Floating-Point-Roundoff Accumulation in Digital-Filter Realizations

By I. W. SANDBERG

(Manuscript received June 20, 1967)

In this paper, several results are presented concerning the effects of roundoff in the floating-point realization of a general discrete filter governed ideally by a stable difference equation of the form

$$w_n = \sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k w_{n-k} , \qquad n \geqq N \tag{1}$$

in which $\{w_n\}$ and $\{x_n\}$ are output and input sequences, respectively.

In particular, for a large class of filters it is proved that there is a function $f(K)$ with $f(K) \to 0$ as $K \to \infty$ and a constant c, both dependent on the b_k, the a_k, the order in which the products on the right side of (1) are summed in the machine, and t, the number of bits allotted to the mantissa, such that

$$\langle e \rangle_K \leqq c \langle y \rangle_K + f(K)$$

for all $K \geqq N$, in which, with $\{y_n\}$ the computed output sequence of the realized filter,

$$\langle y \rangle_K = \left(\frac{1}{K+1} \sum_{n=0}^{K} | y_n |^2 \right)^{\frac{1}{2}}$$

and

$$\langle e \rangle_K = \left(\frac{1}{K+1} \sum_{n=0}^{K} | w_n - y_n |^2 \right)^{\frac{1}{2}}.$$

Bounds on $f(K)$ and c are given that are not difficult to evaluate, and which, in many realistic cases, are informative. For example, for the second-order bandpass filter:

$$w_n = x_n - a_1 w_{n-1} - a_2 w_{n-2} , \qquad n \geqq 2 \tag{2}$$

with a_1 and a_2 chosen so that its poles are at approximately $\pm 45°$ and at distance approximately (but not less than) 0.001 from the unit circle,

we find that c, an upper bound on the "asymptotic output error-to-signal ratio", is not greater than 0.58 × 10^{-4}, assuming that t = 27, that the terms on the right side of (2) are summed in the machine in the order indicated (from right to left), and that the x_n in (2) are machine numbers. If the x_n are not machine numbers, and hence must be quantized before processing, then $c \leqq 0.76 \times 10^{-4}$.

In addition to error bounds, an inequality is derived which, if satisfied, rules out certain types of generally undesirable behavior such as self-sustained output limit cycles due to roundoff effects. This inequality is satisfied for the example described above.

I. INTRODUCTION

The difference equation

$$w_n = \sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k w_{n-k}, \qquad n \geqq N \tag{1}$$

with $M \leqq N$ defines the behavior of a general time-invariant discrete filter which acts on an input sequence $x_0, x_1, x_2, \cdots$ to produce an output sequence $w_N, w_{N+1}, w_{N+2}, \cdots$ that depends on the starting values $w_0, w_1, \cdots, w_{N-1}$.

There is a vast literature concerned with techniques for designing discrete filters [i.e., for determining the a_k and the b_k in (1)] to meet specifications of various types (see, for example, Refs. 1, 2, and 3), and a good deal of material is available on the subject of roundoff effects in fixed-point realizations of discrete filters (see, for instance, Refs. 4 and 5). In this paper, we derive some bounds on a meaningful measure of the overall effect of roundoff errors for discrete filters realized as digital filters on a machine employing floating-point arithmetic operations. This type of realization, as opposed to the fixed-point kind, is of particular importance in connection with, for example, digital computer simulations of systems, as a result of the large dynamic range afforded by the floating-point mode.

There are basic differences concerning fixed-point and floating-point error estimation problems which stem from the fact that the modulus of every individual arithmetic error in the fixed-point mode is bounded by a constant determined by the machine, whereas the maximum modulus of the error in forming, for example, the floating-point sum of two floating-point numbers is proportional to the magnitude of the true sum. For this reason, the approach* presented here, as well as the

* The approach can be extended in several different directions. For example, it can be used to obtain statistical error estimates based on the assumption that each roundoff error is an independent random variable.

character of the results, are quite different from those of earlier writers concerned with fixed-point realizations.

In addition to error bounds, an inequality is derived which, if satisfied, rules out certain types of generally undesirable behavior such as self-sustained output limit cycles due to roundoff effects.

II. ASSUMPTIONS AND RESULTS

2.1 *Assumptions*

It is assumed that:

(*i*) each machine number q is equal to sgn (q) a 2^b in which the exponent b is an integer, and a, the mantissa, is a t-bit number contained in $[\frac{1}{2}, 1]$ or $[\frac{1}{2}, 1] \cup \{0\}$;

(*ii*) the range of values of b is adequate to ensure that all computed numbers lie within the permissible range;

(*iii*) the machine operations of addition and multiplication are performed in accordance with standard rounding conventions* (described, for example, by Wilkinson[6]); and

(*iv*) the coefficients a_k and b_k in (1) are machine numbers.†

2.2 *Results: x_n Machine Numbers*

It is assumed throughout Section 2.2 that the x_n of (1) are floating-point machine numbers.

If the discrete filter (1) is realized on a floating-point machine, then

$$y_n = fl\left(\sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k y_{n-k}\right), \qquad n \geqq N \tag{2}$$

in which the y_n are approximations to the infinite precision numbers w_n, and $fl(\Sigma - \Sigma)$ denotes the machine number corresponding to $(\Sigma - \Sigma)$ with the understanding that the floating-point numbers corresponding to the products $b_k x_{n-k}$ and $a_k y_{n-k}$ are to be machine-added in some specified order.

