01 Jan 1985

Stability Analysis of Fixed-Point Digital Filters using Computer Generated Lyapunov Functions- Part I: Direct Form and Coupled Form Filters

A. Michel

Kelvin T. Erickson
Missouri University of Science and Technology, kte@mst.edu

Recommended Citation

A. Michel and K. T. Erickson, "Stability Analysis of Fixed-Point Digital Filters using Computer Generated Lyapunov Functions- Part I: Direct Form and Coupled Form Filters," *IEEE Transactions on Circuits and Systems*, Institute of Electrical and Electronics Engineers (IEEE), Jan 1985.
The definitive version is available at https://doi.org/10.1109/TCS.1985.1085676

Stability Analysis of Fixed-Point Digital Filters Using Computer Generated Lyapunov Functions—Part I: Direct Form and Coupled Form Filters

KELVIN T. ERICKSON, MEMBER, IEEE, AND ANTHONY N. MICHEL, FELLOW, IEEE

Abstract—**We demonstrate the applicability of the *constructive stability algorithm* of Brayton and Tong in the stability analysis of fixed-point digital filters. In the present paper, we consider direct form and coupled form filters while in a companion paper we treat wave digital filters and lattice filters.**

We compare our results with existing ones which deal with either the global asymptotic stability of digital filters or with existence (resp., nonexistence) of limit cycles in digital filters. Several of the present results are new while some of the present results constitute improvements over existing results. In a few cases, the present results are more conservative than existing results.

It is emphasized that whereas the existing results are obtained by *several diverse methods*, the present results are determined by *one unified approach.*

I. Introduction

IN TWO RECENT papers, Brayton and Tong [3], [4] established some significant results which make it possible to construct computer generated Lyapunov functions to analyze the stability of nonlinear systems (by means of a *constructive algorithm*). These results were subsequently extended in several ways making it possible to estimate the domain of attraction of an equilibrium (see Michel *et al.* [22]) and to apply the constructive algorithm to high-dimensional systems (see Michel *et al.* [20], [21]). Also, the results in [21] are applied in the stability analysis of interconnected power systems.

In the present paper and in a companion paper [13], we apply the constructive algorithm of Brayton and Tong to the stability analysis of several classes of second-order fixed-point digital filters. Specifically, in the present paper we consider direct form digital filters and coupled form digital filters while in [13] we consider wave digital filters and lattice digital filters. Nonlinearities which we encounter in our analysis of these filters include several types of fixed-point quantization effects and overflow effects.

The results which we obtained yield conditions (in the parameter plane for a given filter) under which the digital filters which we consider are globally asymptotically stable and as such, do not possess zero-input limit cycles. We compare these results with several corresponding existing results [1], [2], [5]–[8], [10], [15], [16], [28] which are concerned either with the existence or nonexistence of limit cycles or with the global asymptotic stability of digital filters. For additional references on qualitative analysis of digital filters, the reader is referred to the recent survey by Fettweis [13a]. For related works, refer also to the paper by Parker and Hess [24a] and to the more recent work by Mitra and Lawrence [23a].

This paper consists of five sections and an appendix. In Section II we establish the essential notation, we present certain aspects dealing with the Lyapunov stability of systems described by difference equations, and we provide a summary of the constructive algorithm of Brayton and Tong. In Section III we first discuss the types of nonlinearities that arise in fixed-point digital filters and then we show how the constructive algorithm can be applied in the stability analysis of digital filters in general and how it can be applied to the specific classes of filters mentioned above. In Section IV we discuss the results which we obtained for the specific filters considered herein, and we compare these results with existing ones [1], [2], [5]–[8], [10], [15], [16], [28]. In Section V, several pertinent concluding remarks are made while in the Appendix, a brief description of the computer programs that were used is given.

II. Preliminaries

The present section consists of four subsections. In Section II-A we establish essential notation, in Section II-B we provide certain aspects of stability analysis of general systems described by ordinary difference equations, in Section II-C we present some facts concerning extreme matrices, and in Section II-D we give a brief summary of the constructive stability algorithm of Brayton and Tong.

A. Notation

Let U and V be arbitrary sets. If u is an element of U, we write $u \in U$. We let $U \cup V$, $U \cap V$, and $U \times V$ denote the union, intersection and cross product of U and V,

Manuscript received August 18, 1983. This work was supported in part by the National Science Foundation under Grant ECS-8100690 and the Engineering Research Institute, Iowa State University, Ames, IA 50011.

K. T. Erickson is with Fisher Controls International, Inc., Marshalltown, IA 50158.

A. N. Michel was with Iowa State University, Ames, IA 50011. He is now with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556.

0098-4094/85/0200-0113$01.00

respectively. The boundary of U is denoted by ∂U. If U is a subset of V, we write $U \subseteq V$.

Let R denote the real line, let $R^+ = [0, \infty)$ and let R^n denote the set of real valued n-tuples. The symbol $|\cdot|$ denotes any one of the equivalent norms on R^n. If f is a mapping of a set X into a set Y, we write $f: X \to Y$. Also, $B(r) = \{x \in R^n: |x| < r\}$.

Unless explicitly stated, we will assume that matrices are real and square matrices. We let A^T denote the transpose of the matrix $A = [a_{ij}]$, and we let $\|A\|$ denote the matrix norm of A induced by some vector norm. Sets of matrices are denoted by bold faced upper case letters (e.g., $\boldsymbol{S}$).

A continuous function $\phi: R^+ \to R^+$ is said to belong to *class K* (i.e., $\phi \in K$) if $\phi(0) = 0$ and if ϕ is strictly increasing on R^+. If $\phi: R^+ \to R^+$, if $\phi \in K$, and if $\lim_{r \to \infty} \phi(r) = \infty$, then ϕ is said to belong to *class KR* (i.e., $\phi \in KR$). Also, a function $f: R \to R$ is said to belong to a *sector* $[k_1, k_2]$, where $k_1, k_2 \in R$, if (i) $f(0) = 0$, and (ii) $k_1\sigma^2 \leqslant \sigma f(\sigma) \leqslant k_2\sigma^2$ for all $\sigma \in R$.

Let $I \triangleq \{t_0 + k\}$, $t_0 \in R^+$, $k = 0, 1, 2, \cdots$ and let $j = \sqrt{-1}$. Finally, let z^{-1} represent one unit delay in the block diagram of a digital filter structure.

B. Systems Described by Difference Equations

We consider systems described by ordinary autonomous difference equations of the form

$$x(\tau + 1) = g[x(\tau)] \tag{1}$$

where $x(\tau) \in R^n$ for every $\tau \in I$ and $g: R^n \to R^n$. We denote the unique solutions of (1) by $x(\tau; x_0, \tau_0)$, where $x(\tau_0; x_0, \tau_0) = x_0$. Since we are dealing with autonomous equations, we shall assume without loss of generality that $\tau_0 = 0$. Any point $x_e \in R^n$ for which it is true that $x_e = g(x_e)$ is called an *equilibrium point* of (1). We will henceforth assume that $x = 0$ is an isolated equilibrium of (1), i.e., that there exists a constant $r > 0$ such that $B(r)$ contains no equilibrium points of (1) other than the origin. Thus we have in particular $g(0) = 0$.

We will call any nontrivial periodic solution of (1) a limit cycle. It is customary in the study of digital filters to view nonzero equilibrium points as limit cycles. Unless otherwise stated, we will follow this practice.

Definition 1: (a) The equilibrium $x = 0$ of (1) is said to be *stable* (in the sense of Lyapunov) if for every $\epsilon > 0$ there exists a $\delta = \delta(\epsilon) > 0$ such that $|x(\tau; x_0, 0)| < \epsilon$ for all $\tau \geqslant 0$ whenever $|x_0| < \delta$.

(b) The equilibrium $x = 0$ of (1) is said to be *asymptotically stable* (in the sense of Lyapunov) if (i) it is stable, and (ii) there exists a number $\eta > 0$ having the property that $\lim_{\tau \to \infty} x(\tau; x_0, 0) = 0$ whenever $|x_0| < \eta$. If in particular condition (ii) is true for all $x_0 \in R^n$, then the equilibrium $x = 0$ of (1) is said to be *asymptotically stable in the large* (a.s.i.l.) or *globally asymptotically stable* (g.a.s).

(c) The equilibrium $x = 0$ of (1) is *unstable* if it is not stable. ■

The principal Lyapunov results which yield conditions for stability, asymptotic stability or instability in the sense of Definition 1 involve the existence of functions (Lyapunov functions) $v: R^n \to R$. Such functions are required to have certain definiteness properties which we enumerate next.

Definition 2: (a) A function $v: R^n \to R$ is said to be *positive definite* if there exists a function $\psi \in K$ such that $v(0) = 0$ and $v(x) \geqslant \psi(|x|)$ for all $x \in B(r)$ for some $r > 0$.

(b) A function v is said to be *negative definite* if $-v$ is positive definite.

(c) A function $v: R^n \to R$ is said to be *radially unbounded* if there exists a function $\psi \in KR$ such that $v(0) = 0$ and $v(x) \geqslant \psi(|x|)$ for all $x \in R^n$. ■

The first forward difference of a function $v: R^n \to R$ along the solutions of (1) is given by

$$Dv_{(1)}(x) = v[g(x)] - v(x). \tag{2}$$

Henceforth, we shall assume that v is continuous and that it satisfies a Lipschitz condition in x.

Theorem 1: (a) The equilibrium $x = 0$ of (1) is *stable* if there exists a function $v: R^n \to R$ such that (i) v is positive definite, and (ii) $Dv_{(1)}(x) \leqslant 0$ for all $x \in B(r)$ for some $r > 0$.

(b) The equilibrium $x = 0$ of (1) is *asymptotically stable* if there exists a function $v: R^n \to R$ such that (i) v is positive definite, and (ii) $Dv_{(1)}(x)$ is negative definite.

(c) The equilibrium $x = 0$ of (1) is *asymptotically stable in the large* if there exists a function $v: R^n \to R$ such that (i) v is radially unbounded, and (ii) $Dv_{(1)}(x)$ is negative definite for all $x \in R^n$. ■

For further aspects of the Lyapunov theory, refer to Miller and Michel [23].

We emphasize that if it is possible to find a v-function for (1) which satisfies the conditions of Theorem 1(c), then (a) system (1) has only one equilibrium point, (b) this equilibrium will be the origin, (c) this equilibrium will be asymptotically stable in the large, and (d) no limit cycles will exist for system (1).

In the last part of the present section, we present an algorithm of Brayton and Tong [3], [4] which enables us to construct Lyapunov functions of the norm type for system (1) which satisfy Theorem 1(c) (if $x = 0$ is a.s.i.l.). In the subsequent sections we use such Lyapunov functions in the stability analysis of several classes of second-order fixed-point digital filters. This analysis will yield conditions under which such filters are asymptotically stable and cannot possess limit cycles.

C. Extreme Matrices of a Convex Set of Matrices

We shall require the concepts of a convex set of matrices, an extreme subset of matrices, and an extreme matrix. We phrase our definitions in terms of a linear vector space of real $n \times n$ matrices over the field R. For general definitions of these concepts, refer to Dunford and Schwartz [9].

Definition 3: (a) Let $(R^{n \times n}, R)$ denote the real linear space of real $n \times n$ matrices. A set $\boldsymbol{A} \subset R^{n \times n}$ is said to be *convex* if $X, Y \in \boldsymbol{A}$, $k \in R$, and $0 \leqslant k \leqslant 1$, imply that $kX + (1-k)Y \in \boldsymbol{A}$.

(b) Let $A_1, A_2 \in \boldsymbol{A}$ and let $k \in R$. A nonvoid subset $\boldsymbol{B} \subseteq \boldsymbol{A}$ is said to be an *extreme subset* of $\boldsymbol{A}$ if a proper

convex combination $kA_1+(1-k)A_2$, $0<k<1$, is in $\boldsymbol{B}$ only if A_1, $A_2 \in \boldsymbol{B}$. An extreme subset of $\boldsymbol{A}$ consisting of only one matrix is called an *extreme matrix* of $\boldsymbol{A}$. The *set of extreme matrices* of $\boldsymbol{A}$ is denoted by $E(\boldsymbol{A})$. ■

Consider now in particular the set of 2×2 real matrices given by

$$\boldsymbol{A}_1=\left\{\begin{bmatrix} a & b \\ c & d \end{bmatrix},\ \alpha_1 \leqslant a \leqslant \alpha_2,\ b, c, d \text{ are constants}\right\} \quad (3)$$

where α_1 and α_2 are constants. It is an easy matter to show that $\boldsymbol{A}_1$ is a convex set and that

$$E(\boldsymbol{A}_1)=\{B_{11}, B_{12}\} \quad (4)$$

where

$$B_{11}=\begin{bmatrix} \alpha_1 & b \\ c & d \end{bmatrix} \qquad B_{12}=\begin{bmatrix} \alpha_2 & b \\ c & d \end{bmatrix}. \quad (5)$$

Similarly, if we let

$$\boldsymbol{A}_2=\begin{bmatrix} a & b \\ c & d \end{bmatrix}, \qquad \alpha_1 \leqslant a \leqslant \alpha_2, \quad \gamma_1 \leqslant c \leqslant \gamma_2, \quad b \text{ and } d \text{ are constants} \quad (6)$$

where α_1, α_2, γ_1, and γ_2 are constants, then it can easily be shown that $\boldsymbol{A}_2$ is convex and that

$$E(\boldsymbol{A}_2)=\{B_{21}, B_{22}, B_{23}, B_{24}\} \quad (7)$$

where

$$B_{21}=\begin{bmatrix} \alpha_1 & b \\ \gamma_1 & d \end{bmatrix} \qquad B_{22}=\begin{bmatrix} \alpha_2 & b \\ \gamma_1 & d \end{bmatrix}$$
$$B_{23}=\begin{bmatrix} \alpha_1 & b \\ \gamma_2 & d \end{bmatrix} \qquad B_{24}=\begin{bmatrix} \alpha_2 & b \\ \gamma_2 & d \end{bmatrix}. \quad (8)$$

Finally, it is readily shown that if

$$\boldsymbol{A}=\left\{\begin{bmatrix} a & b \\ c & d \end{bmatrix},\ \alpha_1 \leqslant a \leqslant \alpha_2,\ \beta_1 \leqslant b \leqslant \beta_2,\ \gamma_1 \leqslant c \leqslant \gamma_2,\ \delta_1 \leqslant d \leqslant \delta_2\right\} \quad (9)$$

where α_i, β_i, γ_i, and δ_i, $i=1,2$ are constants, then $\boldsymbol{A}$ is convex and

$$E(\boldsymbol{A})=\left\{\begin{bmatrix} \alpha_i & \beta_j \\ \gamma_k & \delta_l \end{bmatrix},\ i, j, k, l=1,2\right\}. \quad (10)$$

We will have occasion to make use of the sets of matrices $E(\boldsymbol{A}_1)$, $E(\boldsymbol{A}_2)$, and $E(\boldsymbol{A})$ given in (4), (7), and (10), respectively.

