THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 4, April, 1972
Printed in U.S.A.

Some Effects of Quantization and Adder Overflow on the Forced Response of Digital Filters

By A. N. WILLSON, JR.

(Manuscript received December 28, 1971)

The effects of quantization (i.e., roundoff, truncation, etc.) and adder overflow, which are present in any special-purpose computer type realization of a digital filter, cause an otherwise linear system to become quite nonlinear. Moreover, the presence of such nonlinearities can cause the system's response to differ drastically from the ideal response (that is, from the response of the linear model of the filter) even when the level of the filter's input signal is, in a certain reasonable sense, small, and when the quantization effects are made arbitrarily small.

In this paper we derive a criterion for the satisfactory behavior of second-order digital filters in the presence of such nonlinear effects. The criterion is shown to be sharp, in that we also present a procedure for constructing counterexamples which show that, for most filters which violate the criterion, the response to some "small" nonzero input signal is not always even asymptotically close to the ideal response.

I. INTRODUCTION

The effects of quantization (i.e., roundoff, truncation, etc.) and adder overflow are present in any special-purpose computer type realization of a digital filter. When taken into account, these effects cause an otherwise linear system to become quite nonlinear. To date, the analysis of limit cycle phenomena in such nonlinear digital filters has been concerned with the study of the zero-input response of second-order filters.[1-3] A more fundamental problem is that of determining whether or not a filter's response to a nonzero input (the forced response) is in some meaningful sense close to the ideal response. This problem seems to have been ignored.

If we consider input sequences, the levels of which are sufficiently small (in the sense that when the input sequence is applied to the linear

model of the filter, the response *eventually* lies within the open interval determined by the most positive and the most negative machine numbers), then it is tempting to conjecture, as if the system were linear, that when the filter's zero-input response can be made to admit only limit cycles of small amplitude by using sufficiently many bits in the representation of the data so that the quantization errors are made sufficiently small, then the deviation of the filter's forced response from the ideal can also be made small in the same manner. As will be shown by counterexamples, however, this conjecture is false. Thus, since the usual purpose of a digital filter is the processing of nonzero signals, a question of major importance becomes: How can it be determined that, in the presence of quantization and adder overflow, a digital filter's forced response will be satisfactory?

In this paper we analyze the forced response of second-order digital filters which employ a type of arithmetic that has been called *saturation arithmetic.*† The essential structure of a second-order digital filter is shown in Fig. 1 where, for given real numbers a, b the filter's output sequence‡ $v^{(k)}$, $k = 1, 2, \cdots$, is uniquely determined by the input sequence $u^{(k)}$, $k = 1, 2, \cdots$, and by $v^{(-1)}$, $v^{(0)}$, the initial values of the filter's state variables. We develop a criterion by which satisfactory behavior of the filter can be determined. The criterion is shown to be sharp, in the sense that our counterexamples show that for most filters which violate the criterion, the forced response is not always close to the ideal response.

More precisely, we show that when the filter's coefficients a, b are determined by any point lying within the open crosshatched region of Fig. 2, and for any input sequence whose level is small (in the sense mentioned earlier), then the response of the nonlinear filter will be asymptotically close to the ideal response. On the other hand, we show that when the filter's coefficients are determined by any point lying within the shaded regions in the lower corners of the triangle of Fig. 2, and when certain very reasonable assumptions are satisfied concerning the nature of the quantization, then there exist input sequences the levels of which are also small, but for which the filter's response is not asymptotically close to the ideal response.

---

† The definition of this term is given in Section II.

‡ In many applications some linear combination of the quantities $v^{(k)}$, $v^{(k-1)}$, $v^{(k-2)}$ is taken to be the filter's output at the kth time instant. This additional complication has no bearing on the matters considered here. For simplicity, therefore, we consider the sequence $v^{(k)}$ to be the filter's output.

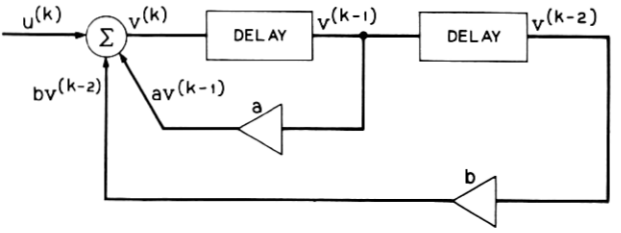

Fig. 1—Second-order digital filter.

II. SECOND-ORDER FILTERS

The usual method of designing digital filters[4] employs the interconnection of many second-order filters. The analysis and design of second-order digital filters is therefore a problem of considerable practical importance.

The behavior of the digital filter of Fig. 1 is characterized by the linear difference equation

$$w^{(k+1)} = Aw^{(k)} + \begin{bmatrix} 0 \\ u^{(k+1)} \end{bmatrix}, \qquad k = 0, 1, 2, \cdots, \tag{1a}$$

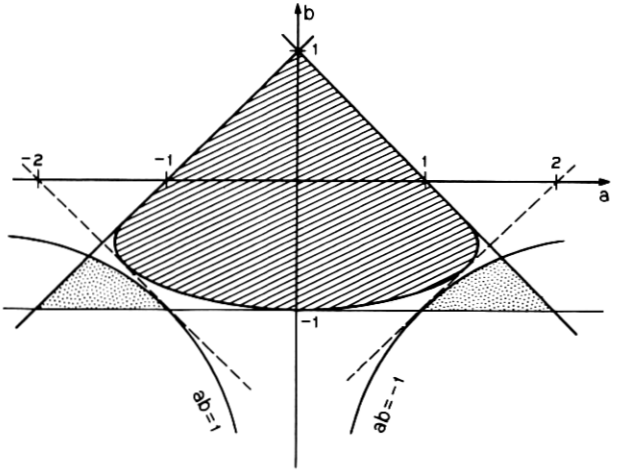

Fig. 2—Region determining filter coefficients for which the filter's forced response can be made asymptotically close to the ideal response (crosshatched region), and region determining coefficients for which the forced response will not always be close to the ideal response (shaded region).

where A denotes the 2×2 matrix

$$A = \begin{bmatrix} 0 & 1 \\ b & a \end{bmatrix}, \tag{1b}$$

and $w^{(k)}$ is a two-dimensional vector (specifying the state of the system at the kth time instant) the second component of which, $w_2^{(k)}$, corresponds to the digital filter's output sequence $v^{(k)}$.

In any special-purpose computer type realization of the digital filter of Fig. 1 the ideal behavior specified by (1) can be only approximated. At each time instant, the output of the summation point can assume only one of a finite number of values. Therefore, the actual value of the summation point's output is given by an expression such as

$$v^{(k)} = f(av^{(k-1)} + bv^{(k-2)} + u^{(k)}) + e^{(k)},$$

where the function f accounts for adder overflow and the sequence $e^{(k)}$ accounts for the quantization error that is inherently present. The equality $f(\xi) = \xi$ is satisfied only in a certain neighborhood of the origin which we take to be the interval $-1 \leqq \xi \leqq 1$. We consider filters employing *saturation arithmetic*; that is, we define $f(\xi) = -1$ for $\xi < -1$ and $f(\xi) = 1$ for $\xi > 1$.

When the effects of quantization and adder overflow are taken into consideration, the digital filter of Fig. 1 is then characterized by the nonlinear difference equation

$$r^{(k+1)} = F\left(Ar^{(k)} + \begin{bmatrix} 0 \\ u^{(k+1)} \end{bmatrix}\right) + \begin{bmatrix} 0 \\ e^{(k+1)} \end{bmatrix}, \qquad k = 0, 1, 2, \cdots, \tag{2}$$

where the state of the system at the kth time instant is now specified by the two-dimensional vector $r^{(k)}$. The mapping F is defined by the relation

$$F\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} x \\ f(y) \end{bmatrix}. \tag{3}$$

Since the purpose of our study is to examine the effects of quantization and adder overflow on the forced response of digital filters, we are interested in comparing the solutions of (1) and (2) when the equations are given identical input sequences and identical initial conditions. We make the reasonable assumption that we are concerned only with digital filters whose linear model, i.e., eq. (1), is asymptotically stable. It is well known that (1) characterizes an asymptotically stable linear

system if and only if each eigenvalue of the matrix A has magnitude less than unity. The eigenvalues of A (the roots of the polynomial $\lambda^2 - a\lambda - b$) are known to have magnitude less than unity if and only if the coefficients a, b have values determined by points that lie within the large open triangular region shown in Fig. 2 (determined by the straight lines: $b \pm a = 1$, $b = -1$).

