left(\frac{i2\pi l}{K+1}\right)\right] \hat{s}_l \exp\left(\frac{i2\pi ln}{K+1}\right). \end{aligned}$$

Therefore, since

$$\| u \| = \| Ws \| \leqq \max_{q \in \mathcal{R}} \left| W\left[\exp\left(\frac{i2\pi q}{K+1}\right)\right] \right| \| s \|,$$

we have

$$\| W \| \leqq \max_{q \in \mathcal{R}} \left| W\left[\exp\left(\frac{i2\pi q}{K+1}\right)\right] \right|$$

and $\| WGs_a - WGs_b \| \leqq \rho_K \| s_a - s_b \|$ for all $s_a \in \mathcal{K}$ and all $s_b \in \mathcal{K}$. This completes the proof of Lemma 4.

By Lemma 1,

$$y_n = \sum_{k=0}^{n} w_{n-k} g(y_k, kh) + \sum_{k=0}^{n} w_{n-k} f(0, kh) + \sum_{k=0}^{n} v_{n-k} \varphi_k, \qquad n \geqq 0$$

with $\varphi_k = 0$ for $k \geqq (p + 1)$. Here, since both $\{w_n\}$ and $\{v_n\}$ belong to l_1, we have

$$\sum_{k=0}^{n} w_{n-k} f(0, kh) + \sum_{k=0}^{n} v_{n-k}\varphi_k = p_n + c_n , \qquad n \geqq 0$$

with $p \varepsilon \mathcal{K}$ and $c \varepsilon l_2$. In fact, with y_a^* as defined in Theorem 2,

$$p_n = \sum_{k=-\infty}^{n} w_{n-k} y_{ak}^* , \qquad n = 0, \pm 1, \pm 2, \cdots .$$

Let $g^*(x, nh)$ be defined by the conditions: $g^*(x, nh) = g^*[x, (n + K + 1)h]$ for all x and $n = 0, \pm 1, \pm 2, \cdots$, and $g^*(x, nh) = g(x, nh)$ for all x and $n = 0, 1, \cdots, K$. Then, since $\rho_K \leqq \rho < 1$, by Lemma 4 there exists a $y_b^* \varepsilon \mathcal{K}$ such that

$$y_{bn}^* = \sum_{k=-\infty}^{n} w_{n-k} g^*(y_{bk}^* , kh) + p_n$$

for $n \geqq 0$. Therefore,

$$y_n - y_{bn}^* = \sum_{k=0}^{n} w_{n-k}[g^*(y_k , kh) - g^*(y_{bk}^* , kh)] + d_n , \qquad n \geqq 0$$

in which

$$d_n = c_n - \sum_{k=-\infty}^{-1} w_{n-k} g^*(y_{bk}^* , kh), \qquad n \geqq 0.$$

But

$$\left| \sum_{k=-\infty}^{-1} w_{n-k} g^*(y_{bk}^* , kh) \right| \leqq \sup_{n \geqq 0} | g^*(y_{bn}^* , nh) | \sum_{k=-\infty}^{-1} | w_{n-k} |$$

and, using the fact that there exist constants $\eta > 0$ and $\zeta > 0$ such that $| w_n | \leqq \eta \exp(-\zeta n)$ for $n \geqq 0$,

$$\sum_{k=-\infty}^{-1} | w_{n-k} | = \sum_{m=(n+1)}^{\infty} | w_m | \leqq \eta \exp[-\zeta(n + 1)] \sum_{m=0}^{\infty} \exp(-\zeta m).$$

We see that

$$\sum_{k=-\infty}^{-1} w_{n-k} g^*(y_{bk}^* , kh) \ \varepsilon \ l_2 ,$$

and consequently $d \varepsilon l_2$.

Let

$$\begin{aligned} a(k) &= \frac{g^*(y_k , kh) - g^*(y_{bk}^* , kh)}{y_k - y_{bk}^*} , \qquad & y_k \neq y_{bk}^* \\ &= 0 , \qquad & y_k = y_{bk}^* . \end{aligned}$$

Then $-\frac{1}{2}(\beta - \alpha) \leqq a(k) \leqq \frac{1}{2}(\beta - \alpha)$, and

$$y_n - y_{bn}^* = \sum_{k=0}^{n} w_{n-k} a(k)(y_k - y_{bk}^*) + d_n , \qquad n \geqq 0.$$

By Lemma 2, we have $(y - y_b^*) \varepsilon\, l_2$, and since it is clear that y_b^* depends on y_a^*, but not on $[f(0, nh) - y_a^*]$, this completes the proof of Theorem 2.