D. Constructive Stability Algorithm

We begin by rewriting the system of equations (1) as

$$x(k+1)=M(x(k))x(k) \quad (11)$$

where $M(x(k))$ is chosen so that $M(x(k))x(k)=g[x(k)]$. For every $x \in R^n$, $M(x)$ will be a real $n\times n$ matrix. If we let $\boldsymbol{M}$ denote the set of all matrices obtained by varying x in $M(x)$ over all allowable values, then we can rewrite (11) equivalently as

$$x(k+1)=M_k x(k), \qquad M_k \in \boldsymbol{M}. \quad (12)$$

Brayton and Tong [3], [4] show that the equilibrium $x=0$ of (1) is stable (globally asymptotically stable) if the set of matrices $\boldsymbol{M}$ is stable (asymptotically stable). (The precise definitions of these two terms are given in the next two paragraphs.) In the following, we give a short summary of the results of Brayton and Tong. The reader is referred to [3] and [4] for further details concerning these results.

We call a set $\boldsymbol{A}$ of $n\times n$ real matrices *stable* if for every neighborhood of the origin $U \subset R^n$, there exists another neighborhood of the origin $V \subset R^n$ such that for every $M \in \boldsymbol{A}'$, we have $MV \subseteq U$. Here $\boldsymbol{A}'$ denotes the multiplicative semigroup generated by $\boldsymbol{A}$ and $MV=\{u \in R^n:\ u=Mv,\ v \in V\}$.

In [3] it is shown that the following statements (which characterize the properties of a class of stable matrices) are equivalent:

a) $\boldsymbol{A}$ is stable.

b) $\boldsymbol{A}'$ is bounded.

c) There exists a bounded neighborhood of the origin $W \subset R^n$ such that $MW \subseteq W$ for every $M \in \boldsymbol{A}$. Furthermore, W can be chosen to be convex and balanced.

d) There exists a vector norm $|\cdot|_w$ such that $|Mx|_w \leqslant |x|_w$ for all $M \in \boldsymbol{A}$ for all $x \in R^n$.

Now let $\alpha \in R$ and let $W \subset R^n$. Let $\alpha W=\{u \in R^n:\ u=\alpha w,\ w \in W\}$. Since statements c) and d) above are related by

$$|x|_w=\inf\{\alpha:\ \alpha \geqslant 0,\ x \in \alpha W\}$$

it follows that $|x|_w$ defines a Lyapunov function for $\boldsymbol{A}$, i.e., it defines a function v with the property

$$v(Mx) \leqslant v(x), \qquad \text{for all } M \in \boldsymbol{A} \text{ and } x \in R^n.$$

Next, we call a set of matrices $\boldsymbol{A}$ *asymptotically stable* if there exists a number $\rho>1$ such that $\rho\boldsymbol{A}$ is stable. (The set $\rho\boldsymbol{A}$ is obtained by multiplying every member of $\boldsymbol{A}$ by ρ.) In [4] it is shown that the following statements (which characterize the properties of a class of asymptotically stable matrices) are equivalent:

a) $\boldsymbol{A}$ is asymptotically stable.

b) There exists a convex, balanced, and polyhedral neighborhood of the origin W and a positive number $\gamma<1$ such that for each $M \in \boldsymbol{A}$, we have $MW \subseteq \gamma W$. (Here $\gamma W=\{u \in R^n:\ u=\gamma w,\ w \in W\}$.)

c) $\boldsymbol{A}$ is stable and there exists a positive constant K such that for all $M \in \boldsymbol{A}'$, $|\lambda_i(M)| \leqslant K<1$, $i=1,\cdots,n$, where $\lambda_i(M)$ denotes the ith eigenvalue of M.

Note that if $\boldsymbol{A}$ is stable, then $\gamma\boldsymbol{A}$ is asymptotically stable for all positive $\gamma<1$. Note also that if $\boldsymbol{A}$ is asymptotically stable ($\rho\boldsymbol{A}$ is stable), then there exists a vector norm $|\cdot|_w$ such that

$$|Mx|_w<|\rho Mx|_w \leqslant |x|_w, \qquad \text{for all } M \in \boldsymbol{A} \text{ and } x \in R^n.$$

In [3] and [4] a *constructive algorithm* is presented which determines whether a set of m $n\times n$ real matrices $\boldsymbol{A}=\{M_0,\cdots,M_{m-1}\}$ is stable. In this algorithm, one starts with an initial polyhedral neighborhood of the origin W_0

and one defines a sequence of sets $\{W_k\}$ by

$$W_{k+1} \triangleq \mathscr{K}\left[\bigcup_{j=0}^{\infty} M_{k'}^{j} W_k\right], \qquad k' = (k-1) \bmod m$$

where $\mathscr{K}[\cdot]$ denotes the convex hull of a set. Now A is stable if and only if the final set

$$W^* = \bigcup_{k=0}^{\infty} W_k$$

is bounded. Note that W^* is also given by

$$W^* = \mathscr{K}\left[\cup MW_0,\ M \in A'\right].$$

Since all extreme points z of W_{k+1} are of the form $z = M_i^j u$, where u is an extreme point of W_k, we need only deal with the extreme points of W_k in order to obtain

$$W_{k+1} = \mathscr{K}\left[M_{k'}^{j} u:\ u \in E(W_k)\right]$$

where $E(W_k)$ denotes the set of extreme points of W_k. Clearly, the new extreme points $E(W_{k+1})$ are images of $E(W_k)$. If $|\lambda(M_{k'})| < 1$ for $M_{k'} \in A$, then there exists an integer $J_{k'}$ such that

$$\mathscr{K}\left[\bigcup_{j=0}^{\infty} M_{k'}^{j} W_k\right] = \mathscr{K}\left[\bigcup_{j=0}^{J_{k'}} M_{k'}^{j} W_k\right]$$

since W_k is a bounded neighborhood of the origin. Notice that $J_{k'}$ can be recognized since it is the smallest J_k to satisfy

$$M_{k'} \mathscr{K}\left[\bigcup_{j=0}^{J_k} M_{k'}^{j} W_k\right] \subseteq \mathscr{K}\left[\bigcup_{j=0}^{J_{k'}} M_{k'}^{j} W_k\right].$$

Thus W_{k+1} will be formed in a *finite* number of steps, since W_k is expressed as the convex hull of a finite set of points.

In practice, W_0 above is usually chosen as simple as possible, i.e., it is chosen as the region defined by

$$E(W_0) = \left\{w_i \in R^n:\ x_{ii} = 1,\ x_{ij} = 0,\ j \neq i,\ i = 1, \cdots, n\right\}$$

where $w_i = (x_{i1}, x_{i2}, \cdots, x_{in}) \in R^n$. Note that W_0 determined in this way is symmetric, and of all symmetric polyhedral regions, it possesses a minimal number of extreme points, namely $2n$.

We call a set of matrices A *unstable* if A is not stable. In [3] the following instability criterion is established: A is unstable if there exists a k such that $\partial W_0 \cap \partial W_k = \varnothing$ where $\varnothing$ denotes the null set. For additional (and improved) instability criteria, refer to [4].

In [4] it is also shown that if a set A of matrices with $E(A)$ finite, is asymptotically stable, then the *constructive algorithm* given above will terminate "stable" in a finite number of steps. Thus a set A can be determined stable in a finite numbers of steps if A is asymptotically stable. We have no way of knowing, by means of the constructive algorithm alone, that A is asymptotically stable at the termination of the algorithm. However, we can show that A is asymptotically stable by choosing a $\rho > 1$ sufficiently small and then showing that ρA is stable by using the constructive algorithm.

Next, we observe that the set M given in (12) consists in general of infinitely many matrices. However, the following result, established in [3], reduces the stability analysis of the equilibrium $x = 0$ of (12) to a finite set of matrices: let A be a set of matrices in the linear space of $n \times n$ matrices and let $E(A)$ be the set of extreme matrices of A. Then $\mathscr{K}(A)$ is stable if and only if $E(A)$ is stable. Thus if $E(A)$ happens to be finite, then the stability analysis of A (and hence of (12)) can be accomplished in a finite number of steps.

III. Application of the Constructive Algorithm to the Stability Analysis of Digital Filters

In this section, we show how to apply the constructive algorithm of Brayton and Tong to the stability analysis of digital filters. This section consists of four parts. In Section III-A the types of nonlinearities that occur in fixed-point digital filters will be presented. In Section III-B, we present the procedure used to determine the extreme matrices for a general second-order digital filter. In Section III-C, this procedure is applied to two types of second-order digital filter structures: direct form and coupled form. (In a companion paper [13] we consider wave filters and lattice filters.) In Section III-D we briefly discuss the implementation of the constructive stability algorithm.

In Section IV, the stability results obtained by the constructive algorithm for these two filter structures are compared with existing stability results.

A. Nonlinearities in Digital Filters

In digital filters, the representation of signals must by necessity have finite precision. This is a consequence of the encoding of the signals in a particular format (e.g., fixed or floating point) and of the storage of these signals in registers which have finite wordlength. Multiplications and additions performed in the digital filter generally lead to an increase in the wordlength required for the result of the operation. If the number of operations performed on a signal remains finite, as in a nonrecursive filter, the increasing wordlength can be handled by using larger registers for storing the results of the arithmetic operations. However, in a recursive digital filter, a wordlength reduction is necessary to prevent the wordlength of the signals from increasing indefinitely.

In the present paper, we assume that the digital filters use fixed-point arithmetic. In fixed-point arithmetic, each number is represented by a sign bit and a magnitude. Thus the magnitude of any number is represented by a string of binary digits of fixed length B. When two B-bit numbers are multiplied, the result is a $2B$-bit number. A *quantization* nonlinearity is produced when the $2B$-bit number is reduced in wordlength to B bits. Quantization only affects the least significant bits. Addition also poses a problem when the sum of two numbers falls outside the representable range. An *overflow* nonlinearity results when this number is modified so that it falls back within the representable range. In general, the overflow nonlinearity

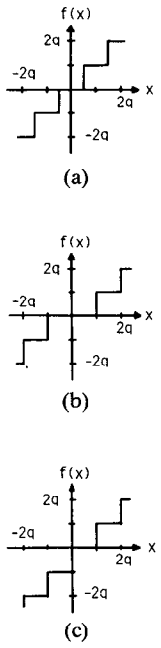


Fig. 1. Fixed-point quantization characteristics. (a) Roundoff. (b) Magnitude truncation. (c) Value truncation.

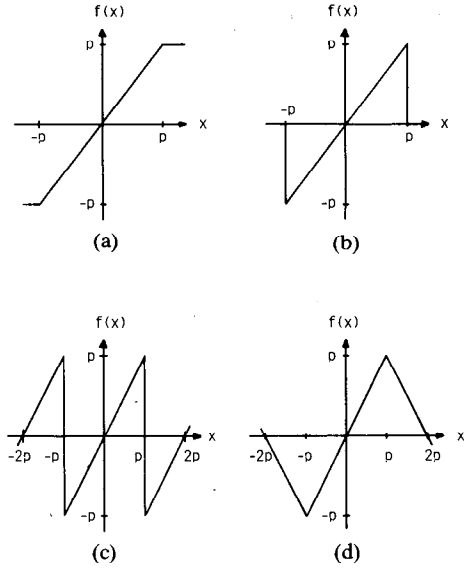


Fig. 2. Overflow characteristics. (a) Saturation. (b) Zeroing. (c) Two's complement. (d) Triangular.

changes the most significant bits as well as the least significant bits of a fixed-point number. These two types of nonlinearities are well described in the literature (see e.g., [24], [25]) and, therefore, will only be briefly discussed here.

Quantization can be performed by substituting the nearest possible number that can be represented by the limited number of bits. This type of nonlinear operation is called a *roundoff quantizer* and its characteristic is shown in Fig. 1(a). Another possibility consists of discarding the least significant bits in the number. If the signals are represented by sign and magnitude then we have a *magnitude truncation quantization* characteristic, as depicted in Fig. 1(b). If the signals are represented in a two's complement format, the nonlinearity is a *two's complement* or *value truncation quantization*, as shown in Fig. 1(c). In this paper, value truncation is not considered. Thus the term *truncation* will always refer to *magnitude truncation* in the sequel.