It is clear that so long as the filter's input sequence is such that the solution of the linear equation (1) is continually being driven into the region† $|| w^{(k)} || > 1$, then there is little point in trying to compare the solutions of (1) and (2); it being clear at the start that at each such time instant, they will differ by at least the amount by which $|| w^{(k)} ||$ exceeds unity (plus or minus the quantization error $e^{(k)}$ which, presumably, will be small). At the other extreme, if it is known in advance that the initial conditions and the input sequence are (small enough in magnitude) such that the solution of the linear equation (1) is within the range $|| w^{(k)} || \leqq 1 - \delta$, for some $\delta > 0$, and for all $k = 1, 2, \cdots$, then there is no problem. That is, it is clear at the outset (due to the assumption that the linear system is asymptotically stable) that the solutions of (1) and (2) will be made arbitrarily close for all such inputs, by simply causing the magnitude of the quantization error $e^{(k)}$ to be bounded by a sufficiently small number. In effect, the nonlinear function f is then not present; we are simply comparing the responses of the same stable linear system to two slightly different inputs.

The interesting question which we shall consider is the one which follows. Suppose we assume only that the filter's input is such that the ideal response, the solution of the linear system (1), *eventually* (i.e., for all k sufficiently large) satisfies $|| w^{(k)} || \leqq 1 - \delta$, for some $\delta > 0$.‡ Then, when is the same thing (i.e., $|| r^{(k)} || \leqq 1 - \delta$ for some $\delta > 0$, and all k sufficiently large) true for the solution of eq. (2)? Thus, we are interested in knowing when the gross effects of the nonlinearity are simply of a transient nature and hence, aside from such transient effects, when can the filter's response be made as close to the ideal as desired by simply causing the quantization error to be sufficiently small (i.e., by using a sufficient number of bits in the representation of the data). Unfortunately, as our counterexamples will show, it is *not* always the case

† For each $w = (w_1, w_2)^T$ we define $||w|| = \max \{|w_1|, |w_2|\}$.

‡ The inequality $||w^{(k)}|| \leqq 1$ might seem more reasonable here. The necessity to write $1 - \delta$ on the right-hand side is the small price that we must pay for the freedom to treat the quantization error in the relatively simple manner that we have chosen. By considering the quantization error at each step to be simply a "small" input $e^{(k)}$, we do not admit to the knowledge that, for example, in all sufficiently small neighborhoods of the points $\xi = \pm 1$, the quantization (be it roundoff, truncation, or whatever) will be done in such a manner that $|\xi + e^{(k)}| \leqq 1$.

that this will occur in the nonlinear system whenever the linear system's response satisfies $\| w^{(k)} \| \leqq 1 - \delta$ for some $\delta > 0$ and all sufficiently large k.

With our objective thus being to compare the asymptotic behavior of the solutions of (1) and (2), and since the linear system (1) is assumed to be asymptotically stable, it is clear that we may drop the requirement that the equations have the same initial conditions. This follows, of course, from the fact that the initial conditions of (1) do not affect the solution's asymptotic behavior.

By including the quantization effects in the linear model of the filter, the system is then described by the equation

$$s^{(k+1)} = As^{(k)} + \begin{bmatrix} 0 \\ u^{(k+1)} \end{bmatrix} + \begin{bmatrix} 0 \\ e^{(k+1)} \end{bmatrix}, \qquad k = 0, 1, 2, \cdots, \tag{4}$$

whose solution can be made arbitrarily close to the solution of (1) by simply requiring that all $| e^{(k)} |$ be sufficiently small. Let us assume, therefore, that the $| e^{(k)} |$ are at least small enough that there exists $\delta' > 0$ and a nonnegative integer K such that, for all nonnegative integers $k \geqq K$,

$$\| s^{(k+1)} \| + | e^{(k+1)} | \leqq 1 - \delta'. \tag{5}$$

Letting

$$z^{(k)} = r^{(k)} - s^{(k)}, \qquad k = 0, 1, 2, \cdots, \tag{6}$$

we find, from (2) and (4), that the sequence $z^{(k)}$ is determined by the equation

$$z^{(k+1)} = F\left(Az^{(k)} + \begin{bmatrix} 0 \\ v^{(k+1)} \end{bmatrix} \right) - \begin{bmatrix} 0 \\ v^{(k+1)} \end{bmatrix}, \qquad k = 0, 1, 2, \cdots, \tag{7}$$

where $z^{(0)} = r^{(0)} - s^{(0)}$ and, for $k = 0, 1, 2, \cdots, v^{(k+1)} = s_2^{(k+1)} - e^{(k+1)}$ which, according to (5), with $\epsilon = 1 - \delta'$, satisfies

$$| v^{(k+1)} | \leqq \epsilon, \qquad \text{for } k \geqq K. \tag{8}$$

We take as our objective, therefore: To determine when, for any sequence $v^{(k+1)}$, $k = 0, 1, 2, \cdots$, satisfying (8) for some ϵ in the interval $0 \leqq \epsilon < 1$, and some nonnegative integer K, the solution of (7) satisfies $\lim_{k \to \infty} \| z^{(k)} \| = 0$.

We note at this point that our objective stated in the preceding paragraph is similar to the objective in Ref. 3 (see the paragraph immediately following eq. (8) of that paper) where the control of

limit cycles in the zero-input response of second-order digital filters is considered. The important difference between the two objectives is that here we must accommodate any value of ϵ in the interval [0, 1). In Ref. 3, however, it was only necessary to consider bounds on the sequence $\nu^{(k+1)}$ that were "sufficiently small". The consequences of this difference are great. It will be clear that a much more delicate analysis is required here than that in Ref. 3.

III. ANALYSIS OF THE FORCED RESPONSE

We now determine, in accordance with the objective explained in Section II, a criterion for the satisfactory behavior of the forced response of second-order digital filters in the presence of quantization and adder overflow. We consider filters employing saturation arithmetic; that is, we define the function f of Section II by

$$f(\xi) = \begin{cases} -1 & \text{for } \xi < -1 \\ \xi & \text{for } -1 \leqq \xi \leqq 1 \\ 1 & \text{for } \xi > 1. \end{cases} \tag{9}$$

The following theorem is fundamental to our analysis.

Theorem 1: Let the matrix A be defined by (1b) *in which the values of a, b are specified by some point lying within the open triangular region of Fig.* 2 *(determined by the straight lines: $b \pm a = 1$, $b = -1$). Let the mapping F be defined by* (3) *in which the function f is specified by* (9). *Then, for any sequence $\nu^{(k+1)}$, $k = 0, 1, 2, \cdots$, satisfying* (8) *for some ϵ in the interval $0 \leqq \epsilon < 1$ and some nonnegative integer K, the solution of* (7) *satisfies $\lim_{k\to\infty} \| z^{(k)} \| = 0$ provided that there exists a real number σ such that*

$$1 - \sigma^2 a^2 > 0, \tag{10}$$

$$[1 - b^2 - (1 - \sigma)a^2]^2 - a^2[\sigma + (2 - \sigma)b]^2 > 0, \tag{11}$$

and

$$\sqrt{1 - \sigma^2 a^2} + \sqrt{[1 - b^2 - (1 - \sigma)a^2]^2 - a^2[\sigma + (2 - \sigma)b]^2} > |b^2 - (1 - \sigma)a^2|. \tag{12}$$

The proof of Theorem 1 is given in the Appendix. We now seek to determine those points lying within the triangular region of Fig. 2 which specify values of a, b such that the inequalities (10), (11), (12) are satisfied for some value of σ.