Let

$$D(z) \triangleq 1 + \sum_{k=1}^{N} a_k z^{-k}, \tag{3}$$

---

* That is, conventions for which the first two equations of Section III are satisfied.

† It is certainly true that *preliminary* design considerations may lead to coefficients that are not machine numbers, and one may then be interested also in the overall effect of approximating the coefficients by machine numbers. That problem also can be treated with the approach used here.

let

$$\langle q \rangle_K \triangleq \left(\frac{1}{K+1} \sum_{k=0}^{K} | q_k |^2 \right)^{\frac{1}{2}}$$

for every sequence $\{q_k\}$ and all $K \geqq 0$, and let e_n denote the nth error $(y_n - w_n)$ for $n \geqq 0$.

Our first result (all proofs are given in Section III) is as follows. If $D(z) \neq 0$ for $| z | \geqq 1$ [i.e., if the discrete filter (1) is stable], then

$$\begin{aligned} \langle e \rangle_K \leqq & \max_{0 \leq \omega \leq 2\pi} | D(e^{i\omega})^{-1} | \left(\frac{1}{K+1} \sum_{n=0}^{N-1} | \eta_n |^2 \right)^{\frac{1}{2}} \\ & + 2^{-t} \left(\sum_{k=0}^{M} | b_k | \beta_k \right) \max_{0 \leq \omega \leq 2\pi} | D(e^{i\omega})^{-1} | \left(\frac{1}{K+1} \sum_{n=N-M}^{K} | x_n |^2 \right)^{\frac{1}{2}} \\ & + 2^{-t} \left(\sum_{k=1}^{N} | a_k | \alpha_k \right) \max_{0 \leq \omega \leq 2\pi} | D(e^{i\omega})^{-1} | \langle y \rangle_K \end{aligned} \tag{4}$$

for all $K \geqq N$, in which, with $y_n = w_n = 0$ for $n < 0$,

$$\eta_n = \sum_{k=0}^{N} a_k (y_{n-k} - w_{n-k}) \qquad n = 0, 1, 2, \cdots, (N-1)$$

and the α_k and β_k are easily evaluated nonnegative numbers which depend on the order in which the products in (2) are summed.

Since the first term on the right side of (4), which arises as a result of the possibility of differences in the starting values, approaches zero as $K \to \infty$, we see that, after a reasonable number of evaluations of the successive y_n, $\langle e \rangle_K$ is bounded essentially by a constant times the root-mean-squared value of the input sequence, plus another constant times the root-mean-squared value of the output sequence.

In order to determine the α_k and β_k, we draw a signal-flow graph that indicates the ordering of the operations that would be used to compute

$$fl\left(\sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k y_{n-k} \right) \tag{5}$$

if x_n and y_n were unity for all n. This graph is to contain an input node with input b_k' for each $b_k \neq 0$, an input node with input a_k' for each $a_k \neq 0$, no other input nodes, and a single output node θ which is associated with

$$\sum_{k=0}^{M} b_k' - \sum_{k=1}^{N} a_k' .$$

All other nodes represent an addition or subtraction of two signals to produce a third signal. Exactly one branch is connected to each of the input nodes and to the output node. We assign the value ρ to all of the branch transmissions with the exception of those branches, if any, which terminate on an input b_k' or a_k' for which b_k or a_k, respectively, is equal to unity. These branches are assigned unity transmission. Then, by inspection, we evaluate the signal at θ, which must clearly be of the form

$$\sum_{k=0}^{M} b_k' \rho^{\varphi_\beta(k)} + \sum_{k=1}^{N} a_k' \rho^{\varphi_\alpha(k)} \tag{6}$$

in which $\varphi_\beta(k)$ and $\varphi_\alpha(k)$ are positive-integer valued functions. In terms of these functions*

$$\beta_k = (1.06)\varphi_\beta(k)$$

$$\alpha_k = (1.06)\varphi_\alpha(k).$$

For example, if the right side of (2) is computed as the floating-point difference of the machine sums

$$fl(b_0 x_n + b_1 x_{n-1} + \cdots + b_M x_{n-M})$$

and

$$fl(a_1 y_{n-1} + a_2 y_{n-2} + \cdots + a_N y_{n-N}),$$

each obtained by performing machine summations in the order indicated (from left to right), if all of the b_k and a_k are nonzero and not unity, and if $M \geqq 1$ and $N \geqq 2$, then the relevant flow graph is shown in Fig. 1, from which it follows that

$$\beta_0 = (1.06)(M + 2)$$

$$\beta_1 = (1.06)(M + 2)$$

$$\beta_k = (1.06)(3 + M - k); \qquad k = 2, 3, \cdots, M$$

$$\alpha_1 = (1.06)(N + 1)$$

$$\alpha_2 = (1.06)(N + 1)$$

$$\alpha_k = (1.06)(3 + N - k); \qquad k = 3, 4, \cdots, N.$$

The bound (4), although revealing, requires a knowledge of both $\langle x \rangle_K$ and $\langle y \rangle_K$ and is, therefore, not as explicit as we would like.