If an overflow occurs, a number of different actions may be taken. If the number that caused the overflow is replaced by a number having the same sign, but with a magnitude corresponding to the overflow level, a *saturation overflow* characteristic shown in Fig. 2(a) is obtained. *Zeroing overflow* substitutes the number zero in case of an overflow (see Fig. 2(b)). In two's complement arithmetic, the most significant bits that caused the overflow are discarded. In this case, overflows in intermediate results do not cause errors, as long as the final result does not have overflow. This *two's complement overflow* characteristic is illustrated in Fig. 2(c). Another way of dealing with overflow is the *triangular overflow* characteristic (see Fig. 2(d)) as proposed by Eckhardt and Winkelnkemper [11].

It is possible to have different wordlengths for the various signals in the filter, resulting in different quantization step sizes and/or different overflow levels. We will assume throughout this paper that all quantizers in a filter have the same quantization step size, q, and are of the same type (e.g., roundoff or truncation). Similarly, we will assume that all overflow nonlinearities in a filter have the same overflow level, p, and are of the same type.

The above nonlinearities may be viewed as belonging to a sector $[k_m, k_M]$. Thus if $f(\cdot)$ denotes a given nonlinearity, then

$$k_m\sigma^2 \leqslant \sigma f(\sigma) \leqslant k_M\sigma^2, \qquad \text{for all } \sigma \in R$$

where k_m, k_M are constants such that $-\infty < k_m \leqslant k_M < \infty$.

Under the above assumptions, we view the quantization nonlinearities as belonging to the sector $[0, k_q]$ where

$$k_q = \begin{cases} 1, & \text{for truncation} \\ 2, & \text{for roundoff} \end{cases}. \tag{13}$$

Henceforth, k_q will represent the upper slope of the sector that contains the quantization nonlinearity. The overflow nonlinearities are viewed as belonging to the sector $[k_0, 1]$ where

$$k_0 = \begin{cases} 0, & \text{for saturation or zeroing} \\ -\dfrac{1}{3}, & \text{for triangular} \\ -1, & \text{for two's complement.} \end{cases} \tag{14}$$

Henceforth, k_0 will represent the lower slope of the sector which contains the overflow nonlinearity.

When the above two nonlinear operations are combined (i.e., quantization and overflow functions are executed simultaneously), then the composite nonlinear operation may be viewed as belonging to the sector $[k_0, k_q]$. The constant k_0 is determined by the type of overflow being performed and the constant k_q is determined by the type of quantization operation.

Our representation of a fixed-point digital filter is not an exact description of an actual realization of such a filter.

Due to the finite number of values that a signal in a digital filter can assume, actual realizations of digital filters constitute finite state machines. The digital filters which we analyze are still idealizations in the sense that they are not finite state machines. This difficulty does not pose a serious problem since we assume that a filter operates in its intended range.

B. General Digital Filter

In order to apply the constructive algorithm, we represent a digital filter by a system of difference equations,

$$x(k+1) = g[x(k)] \tag{15}$$

where $k = 0,1,2,\cdots$. Following the procedure outlined in Section II, we rewrite the given system equations as

$$x(k+1) = M(x(k))x(k) \tag{16}$$

where $M(x(k))$ is chosen so that $M(x(k))x(k) = g[x(k)]$ for all allowable x. Since we consider only second-order systems in this paper, the matrix M may be rewritten as

$$M(x(k)) = \begin{bmatrix} a(x(k)) & b(x(k)) \\ c(x(k)) & d(x(k)) \end{bmatrix}.$$

We assume that the elements of M satisfy the inequalities

$$\alpha_1 \leqslant a(x(k)) \leqslant \alpha_2 \qquad \beta_1 \leqslant b(x(k)) \leqslant \beta_2$$
$$\gamma_1 \leqslant c(x(k)) \leqslant \gamma_2 \qquad \delta_1 \leqslant d(x(k)) \leqslant \delta_2$$

where α_i, β_i, γ_i, and δ_i, $i = 1,2$ are constants.

Let $\boldsymbol{M}$ be the set of all matrices obtained by varying $x(k)$ in $M(x(k))$ over all allowable values. The extreme matrices of $\boldsymbol{M}$ are given by (see Section II-C)

$$E(\boldsymbol{M}) = \left\{ \begin{bmatrix} \alpha_i & \beta_j \\ \gamma_k & \delta_l \end{bmatrix}, i,j,k,l = 1,2 \right\}.$$

By the results of Section II-D, the set $\boldsymbol{M}$ is stable (asymptotically stable) if and only if $E(\boldsymbol{M})$ is stable (asymptotically stable). Therefore, we need only determine the stability (asymptotic stability) properties of $E(\boldsymbol{M})$ to determine stability (global asymptotic stability) of the digital filter described by (15). If the set $\boldsymbol{M}$ is unstable then we can draw no conclusion about the stability of (15).

Using the results of Section II-D, we show that $\boldsymbol{M}$ is asymptotically stable by choosing a $\rho > 1$ sufficiently small and then showing that $\rho\boldsymbol{M}$ is stable by applying the constructive algorithm. For the digital filters treated in the present paper, the choice of $\rho = 1.0000001$ was satisfactory to ascertain asymptotic stability in all cases considered.

Since the constructive algorithm shows that the equilibrium $x = 0$ of a given digital filter (15) is *globally asymptotically stable*, then in particular, no limit cycles will exist in such a filter.

In the Appendix, a brief description is given of the computer programs used in our stability analysis of digital filters by the constructive algorithm.

C. Specific Digital Filters

In this part, we give details of the application of the constructive algorithm to the stability analysis of the following filter structures: a) Direct Form digital filter, and b) Coupled Form digital filter. A stability analysis of wave digital filters and lattice digital filters (by the constructive algorithm) will be given in a companion paper [13].

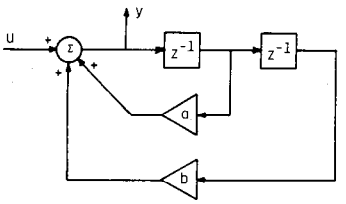


Fig. 3. Linear second-order direct form digital filter.

For each of the digital filter structures considered, the region, in terms of the filter parameters, where the linear filter (i.e., the filter without quantization or overflow) is globally asymptotically stable is known precisely. Since a linear filter is either only stable or is unstable outside of this region, we are not interested in nonlinear filters whose parameters fall outside of this region.

Next, we present the particular nonlinear structures for each type of filter which we consider. In addition, for each nonlinear filter structure we derive the set of extreme matrices used by the constructive algorithm.

1) Direct Form Digital Filter:

The second-order direct form digital filter has been investigated extensively [8]. Since we only consider filters with zero input, the recursive parts of the direct form 1 structure and the direct form 2 structure are equivalent. The linear recursive part of this digital filter is shown in Fig. 3.

The region where this linear filter is globally asymptotically stable in terms of the parameters a and b is obtained by considering the transfer function of the linear filter,

$$H(z) = \frac{z^2}{z^2 - az - b}.$$

Using Jury's criterion [17], it follows that the ideal second-order digital filter is globally asymptotically stable if and only if

$$|b| < 1$$
$$|a| + b < 1.$$

This stability region corresponds to the triangular region shown in Fig. 4. The linear filter is globally asymptotically stable for all coefficients inside this region.

When the linear second-order direct form digital filter is implemented in fixed-point arithmetic, there are two possible ways of placing the quantization nonlinearity. Quantization can be performed immediately after each multiplication. This nonlinear second order digital filter structure is shown in Fig. 5, with Q representing a quantizer. Alternatively, the results of the two multiplications may be added with full precision and only one quantization is needed. This structure is shown in Fig. 6. For both possible quantizer configurations, the overflow nonlinearity, P, is placed after the adder as shown. We next develop the set of extreme matrices for each structure.

a) One quantizer: The structure for the second-order direct form digital filter with one quantizer is shown in Fig. 6. We will consider the quantization and overflow nonlin-

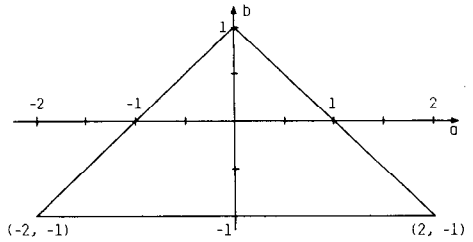


Fig. 4. Region in the parameter plane where a linear second-order direct form filter is globally asymptotically stable.

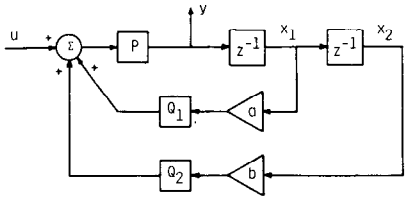


Fig. 5. Direct form digital filter with two quantizers.

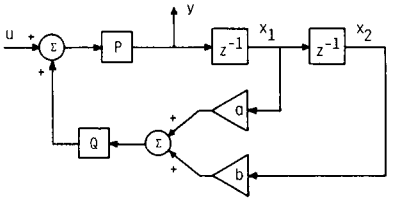


Fig. 6. Direct form digital filter with one quantizer.

earities together. With this assumption, the state equations are

$$x_1(k+1) = f[ax_1(k) + bx_2(k)]$$
$$x_2(k+1) = x_1(k) \tag{17}$$

where $f(\cdot)$ is the combined quantization and overflow nonlinearity.

Following the technique outlined in Section III-B, the state equations are expressed as

$$x(k+1) = M(x(k))x(k)$$

where $M(x(k))$ is given by

$$M(x) = \begin{bmatrix} \Phi(x)a & \Phi(x)b \\ 1 & 0 \end{bmatrix} \tag{18}$$

with

$$\Phi(x) = \frac{f[ax_1 + bx_2]}{ax_1 + bx_2}. \tag{19}$$

Since we view the quantization and overflow nonlinearities as belonging to a sector, the function $\Phi(x)$ is bounded by constants α_1 and α_2 such that

$$\alpha_1 \leqslant \Phi(x) \leqslant \alpha_2.$$

For the particular nonlinearities which we consider, we have

$$\alpha_1 = k_0 \quad \text{and} \quad \alpha_2 = k_q$$

where k_0 and k_q are defined by (14) and (13), respectively. The extreme matrices of the set $\boldsymbol{M}$ are

$$E(\boldsymbol{M}) = \left\{ \begin{bmatrix} \alpha_i a & \alpha_i b \\ 1 & 0 \end{bmatrix}, i = 1,2 \right\}. \tag{20}$$

In this case, for each point (a, b) (in the $a-b$ parameter plane), there are two extreme matrices given by A_1 and A_2, where

$$A_1 = \begin{bmatrix} k_0 a & k_0 b \\ 1 & 0 \end{bmatrix} \qquad A_2 = \begin{bmatrix} k_q a & k_q b \\ 1 & 0 \end{bmatrix}. \tag{21}$$

If the overflow nonlinearity is absent, then $\alpha_1 = 0$ and the set of extreme matrices in this case is the same as for one saturation or zeroing overflow nonlinearity.

b) Two quantizers: The structure of the second order direct form digital filter with two quantizers is shown in Fig. 5. We cannot combine the quantization and overflow nonlinearities in this case. The state equations are given by

$$x_1(k+1) = P\{Q_1[ax_1(k)] + Q_2[bx_2(k)]\}$$
$$x_2(k+1) = x_1(k) \tag{22}$$

and can be rewritten as

$$x(k+1) = M(x(k))x(k)$$

where

$$M(x) = \begin{bmatrix} \Phi_1(x)\Phi_3(x)a & \Phi_2(x)\Phi_3(x)b \\ 1 & 0 \end{bmatrix} \tag{23}$$

and

$$\Phi_1(x) = \frac{Q_1[ax_1]}{ax_1} \qquad \Phi_2(x) = \frac{Q_2[bx_2]}{bx_2}$$
$$\Phi_3(x) = \frac{P\{Q_1[ax_1] + Q_2[bx_2]\}}{Q_1[ax_1] + Q_2[bx_2]}. \tag{24}$$

When the $M(x(k))$ given by (23) and (24) is multiplied by $x(k) = [x_1(k) \quad x_2(k)]^T$, the state equations (22) are obtained.