We begin by examining the case in which $\sigma = 0$. In this case (10) is satisfied for all a, b and, as shown in the Appendix, (11) is satisfied for only those values of a, b specified by points lying within the open crosshatched region of Fig. 5 which, for $\sigma = 0$, is the open crosshatched region of Fig. 3. By squaring each side of the inequality (12), it is easily shown that that inequality, with $\sigma = 0$, is equivalent to

$$(1 - a^2 - b^2) + \sqrt{(1 - a^2 - b^2)^2 - 4a^2b^2} > 0.$$

Since values of a, b specified by points lying within the crosshatched region of Fig. 3 satisfy $1 - a^2 - b^2 > 0$, it is clear that all such values (and only those values) of a, b satisfy (10), (11), and (12) for $\sigma = 0$.

For negative values of σ and for $\sigma \geqq 2$ it is clear that the crosshatched region of Fig. 5 lies interior to the crosshatched region of Fig. 3. Thus, consideration of such values of σ can determine no values of a, b that are not already determined in Fig. 3 by consideration of the $\sigma = 0$ case.

We now show that values of σ in the interval $1 \leqq \sigma < 2$ yield no values of a, b satisfying (10), (11), and (12) that cannot also be determined by considering some value of σ in the interval $0 < \sigma \leqq 1$. Let $\hat{\sigma}$ satisfy $1 \leqq \hat{\sigma} < 2$ and then define $\tilde{\sigma} = 2 - \hat{\sigma}$. Clearly $0 < \tilde{\sigma} \leqq 1$. Now, if (10) is satisfied for $\sigma = \hat{\sigma}$, then, clearly, (10) is also satisfied for $\sigma = \tilde{\sigma}$. The expression on the left-hand side of (11) can be rewritten as $(1 - b^2)^2 + a^2\{[a^2 - 2(1 + b^2)] - [a^2 - (1 - b)^2]\sigma(2 - \sigma)\}$. The form of this expression shows that it has the same value for $\sigma = \hat{\sigma}$ and $\sigma = \tilde{\sigma}$. Finally, it is clear that if (12) is satisfied for $\sigma = \hat{\sigma}$, then (12) is also satisfied for $\sigma = \tilde{\sigma}$ since [using our observations regarding (10) and (11)] the left-hand side of that inequality is not decreased by

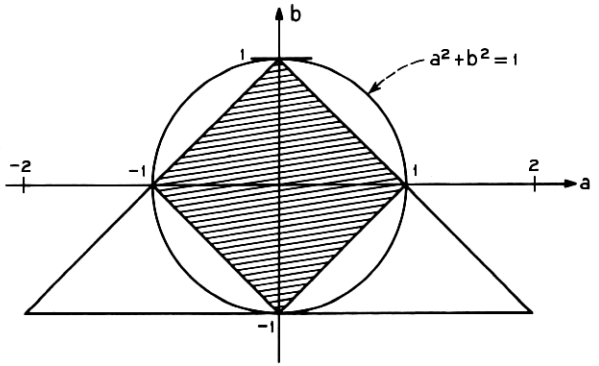

Fig. 3—Region in which inequality (11) is satisfied for $\sigma = 0$.

replacing $\hat{\sigma}$ with $\tilde{\sigma}$, and since

$$| b^2 - (1 - \hat{\sigma})a^2 | = | b^2 + (1 - \tilde{\sigma})a^2 | \geqq | b^2 - (1 - \tilde{\sigma})a^2 |.$$

There remains to consider only those values of σ in the interval $0 < \sigma \leqq 1$. Thus, for each such value of σ we wish to determine the values of the parameters a, b specified by points lying within the open crosshatched region of Fig. 5 and, from (10), within the open region specified by $| a | < 1/\sigma$, for which the inequality (12) is satisfied. It is not difficult to show that for each value of σ satisfying $0 < \sigma \leqq 1$ the function

$$\varphi_\sigma(a, b) \equiv \sqrt{1 - \sigma^2 a^2} + \sqrt{[1 - b^2 - (1 - \sigma)a^2]^2 - a^2[\sigma + (2 - \sigma)b]^2} - |b^2 - (1 - \sigma)a^2|,$$

whose domain is that portion of the open crosshatched region of Fig. 5 where $| a | < 1/\sigma$, is monotone decreasing in $| a |$ for each value of b in the interval $-1 < b \leqq 0$. Thus, the region in which the inequality (12) is satisfied is easily located by determining the curves $\varphi_\sigma(a, b) = 0$. Moreover, because of the above observation concerning the monotonicity of φ_σ, it is easy to determine these curves numerically. Several such curves, for various values of σ, are shown in Fig. 4.

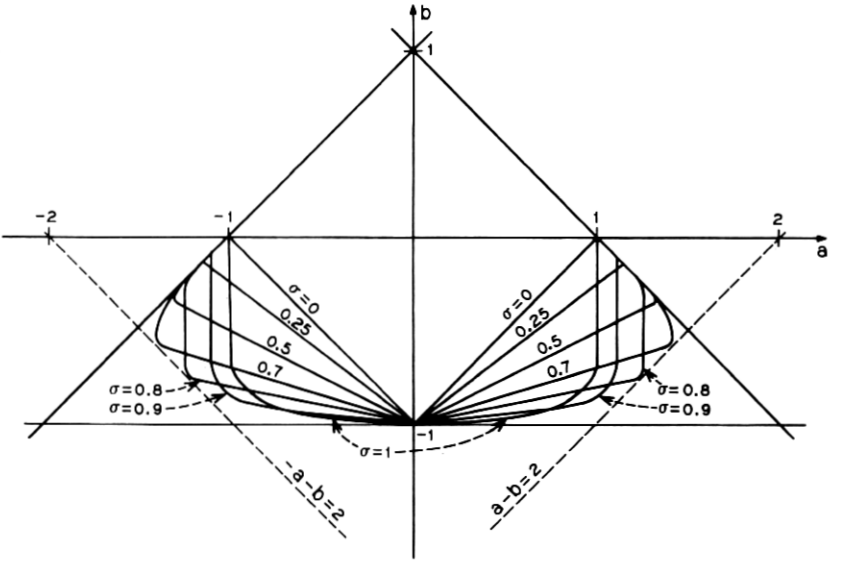

Fig. 4—Location of the $\varphi_\sigma(a, b) = 0$ curves for several values of σ.

The region in which the inequalities (10), (11), (12) are satisfied for *some* real number, σ, is the union of the regions determined by the $\varphi_\sigma(a, b) = 0$ curves for $0 \leqq \sigma \leqq 1$. The numerical results show that this region has the shape indicated by the crosshatched area in Fig. 2. The boundary of this region appears to be determined by the straight lines $b \pm a = 1$ for $b \geqq -\frac{1}{3}$, and by the ellipse $a^2 + 8b(1 + b) = 0$ for $b \leqq -\frac{1}{3}$.

It is clear that there are several ways in which our analysis could be refined in order to provide the possibility of improving upon the result of Theorem 1. In the next section, however, we show a fundamental limitation on the extent of any such improvement. We show there, how to construct counterexamples which demonstrate that for a certain large portion of the uncrosshatched area of the triangular region of Fig. 2 (in particular, the shaded areas in each lower corner) the conclusion of Theorem 1 is, in fact, false.

IV. COUNTEREXAMPLES

We now show how to construct the counterexamples which have been referred to in the preceding sections. We begin by showing that, when the function f is defined by (9), and when the values of the filter's coefficients are determined by any point lying within the open shaded regions in each lower corner of the triangle shown in Fig. 2, then there exist nonzero initial conditions and, for some $\epsilon < 1$, a periodic input sequence $v^{(k+1)}$ satisfying (8) such that the solution of (7) is periodic (and thus does not satisfy $\lim_{k\to\infty} \| z^{(k)} \| = 0$).