* We are assuming here only that $\max_k |\varphi_\beta(k)|2^{-t} < 0.1$ and $\max_k |\varphi_\alpha(k)|2^{-t} < 0.1$. Also if $\varphi_\beta(k) = 1$, then we can take $\beta_k = 1$, and similarly for $\varphi_\alpha(k)$.

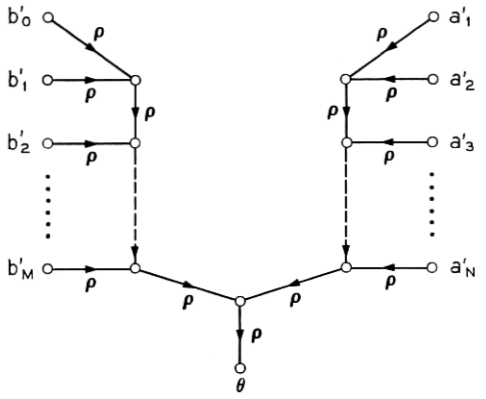

Fig. 1 — Flow graph for the example.

For the important case in which $b_0 \neq 0$ and $N(z) \triangleq \sum_{k=0}^{M} b_k z^{-k} \neq 0$ for $| z | \geqq 1$ (i.e., for the minimum-phase filter case) we prove that if the filter (1) is stable and if

$$\min_{0 \leqq \omega \leqq 2\pi} | N(e^{i\omega}) | > 2^{-t} \sum_{k=0}^{M} | b_k | \beta_k , \tag{7}$$

then there exists a constant c, independent of K, and a function $f(K)$ with the property that $f(K) \to 0$ as $K \to \infty$ such that

$$\langle e \rangle_K \leqq c \langle y \rangle_K + f(K) \tag{8}$$

for all $K \geqq N$. Moreover, it is proved that

$$c \leqq 2^{-t} \max_{0 \leqq \omega \leqq 2\pi} | D(e^{i\omega})^{-1} | \left\{ \sum_{k=1}^{N} | a_k | \alpha_k + \sum_{k=0}^{M} | b_k | \beta_k \cdot \frac{\max_{\omega} | D(e^{i\omega})/N(e^{i\omega}) | + \max_{\omega} | N(e^{i\omega})^{-1} | 2^{-t} \sum_{k=1}^{N} | a_k | \alpha_k}{1 - 2^{-t} \sum_{k=0}^{M} | b_k | \beta_k \max_{\omega} | N(e^{i\omega})^{-1} |} \right\} \tag{9}$$

and

$$f(K) \leqq \max_{\omega} | D(e^{i\omega})^{-1} | \left(\frac{1}{K+1} \sum_{n=0}^{N-1} | \eta_n |^2 \right)^{\frac{1}{2}} + \max_{\omega} | D(e^{i\omega})^{-1} | 2^{-t}$$

$$\cdot\left(\sum_{k=0}^{M} \mid b_k \mid \beta_k\right) \frac{\max \mid N(e^{i\omega})^{-1} \mid \left(\frac{1}{K+1}\sum_{n=0}^{N-1} \mid q_n \mid^2\right)^{\frac{1}{2}}}{1 - 2^{-t}\sum_{k=0}^{M} \mid b_k \mid \beta_k \max_{\omega} \mid N(e^{i\omega})^{-1} \mid} \tag{10}$$

for all $K \geqq N$, in which, with $a_0 = 1$ and $x_n = y_n = 0$ for $n < 0$,

$$q_n = \sum_{k=0}^{N} a_k y_{n-k} - \sum_{k=0}^{M} b_k x_{n-k}$$

for $n = 0, 1, 2, \cdots, (N-1)$.

Since $\langle y \rangle_K$ is the root-mean-squared value of the *computed output*, and since $f(K) \to 0$ fairly rapidly as $K \to \infty$, we may interpret the smallest value of c for which (8) is satisfied (for all input sequences) as an "output error-to-signal ratio" of the realized digital filter. Note that the bound (9) on c is not difficult to evaluate.

2.2.1. *Stability in the Presence of Roundoff*

If roundoff effects are ignored, it is well known that the discrete filter is stable in several different senses of the word if $D(z) \neq 0$ for $\mid z \mid \geqq 1$. In Section III it is proved that, with roundoff effects taken into account, the digital filter is stable in the sense that there is a constant c_1 and a function $f_1(K)$, with $f_1(K)$ independent of the values of x_n for $n \geqq N$ and $f_1(K) \to 0$ as $K \to \infty$, such that

$$\langle y \rangle_K \leqq c_1 \langle x \rangle_K + f_1(K) \tag{11}$$

for all $K \geqq N$, provided that $D(z) \neq 0$ for $\mid z \mid \geqq 1$, and

$$\min_{\omega} \mid D(e^{i\omega}) \mid > 2^{-t} \sum_{k=1}^{N} \mid a_k \mid \alpha_k \, . \tag{12}$$

Roughly speaking, inequality (12) is satisfied if the damping of the infinite precision counterpart of the digital filter is sufficiently large relative to the number of bits allotted to the mantissa. Stability in the sense of (11) rules out, for example, the possibility, due to roundoff effects, of a limit-cycle response to a zero input sequence or to an input sequence $\{x_n\}$ that approaches zero as $n \to \infty$.*