3.3 *Proof of Theorem 3*

We need the following lemma.

Lemma 5: If $y_n = y_n^* + \eta_n$ with $y^* \varepsilon\, \mathcal{K}$ and $\eta_n \to 0$ as $n \to \infty$, if $g(x, kh) = g[x, (k + K + 1)h]$ for all $k \geqq 0$ and all x, if there exists a positive constant c such that $| g(u_1 , kh) - g(u_2 , kh) | \leqq c \,| u_1 - u_2 |$ for all real u_1 and u_2 and all $k \geqq 0$, and if

$$y_n = \sum_{k=0}^{n} w_{n-k} g(y_k , kh) + p_n + \delta_n , \qquad n \geqq 0$$

with $p \varepsilon\, \mathcal{K}$ and $\delta_n \to 0$ as $n \to \infty$, then

$$y_n^* = \sum_{k=-\infty}^{n} w_{n-k} g^*(y_k^* , kh) + p_n$$

for all $n = 0, \pm 1, \pm 2, \cdots$, in which $g^*(x, kh)$ is defined by the conditions:

$$g^*(x, kh) = g^*[x, (k + K + 1)h]$$

for all k and all x, and

$$g^*(x, kh) = g(x, kh)$$

for all x and $k = 0, 1, 2, \cdots , K$.

Proof of Lemma 5:

For $n \geqq 0$:

$$\begin{aligned} y_n^* + \eta_n &= \sum_{k=0}^{n} w_{n-k} g[y_k^* + \eta_k , kh] + p_n + \delta_n \\ &= \sum_{k=0}^{n} w_{n-k} g(y_k^* , kh) + \sum_{k=0}^{n} w_{n-k}[g(y_k^* + \eta_k , kh) - g(y_k^* , kh)] \\ &\qquad + p_n + \delta_n \\ & \sum_{k=-\infty}^{n} w_{n-k} g^*(y_k^* , kh) + \sum_{k=0}^{n} w_{n-k}[g(y_k^* + \eta_k , kh) - g(y_k^* , kh)] \\ &\qquad - \sum_{k=-\infty}^{-1} w_{n-k} g^*(y_k^* , kh) + p_n + \delta_n . \end{aligned}$$

Therefore,

$$\begin{aligned} y_n^* - \sum_{k=-\infty}^{n} w_{n-k} g^*(y_k^*, kh) - p_n = &-\eta_n \\ &+ \sum_{k=0}^{n} w_{n-k}[g(y_k^* + \eta_k, kh) - g(y_k^*, kh)] \\ &- \sum_{k=-\infty}^{-1} w_{n-k} g^*(y_k^*, kh) + \delta_n, \qquad n \geqq 0. \end{aligned}$$

Since $\{w_n\} \varepsilon l_1$, both sums on the right-side approach zero as $n \to \infty$. Thus, the left side also approaches zero as $n \to \infty$. But the values of the left side are periodic. Therefore,

$$y_n^* - \sum_{k=-\infty}^{n} w_{n-k} g^*(y_k^*, kh) - p_n = 0 \tag{13}$$

for all $n \geqq 0$, and since $y^* \varepsilon \mathcal{K}$ and $p \varepsilon \mathcal{K}$, (13) holds for all n. This proves Lemma 5.