We first consider values of a, b determined by any point lying within the shaded open region in the lower *left* corner of the triangle of Fig. 2. In particular, we assume that

$$ab > 1. \tag{13}$$

It is also clear that the inequality

$$a < -b^2 \tag{14}$$

holds for any such point. Denoting the initial condition $z^{(0)}$ by $z^{(0)} = (x^{(0)}, y^{(0)})^T = (p, q)^T$, and considering an input sequence having period *three*, specified by $v^{(1)} = 0$, $v^{(2)} = 1 - p$, $v^{(3)} = -1 - q$, it is clear from Table I that (7) has a nonzero periodic solution provided that values of p, q can be found such that the inequalities specified in parentheses in Table I are satisfied.

The inequalities in the $v^{(k+1)}$ column [which must hold in order that

the input sequence satisfy (8) for some $\epsilon < 1$] and the inequalities on the first line of the column labeled "$bx^{(k)} + ay^{(k)} + \nu^{(k+1)}$" are equivalent to:

$$0 < p < 2,$$
$$-2 < q < 0,$$
$$-1 < bp + aq < 1. \tag{15}$$

Thus, so long as we consider only positive values of p and negative values of q, these inequalities will always hold whenever p and q have sufficiently small magnitude. The remaining two inequalities specified in Table I will be satisfied provided that

$$(1 - ab)p - (a^2 + b)q < 0$$

and

$$(a + b^2)p - (1 - ab)q < 0. \tag{16}$$

In view of the inequalities (13), (14), it is clear that there exist values of $p > 0$ and $q < 0$ such that (16) is satisfied. Moreover, it is clear that the magnitudes of p and q can be scaled such that the inequalities (15) are also satisfied. Thus, there exists a nonzero periodic (of period three) solution of (7).

For values of a, b determined by any point lying within the open shaded region in the lower *right* corner of the triangle of Fig. 2 a similar line of reasoning shows that a nonzero solution of period *six* can be obtained. The existence of such a solution is easy to demonstrate by noting the odd-symmetry of the function f and, with the initial condition $z^{(0)} = (p, q)^T$, showing that, with $\nu^{(1)} = 0$, $\nu^{(2)} = 1 + p$, $\nu^{(3)} = -1 + q$, there exist values of p, q such that $z^{(3)} = (-p, -q)^T$. We omit the details.

The above procedure for constructing counterexamples is concerned explicitly with the solutions of (7). The simple relationships between (7) and the original equations of interest, i.e., eqs. (1) and (2), are described

TABLE I—CONSTRUCTION OF A PERIODIC SOLUTION FOR EQUATION (7)

k	$x^{(k)}$	$y^{(k)}$	$\nu^{(k+1)}$	$bx^{(k)} + ay^{(k)} + \nu^{(k+1)}$	$f - \nu^{(k+1)}$
0	p	q	0	$(-1<)\ bp + aq\ (<1)$	$bp + aq$
1	q	$bp + aq$	$(-1<)\ 1 - p\ (<1)$	$bq + abp + a^2q + 1 - p\ (>1)$	p
2	$bp + aq$	p	$(-1<)\ -1 - q\ (<1)$	$b^2p + abq + ap - 1 - q\ (<-1)$	q
3	p	q	0	...	...

in Section II. It is instructive, however, to consider explicitly the implications of such counterexamples concerning the solutions of (1) and (2).

Let any values of the parameters a, b, determined by some point lying within the open shaded regions in the lower corners of the triangle of Fig. 2, be given. Consider *any* counterexample constructed according to the above procedure. Then, assuming that the quantization occurs in an appropriate manner (or, assuming that there is no quantization) it is a straightforward matter to use the relationships between the variables of (1), (2), and (7) to demonstrate a periodic input sequence $u^{(k)}$ and appropriate values of the initial conditions $v^{(-1)}$, $v^{(0)}$ such that the response of the linear model of the filter of Fig. 1 [i.e., $w^{(k)}$, the solution of (1)] is asymptotic to a periodic sequence, and satisfies $\| w^{(k)} \| < 1$ for all sufficiently large k, while the response of the nonlinear filter [i.e., $r^{(k)}$, the solution of (2)], although also periodic, is such that $\| w^{(k)} - r^{(k)} \|$ does not approach zero as $k \to \infty$.

These counterexamples, while clearly demonstrating that there exists *potential* trouble whenever a filter's coefficients are assigned such "bad" values, do not show that such behavior will necessarily be possible for some *particular* filter. They do not demonstrate, for example, that with a particular (specified) kind of quantization, and with a particular set of permissible values for the filter's input sequence, there will necessarily exist a periodic input sequence for which the linear, and the nonlinear digital filters have asymptotically different responses. It is possible, however, by considering at the outset the details of the quantization and thereby imposing somewhat different constraints (to those of Table I) on the values chosen for p, q, $v^{(1)}$, to construct certain counterexamples which show just that.

We assume that the values specified for the parameters a, b are determined by a point lying within the open shaded region in the lower *left* corner of the triangle of Fig. 2. (A similar development could, of course, be considered for the other shaded region.) We also assume that a certain *finite* set Q of allowable *machine numbers*, satisfying $x \in Q \Rightarrow | x | \leqq 1$, is specified. Thus, we assume that for the nonlinear digital filter with quantization the variables $u^{(k)}$, $v^{(k)}$, $v^{(k-1)}$, $v^{(k-2)}$ of Fig. 1 can assume only those values specified by the set Q. Furthermore, we assume that the filter employs saturation arithmetic with the overflow and quantization effects both specified by a certain function f_q; that is, given any values for $u^{(k)}$, $v^{(k-1)}$, $v^{(k-2)}$ taken from the set Q, the value for $v^{(k)}$ appearing at the output of the summation point in Fig. 1 is specified by

$$v^{(k)} = f_q(av^{(k-1)} + bv^{(k-2)} + u^{(k)}). \tag{17}$$

If, for example, with $a = -1.3$ and $b = -0.9$, the values $u^{(k)} = 0.0$, $v^{(k-1)} = -0.9$, $v^{(k-2)} = 1.0$ are considered; and if the quantization is accomplished by simply rounding the ideal output of the summation point to the nearest tenth, the value $v^{(k)}$ specified by (17) is $v^{(k)} = 0.3$.

Clearly, if the set Q imposes sufficiently severe (indeed, for practical purposes, *unreasonable*) restrictions on the values that the input sequence and the initial conditions may assume, then it will be impossible to construct a counterexample. It is no surprise, therefore, that the success of the process to be described depends upon the assumption that the quantization is "sufficiently fine" (that is, that there are sufficient quantization levels distributed throughout the interval $[-1, 1]$), and that when $| av^{(k-1)} + bv^{(k-2)} + u^{(k)} | \leqq 1$, the actual output of the summation point is reasonably close to the ideal value, that is,

$$f_q(av^{(k-1)} + bv^{(k-2)} + u^{(k)}) \approx av^{(k-1)} + bv^{(k-2)} + u^{(k)}. \tag{18}$$