* There are simple examples which illustrate that instability may result with $D(z) \neq 0$ for $\mid z \mid \geqq 1$ if (12) is not satisfied. For instance, suppose that each machine number is represented in the form $(-m_0 2^0 + m_1 2^{-1} + m_2 2^{-2} + \cdots + m_t 2^{-t})2^b$ with the m_i zeros or ones, and $t > 1$. Let

$$w_n = (1 - 2^{-t})w_{n-1} + (1 - 2^{-t})2^{-t}w_{n-2} \text{ for } n \geqq 2, \text{ with } w_0 = w_1 = -1.$$

Then $fl[(1 - 2^{-t})w_1] = -(1 - 2^{-t})$, $fl[(1 - 2^{-t})2^{-t}w_0] = -(1 - 2^{-t})2^{-t}$, and $fl[-(1 - 2^{-t}) - (1 - 2^{-t})2^{-t}] = -1$, which shows that the computed approximation y_n to w_n satisfies $y_n = -1$ for *all* $n \geqq 0$. This example is a slight modification of one suggested by S. Darlington.

2.3 *A Result Concerning the Overall Effect of Input Quantization Errors*

In many applications the sequence $\{x_n\}$ of (1) is obtained by quantizing an input sequence $\{\bar{x}_n\}$ [i.e., by replacing each $\bar{x}_n$ with the machine number (or one of the possibly two machine numbers) of closest value]. The infinite precision response $\bar{w}_N$, $\bar{w}_{N+1}$, $\cdots$ to the sequence $\{\bar{x}_N\}$ satisfies

$$\bar{w}_n = \sum_{k=0}^{M} b_k \bar{x}_{n-k} - \sum_{k=1}^{N} a_k \bar{w}_{n-k}, \qquad n \geqq N \tag{13}$$

with $\bar{w}_0$, $\bar{w}_1$, $\cdots$, $\bar{w}_{N-1}$ some set of starting values. Let w_N, w_{N+1}, $\cdots$ be defined by (1) with $w_n = \bar{w}_n$ for $n = 0, 1, 2, \cdots, (N-1)$. It is clear that $\langle y - \bar{w} \rangle_K$, the root-mean-squared value of the difference of the computed output and the infinite precision response to $\{\bar{x}_n\}$, satisfies

$$\langle y - \bar{w} \rangle_K \leqq \langle y - w \rangle_K + \langle w - \bar{w} \rangle_K . \tag{14}$$

Bounds on the first term on the right side of (14) are given in Section 2.2. In Section III it is proved that if both $N(z)$ and $D(z)$ have no zeros on or outside the unit circle, $b_0 \neq 0$, and

$$\min_{\omega} | N(e^{i\omega}) | > 2^{-t} \sum_{k=0}^{M} | b_k | \beta_k ,$$

then* there is a constant c_2 and a function $f_2(K)$ such that $f_2(K) \to 0$ as $K \to \infty$, and

$$\langle w - \bar{w} \rangle_K \leqq c_2 \langle y \rangle_K + f_2(K) \tag{15}$$

for all $K \geqq N$. It is proved also that

$$c_2 \leqq 2^{-t} \sum_{k=0}^{M} | b_k | \max_{\omega} | D(e^{i\omega})^{-1} | \cdot \frac{\max_{\omega} | D(e^{i\omega})/N(e^{i\omega}) | + \max_{\omega} | N(e^{i\omega})^{-1} | 2^{-t} \sum_{k=1}^{N} | a_k | \alpha_k}{1 - 2^{-t} \sum_{k=0}^{M} | b_k | \beta_k \max_{\omega} | N(e^{i\omega})^{-1} |} . \tag{16}$$

2.4 *A Realistic Example*

For the ideally stable second-order bandpass filter

$$w_n = x_n - a_1 w_{n-1} - a_2 w_{n-2}, \qquad n \geqq 2$$

* It is assumed here that the range of values assigned to the mantissa includes the number zero.

with poles in the z-plane at angles $\approx \pm 45°$ and at distance ≈ 0.001 (but not less than 0.001) from the unit circle, we have $a_1 \approx -1.41$, $a_2 \approx 1$, and $\min_\omega \mid D(e^{i\omega})^{-1} \mid \approx (0.00141)^{-1}$. We assume that the operations are performed as indicated in Fig. 2, so that $\beta_0 = 1$, $\alpha_1 = 3(1.06)$, and $\alpha_2 = 3(1.06)$. Assuming that $t = 27$, we find that c our bound on the "asymptotic output error-to-signal ratio," ignoring input quantization effects, is approximately 0.584×10^{-4}. For this problem, our bound on c_2 is approximately 0.18×10^{-4}. Thus, even taking into account input quantization effects, the error-to-signal ratio is not more than 0.764×10^{-4}. Finally, a simple calculation shows that this filter is stable in the presence of roundoff, in the sense of inequality (11).