Since the quantization and overflow nonlinearities belong to a sector, the functions $\Phi_1(x)$, $\Phi_2(x)$, and $\Phi_3(x)$ are bounded by constants

$$\alpha_{i1} \leqslant \Phi_i(x) \leqslant \alpha_{i2}, \qquad i = 1,2,3$$

where

$$\alpha_{11} = \alpha_{21} = 0 \quad \text{and} \quad \alpha_{12} = \alpha_{22} = k_q$$

and

$$\alpha_{31} = k_0, \quad \alpha_{32} = 1.$$

The functions $\Phi_1(x)\Phi_3(x)$ and $\Phi_2(x)\Phi_3(x)$ are also bounded by constants, β_i and γ_i, such that

$$\beta_1 \leqslant \Phi_1(x)\Phi_3(x) \leqslant \beta_2 \quad \text{and} \quad \gamma_1 \leqslant \Phi_2(x)\Phi_3(x) \leqslant \gamma_2$$

where

$$\beta_1 = \min(\alpha_{11}\alpha_{31}, \alpha_{11}\alpha_{32}, \alpha_{12}\alpha_{31}, \alpha_{12}\alpha_{32}) = k_q k_0$$
$$\beta_2 = \max(\alpha_{11}\alpha_{31}, \alpha_{11}\alpha_{32}, \alpha_{12}\alpha_{31}, \alpha_{12}\alpha_{32}) = k_q$$
$$\gamma_1 = \min(\alpha_{21}\alpha_{31}, \alpha_{21}\alpha_{32}, \alpha_{22}\alpha_{31}, \alpha_{22}\alpha_{32}) = k_q k_0$$
$$\gamma_2 = \max(\alpha_{21}\alpha_{31}, \alpha_{21}\alpha_{32}, \alpha_{22}\alpha_{31}, \alpha_{22}\alpha_{32}) = k_q.$$

The extreme matrices of the set $\boldsymbol{M}$ are

$$E(\boldsymbol{M}) = \left\{ \begin{bmatrix} \beta_i a & \gamma_j b \\ 1 & 0 \end{bmatrix}, i, j = 1,2 \right\}. \tag{25}$$

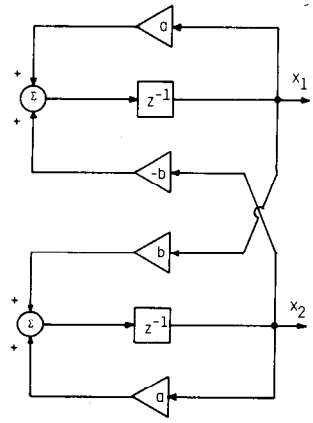


Fig. 7. Linear second-order coupled form digital filter.

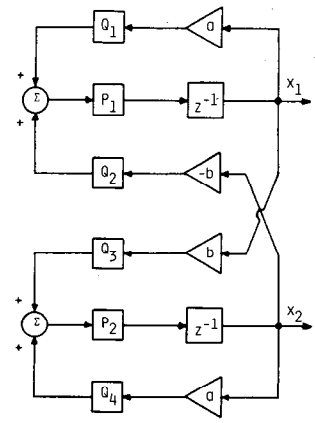


Fig. 8. Coupled form digital filter with four quantizers.

For each point (a, b) in the parameter plane, there are four extreme matrices used in the constructive algorithm. For this example, these extreme matrices are

$$A_1 = \begin{bmatrix} k_q k_0 a & k_q k_0 b \\ 1 & 0 \end{bmatrix} \quad A_2 = \begin{bmatrix} k_q k_0 a & k_q b \\ 1 & 0 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} k_q a & k_q k_0 b \\ 1 & 0 \end{bmatrix} \quad A_4 = \begin{bmatrix} k_q a & k_q b \\ 1 & 0 \end{bmatrix}. \tag{26}$$

If the overflow nonlinearity is absent, then $\alpha_{31} = \beta_1 = \gamma_1 = 0$ and the set of extreme matrices in this case is the same as for the filter with a saturation or zeroing overflow nonlinearity.

2) Coupled Form Digital Filter:

The coupled or normal form digital filter was first proposed by Rader and Gold [26] as a digital filter structure whose pole locations are less sensitive than the direct form structure to parameter errors. However, the coupled form can only realize complex-conjugate poles. With finite wordlength parameters this structure also has a uniform grid of possible pole locations [24]. The linear recursive part of a coupled form digital filter whose poles are at $a \pm jb$ and which has zero input is shown in Fig. 7.

The linear filter is globally asymptotically stable if and only if its poles lie within the unit circle. Equivalently, the parameters a and b must satisfy

$$a^2 + b^2 < 1.$$

This region corresponds to the interior of the unit circle in the $a - b$ parameter plane.

As in the direct form digital filter, there are two possible ways of placing the quantization nonlinearity. Quantization can be performed immediately after each multiplication and thus four quantizers will be needed. This filter structure is shown in Fig. 8 with Q_i, $i = 1, \cdots, 4$ representing the quantizers. Alternatively, the results of two multiplications may be added with full precision and then quantized. This implementation uses two quantizers and is shown in Fig. 9. For both possible placements of the quantization nonlinearity, the overflow nonlinearities, P_1 and P_2, must be placed after each addition as shown. We next develop the set of extreme matrices for each structure which will be used by the constructive algorithm.

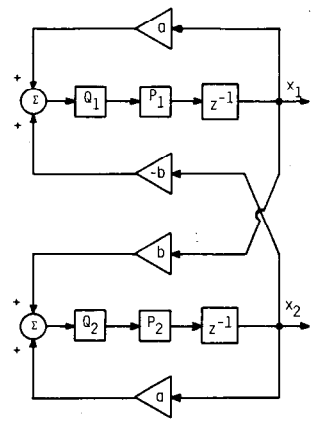


Fig. 9. Coupled form digital filter with two quantizers.

a) Two quantizers: The coupled form digital filter structure to be analyzed is shown in Fig. 9. As in the direct form digital filter, we assume that the overflow and quantization nonlinearities before each delay are combined. With this assumption, the state equations for the filter are

$$\begin{aligned} x_1(k+1) &= f_1[ax_1(k) - bx_2(k)] \\ x_2(k+1) &= f_2[bx_1(k) + ax_2(k)] \end{aligned} \tag{27}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are the combined quantization and overflow nonlinearities.

Following the technique outlined in Section III-B, the state equations are written as

$$x(k+1) = M(x(k))x(k).$$

Defining

$$\Phi_1(x) = \frac{f_1[ax_1 - bx_2]}{ax_1 - bx_2}$$

$$\Phi_2(x) = \frac{f_2[bx_1 + ax_2]}{bx_1 + ax_2} \tag{28}$$

the matrix $M(x)$ is given by

$$M(x) = \begin{bmatrix} \Phi_1(x)a & -\Phi_1(x)b \\ \Phi_2(x)b & \Phi_2(x)a \end{bmatrix}. \tag{29}$$

The functions $\Phi_1(x)$ and $\Phi_2(x)$ are bounded by constants

$$\alpha_1 \leqslant \Phi_1(x) \leqslant \alpha_2$$
$$\beta_1 \leqslant \Phi_2(x) \leqslant \beta_2.$$

For the particular nonlinearities which we consider, these constants are

$$\alpha_1 = \beta_1 = k_0$$
$$\alpha_2 = \beta_2 = k_q.$$

The extreme matrices of the set $\boldsymbol{M}$ are

$$E(\boldsymbol{M}) = \left\{ \begin{bmatrix} \alpha_i a & -\alpha_i b \\ \beta_j b & \beta_j a \end{bmatrix}, \ i, j = 1,2 \right\}. \tag{30}$$

Therefore, for a given point in the $a-b$ parameter plane, the constructive algorithm uses four extreme matrices. If the overflow nonlinearities are absent, then $\alpha_1 = \beta_1 = 0$ and the set of extreme matrices in this case is the same as for the filter with two saturation or zeroing overflow nonlinearities.

b) Four quantizers: The structure of the coupled form digital filter with four quantizers is shown in Fig. 8. The state equations are given by

$$x_1(k+1) = P_1\{Q_1[ax_1(k)] + Q_2[-bx_2(k)]\}$$
$$x_2(k+1) = P_2\{Q_3[bx_1(k)] + Q_4[ax_2(k)]\} \tag{31}$$

or, equivalently, by

$$x(k+1) = M(x(k))x(k)$$

where

$$M(x) = \begin{bmatrix} \Phi_1(x)\Phi_3(x)a & -\Phi_1(x)\Phi_4(x)b \\ \Phi_2(x)\Phi_5(x)b & \Phi_2(x)\Phi_6(x)a \end{bmatrix} \tag{32}$$

and

$$\Phi_1(x) = \frac{P_1\{Q_1[ax_1] + Q_2[-bx_2]\}}{Q_1[ax_1] + Q_2[-bx_2]}$$
$$\Phi_2(x) = \frac{P_2\{Q_3[bx_1] + Q_4[ax_2]\}}{Q_3[bx_1] + Q_4[ax_2]}$$
$$\Phi_3(x) = \frac{Q_1[ax_1]}{ax_1} \qquad \Phi_4(x) = \frac{Q_2[-bx_2]}{-bx_2}$$
$$\Phi_5(x) = \frac{Q_3[bx_1]}{bx_1} \qquad \Phi_6(x) = \frac{Q_4[ax_2]}{ax_2}. \tag{33}$$

The functions $\Phi_i(x)$, are bounded by constants α_{ij} such that

$$\alpha_{i1} \leqslant \Phi_i(x) \leqslant \alpha_{i2}, \qquad i = 1,2,3,4,5,6$$

where

$$\alpha_{11} = \alpha_{21} = k_0, \qquad \alpha_{12} = \alpha_{22} = 1$$
$$\alpha_{31} = \alpha_{41} = \alpha_{51} = \alpha_{61} = 0$$

and

$$\alpha_{32} = \alpha_{42} = \alpha_{52} = \alpha_{62} = k_q.$$

Therefore, the functions $\Phi_1(x)\Phi_3(x)$, $\Phi_1(x)\Phi_4(x)$, $\Phi_2(x)\Phi_5(x)$, and $\Phi_2(x)\Phi_6(x)$ are bounded by constants β_i, γ_i, δ_i, ϵ_i, $i = 1,2$ such that

$$\beta_1 \leqslant \Phi_1(x)\Phi_3(x) \leqslant \beta_2 \qquad \gamma_1 \leqslant \Phi_1(x)\Phi_4(x) \leqslant \gamma_2$$
$$\delta_1 \leqslant \Phi_2(x)\Phi_5(x) \leqslant \delta_2$$

and

$$\epsilon_1 \leqslant \Phi_2(x)\Phi_6(x) \leqslant \epsilon_2$$

where

$$\beta_1 = \gamma_1 = \delta_1 = \epsilon_1 = k_q k_0$$

and

$$\beta_2 = \gamma_2 = \delta_2 = \epsilon_2 = k_q.$$

The extreme matrices of the set $\boldsymbol{M}$ are

$$E(\boldsymbol{M}) = \left\{ \begin{bmatrix} \beta_i a & -\gamma_j b \\ \delta_k b & \epsilon_l a \end{bmatrix}, \ i, j, k, l = 1,2 \right\}. \tag{34}$$

For this filter, there are sixteen extreme matrices for a given point in the $a-b$ parameter plane. If the overflow nonlinearities are absent, then $\beta_1 = \gamma_1 = \delta_1 = \epsilon_1 = 0$ and the set of extreme matrices in this case is the same as for the filter with two saturation or zeroing overflow nonlinearities.

D. Implementation of the Constructive Algorithm

In the next section, we apply the constructive algorithm to the extreme matrices given in (20), (25), (30), and (34) to determine regions in the parameter plane for which the various filter structures under discussion are globally asymptotically stable. To simplify matters, we phrase the following discussion in terms of the extreme matrix (20),

$$E(\boldsymbol{M}) = \left\{ \begin{bmatrix} \alpha_i a & \alpha_i b \\ 1 & 0 \end{bmatrix}, \ i = 1,2 \right\}. \tag{20}$$

There are several ways of estimating regions in the $a-b$ parameter plane for which the filter given by (17) is globally asymptotically stable. We comment on two:

Method 1: Equation (20) is evaluated on a sufficiently fine grid in a subset of the $a-b$ parameter plane and the constructive algorithm is then applied to each of the resulting sets of extreme matrices. In the Appendix, a brief description of the computer programs which accomplish this is given.

Method 2: We can modify the extreme matrices in (20) by incorporating intervals for the parameters a and b. For example, suppose we want to determine whether the filter given by (17) is globally asymptotically stable for *all* points in the rectangle

$$D_k = \{(a,b) \in R^2 : A_{k1} \triangleq a_k - \delta_k \leqslant a \leqslant a_k + \delta_k \triangleq A_{k2}$$
$$\text{and } B_{k1} \triangleq b_k - \epsilon_k \leqslant b \leqslant b_k + \epsilon_k \triangleq B_{k2}\}$$

for some $\epsilon_k > 0$, $\delta_k > 0$. In this case, the constructive algorithm is applied to the set of extreme matrices $E_k(\boldsymbol{M})$ given by

$$E_k(\boldsymbol{M}) = \left\{ \begin{bmatrix} \alpha_i A_{kj} & \alpha_i B_{kl} \\ 1 & 0 \end{bmatrix}, \ i, j, l = 1,2, \right\}. \quad \blacksquare$$

Remark: In Method 2, the set $E_k(M)$ can be further simplified by recognizing that for $i, j, l = 1,2$ we have the estimates

$$\beta_{k1} \leqslant \alpha_i A_{kj} \leqslant \beta_{k2}, \qquad \gamma_{k1} \leqslant \alpha_i B_{kl} \leqslant \gamma_{k2}$$

where β_{k1}, β_{k2}, γ_{k1}, and γ_{k2} are appropriate constants (as discussed in the next paragraph). Thus for Method 2, the set of extreme matrices $E_k(M)$ assumes the form

$$E_k(M) = \left\{ \begin{bmatrix} \beta_{ki} & \gamma_{kj} \\ 1 & 0 \end{bmatrix}, \; i, j = 1,2 \right\}.$$

For rectangles located in the first quadrant of the parameter plane, we have $\beta_{k1} = k_0 A_{k2}$, $\beta_{k2} = k_q A_{k2}$, $\gamma_{k1} = k_0 B_{k2}$, and $\gamma_{k2} = k_q B_{k2}$. Therefore, if the filter (17) has been determined to be globally asymptotically stable for the point (A_{k2}, B_{k2}) in the $a-b$ parameter plane, then it turns out that the filter (17) will actually be globally asymptotically stable for parameters located in any rectangle which lies in the first quadrant of the parameter plane and whose upper right-hand corner is (A_{k2}, B_{k2}).