We first note that the values of a, b determined by any point lying within the open shaded region in the lower left corner of the triangle of Fig. 2 are such that $-1 < a - b < 1$. Thus, since $b/a > 0$, we also have $-1 < a - b + b/a$ and $a - b < 1$. This ensures that the open intervals $(-1, 1)$ and $(a - b, a - b + b/a)$ overlap. Hence, if the quantization is sufficiently fine, there exists $u^{(1)} \in Q$ such that

$$-a/b < 0 < b - a + u^{(1)} < b/a < 1. \tag{19}$$

Thus, for such a value of $u^{(1)}$,

$$b - a(b - a + u^{(1)}) < 0$$

and

$$a + b(b - a + u^{(1)}) < 0.$$

Hence, for sufficiently fine quantization, there exist $u^{(2)}, u^{(3)} \in Q$ such that

$$1 + b - a(b - a + u^{(1)}) - u^{(2)} < 0, \tag{20}$$

and

$$1 + a + b(b - a + u^{(1)}) + u^{(3)} < 0. \tag{21}$$

We let $r_{\max}$ denote the most positive value in the set Q and let $r_{\min}$ denote the most negative value in Q. We also let

$$r_2^{(1)} = f_q(ar_{\min} + br_{\max} + u^{(1)}),$$

$$r_2^{(2)} = r_{\max},$$

$$r_2^{(3)} = r_{\min}.$$

Now, assuming that (18) holds, it can be expected, due to (20) and (21), that there exist p, q such that

$$(1 - ab)p - (a^2 + b)q = 1 - br_{\min} - ar_2^{(1)} - u^{(2)} < 0, \quad (22)$$

$$(a + b^2)p - (1 - ab)q = 1 + ar_{\max} + br_2^{(1)} + u^{(3)} < 0. \quad (23)$$

Moreover, if the quantization is sufficiently fine, the values for $u^{(2)}$ and $u^{(3)}$ can be chosen such that $1 - br_{\min} - ar_2^{(1)} - u^{(2)} \approx 0$ and $1 + ar_{\max} + br_2^{(1)} + u^{(3)} \approx 0$, and such that the values of these expressions are in the proper ratio that, in fact, *small* values of $p > 0$ and $q < 0$ are determined by the equations in (22), (23). Thus, since for sufficiently fine quantization

$$ar_{\min} + br_{\max} + u^{(1)} \approx b - a + u^{(1)}, \quad (24)$$

and, due to (19), it is reasonable to expect that there exists $\nu^{(1)}$ such that

$$-1 < bp + aq + \nu^{(1)} = ar_{\min} + br_{\max} + u^{(1)} < 1. \quad (25)$$

Furthermore, for $p > 0$, $q < 0$ small, we expect that the following inequalities also hold:

$$-1 < r_2^{(1)} - bp - aq < 1, \quad (26)$$

$$-1 < r_{\max} - p < 1, \quad (27)$$

$$-1 < r_{\min} - q < 1. \quad (28)$$

Assuming therefore that the values of $u^{(1)}$, $u^{(2)}$, $u^{(3)}$, p, q, $\nu^{(1)}$ are such that (22), (23), (25), (26), (27), and (28) hold, we proceed with the construction of a counterexample by simply assigning the values to the remaining variables that are dictated by the relationships specified in Section II. In particular, we let

$$s_2^{(1)} = r_2^{(1)} - bp - aq, \qquad e^{(1)} = s_2^{(1)} - \nu^{(1)},$$

$$s_2^{(2)} = r_{\max} - p, \qquad e^{(2)} = -1 + r_{\max},$$

$$s_2^{(3)} = r_{\min} - q, \qquad e^{(3)} = 1 + r_{\min},$$

$$\nu^{(2)} = 1 - p,$$

$$\nu^{(3)} = -1 - q,$$

and

$$r_1^{(1)} = r_2^{(3)}, \qquad s_1^{(1)} = s_2^{(3)},$$

$$r_1^{(2)} = r_2^{(1)}, \qquad s_1^{(2)} = s_2^{(1)},$$

$$r_1^{(3)} = r_2^{(2)}, \qquad s_1^{(3)} = s_2^{(2)}.$$

At this point, one final step remains in our construction of a counterexample. We have obtained periodic solutions of eqs. (4) and (2), with the solution of (4) satisfying $\| s^{(k)} \| < 1$. We would like to obtain the corresponding periodic sequence to which the solution of (1) is asymptotic. This sequence, which we shall call $\hat{w}^{(k)}$, is easily determined by the equations

$$\begin{bmatrix} \hat{w}_2^{(1)} \\ \hat{w}_2^{(2)} \\ \hat{w}_2^{(3)} \end{bmatrix} = \begin{bmatrix} 1 & -b & -a \\ -a & 1 & -b \\ -b & -a & 1 \end{bmatrix}^{-1} \begin{bmatrix} u^{(1)} \\ u^{(2)} \\ u^{(3)} \end{bmatrix},$$

$$\hat{w}_1^{(1)} = \hat{w}_2^{(3)},$$

$$\hat{w}_1^{(2)} = \hat{w}_2^{(1)},$$

$$\hat{w}_1^{(3)} = \hat{w}_2^{(2)}.$$

We have now found a true counterexample only if the values of $\hat{w}^{(k)}$ are also such that $\| \hat{w}^{(k)} \| < 1$. It is reasonable to expect that this inequality will hold, however, since $\| \hat{w}^{(k)} - s^{(k)} \|$ is known to be small provided that the values of $e^{(1)}$, $e^{(2)}$, $e^{(3)}$ are small, and these terms will be small whenever the quantization is sufficiently fine [note that $e^{(1)} = r_2^{(1)} - bp - aq - \nu^{(1)}$, and recall the equality expressed in (25)].

Computer programs have been written which use the above process for constructing counterexamples, and which simulate the behavior of linear and nonlinear digital filters. It has been our experience, based upon experimentation with these programs, that counterexamples of the type described above can easily be found for values of the coefficients a, b determined by points lying within the shaded region in the lower left-hand corner of the triangle in Fig. 2 even when the quantization is extremely coarse, much coarser than the quantization occurring in current practical digital filter realizations. We give, for example, the following numerical counterexample, constructed according to the above procedure, in which we have intentionally considered very coarse quantization, and have also made the task even more difficult by choosing $u^{(1)}$, $u^{(2)}$, $u^{(3)}$ in, obviously, a somewhat less-than-optimum manner, with the result that $| p |$ and $| q |$ are larger than necessary. This does, however, cause the resulting sequences $r^{(k)}$, $w^{(k)}$ to be quite different.

We assume that the coefficient values $a = -1.3$, $b = -0.9$ have been specified. We also assume that

$$Q = \{-0.9, -0.8, \cdots, 0.9, 1\}.$$

We assume that the quantization is performed by simple rounding, at the output of the summation point, of the ideal sum to the nearest tenth. We then have $r_{\max} = 1$, $r_{\min} = -0.9$, and therefore, choosing

$$u^{(1)} = 0, \qquad u^{(2)} = 0.6, \qquad u^{(3)} = 0.2,$$

it follows that

$$r_2^{(1)} = 0.3, \qquad r_2^{(2)} = 1, \qquad r_2^{(3)} = -0.9. \tag{29}$$

We find that the approximate values of p, q, specified by (22), (23) are: $p = 0.711$, $q = -0.128$. Following the above outlined procedure, we find that all of the required relationships hold. The resulting periodic sequence $\hat{w}^{(k)}$ to which the sequence $w^{(k)}$ is asymptotic is specified by the following approximate values:

$$\hat{w}_2^{(1)} = 0.905, \qquad \hat{w}_2^{(2)} = 0.135, \qquad \hat{w}_2^{(3)} = -0.790. \tag{30}$$

Note that quite different solutions are specified by (29) and (30).

V. THE FORCED RESPONSE AND INPUT SCALING

We have shown in Section IV that the forced response of a stable second-order digital filter employing saturation arithmetic might not, for some inputs, be even asymptotically close to the filter's ideal response (the response of the linear filter) if the coefficients a, b are specified by a point lying outside the crosshatched region of the triangle in Fig. 2. More precisely, we have shown that this certainly happens for coefficient values determined by points lying within the shaded regions in each lower corner of that triangle (so long as certain reasonable assumptions hold concerning the nature of the quantization). Thus one concludes that, when designing a filter, it is desirable to avoid choosing such coefficient values. In practical applications, however, it might be the case that due to other considerations such a choice cannot be avoided. Then it is clear that the designer must be careful to impose appropriate restrictions on the filter's input sequence and on its initial conditions. He might, for example, scale the input sequence such that it is always small enough. The question thus arises: How small is "small enough"? One obvious answer to this question is that the input and the initial conditions be required to be small enough that the response of the *linear* filter [described by (1)] satisfies, for some $\delta > 0$ and all $k = 0, 1, 2, \cdots$, the inequality $\| w^{(k)} \| \leqq 1 - \delta$. Then, by using sufficiently many bits in the representation of the data, the quantization

error can always be made sufficiently small that the adder overflow nonlinearity is not encountered.