III. PROOFS

3.1 *Derivation of Inequality (4)*

If a and b are floating-point machine numbers, then the floating-point product and sum $fl(ab)$ and $fl(a + b)$, respectively, satisfy[6]

$$fl(ab) = ab(1 + \epsilon)$$

$$fl(a + b) = (a + b)(1 + \delta)$$

with $\mid \epsilon \mid \leqq 2^{-t}$ and $\mid \delta \mid \leqq 2^{-t}$. Thus,

$$fl\left(\sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k y_{n-k}\right)$$

is equal to the value of the output signal θ of the flow graph described in Section II with

(*i*) $b'_k = b_k x_{n-k}$

$a'_k = a_k y_{n-k}$

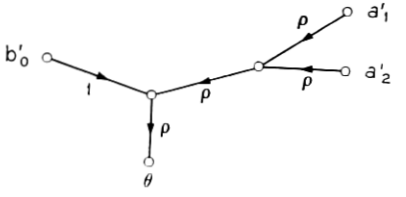

Fig. 2 — Flow graph for the second-order band-pass filter.

and

(*ii*) each of the branch transmissions of the form: $(1 + \epsilon)$ with $| \epsilon | \leqq 2^{-t}$ (recall that in certain special cases ϵ is *taken* to be zero), or $-(1 + \epsilon)$ with $| \epsilon | \leqq 2^{-t}$. Therefore,

$$fl\left(\sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k y_{n-k} \right)$$

is equal to

$$\sum_{k=0}^{M} b_k x_{n-k} q_k - \sum_{k=1}^{N} a_k y_{n-k} r_k$$

in which

$$(1 - 2^{-t})^{\varphi_\beta(k)} \leqq q_k \leqq (1 + 2^{-t})^{\varphi_\beta(k)} \tag{17}$$

and

$$(1 - 2^{-t})^{\varphi_\alpha(k)} \leqq r_k \leqq (1 + 2^{-t})^{\varphi_\alpha(k)}. \tag{18}$$

Inequalities (17) and (18) imply[6]

$$1 - (1.06)\varphi_\beta(k)2^{-t} \leqq q_k \leqq 1 + (1.06)\varphi_\beta(k)2^{-t}$$

$$1 - (1.06)\varphi_\alpha(k)2^{-t} \leqq r_k \leqq 1 + (1.06)\varphi_\alpha(k)2^{-t}$$

provided that $2^{-t} \max_k \varphi_\beta(k) < 0.1$ and $2^{-t} \max_k \varphi_\alpha(k) < 0.1$.

Thus, for $n \geqq N$

$$\begin{aligned} y_n &= fl\left(\sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k y_{n-k} \right) \\ &= \sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k y_{n-k} + \eta_n \end{aligned} \tag{19}$$

with

$$| \eta_n | \leqq 2^{-t} \sum_{k=0}^{M} | b_k | \cdot | x_{n-k} | \beta_k + 2^{-t} \sum_{k=1}^{N} | a_k | \cdot | y_{n-k} | \alpha_k \tag{20}$$

and

$$\beta_k = (1.06)\varphi_\beta(k), \qquad \alpha_k = (1.06)\varphi_\alpha(k).$$

Using (1) and (19),

$$\sum_{k=0}^{N} a_k e_{n-k} = \eta_n , \qquad n \geqq 0$$

in which, with $y_n = w_n = 0$ for $n < 0$,

$$\eta_n = \sum_{k=0}^{N} a_k(y_{n-k} - w_{n-k})$$

for $n = 0, 1, \cdots, (N-1)$. By Propositions 1 and 2 (see Sections 3.5 and 3.6)

$$\langle e\rangle_K \leqq \max_{0\leqq\omega\leqq 2\pi} | D(e^{i\omega})^{-1} | \langle\eta\rangle_K , \qquad K \geqq 0. \tag{21}$$

By Proposition 3 (Section 3.7), inequality (20), and Minkowski's inequality

$$\langle\eta\rangle_K \leqq \left(\frac{1}{K+1}\sum_{n=0}^{N-1} | \eta_n |^2\right)^{\frac{1}{2}} + 2^{-t}\sum_{k=0}^{M} | b_k | \beta_k\left(\frac{1}{K+1}\sum_{n=N-M}^{K} | x_n |^2\right)^{\frac{1}{2}} + 2^{-t}\sum_{k=1}^{N} | a_k | \alpha_k\langle y\rangle_K \tag{22}$$

for all $K \geqq N$. This proves inequality (4).

3.2 *Inequality (8)*

Here we assume that both $D(z)$ and $N(z)$ are zero free for $| z | \geqq 1$, that $b_0 \neq 0$, and that

$$\min_{0\leqq\omega\leqq 2\pi} | N(e^{i\omega}) | > 2^{-t}\sum_{k=0}^{M} | b_k | \beta_k . \tag{23}$$

From (19), we have, with $a_0 \triangleq 1$,

$$\sum_{k=0}^{N} a_k y_{n-k} = \sum_{k=0}^{M} b_k x_{n-k} + q_n , \qquad n \geqq 0, \tag{24}$$

where

$$q_n = \eta_n , \qquad n \geqq N$$
$$= \sum_{k=0}^{N} a_k y_{n-k} - \sum_{k=0}^{M} b_k x_{n-k} , \qquad n = 0, 1, 2, \cdots, (N-1)$$

with $x_n = y_n = 0$ for $n < 0$. Therefore, by Propositions 1 and 2,

$$\langle x\rangle_K \leqq \max_{\omega} | D(e^{i\omega})/N(e^{i\omega}) | \langle y\rangle_K + \max_{\omega} | N(e^{i\omega})^{-1} | \langle q\rangle_K , \quad K \geqq 0. \tag{25}$$