For rectangles located in the other three quadrants of the parameter plane, similar statements apply. ■

Methods 1 and 2 can be combined in an effective manner. In this approach, one first uses Method 1 to determine a region of stability G for (17) via a grid; and then, one attempts to cover as much of G as possible with an appropriate set of rectangles, obtained using Method 2, to ensure that filter (17) is globally asymptotically stable for *all* parameters corresponding to the subset of G covered by the rectangles.

We found that Method 1 by itself yields quite satisfactory results and is easily implemented. The results which are presented in the next section were obtained by Method 1. In general, results obtained by Method 2 will be more conservative than results obtained by Method 1.

The above discussion is modified for the extreme matrices (25), (30), and (34) in the obvious way.

IV. Comparison of Stability Results by the Constructive Algorithm with Existing Stability Results

In this section, we present results obtained by applying the constructive algorithm to different nonlinear digital filter structures. For purposes of comparison, we also summarize existing qualitative results for corresponding filter structures. There are two categories of existing qualitative results for fixed-point digital filters. The results of one category constitute sufficient conditions for the absence of limit cycles in a digital filter, while the results in the second category provide sufficient conditions for the global asymptotic stability of a digital filter. Of course, the results in the latter category yield also sufficient conditions for the absence of limit cycles. We compare these existing results with the stability results obtained via the constructive algorithm. Specifically, we will use the constructive algorithm to ascertain the global asymptotic stability of the equilibrium $x = 0$ of a digital filter; this also guarantees the absence of limit cycles in the digital filter in question.

A. Direct Form Digital Filters

For a direct form digital filter, we consider the filter implemented with one or two quantizers. For both of these structures, we summarize existing qualitative results for the direct form filter and compare these results with the global asymptotic stability results obtained using the constructive algorithm.

1) One Quantizer:

a) Truncation quantizer: For a second-order direct form digital filter with one truncation quantizer and no overflow nonlinearity, the largest region in the $a-b$ parameter plane where zero-input limit cycles are proven to be absent has been reported in Claasen [5]. This result has also been reported in [8]. This region, where limit cycles do not exist, is the same region in the parameter plane where it is shown by Claasen and Kristiansson [6] that the direct form digital filter with one saturation overflow nonlinearity is asymptotically stable.

Theorem 2[5]: No zero-input limit cycles exist in the second-order direct form digital filter of Fig. 6 with one truncation quantizer and no overflow nonlinearity if the following conditions are satisfied:

$$\max_{n>0} |q(n)| < 1.$$

If $(a^2/4) + b > 0$, $q(n)$ is defined as

$$q(n) = -\frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2}(\lambda_1^n - \lambda_2^n)$$

and

$$\lambda_1, \lambda_2 = \frac{a}{2} \pm \sqrt{\frac{a^2}{4} + b}\,.$$

If $(a^2/4) + b \leqslant 0$, the $q(n)$ is rewritten as

$$q(n) = \frac{-r^{n+1}}{\sin(\beta)} \sin(\beta n)$$

where

$$r = \sqrt{-b}\,, \qquad \beta = \arccos \frac{a}{2r}\,. \qquad ■$$

The region in the $a-b$ parameter plane where no limit cycles exist for a direct form filter with one truncation quantizer is represented by the unhatched region of Fig. 10. Only half of the region is shown since it is symmetric with respect to the b axis.

Limit cycles in a second-order direct form digital filter with only an overflow nonlinearity have been studied by Ebert, Mazo, and Taylor [10]. They show that no overflow oscillations exist in the digital filter when saturation overflow or triangular overflow is used. For two's complement overflow, they show that a necessary and sufficient condition for the absence of limit cycles in the filter is given by

$$|a| + |b| < 1. \tag{35}$$

This region in the $a-b$ parameter plane is depicted as the unhatched region in Fig. 11. Ebert, Mazo, and Taylor also state that zeroing overflow also leads to oscillations, but no analysis is presented in [10] to justify this assertion. How-

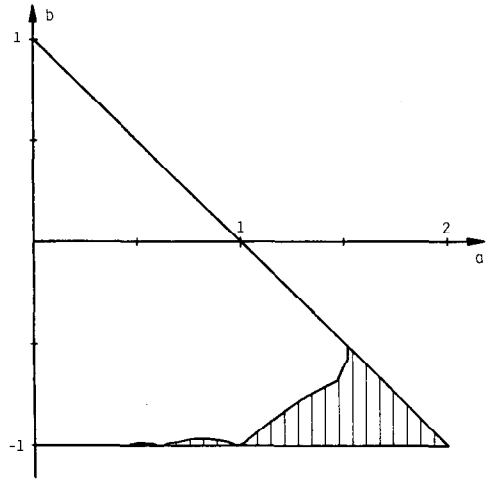


Fig. 10. Region where a direct form filter with one truncation quantizer is free of limit cycles by Theorem 2.

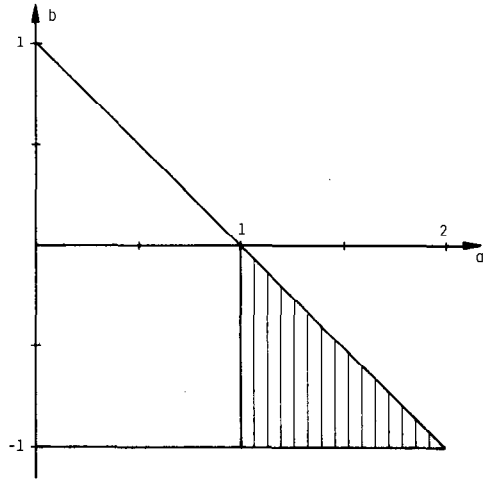


Fig. 10(a). Region where a direct form filter with one zeroing overflow nonlinearity is free of limit cycles by Willson [28].

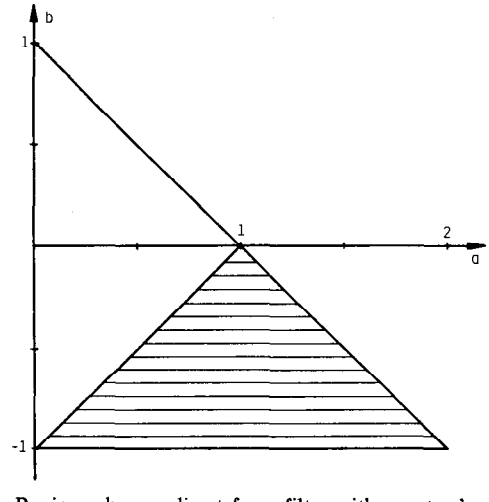


Fig. 11. Region where a direct form filter with one two's complement overflow nonlinearity is free of limit cycles by (35).

ever, Willson [28] obtains an estimate in the $a - b$ parameter plane where limit cycles do not exist in second-order direct form filters with no quantization and zeroing overflow. This region is depicted as the unhatched region in Fig. 10(a). Willson also shows that in the case of second-order direct form filters with quantization and zeroing overflow and with parameter pairs (a, b) belonging to the unhatched region of Fig. 10(a), the amplitudes of limit cycles can be made arbitrarily small by sufficiently decreasing the quantization step size. For earlier results dealing with the problem addressed by Willson [28], see Sandberg [27].

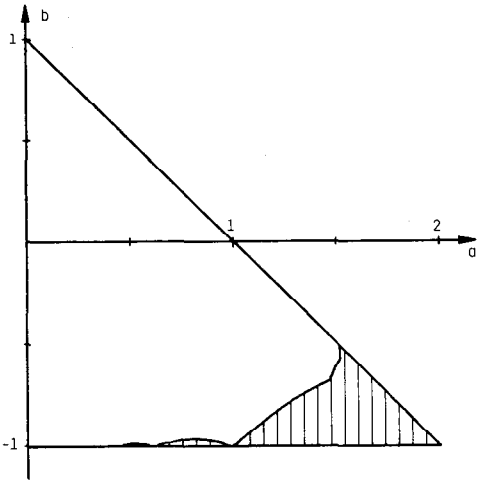


Fig. 12. Region where a direct form filter with one truncation quantizer and saturation, zeroing or no overflow is g.a.s. by the constructive algorithm.

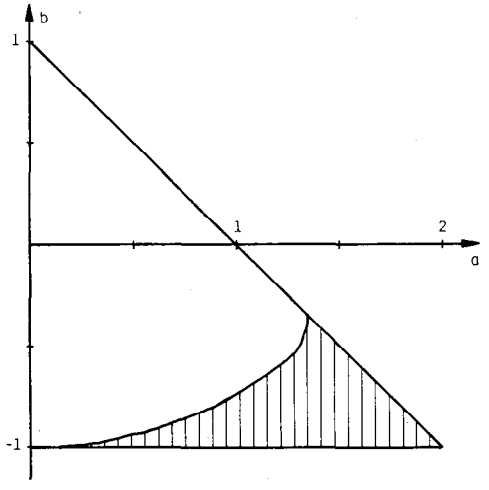


Fig. 13. Region where a direct form filter with one truncation quantizer and triangular overflow is g.a.s. by the constructive algorithm.

To apply the constructive algorithm to the direct form filter with one truncation quantizer, we use the extreme matrices determined by (20). The region of global asymptotic stability in the $a - b$ parameter plane obtained by the constructive algorithm for this filter with a truncation quantizer and saturation, zeroing or no overflow is shown in Fig. 12. The regions in the parameter plane where this digital filter is globally asymptotically stable with triangular and two's complement overflow are shown in Figs. 13 and 14, respectively. The stability results obtained by the constructive algorithm for the overflow nonlinearity without quantization are the same as the results obtained for the

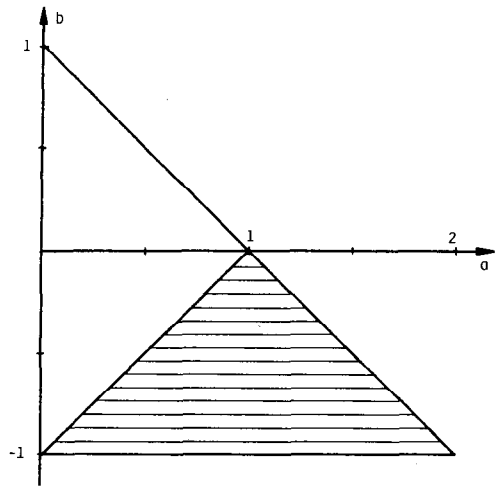


Fig. 14. Region where a direct form filter with one truncation quantizer and two's complement overflow is g.a.s. by the constructive algorithm.

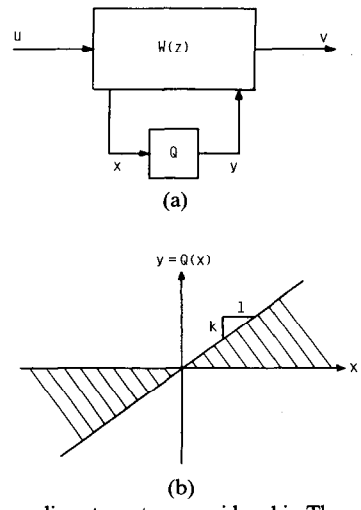


Fig. 15. Nonlinear discrete system considered in Theorem 3. (a) Nonlinear discrete system. (b) Sector in which Q must lie.

overflow and truncation quantization nonlinearities combined.

When quantization is considered separately, the constructive algorithm yields the same region in the parameter plane where the filter is globally asymptotically stable as the result of Claasen [5] which deals with the absence of limit cycles. If we consider the overflow nonlinearity only, then the stability results by the constructive method are more conservative than those of Ebert, Mazo and Taylor [10] for saturation or triangular overflow. However, the constructive algorithm yields the same region where limit cycles are absent for the filter with truncation quantization and two's complement overflow as Ebert, Mazo and Taylor for a single two's complement overflow nonlinearity.

Results contained in Willson [28] pertaining to the case of no quantization and zeroing overflow (see Fig. 10(a)) can be compared to our results pertaining to the same conditions (see Fig. 12). An examination of Figs. 10(a) and 12 shows that neither result implies the other (i.e., the unhatched region of Fig. 12 does not contain the unhatched region of Fig. 10(a), and vice versa).

Those results in Willson [28] which are concerned with the case of quantization and zeroing overflow (where limit cycle amplitude can be controlled by quantization step size) and the present results complement each other. Indeed, our results indicate that for most of the unhatched region in Fig. 10(a), limit cycles do not exist at all.

The present results obtained by the constructive algorithm for any overflow combined with a truncation quantizer seem to be new.

b) Roundoff quantizer: For the direct form digital filter with one roundoff quantizer, Claasen *et al.* [7] have derived a sufficient condition for the absence of zero-input limit cycles. To develop sufficient conditions for the absence of limit cycles in the filter structure shown in Fig. 6 with just the quantization nonlinearity, consider a nonlinear discrete system with one nonlinear element, Q, depicted in Fig. 15(a). In considering zero-input limit cycles, the linear part of the system, W, is described by the transfer function $W(z)$ where $X(z) = W(z)Y(z)$. For Q, we assume that

$$Q(0) = 0$$

$$0 \leqslant \frac{Q(x)}{x} \leqslant k, \qquad x \neq 0$$

$$[Q(x+h) - Q(x)]\,h \geqslant 0, \qquad \text{for all } x \text{ and } h$$

$$Q(-x) = -Q(x). \tag{36}$$

These assumptions imply that the nonlinear characteristic lies in the sector shown in Fig. 15(b) and is a nondecreasing, odd and symmetric function of x.