The results contained in a paper[3] on limit cycles can provide another answer to this question. This answer requires consideration of only the *asymptotic* nature of the input sequence, and applies to filters using a variety of kinds of arithmetic including, in particular, saturation arithmetic. It is clear from the analysis presented in Ref. 3, that it is sufficient that the input sequence $u^{(k+1)}$ and the quantization error sequence $e^{(k+1)}$ be such that the solution of (4) satisfy, for some nonnegative integer K, the inequality

$$|| s^{(k)} || + | e^{(k+1)} | < \delta, \qquad \text{for } k \geqq K,$$

where δ is one of the bounds specified in Theorem 1 of Ref. 3 for the sequence $\nu^{(k+1)}$. In the case of saturation arithmetic we have

$$\delta = \max \left\{ \frac{2 - |a|}{2 + |a|}, \frac{1 - |b|}{1 + |b|} \right\}.$$

Then, it is clear (by Theorem 1 of Ref. 3) that the solution of (7) satisfies $\lim_{k \to \infty} || z^{(k)} || = 0$, which therefore ensures proper asymptotic behavior of the forced response of the nonlinear filter.

VI. ACKNOWLEDGMENT

The author is pleased to acknowledge the helpful comments of his colleagues J. F. Kaiser and I. W. Sandberg concerning this work.

APPENDIX

The proof of Theorem 1 follows. The proof uses the following well-known result concerning the application of Liapunov's "second method" to the study of the stability of difference equations.[5-7]

Lemma 1: Let G denote a subset of the n-dimensional Euclidean space E^n containing the origin θ. If there exist continuous functions $W: G \to E^1$, $V: G \to E^1$, and if there exists a nonnegative integer K such that:

(i) $W(z) > 0$ *for all* $z \in G$, $z \neq \theta$,
(ii) $W(\theta) = 0$,
(iii) $V(z) \geqq 0$ *for all* $z \in G$,
(iv) $\Delta V(k, z) = V(g(k, z)) - V(z) \leqq -W(z)$ *for all* $k \geqq K$ *and all* $z \in G$,

then each solution of the difference equation $z^{(k+1)} = g(k, z^{(k)})$ which remains in G for all $k \geqq K$ approaches the origin as $k \to \infty$.

For any particular application, the effectiveness of Liapunov's method is of course highly dependent upon the appropriateness of the particular Liapunov function V that is chosen. The quadratic form that will now be described is quite useful for our purposes.

For any given values of the parameters a, b which specify an asymptotically stable linear digital filter, and with the eigenvalues of the matrix A of (1b) being denoted by λ_1, λ_2, let the Liapunov function V be defined by

$$V(z) = z^T B z, \tag{31}$$

with

$$B = \begin{bmatrix} |\lambda_1|^2 + |\lambda_2|^2 + 2\mu & -\sigma a \\ -\sigma a & 2 \end{bmatrix}, \tag{32}$$

where the values of σ and μ are yet to be determined.

In the following lemma we determine, for any given value of σ, those values of μ for which the matrices B and $B - A^T B A$ are positive definite.

Lemma 2: Let σ be a given real number. Then, necessary and sufficient conditions for the matrices B and $B - A^T B A$ both to be positive definite for values of a, b which specify an asymptotically stable linear digital filter are: that the values of the parameters a, b be restricted to those values specified by points lying within the open crosshatched region of Fig. 5, and that, with $\mu_1 < \mu_2$ specified by

$$\mu_{1,2} = \tfrac{1}{2}\{1 + b^2 - (1 - \sigma)a^2 - (|\lambda_1|^2 + |\lambda_2|^2) \pm \sqrt{[1 - b^2 - (1 - \sigma)a^2]^2 - a^2[\sigma + (2 - \sigma)b]^2}\}, \tag{33}$$

a value be assigned to μ such that

$$\mu_1 < \mu < \mu_2 .$$

Proof: It is clear that a necessary and sufficient condition for the matrix B of (32) to be positive definite is that $\det B > 0$, which is equivalent to the inequality

$$\mu > -\tfrac{1}{2}(|\lambda_1|^2 + |\lambda_2|^2) + \tfrac{1}{4}\sigma^2 a^2. \tag{34}$$

The matrix $B - A^T B A$, which has the form

$$\begin{bmatrix} (|\lambda_1|^2 + |\lambda_2|^2) - 2b^2 + 2\mu & -a[\sigma + (2 - \sigma)b] \\ -a[\sigma + (2 - \sigma)b] & 2 - (|\lambda_1|^2 + |\lambda_2|^2) - 2(1 - \sigma)a^2 - 2\mu \end{bmatrix},$$

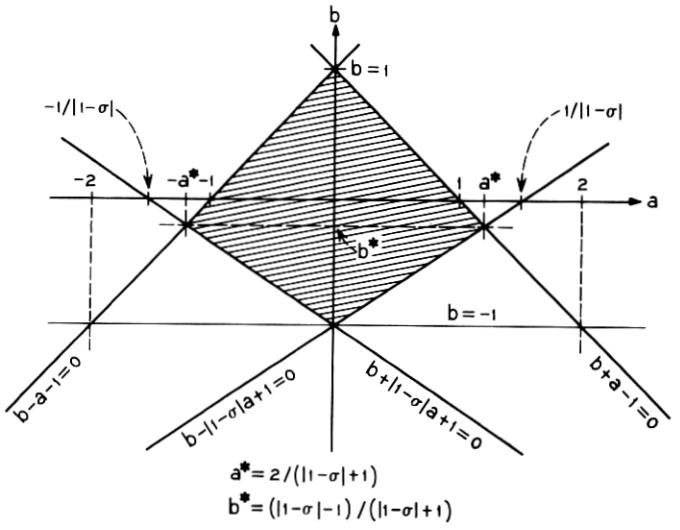

Fig. 5—Region of the a-b plane in which the matrices B and $B - A^TBA$ both may be positive definite.

is positive definite if and only if det $(B - A^TBA) > 0$ and

$$\mu > b^2 - \tfrac{1}{2}(|\lambda_1|^2 + |\lambda_2|^2). \tag{35}$$

As is easily verified, the inequality det $(B - A^TBA) > 0$ is equivalent to

$$\begin{aligned} &-4\mu^2 + 4[1 + b^2 - (1 - \sigma)a^2 - (|\lambda_1|^2 + |\lambda_2|^2)]\mu \\ &+ \{-(|\lambda_1|^2 + |\lambda_2|^2)^2 + 2[1 + b^2 - (1 - \sigma)a^2](|\lambda_1|^2 + |\lambda_2|^2) \\ &- a^2[\sigma + (2 - \sigma)b]^2 - 4b^2[1 - a^2(1 - \sigma)]\} > 0. \end{aligned} \tag{36}$$

We view the left-hand side of the inequality (36) as a quadratic function in the variable μ whose coefficients depend upon the values of the parameters σ, a, b. Clearly, for any choice of these parameter values, (36) will not be satisfied for all large values of $|\mu|$. Thus, a necessary and sufficient condition for the existence of real values of μ satisfying (36) is that the quadratic function on the left-hand side of (36) have distinct real zeros $\mu_1 < \mu_2$. Moreover, if such is the case, (36) will be satisfied if and only if $\mu_1 < \mu < \mu_2$. The zeros μ_1, μ_2 are given by (33) and, therefore, they are real and distinct if and only if

$$|1 - b^2 - (1 - \sigma)a^2| > |a| \cdot |\sigma + (2 - \sigma)b|. \tag{37}$$

We now prove that for any given value of σ, the values of the param-

eters a, b specified by points lying within the open triangular region of Fig. 2, and which satisfy (37), are those (and only those) values of a, b specified by points lying within the open crosshatched region of Fig. 5.