Using Proposition 3, Minkowski's inequality, and (20),

$$\langle q\rangle_K \leqq \left(\frac{1}{K+1}\sum_{n=0}^{N-1} | q_n |^2\right)^{\frac{1}{2}} + 2^{-t}\sum_{k=0}^{M} | b_k | \beta_k\langle x\rangle_K + 2^{-t}\sum_{k=1}^{N} | a_k | \alpha_k\langle y\rangle_K , \qquad K \geqq N. \tag{26}$$

Therefore,

$$\langle x\rangle_K \leqq \frac{\max_\omega \mid D(e^{i\omega})/N(e^{i\omega}) \mid + \max_\omega \mid N(e^{i\omega})^{-1} \mid 2^{-t} \sum_{k=1}^{N} \mid a_k \mid \alpha_k}{1 - 2^{-t} \sum_{k=0}^{M} \mid b_k \mid \beta_k \max_\omega \mid N(e^{i\omega})^{-1} \mid} \langle y\rangle_K$$

$$+ \frac{\max_\omega \mid N(e^{i\omega})^{-1} \mid \left(\frac{1}{K+1} \sum_{n=0}^{N-1} \mid q_n \mid^2\right)^{\frac{1}{2}}}{1 - 2^{-t} \sum_{k=0}^{M} \mid b_k \mid \beta_k \max_\omega \mid N(e^{i\omega})^{-1} \mid} \tag{27}$$

for all $K \geqq N$, which together with (21) and (22) yields

$$\langle e\rangle_K \leqq 2^{-t} \max_{0\leq\omega\leq 2\pi} \mid D(e^{i\omega})^{-1} \mid \left\{ \sum_{k=1}^{N} \mid a_k \mid \alpha_k + \sum_{k=0}^{M} \mid b_k \mid \beta_k \right.$$

$$\left. \cdot \frac{\max_\omega \mid D(e^{i\omega})/N(e^{i\omega}) \mid + \max_\omega \mid N(e^{i\omega})^{-1} \mid 2^{-t} \sum_{k=1}^{N} \mid a_k \mid \alpha_k}{1 - 2^{-t} \sum_{k=0}^{M} \mid b_k \mid \beta_k \max_\omega \mid N(e^{i\omega})^{-1} \mid} \right\} \langle y\rangle_K$$

$$+ \max_\omega \mid D(e^{i\omega})^{-1} \mid \left(\frac{1}{K+1} \sum_{n=0}^{N-1} \mid \eta_n \mid^2\right)^{\frac{1}{2}} + \max_\omega \mid D(e^{i\omega})^{-1} \mid 2^{-t}$$

$$\cdot \sum_{k=0}^{M} \mid b_k \mid \beta_k \frac{\max_\omega \mid N(e^{i\omega})^{-1} \mid \left(\frac{1}{K+1} \sum_{n=0}^{N-1} \mid q_n \mid^2\right)^{\frac{1}{2}}}{1 - 2^{-t} \sum_{k=0}^{M} \mid b_k \mid \beta_k \max_\omega \mid N(e^{i\omega})^{-1} \mid}. \tag{28}$$

This proves that there exists a constant c and a function $f(K)$ with the property that $f(K) \to 0$ as $K \to \infty$ such that (8) is satisfied for all $K \geqq N$, and of course it also proves that c and $f(K)$ are bounded as stated in Section 2.2.

3.3 *Proof of (11) Under the Conditions Stated*

From (24) and Propositions 1 and 2,

$$\langle y\rangle_K \leqq \max_\omega \mid N(e^{i\omega})/D(e^{i\omega}) \mid \langle x\rangle_K + \max_\omega \mid D(e^{i\omega})^{-1} \mid \langle q\rangle_K ,$$

and using (26)

$$\langle y\rangle_K \leqq \frac{\max_\omega |N(e^{i\omega})/D(e^{i\omega})| + \max_\omega |D(e^{i\omega})^{-1}|\, 2^{-t} \sum_{k=0}^{M} |b_k|\beta_k}{1 - 2^{-t}\sum_{k=1}^{N} |a_k|\alpha_k \max_\omega |D(e^{i\omega})^{-1}|}\langle x\rangle_K + \frac{\max_\omega |D(e^{i\omega})^{-1}| \left(\frac{1}{K+1}\sum_{n=0}^{N-1} |q_n|^2\right)^{\frac{1}{2}}}{1 - 2^{-t}\sum_{k=1}^{N} |a_k|\alpha_k \max_\omega |D(e^{i\omega})^{-1}|},$$

which completes the proof.

3.4 *Derivation of Inequalities (15) and (16)*

We have, from (1) and (13),

$$\sum_{k=0}^{N} a_k(w_{n-k} - \bar{w}_{n-k}) = \xi_n, \qquad n \geqq 0 \tag{29}$$

in which $a_0 \triangleq 1$,

$$\xi_n = \sum_{k=0}^{M} b_k(x_{n-k} - \bar{x}_{n-k}), \qquad n \geqq N$$

and

$$\xi_n = 0, \qquad n = 0, \quad 2, \cdots, (N-1).$$

Since $\bar{x}_n = \operatorname{sgn}(\bar{x}_n)h2^b$ for some integer b and some $h \in [\frac{1}{2}, 1]$ (assuming that $\bar{x}_n \neq 0$), the magnitude of the error in approximating $\bar{x}_n$ by the closest machine number $x_n = \operatorname{sgn}(\bar{x}_n)a2^b$ is at most $\frac{1}{2}2^{-t}2^b = \frac{1}{2}2^{-t}a^{-1}|x_n| \leqq 2^{-t}|x_n|$. Therefore, for $n \geqq N$