Theorem 3[7]: Let the discrete system be modeled as shown in Fig. 15(a), containing a linear part described by the transfer function $W(z)$, which must be finite for $|z| = 1$, and a nonlinearity satisfying (36). Limit cycles of length N are absent from the discrete system if there exist $\alpha_p, \beta_p \geqslant 0$ such that for $l = 0, 1, \cdots, [N/2]$,

$$\operatorname{Re}\left[W(z_l)\left[1 + \sum_{p=1}^{N-1}\left\{\alpha_p(1 - z_l^p) + \beta_p(1 + z_l^p)\right\}\right] - \frac{1}{k}\right] < 0 \tag{37}$$

where $z_l = e^{j(2\pi/N)l}$ and $[r]$ denotes the integer part of r.

Claasen *et al.* [7] implement this criterion by transforming it into a linear programming problem and applying existing linear programming algorithms. The region in the parameter plane where no limit cycles exist is approximated by taking a large value of N (e.g., $N = 70$). For roundoff quantization ($k = 2$), the region in the parameter plane where no limit cycles exist by Theorem 3 is identified by the unhatched region in Fig. 16. This criterion can also be applied to the case of one magnitude truncation quantizer, but the region obtained where no limit cycles exist is smaller than the region determined by Theorem 2.

The extreme matrices determined in (20) were used to apply the constructive algorithm to the stability analysis of the direct form digital filter with one roundoff quantizer. The region of global asymptotic stability in the parameter plane obtained by the constructive algorithm for this filter

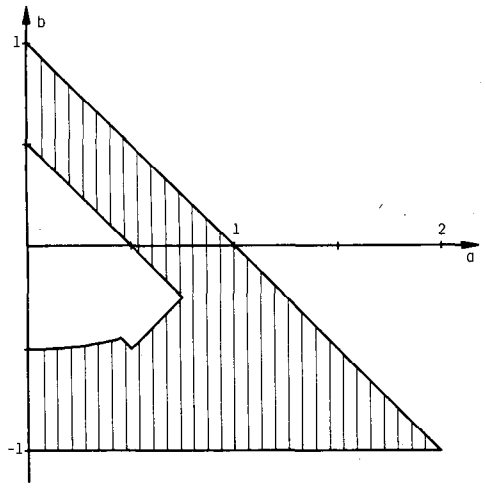


Fig. 16. Region where a direct form filter with one roundoff quantizer and no overflow is free of limit cycles by Theorem 3.

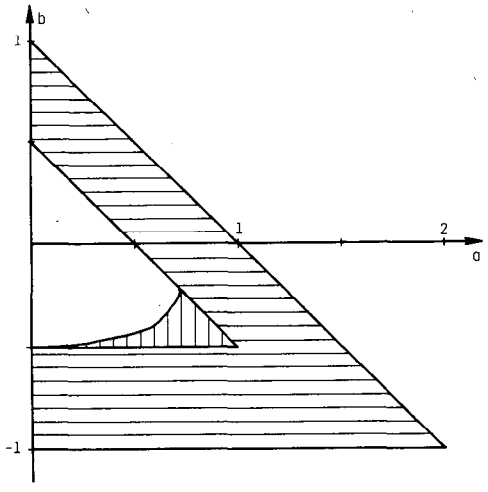


Fig. 18. Region where a direct form filter with one roundoff quantizer and triangular overflow is g.a.s. by the constructive algorithm.

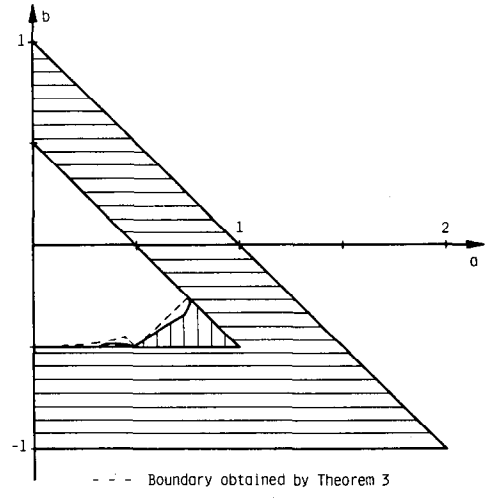


Fig. 17. Region where a direct form filter with one roundoff quantizer and saturation, zeroing or no overflow is g.a.s. by the constructive algorithm.

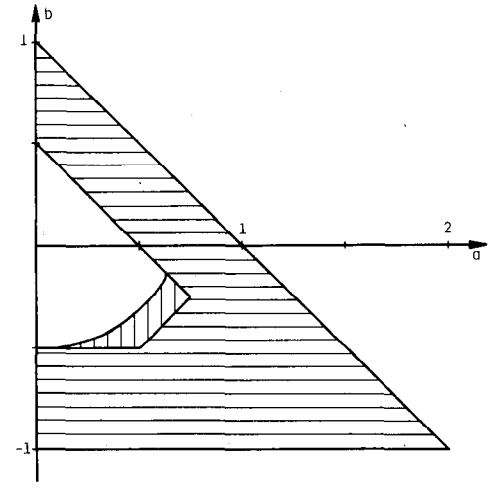


Fig. 19. Region where a direct form filter with one roundoff quantizer and two's complement overflow is g.a.s. by the constructive algorithm.

with a roundoff quantizer and saturation, zeroing or no overflow is shown in Fig. 17. For this case, the region where the filter is determined to be globally asymptotically stable by the constructive algorithm, is slightly larger than the region obtained by applying Theorem 3. The horizontally hatched area indicates the region where at least one of the extreme matrices has an eigenvalue with magnitude greater than one. Limit cycles have been found by others in all of the horizontally hatched region [8]. Vertical hatching indicates that remaining region for which the constructive algorithm is unable to predict global asymptotic stability.

For roundoff quantization and triangular overflow, the region in the parameter plane where the filter is determined to be globally asymptotically stable by the constructive algorithm, is indicated in Fig. 18. The corresponding region when two's complement overflow is used is shown in Fig. 19. Again, horizontal hatching indicates the region where at least one of the extreme matrices has an eigenvalue with magnitude greater than one. Vertical hatching indicates the rest of the uncertain region where we can draw no conclusion about the stability of the system. These results appear to be new.

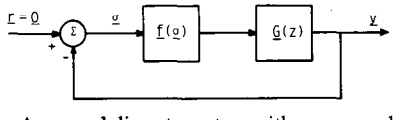


Fig. 20. A general discrete system with many nonlinearities.

2) Two Quantizers:

For two roundoff quantizers or two truncation quantizers, an absolute stability criterion by Jury and Lee [18] can be used to determine sufficient conditions for the global asymptotic stability of the second-order direct form filter with two quantizers. A discrete system with several nonlinearities is represented by the system shown in Fig. 20. The m nonlinear elements are represented by the vector valued function $\boldsymbol{f}(\boldsymbol{\sigma})$ where $f_i(\sigma_i)$ is the output of the ith nonlinear element. The input of this element is the ith

component of the vector σ. The inputs and outputs of the nonlinear elements are interconnected by linear filters with transfer functions $g_{ij}(z)$, which are assumed to be controllable and observable [14]. The functions $g_{ij}(z)$ are the elements of the $m \times m$ transfer matrix $G(z)$. We assume that each element $g_{ij}(z)$ has all of its poles within the unit circle except possibly one pole at $z=1$. The linear filter with the transfer function $g_{ij}(z)$ connects the output of the jth nonlinear element and the input of the ith nonlinear element. We assume that the nonlinearities $f_i(\sigma_i)$ satisfy the following conditions:

$$\text{i)} \quad f_i(0)=0, \qquad i=1,2,\cdots,m$$

$$\text{ii)} \quad 0<\frac{f_i(\sigma_i)}{\sigma_i}<k_{ii}, \qquad \text{for all } \sigma_i \neq 0$$

$$\text{iii)} \quad \sigma(k)\to \mathbf{0} \text{ implies } \mathbf{y}(k)\to 0$$

$$\text{iv)} \quad -\infty<\frac{df_i(\sigma_i)}{d\sigma_i}<\infty \tag{38}$$

where k_{ii} is the ith diagonal element of the $m \times m$ matrix K.

Theorem 4[18]: The system of Fig. 20 satisfying the above conditions for $G(z)$ with nonlinearities described by (38) is globally asymptotically stable if

$$H(z)=2K^{-1}+G(z)+G^*(z)>0, \qquad \text{for all } |z|=1 \tag{39}$$

where $G^*(z)$ denotes the conjugate transpose of $G(z)$ and " $>$ " signifies that the matrix is positive definite. ■

A sufficient condition which guarantees the absence of limit cycles in a direct form filter with two quantizers which is equivalent to Theorem 4 is given in Claasen *et al.* [7].

For the second-order direct form digital filter with two quantization nonlinearities, as shown in Fig. 5, the matrix $G(z)$ may be written as

$$G(z)=\begin{bmatrix} -az^{-1} & -az^{-1} \\ -bz^{-2} & -bz^{-2} \end{bmatrix}.$$

The matrix $H(z)$, given by

$$H(z)=\begin{bmatrix} \dfrac{2}{k_{11}}-az^{-1}-\overline{(az^{-1})} & -az^{-1}-\overline{(bx^{-2})} \\ -bz^{-2}-\overline{(az^{-1})} & \dfrac{2}{k_{22}}-bz^{-2}-\overline{(bz^{-2})} \end{bmatrix}$$

must be positive definite for $|z|=1$. For magnitude truncation quantizers, $k_{11}=k_{22}=1$. The corresponding region in the parameter plane where the filter is globally asymptotically stable is shown as the unhatched region in Fig. 21. Only half of the region is shown, since it is symmetric about the b axis. For two roundoff quantizers, $k_{11}=k_{22}=2$. The region where the filter is globally asymptotically stable for this case is presented in Fig. 22.

We note that Theorem 4 cannot be readily applied to triangular or two's complement overflow nonlinearities, since nonlinearities described by (38) are constrained to lie entirely in the first and third quadrants.

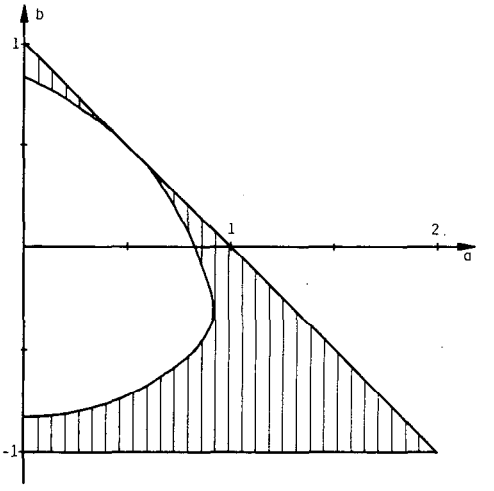


Fig. 21. Region where a direct form filter with two truncation quantizers and no overflow is g.a.s. by Theorem 4.

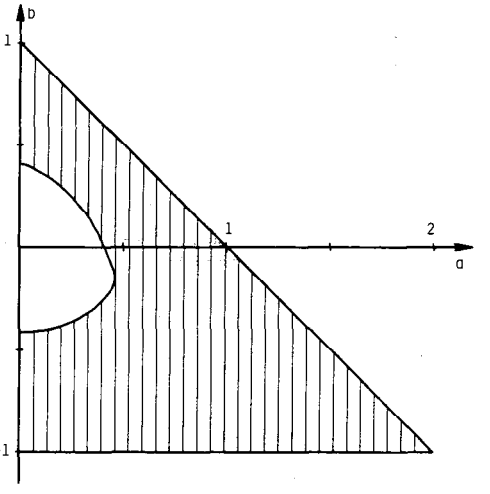


Fig. 22. Region where a direct form filter with two roundoff quantizers and no overflow is g.a.s. by Theorem 4.

To apply the constructive stability algorithm to this filter structure, we use the extreme matrices determined in (25). The regions in the parameter plane, determined by the constructive algorithm, where the digital filter is globally asymptotically stable, are identified for all cases as the unhatched regions in Figs. 23–28. Only half of these regions are shown since they are symmetric about the b-axis. Horizontal hatching indicates the region where at least one extreme matrix has one eigenvalue with a magnitude greater than one. Vertical hatching indicates the rest of the region where we can draw no conclusion about the stability of the system.

As can be eeen from Figs. 23 and 26, the constructive algorithm obtains in this case less conservative results than the application of Theorem 4. Other workers have shown that limit cycles exist in this filter with truncation quantization and no overflow for all of the horizontally hatched region of Fig. 23 [8]. For this filter with roundoff quantization and no overflow, other workers have found that limit cycles exist in most of the horizontally hatched region of Fig. 26 [8]. All of the results obtained for the overflow nonlinearities seem to be new.