By Lemma 1,

$$y_n = \sum_{k=0}^{n} w_{n-k} g(y_k, kh) + \sum_{k=0}^{n} w_{n-k} f(0, kh) + \sum_{k=0}^{n} v_{n-k}\varphi_k, \qquad n \geqq 0$$

in which $g(y_k, kh)$ is defined in Lemma 1, and $\varphi_k = 0$ for $k \geqq (p+1)$. Since $\{w_n\}$ and $\{v_n\} \varepsilon l_1$, and $f(0, kh) \to 0$ as $k \to \infty$, we have

$$\sum_{k=0}^{n} w_{n-k} f(0, kh) + \sum_{k=0}^{n} v_{n-k}\varphi_k \to 0 \quad \text{as} \quad n \to \infty .$$

By Lemma 5 and the hypotheses of Theorem 3,

$$y_n^* = \sum_{k=-\infty}^{n} w_{n-k} g^*(y_k^*, kh)$$

for $n = 0, \pm 1, \pm 2, \cdots$, with $y^* \varepsilon \mathcal{K}$. If ρ_K were less than unity, it would follow from Lemma 4 (in particular the uniqueness property of y^* of Lemma 4) that $y_n^* = 0$ for all n, since $g^*(0, kh) = 0$ for all $k \geqq 0$. Therefore, $\rho_K \geqq 1$, which completes the proof of Theorem 3.

3.4 *Proof of Theorem 4:*

According to Lemma 1,

$$y_n - z_n = \sum_{k=0}^{n} w_{n-k}[g(y_k, kh) - g(z_k, kh)] + \sum_{k=0}^{n} v_{n-k}\psi_k, \qquad n \geqq 0.$$

Therefore, with

$$b_n = \sum_{k=0}^{n} v_{n-k}\psi_k , \qquad n \geqq 0$$

we have, by Lemma 2,

$$\langle y - z \rangle_N \geqq (1 + \rho)^{-1} \langle b \rangle_N$$

and if $\rho < 1$,

$$\langle y - z \rangle_N \leqq (1 - \rho)^{-1} \langle b \rangle_N .$$

Since[2]

$$\langle b \rangle_N \leqq \max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle \psi \rangle_N ,$$

it remains only to prove the following lemma.†

Lemma 6: If

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

then

$$\langle d \rangle_N \geqq \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle c \rangle_N .$$

Proof:

Let $\{e_k\}$ be the inverse z-transform of $V^{-1}(z)$. Clearly, $\{e_k\} \in l_1$. We have

$$\sum_{m=0}^{n} e_{n-m}\, d_m = \sum_{m=0}^{n} e_{n-m} \sum_{k=0}^{m} v_{m-k} c_k = c_n \quad \text{for} \quad n \geqq 0.$$

Thus,[2]

$$\langle c \rangle_N \leqq \max_{0 \leqq \omega \leqq 2\pi} | V^{-1}(e^{i\omega}) | \langle d \rangle_N$$

and, since $F(z) = V^{-1}(z)$,

$$\langle d \rangle_N \geqq (\max_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |)^{-1} \langle c \rangle_N$$

$$\geqq \min_{0 \leqq \omega \leqq 2\pi} | F(e^{i\omega}) |^{-1} \langle c \rangle_N$$

which proves Lemma 6, and completes the proof of Theorem 4.

3.5 *Proof of the Corollary to Theorem 4*

Minkowski's inequality.

† Lemma 6 is proved in Ref. 2. The proof given here is simpler

IV. ACKNOWLEDGMENT

The writer is pleased to acknowledge the discussions held with his colleagues H. Schichman and J. F. Traub on various aspects of Numerical Analysis.

REFERENCES

1. Ralston, A., *First Course in Numerical Analysis,* McGraw-Hill Book Company, Inc., New York, 1965.
2. Sandberg, I. W., Two Theorems on the Accuracy of Numerical Solutions of Systems of Ordinary Differential Equations, B.S.T.J., *46,* July-August, 1967, pp. 1243–1266.
3. Hamming, R. W., *Numerical Methods for Scientists and Engineers,* McGraw-Hill Book Company, Inc., New York, 1962.
4. Wilkinson, J. H., *Rounding Errors in Algebraic Processes,* Prentice Hall, Englewood Cliffs, New Jersey, 1963.
5. Sandberg, I. W., On the Theory of Physical Systems Governed by Nonlinear Functional Equations, B.S.T.J., *44,* May-June, 1965, p. 871.
6. Sandberg, I. W., On the Boundedness of Solutions of Nonlinear Integral Equations, B.S.T.J., *44,* March 1965, p. 439–453.
7. Aizerman, M. A. and Gantmacher, F. R., *Absolute Stability of Regulator Systems,* Holden-Day, San Francisco, 1964, p. 51.