$$|\xi_n| \leqq 2^{-t}\sum_{k=0}^{M} |b_k|\cdot|x_{n-k}|,$$

and by Propositions 1, 2, and 3

$$\langle w - \bar{w}\rangle_K \leqq \max_\omega |D(e^{i\omega})^{-1}|\, 2^{-t} \cdot \sum_{k=0}^{M} |b_k| \left(\frac{1}{K+1}\sum_{n=N-M}^{K} |x_n|^2\right)^{\frac{1}{2}}, \qquad K \geqq N. \tag{30}$$

From (30) and (27)

$$\langle w - \bar{w}\rangle_K \leqq 2^{-t}\sum_{k=0}^{M} |b_k| \max_\omega |D(e^{i\omega})^{-1}|$$

$$\cdot \frac{\max_{\omega} | D(e^{i\omega})/N(e^{i\omega}) | + \max_{\omega} | N(e^{i\omega})^{-1} | 2^{-t} \sum_{k=1}^{N} | a_k | \alpha_k}{1 - 2^{-t} \sum_{k=0}^{M} | b_k | \beta_k \max_{\omega} | N(e^{i\omega})^{-1} |} \langle y \rangle_K$$

$$+ \frac{2^{-t} \sum_{k=0}^{M} | b_k | \max_{\omega} | D(e^{i\omega})^{-1} | \max_{\omega} | N(e^{i\omega})^{-1} | \left(\frac{1}{K+1} \sum_{n=0}^{N-1} | q_n |^2 \right)^{\frac{1}{2}}}{1 - 2^{-t} \sum_{k=0}^{M} | b_k | \beta_k \max_{\omega} | N(e^{i\omega})^{-1} |}$$

for all $K \geqq N$, provided that $N(z) \neq 0$ for $| z | \geqq 1$, $b_0 \neq 0$, and

$$\min_{\omega} | N(e^{i\omega}) | > 2^{-t} \sum_{k=0}^{M} | b_k | \beta_k .$$

This completes the derivation.

3.5 *Proposition 1:*

If

$$\sum_{l=0}^{L} c_l r_{n-l} = \sum_{l=0}^{L'} d_l s_{n-l} + f_n , \qquad n \geqq 0$$

with: $r_n = s_n = 0$ for $n < 0$, $c_0 \neq 0$, and $\sum_{l=0}^{L} c_l z^{-l} \neq 0$ for $| z | \geqq 1$, then

$$r_n = \sum_{k=0}^{n} u_{n-k} s_k + \sum_{k=0}^{n} v_{n-k} f_k , \qquad n \geqq 0$$

in which

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

$$\sum_{n=0}^{\infty} u_n e^{-in\omega} = \sum_{l=0}^{L'} d_l e^{-il\omega} \Big/ \sum_{l=0}^{L} c_l e^{-il\omega} ,$$

and

$$\sum_{n=0}^{\infty} v_n e^{-in\omega} = 1 \Big/ \sum_{l=0}^{L} c_l e^{-il\omega}$$

for $0 \leqq \omega \leqq 2\pi$.

*Proof:**

* The proof of this result, although rather trivial, is included because the writer knows of no reference where it is proved without the assumption that the sequences $\{s_n\}$ and $\{f_n\}$ are z-transformable.

Let $M > 0$, and let

$$\begin{aligned} \hat{s}_n &= s_n \quad \text{for} \quad n \leqq M \\ &= 0 \quad \text{for} \quad n > M \\ \hat{f}_n &= f_n \quad \text{for} \quad n \leqq M \\ &= 0 \quad \text{for} \quad n > M. \end{aligned}$$

Then $r_n = \hat{r}_n$ for $n \leqq M$, with

$$\sum_{l=0}^{L} c_l \hat{r}_{n-l} = \sum_{l=0}^{L'} d_l \hat{s}_{n-l} + \hat{f}_n , \qquad n \geqq 0$$

and with $\{\hat{r}_n\}$, $\{\hat{s}_n\}$, and $\{\hat{f}_n\}$ z-transformable. Therefore, we have

$$\hat{R}(z) = \left(\sum_{l=0}^{L'} d_l z^{-l}\right)\left(\sum_{l=0}^{L} c_l z^{-l}\right)^{-1} \hat{S}(z) + \left(\sum_{l=0}^{L} c_l z^{-l}\right)^{-1} \hat{F}(z)$$

in which

$$\begin{aligned} \hat{R}(z) &= \sum_{n=0}^{\infty} \hat{r}_n z^{-n} \\ \hat{S}(z) &= \sum_{n=0}^{M} s_n z^{-n} \\ \hat{F}(z) &= \sum_{n=0}^{M} f_n z^{-n}. \end{aligned}$$

Thus,

$$\hat{r}_n = \sum_{k=0}^{n} u_{n-k} \hat{s}_k + \sum_{k=0}^{n} v_{n-k} \hat{f}_k , \qquad n \geqq 0$$

and hence

$$r_n = \sum_{k=0}^{n} u_{n-k} s_k + \sum_{k=0}^{n} v_{n-k} f_k , \tag{31}$$

for $n = 0, 1, \cdots, M$. However, since M is arbitrary, (31) is satisfied for all $n \geqq 0$. This proves Proposition 1.