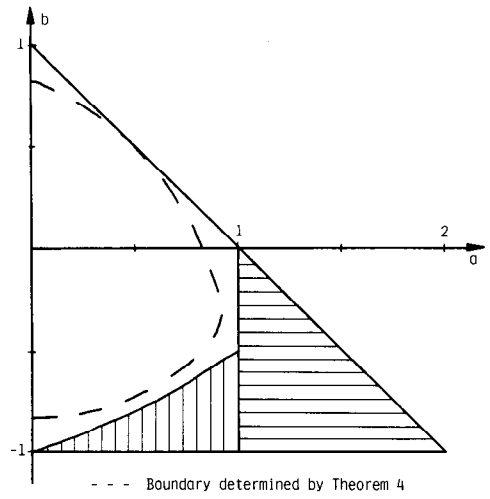


Fig. 23. Region where a direct form filter with two truncation quantizers and saturation, zeroing or no overflow is g.a.s. by the constructive algorithm.

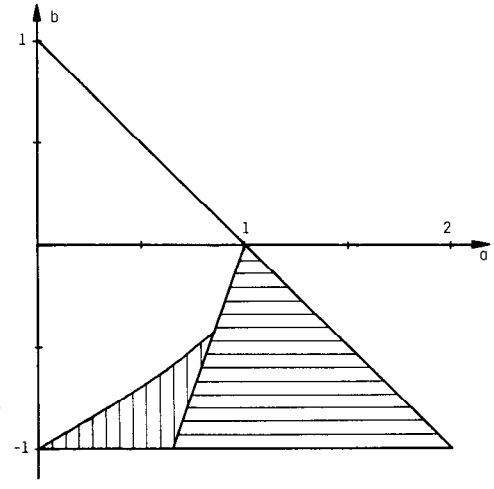


Fig. 24. Region where a direct form filter with two truncation quantizers and triangular overflow is g.a.s. by the constructive algorithm.

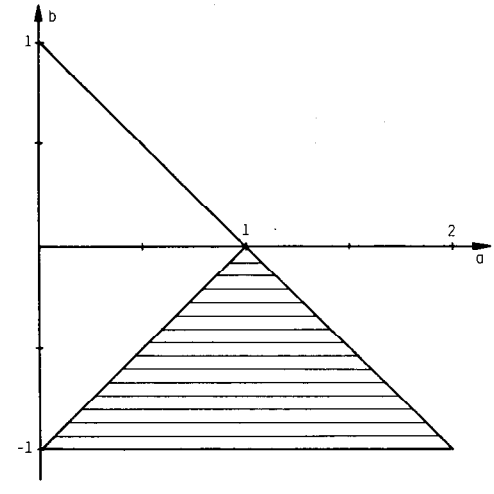


Fig. 25. Region where a direct form filter with two truncation quantizers and two's complement overflow is g.a.s. by the constructive algorithm.

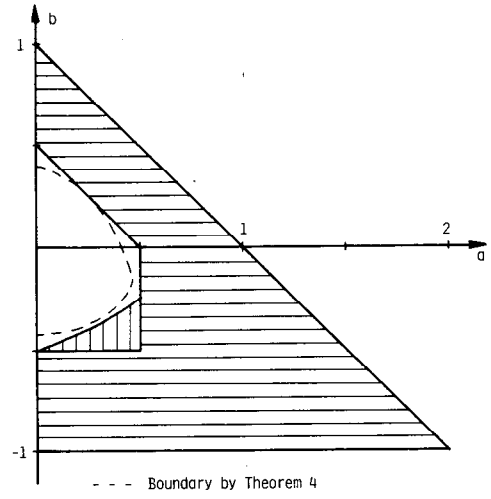


Fig. 26. Region where a direct form filter with two roundoff quantizers and saturation, zeroing or no overflow is g.a.s. by the constructive algorithm.

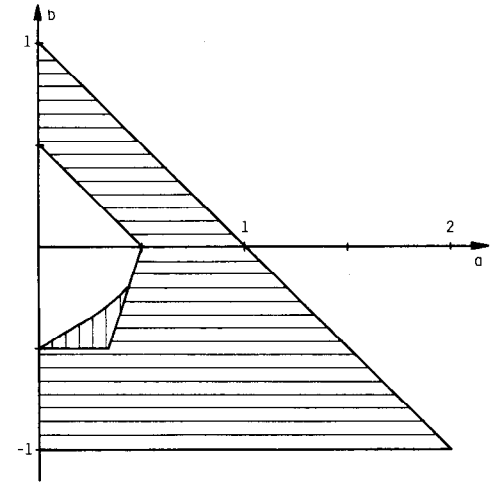


Fig. 27. Region where a direct form filter with two roundoff quantizers and triangular overflow is g.a.s. by the constructive algorithm.

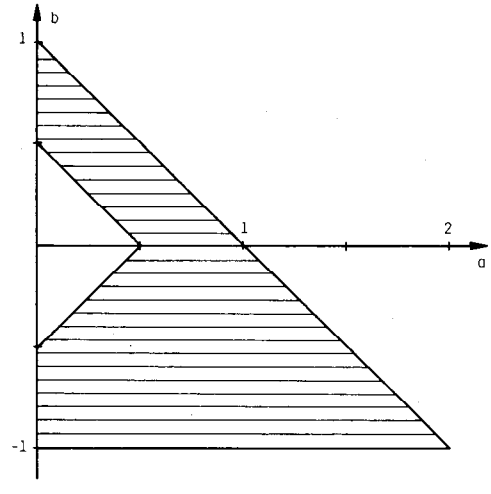


Fig. 28. Region where a direct form filter with two roundoff quantizers and two's complement overflow is g.a.s. by the constructive algorithm.

B. Coupled Form Digital Filters

For coupled form digital filters, we consider the fixed-point filter implemented with two or four quantizers. For both structures considered, we summarize existing results on the stability of the filter and then compare these results with the stability results obtained by the constructive algorithm.

1) Two Quantizers:

For the coupled form digital filter of Fig. 9, previous results indicate that this structure is free of overflow and quantization limit cycles when truncation is used in the quantizer. Barnes and Fam [1] show that the coupled form is free of limit cycles due to overflow nonlinearities. They consider autonomous nonlinear systems of the type

$$x(k+1) = f[Ax(k)]$$

where $f(\cdot)$ is a bounded nonlinear function defined on R^n. Specifically, they assume the existence of a real number $\mu > 0$, such that for every $x \in R^n$

$$|f(x)|_2 \leq \mu |x|_2$$

where $|\cdot|_2$ denotes the Euclidean vector norm on R^n. Let $\|A\|_2$ denote the matrix norm of A induced by the Euclidean norm. They show that if

$$\mu \|A\|_2 < 1 \tag{41}$$

then no autonomous limit cycles will exist in the system described by (40). The coupled form filter of Fig. 9 without the quantizers fulfills condition (41) and thus no limit cycles exist. Jackson [15] extends these results to also include the quantization nonlinearity by noting that the truncation quantization nonlinearity also fulfills the condition (41).

For roundoff quantizers, Barnes and Shinnaka [2] show that quantization limit cycles will not be supported by the coupled form for parameters located within the unit square depicted in Fig. 29. They consider the second-order linear filter of Fig. 7 described by the state equations:

$$x_1(k+1) = ax_1(k) - bx_2(k) \tag{42}$$

$$x_2(k+1) = bx_1(k) + ax_2(k). \tag{42}$$

Letting $f(\cdot)$ denote roundoff quantization, the autonomous system with two roundoff quantizers is given by

$$x_1(k+1) = f[ax_1(k) - bx_2(k)]$$

$$x_2(k+1) = f[bx_1(k) + ax_2(k)]. \tag{43}$$

Their assertion and its proof are given here because we will extend it to the case when overflow nonlinearities are present.

Assertion 1 [2]: For the system given by (43), if the point (a, b) is within the unit square of Fig. 29, then

$$|x(k+1)|_2 < |x(k)|_2.$$

Proof: We consider the construction of embedded squares in Fig. 30. If $x(k)$ is on the boundary of square 1 in Fig. 30, then $x(k+1)$ will be within or on the boundary of square 2. Furthermore, if $x(k)$ is at a midpoint of a side

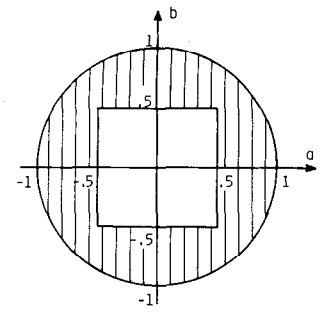


Fig. 29. Region where a coupled form filter with two or four roundoff quantizers and any overflow is free of limit cycles by Assertion 1.

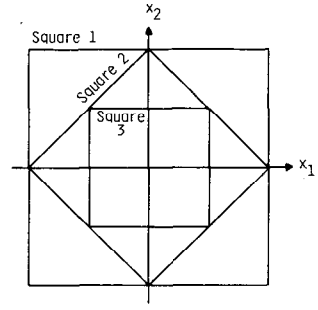


Fig. 30. Imbedded squares in state space of coupled form filter.

of square 1, then $x(k+1)$ will be within or on the boundary of square 3. Thus the desired result follows. ■

Since the norm is decreasing monotonically, they conclude that no limit cycles exist in the filter when (a, b) is within the unit square.

When overflow is considered with roundoff quantization, then the coupled form filter will not support limit cycles when the poles are within the unit square of Fig. 29. This conclusion follows immediately from Assertion 1. The proof of this assertion is the same if the operator $f(\cdot)$ in (43) represents a roundoff and overflow, since

$$|P(x)| \leq |x|$$

where $P(\cdot)$ represents any of the overflow nonlinearities in Fig. 2.

To apply the constructive algorithm to the coupled form digital filter with two quantizers, we use the extreme matrices determined in (30). When truncation quantizers are used with any type of overflow, the constructive algorithm shows that this filter is globally asymptotically stable everywhere that the linear filter is globally asymptotically stable. This result is identical to existing results. For roundoff quantizers with any type of overflow, the constructive algorithm shows that this filter is globally asymptotically stable when the parameters a and b satisfy

$$a^2 + b^2 < 0.25.$$

This region is smaller than the region in the parameter plane where Barnes and Shinnaka [2] show that no limit cycles exist. It is not too surprising that our results are more conservative in this case, since the constructive algorithm essentially shows that a filter is globally asymptotically stable for a *class* of nonlinearities whereas Barnes and Shinnaka consider some *specific* nonlinearities.

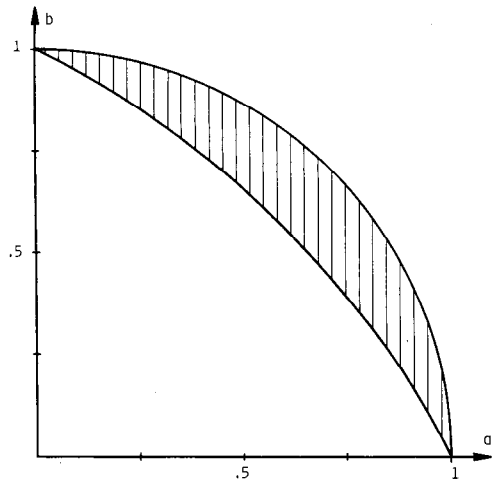


Fig. 31. Region where a coupled form filter with four truncation quantizers and no overflow is free of limit cycles by (44).

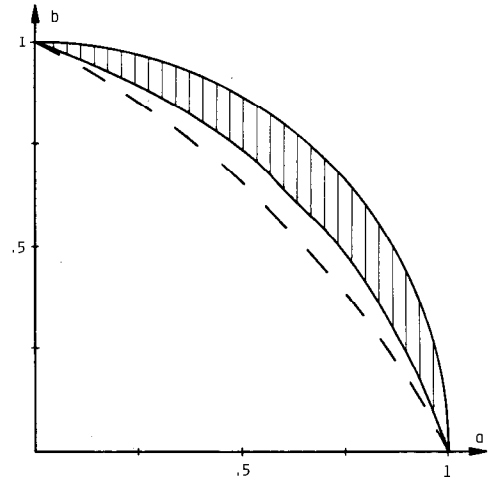


Fig. 32. Region where a coupled form filter with four truncation quantizers and saturation, zeroing or no overflow is g.a.s. by the constructive algorithm.

2) Four Quantizers:

For the coupled form digital filter with quantizers after each multiplication in Fig. 8, the only known result deals with the absence of limit cycles in this structure without overflow. When truncation quantization is used without overflow, Jackson and Judell [16] state that no limit cycles exist if the parameters of the coupled form structure of Fig. 8 satisfy

$$a^2 + |ab| + b^2 < 1. \tag{44}$$

However, no proof of their assertion is given in [16]. This region in the parameter plane where no limit cycles exist is shown as the unhatched region in Fig. 31. (The region is symmetric about both the a and b axes.) For roundoff quantizers, Barnes and Shinnaka [2] prove that limit cycles due to quantization will not exist if the parameters of the coupled form filter are within the unit square shown in Fig. 29. They consider the linear filter of Fig. 7 described by (42). Letting $f(\cdot)$ denote roundoff, the autonomous system with four roundoff quantizers is represented by

$$x_1(k+1) = f[ax_1(k)] + f[-bx_2(k)]$$
$$x_2(k+1) = f[bx_1(k)] + f[ax_2(k)].$$

If the point (a, b) is within the unit square of Fig. 29, then

$$|x(k+1)|_2 < |x(k)|_2$$

and thus no quantization limit cycles exist. The interested reader is referred to [2] for details of the proof. The extension of their proof to overflow nonlinearities does not seem obvious at this time, even though their proof could be extended in the case of the coupled form filters with two roundoff quantizers.