We begin by first showing that there exist no such values of a, b for which $1 - b^2 - (1 - \sigma)a^2 < 0$. Let us assume that this inequality holds for some value of σ. Then, since $1 - b^2 > 0$, it follows that $\sigma < 1$. Now, either

$$-\sigma/(2 - \sigma) \leqq b < 1, \tag{38}$$

or else

$$-1 < b < -\sigma/(2 - \sigma). \tag{39}$$

If (38) holds, then (37) is equivalent to

$$-1 + b^2 + (1 - \sigma) \mid a \mid^2 > \mid a \mid [\sigma + (2 - \sigma)b],$$

or

$$(b - \mid a \mid - 1)[b - (1 - \sigma) \mid a \mid + 1] > 0. \tag{40}$$

If, however, (39) holds, then (37) is equivalent to

$$(b + \mid a \mid - 1)[b + (1 - \sigma) \mid a \mid + 1] > 0. \tag{41}$$

By considering first only nonnegative values of a, and then considering only nonpositive values of a, it is easy to use Fig. 5 and, by inspection, determine that there exist no values of the parameters a, b specified by points lying within the triangular region, such that both (38) and (40) hold. Similarly, it is easy to verify that the same is true regarding inequalities (39) and (41).

We now assume that the parameters σ, a, b are to be chosen such that $1 - b^2 - (1 - \sigma)a^2 \geqq 0$. Then there are three cases to consider:

If $\sigma \geqq 1$, it follows that $\sigma + (2 - \sigma)b > 0$ and hence (37) is easily shown to be equivalent to

$$(b + \mid a \mid - 1)[b - \mid 1 - \sigma \mid \cdot \mid a \mid + 1] < 0. \tag{42}$$

If $\sigma < 1$ and (38) holds, then it follows that (37) is equivalent to

$$(b + \mid a \mid - 1)[b + \mid 1 - \sigma \mid \cdot \mid a \mid + 1] < 0. \tag{43}$$

If $\sigma < 1$ and (39) holds, then it follows that (37) is equivalent to

$$(b - \mid a \mid - 1)[b - \mid 1 - \sigma \mid \cdot \mid a \mid + 1] < 0. \tag{44}$$

By first considering only nonnegative values of a, and then considering only nonpositive values of a, it is easy to use Fig. 5 and, by inspection, determine that the inequality (42) is satisfied if and only if the values of the parameters a, b are determined by points lying within the open crosshatched region of Fig. 5. Similarly, the inequalities (38) and (43), or the inequalities (39) and (44), hold if and only if the values of the parameters a, b are determined by points lying within the open crosshatched region of Fig. 5.

It can easily be shown that for any given value of σ, and any values of the parameters a, b specified by points lying within the open crosshatched region of Fig. 5, it follows from $\mu > \mu_1$ that the inequalities (34) and (35) also hold. We omit the details of the algebra. □

Proof of Theorem 1: Let the Liapunov function V be defined for all $z \in G \equiv E^2$ by (31) and (32) with the values of σ, a, b, μ assumed to be such that both of the matrices B and $B - A^TBA$ are positive definite. It is clear that the equations

$$z^T(B - A^TBA)z = c, \qquad c > 0 \tag{45}$$

define a family of concentric ellipses, centered at the origin θ in the x-y plane [where $z = (x, y)^T$]. The origin also lies between the two parallel straight lines $bx + ay = \pm(1 - \epsilon)$, each of which is tangent to exactly one (in fact, the same one) of the ellipses (45). Thus, there is a unique value of $c^* > 0$ such that

$$c^* = \min \{z^T(B - A^TBA)z : bx + ay = \pm(1 - \epsilon)\}.$$

Let the function W be defined for all $z \in G \equiv E^2$ by

$$W(z) = \min \{z^T(B - A^TBA)z, c^*\}.$$

Thus, $W(z)$ is defined by the positive definite quadratic form $z^T(B - A^TBA)z$ for all points lying within the ellipse $z^T(B - A^TBA)z = c^*$, and $W(z)$ is defined by $W(z) = c^*$ for all other points in the x-y plane.

It is clear that for each value of $\nu^{(k+1)}$ for which (8) holds, the points of the x-y plane determined by $bx + ay + \nu^{(k+1)} \geqq 1$ lie on the opposite side of the line $bx + ay = 1 - \epsilon$ from the ellipse $z^T(B - A^TBA)z = c^*$. The situation is similar regarding the points of the x-y plane determined by $bx + ay + \nu^{(k+1)} \leqq -1$ and the line $bx + ay = -(1 - \epsilon)$. See Fig. 6.

With

$$g(k, z) \equiv F\left(Az + \begin{bmatrix} 0 \\ \nu^{(k+1)} \end{bmatrix}\right) - \begin{bmatrix} 0 \\ \nu^{(k+1)} \end{bmatrix},$$

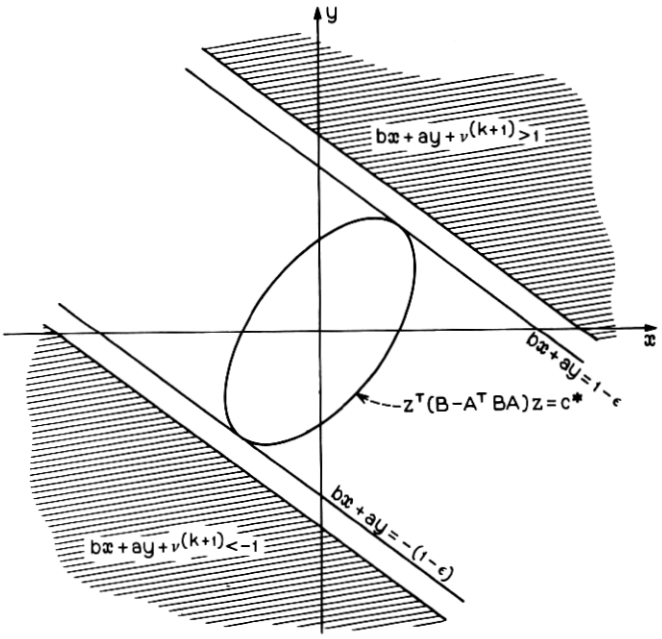

Fig. 6—Location of the ellipse $z^T(B - A^TBA)z = c^*$.

it follows that

$$\Delta V(k, z) = V(g(k, z)) - V(z)$$

$$= \left[F\left(Az + \begin{bmatrix} 0 \\ \nu^{(k+1)} \end{bmatrix}\right) - \begin{bmatrix} 0 \\ \nu^{(k+1)} \end{bmatrix}\right]^T B\left[F\left(Az + \begin{bmatrix} 0 \\ \nu^{(k+1)} \end{bmatrix}\right) - \begin{bmatrix} 0 \\ \nu^{(k+1)} \end{bmatrix}\right] - z^TBz.$$