3.6 *Proposition 2:*

If

$$f_n = \sum_{l=0}^{n} c_{n-l} g_l , \qquad n \geqq 0$$

with $\sum_{l=0}^{\infty} | c_l | < \infty$, then

$$\langle f \rangle_K \leqq \max_{0 \leqq \omega \leqq 2\pi} \left| \sum_{l=0}^{\infty} c_l e^{-il\omega} \right| \langle g \rangle_K$$

for all $K \geqq 0$.

Proof:

$$\sum_{n=0}^{K} | f_n |^2 = \frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{n=0}^{K} e^{-in\omega} \sum_{l=0}^{n} c_{n-l} g_l \right|^2 d\omega$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{n=0}^{K} e^{-in\omega} \sum_{l=0}^{n} c_{n-l} \hat{g}_l \right|^2 d\omega$$

in which

$$\hat{g}_l = g_l , \qquad l = 0, 1, \cdots, K$$

$$= 0, \qquad l > K.$$

Thus,

$$\sum_{n=0}^{K} | f_n |^2 \leqq \frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{n=0}^{\infty} e^{-in\omega} \sum_{l=0}^{n} c_{n-l} \hat{g}_l \right|^2 d\omega$$

$$\leqq \frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{l=0}^{\infty} c_l e^{-il\omega} \sum_{n=0}^{\infty} e^{-in\omega} \hat{g}_n \right|^2 d\omega$$

$$\leqq \max_{\omega} \left| \sum_{l=0}^{\infty} c_l e^{-il\omega} \right|^2 \frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{n=0}^{\infty} e^{-in\omega} \hat{g}_n \right|^2 d\omega$$

$$\leqq \max_{\omega} \left| \sum_{l=0}^{\infty} c_l e^{-il\omega} \right|^2 \sum_{n=0}^{K} | g_n |^2,$$

which proves Proposition 2.

3.7 *Proposition 3:*

If

$$| f_n | \leqq \sum_{l=0}^{L} | g_l | \cdot | h_{n-l} |, \qquad n \geqq N$$

with $L \leqq N$, then

$$\left(\sum_{n=N}^{K} | f_n |^2 \right)^{\frac{1}{2}} \leqq \left(\sum_{l=0}^{L} | g_l | \right) \left(\sum_{n=N-L}^{K} | h_n |^2 \right)^{\frac{1}{2}}$$

for all $K \geqq N$.

Proof:

$$\sum_{n=N}^{K} | f_n |^2 \leqq \sum_{n=N}^{K} \left| \sum_{l=0}^{L} | g_l | \cdot | h_{n-l} | \right|^2$$

$$\leqq \sum_{n=N}^{K} \left| \sum_{l=0}^{L} | g_l |^{\frac{1}{2}} | g_l |^{\frac{1}{2}} | \hat{h}_{n-l} | \right|^2$$

in which

$$\hat{h}_n = h_n \qquad n = 0, 1, 2, \cdots, K$$

$$= 0 \qquad n > K.$$

Therefore, by the Schwarz inequality,

$$\sum_{n=N}^{K} | f_n |^2 \leqq \sum_{n=N}^{K} \sum_{l=0}^{L} | g_l | \sum_{l=0}^{L} | g_l | \cdot | \hat{h}_{n-l} |^2$$

$$\leqq \sum_{l=0}^{L} | g_l | \sum_{l=0}^{L} \left(| g_l | \sum_{n=N}^{K} | \hat{h}_{n-l} |^2 \right)$$

$$\leqq \sum_{l=0}^{L} | g_l | \sum_{l=0}^{L} \left(| g_l | \sum_{m=N-l}^{K-l} | \hat{h}_m |^2 \right)$$

$$\leqq \left(\sum_{l=0}^{L} | g_l | \right)^2 \sum_{m=N-L}^{K} | h_m |^2.$$

This completes the proof.

IV. ACKNOWLEDGMENT

The writer is indebted to his colleague J. F. Kaiser for emphasizing the need for analytical results relating to problems of the type discussed here.

REFERENCES

1. Blackman, R. B. and Tukey, J. W., *The Measurement of Power Spectra from the Point of View of Communication Engineering,* Dover Press, 1959.
2. Blackman, R. B., *Linear Data-Smoothing and Prediction in Theory and Practice,* Addison-Wesley, Reading, Massachusetts, 1965.
3. Kaiser, J. F., Digital Filters, Chapter 7 of *System Analysis by Digital Computer,* edited by F. F. Kuo and J. F. Kaiser, John Wiley & Sons, Inc., New York, 1966.
4. Bennett, W. R., Spectra of Quantized Signals, B.S.T.J., *27,* July, 1948, pp. 446–472.
5. Knowles, J. B. and Edwards, R., Effects of a Finite-Word-Length Computer in a Sampled-Data Feedback System, Proc. IEE (London), *113,* June, 1965, pp. 1197–1207.
6. Wilkinson, J. H., *Rounding Errors in Algebraic Processes,* Prentice-Hall, Englewood Cliffs, New Jersey, 1963.