Thus, whenever $| bx + ay + \nu^{(k+1)} | \leqq 1$, we have

$$\Delta V(k, z) = -z^T(B - A^TBA)z \leqq -W(z).$$

When $bx + ay + \nu^{(k+1)} > 1$,

$$\Delta V(k, z) =$$

$$- \{[(|\lambda_1|^2 + |\lambda_2|^2) + 2\mu]x^2 - 2\sigma axy + [2 - (|\lambda_1|^2 + |\lambda_2|^2) - 2\mu]y^2$$

$$+ 2\sigma a(1 - \nu^{(k+1)})y - 2(1 - \nu^{(k+1)})^2\}; \tag{46}$$

and when $bx + ay + \nu^{(k+1)} < -1$,

$$\Delta V(k, z) =$$

$$- \{[(|\lambda_1|^2 + |\lambda_2|^2) + 2\mu]x^2 - 2\sigma axy + [2 - (|\lambda_1|^2 + |\lambda_2|^2) - 2\mu]y^2$$

$$- 2\sigma a(1 + \nu^{(k+1)})y - 2(1 + \nu^{(k+1)})^2\}. \tag{47}$$

It is an elementary result of analytic geometry that a general second-degree equation of the form $ax^2 + bxy + cy^2 + dx + ey + f = 0$ represents an ellipse if and only if $b^2 - 4ac < 0$. It follows that, if we consider the constant-ΔV loci in the $bx + ay + \nu^{(k+1)} > 1$ region, and in the $bx + ay + \nu^{(k+1)} < -1$ region of the x-y plane, a necessary and sufficient condition for these loci to be arcs of concentric ellipses is:

$$4\mu^2 - 4[1 - (|\lambda_1|^2 + |\lambda_2|^2)]\mu + [\sigma^2 a^2 - 2(|\lambda_1|^2 + |\lambda_2|^2) + (|\lambda_1|^2 + |\lambda_2|^2)^2] < 0. \quad (48)$$

Furthermore, since $\Delta V(k, z)$ is continuous in z, and since the values of $\Delta V(k, z)$ along the lines $bx + ay + \nu^{(k+1)} = \pm 1$ are given by $\Delta V(k, z) = -z^T(B - A^TBA)z$, with $B - A^TBA$ a positive definite matrix, it is clear that when (48) is satisfied, the constant-ΔV curves specified by (46) (temporarily extending the domain of definition of that function to the entire x-y plane) are of the type shown in either Fig. 7a or Fig. 7b; that is, the line $bx + ay + \nu^{(k+1)} = 1$ intersects only certain constant-ΔV curves—in particular, only certain such curves for which the value of ΔV is negative. Thus, the center of the ellipses is situated to one side or the other of the line $bx + ay + \nu^{(k+1)} = 1$ in such a manner that the constant-ΔV ellipses for which ΔV is positive are not intersected by the line. Considering, however, that when $\Delta V(k, z)$ of (46) is evaluated at $z = \theta$ its value is positive, it is clear that Fig. 7b is impossible. Thus [applying exactly the same reasoning to the constant-ΔV curves defined by (47)], it follows that whenever the inequality (48) is satisfied, the

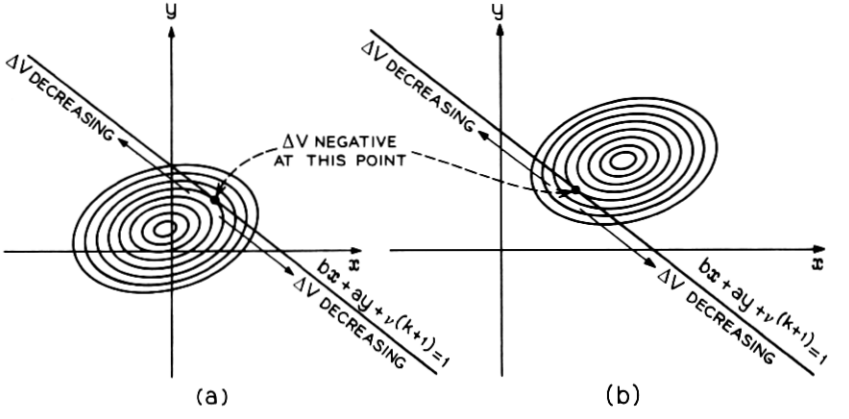

Fig. 7—Possible shape of constant-ΔV curves defined by equation (46).

function $\Delta V(k, z)$ achieves its maximum for $bx + ay + \nu^{(k+1)} \geqq 1$ on the line $bx + ay + \nu^{(k+1)} = 1$, and similarly for the behavior of $\Delta V(k, z)$ in the $bx + ay + \nu^{(k+1)} \leqq -1$ region of the x-y plane. It follows, therefore, that there exists $c' \geqq c^* > 0$ such that for $bx + ay + \nu^{(k+1)} \geqq 1$,

$$\Delta V(k, z) \leqq -c' \leqq -c^* = -W(z). \tag{49}$$

Similarly, there exists $c'' \geqq c^* > 0$ such that for $bx + ay + \nu^{(k+1)} \leqq -1$,

$$\Delta V(k, z) \leqq -c'' \leqq -c^* = -W(z). \tag{50}$$

We have shown that, with the functions V, W defined as specified above, the hypotheses of Lemma 1 are satisfied. Thus, the solution of (7) satisfies $\lim_{k\to\infty} \| z^{(k)} \| = 0$ provided that the values of σ, a, b, μ are such that B and $B - A^T BA$ are positive definite, and provided that (48) holds.

We view the left-hand side of the inequality (48) as a quadratic function in the variable μ whose coefficient values depend upon the values of the parameters σ, a, b. Clearly, for any choice of these parameter values (48) will not hold for all large values of $|\mu|$. Thus, a necessary and sufficient condition for the existence of real values of μ satisfying (48) is that the quadratic function on the left-side of (48) has distinct real zeros $\hat{\mu}_1 < \hat{\mu}_2$; moreover (48) will be satisfied if and only if $\hat{\mu}_1 < \mu < \hat{\mu}_2$. The zeros $\hat{\mu}_1$, $\hat{\mu}_2$ are given by

$$\hat{\mu}_{1,2} = \tfrac{1}{2}[1 - (|\lambda_1|^2 + |\lambda_2|^2) \pm \sqrt{1 - \sigma^2 a^2}]. \tag{51}$$

They are real and distinct if and only if the inequality (10) holds.

According to Lemma 2, for any given value of σ the matrices B and $B - A^T BA$ are positive definite for values of a, b that are specified by some point lying within the open triangular region of Fig. 2 if and only if $\mu_1 < \mu < \mu_2$, where μ_1, μ_2 are specified by (33). Thus, assuming that σ, a, b satisfy (10) and (11), there exists a value of μ such that B and $B - A^T BA$ are positive definite and such that (48) holds if and only if the open intervals (μ_1, μ_2) and $(\hat{\mu}_1, \hat{\mu}_2)$ overlap. That is, if and only if $\mu_1 < \hat{\mu}_2$ and $\hat{\mu}_1 < \mu_2$. Using (33) and (51), these last two inequalities are easily shown to be equivalent to (12). □

REFERENCES

1. Sandberg, I. W., "A Theorem Concerning Limit Cycles in Digital Filters," Proc. Seventh Annual Allerton Conf. on Circuit and System Theory, (October 1969), pp. 63–68.
2. Ebert, P. M., Mazo, J. E., and Taylor, M. G., "Overflow Oscillations in Digital Filters," B.S.T.J., *48*, No. 9 (November 1969), pp. 2999–3020.

3. Willson, A. N., Jr., "Limit Cycles due to Adder Overflow in Digital Filters," to be published in IEEE Trans. on Circuit Theory, *CT-19*, No. 4 (July 1972).
4. Jackson, L. B., Kaiser, J. F., and McDonald, H. S., "An Approach to the Implementation of Digital Filters," IEEE Trans. on Audio and Electroacoustics, *AU-16*, No. 3 (September 1968), pp. 413–421.
5. Hahn, W., *Theory and Application of Liapunov's Direct Method*, Englewood Cliffs, N. J.: Prentice-Hall, 1963, pp. 146ff.
6. Freeman, H., *Discrete-Time Systems*, New York: Wiley, 1965, pp. 158ff.
7. Hurt, J., "Some Stability Theorems for Ordinary Difference Equations," SIAM J. Numer. Anal., *4*, No. 4 (December 1967), pp. 582–596.