To apply the constructive algorithm to this filter structure, we use the extreme matrices determined in (34). The regions in the parameter plane, determined by the constructive algorithm, where the digital filter is globally asymptotically stable, are shown for all cases as the unhatched regions in Figs. 32–37. Horizontal hatching identifies a region where at least one extreme matrix has an eigenvalue with a magnitude greater than one. Vertical hatching identifies the rest of the region where we can draw no conclusion about the stability of the filter. Only the first quadrants of these regions are shown since they are symmetric about both the a and b axes.

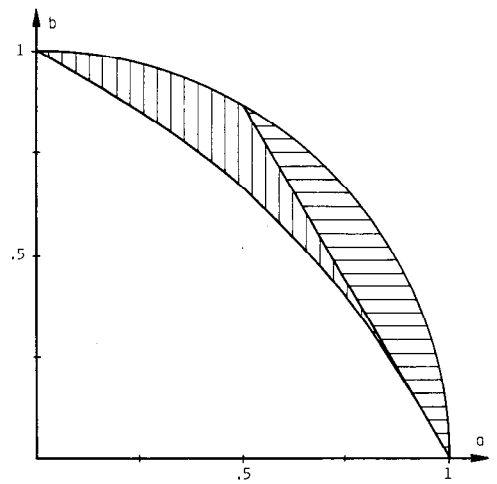


Fig. 33. Region where a coupled form filter with four truncation quantizers and triangular overflow is g.a.s. by the constructive algorithm.

As indicated in Fig. 32, the constructive algorithm determines a region where limit cycles are absent which is larger than the region where Jackson and Judell [16] indicate the absence of limit cycles for four truncation quantizers and no overflow nonlinearities. However, with four roundoff quantizers, the constructive algorithm determines a region where no limit cycles exist that is smaller than the region where Barnes and Shinnaka [2] prove the absence of limit cycles (Fig. 29). Again, this is to be expected, since our result by the constructive algorithm determines the region where the filter is globally asymptotically stable for a class of nonlinearities whereas only a specific nonlinearity (i.e., roundoff quantization) is considered in [2]. All of the results obtained by the constructive algorithm for saturation and two's complement overflow with roundoff or truncation quantization seem to be new results.

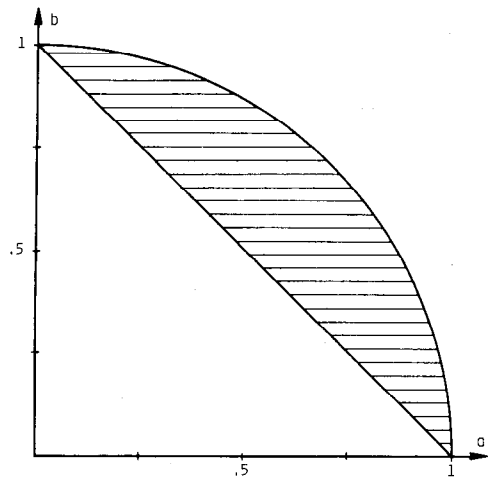


Fig. 34. Region where a coupled form filter with four truncation quantizers and two's complement overflow is g.a.s. by the constructive algorithm.

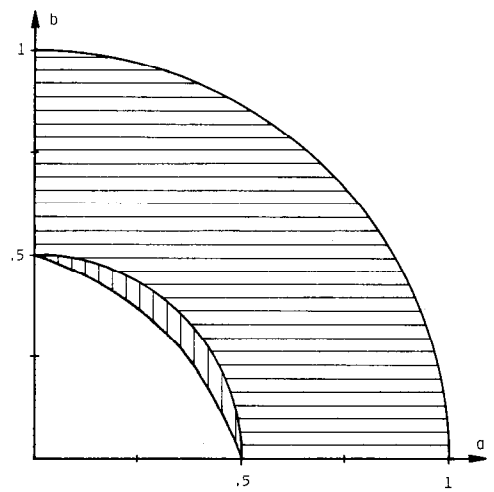


Fig. 35. Region where a coupled form filter with four roundoff quantizers and saturation, zeroing or no overflow is g.a.s. by the constructive algorithm.

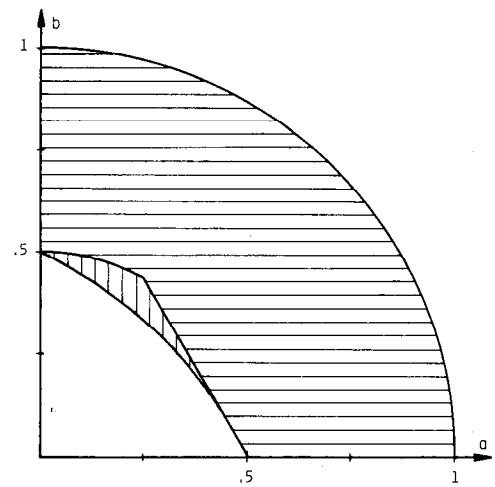


Fig. 36. Region where a coupled form filter with four roundoff quantizers and triangular overflow is g.a.s. by the constructive algorithm.

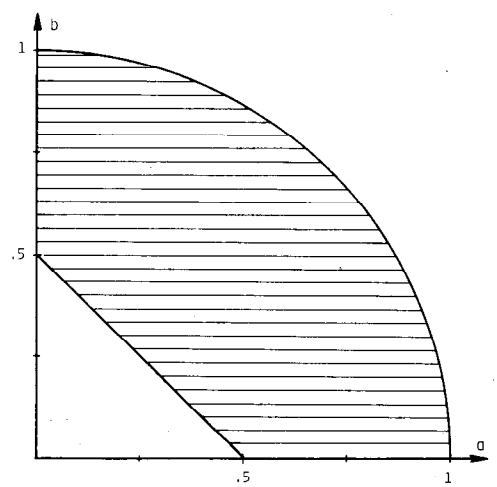


Fig. 37. Region where a coupled form filter with four roundoff quantizers and two's complement overflow is g.a.s. by the constructive algorithm.

V. Concluding Remarks

The fixed-point digital filter structures which we analyzed demonstrate that the Brayton–Tong constructive algorithm is a powerful tool in the stability analysis of fixed-point digital filters. We have obtained *new results* for many of the structures which we analyzed. We also *improved* upon many existing stability results. Our results are *more conservative* only for a few cases. In these cases, the existing results consider a *specific* nonlinearity, whereas the constructive algorithm obtains a stability result which applies to a *class* of nonlinearities.

Following is a summary of the results obtained by the constructive algorithm for the various digital filter structures which we studied. For comparison, the references for the existing results are also listed.

New stability results

1) direct form, one quantizer with overflow,
2) direct form, two quantizers with overflow,
3) coupled form, four quantizers with overflow.

Improvement upon existing results:

1) direct form, one zeroing overflow only [27],
2) direct form, one roundoff quantizer without overflow [7],
3) direct form, two quantizers without overflow [18],
4) coupled form, four truncation quantizers without overflow [16].

Same as existing results:

1) direct form, one truncation quantizer without overflow [5],
2) coupled form, two truncation quantizers with or without overflow [1].

More conservative than existing results:

1) direct form, saturation or triangular overflow only [10],
2) coupled form, two roundoff quantizers with or without overflow [2],
3) coupled form, four roundoff quantizers without overflow [2].

Whereas existing methods of stability analysis are generally different for each particular filter structure, the constructive algorithm allows us to use *one* method to study the stability of nonlinear digital filter structures. This method can be cumbersome for complicated structures, however, it is straightforward. We feel that this method should become a tool to be used in the evaluation of any proposed new filter structure.

In a companion paper [13], we consider wave digital filters and lattice filters. There are many other digital filter structures which one might want to analyze by the constructive algorithm.

We have only considered second-order digital filters in this paper. The constructive method can be applied to higher order filters either directly or by considering the higher order filter as an interconnection or lower order structures (see, e.g., [19]–[21]).

Appendix

We used two computer programs, BGRID and BORDR, in our investigation of digital filter stability by the constructive algorithm. These programs are described in detail in [12] and thus are only briefly discussed here.

The program BGRID uses the constructive algorithm to determine the region in the parameter plane where a second-order digital filter is globally asymptotically stable. BGRID determines this region by checking the filter stability at individual points in a grid of points. The program BORDR identifies points along the boundary of the region where a digital filter is globally asymptotically stable.

In a typical sequence, the program BGRID is used to obtain a general idea of the form of the region where the filter is globally asymptotically stable. Next, the program BORDR is used to determine the points along the boundary of this region.

To provide maximum flexibility, each of the above two programs consist of three parts:

a) the main program (i.e., BGRID or BORDR),

b) the subroutines which implement the constructive algorithm, and

c) the subroutines which characterize a specific filter structure.

The implementation of the constructive algorithm is discussed in detail in [3], [4] and [12] and is, therefore, not repeated here. Of the subroutines which characterize a given digital filter, the most important one is the subroutine which generates the set of extreme matrices. In addition, there are minor subroutines which pertain to the various different filter structures which we considered. Because of this program organization, the task of analyzing a given digital filter structure is relatively simple. Essentially, all that is needed is that the appropriate three parts of the programs BGRID and BORDR be linked for a given filter structure. In [12], detailed documentation of these programs is provided.

All computations were performed in double precision. As indicated before in all cases, ρ was chosen as $\rho = 1.0000001$. In the programs BGRID and BORDR, the distance between points is 0.02 units.

References

[1] C. W. Barnes and A. T. Fam, "Minimum norm recursive digital filters that are free of overflow limit cycles," *IEEE Trans. Circuits Syst.*, vol. CAS-24, pp. 569–574, 1977.

[2] C. W. Barnes and S. Shinnaka, "Stability domains for second-order recursive digital filters in normal form with 'matrix power' feedback," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 841–843, 1980.

[3] R. K. Brayton and C. H. Tong, "Stability of dynamical systems: A constructive approach," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 224–234, 1979.

[4] ——, "Constructive stability and asymptotic stability of dynamical systems," *IEEE Trans. Circuits Syst.* vol. CAS-27, pp. 1121–1130, 1980.

[5] T. A. C. M. Claasen, "Survey of stability concepts of digital filters," Tech. Rep. 108 (Telecommunication Theory), Stockholm, Sweden: The Royal Inst. of Technology, 1976.

[6] T. A. C. M. Claasen and L. O. G. Kristiansson, "Necessary and sufficient conditions for the absence of overflow phenomena in a second-order recursive digital filter," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-23, pp. 509–515, 1975.

[7] T. Claasen, W. F. G. Mecklenbräuker, and J. B. H. Peek, "Frequency domain criteria for the absence of zero-input limit cycles in nonlinear discrete-time systems, with applications to digital filters," *IEEE Trans. Circuits Syst.*, vol CAS-22, pp. 232–239, 1975.

[8] ——, "Effects of quantization and overflow in recursive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 517–529, 1976.

[9] N. Dunford and J. T. Schwartz, *Linear Operators — Part I: General Theory*. New York: Interscience, 1958.

[10] P. M. Ebert, J. E. Mazo, and M. G. Taylor, "Overflow oscillations in digital filters," *Bell System Technical J.*, vol. 48, pp. 2999–3020, 1969.

[11] B. Eckhardt and W. Winkelnkemper, "Implementation of a second order digital filter section with stable overflow behaviour," *Nachrichtentech. Z.*, vol. 26, pp. 282–284, 1973.

[12] K. T. Erickson, "Stability analysis of fixed-point digital filters using a constructive algorithm," Ph.D. dissertation, Iowa State Univ., Ames, IA, 1983.

[13] K. T. Erickson and A. N. Michel, "Stability analysis of fixed-point digital filters using computer generated Lyapunov functions—Part II: Wave digital filters and Lattice Digital Filters," *IEEE Trans. Circuits Syst.*, pp. 00–00, this issue.

[13a] A. Fettweis, "Digital circuits and systems," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 31–48, 1984.

[14] G. F. Franklin and J. D. Powell, *Digital Control of Dynamical Systems*. Reading, MA: Addison-Wesley, 1980.

[15] L. B. Jackson, "Limit cycles in state-space structures for digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 67–68, 1979.

[16] L. B. Jackson and N. K. H. Judell, "Addendum to 'Limit cycles in state-space structures for digital filters'," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 320, 1980.

[17] E. I. Jury, "A simplified stability criterion for linear discrete systems," *Proc. IRE*, vol. 50, pp. 1493–1500, 1962.

[18] E. I. Jury and B. W. Lee, "The absolute stability of systems with many nonlinearities," *Automation Remote Contr.*, vol 26, pp. 943–961, 1965.

[19] A. N. Michel and R. K. Miller, *Qualitative Analysis of Large Scale Dynamical Systems*. New York: Academic, 1977.

[20] A. N. Michel, R. K. Miller, and B. H. Nam, "Stability Analysis of Interconnected systems using computer generated Lyapunov functions," *IEEE Trans. Circuits Syst.*, vol. CAS-29, pp. 431–440, 1982.

[21] A. N. Michel, B. H. Nam, and V. Vittal, "Computer generated Lyapunov functions for interconnected systems: Improved results with applications to power systems." *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 189–198, 1984.

[22] A. N. Michel, N. R. Sarabudla, and R. K. Miller, "Stability analysis of complex dynamical systems: Some computational methods," *Circuits, Systems and Signal Processing*, vol. 1, pp. 171–202, 1982.

[23] R. K. Miller and A. N. Michel, *Ordinary Differential Equations*. New York: Academic, 1982.

[23a] D. Mitra and V. B. Lawrence, "Controlled rounding arithmetics, for second-order direct-form digital filters, That Eliminate All Self-Sustained oscillations," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 894–905, 1981.

[24] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

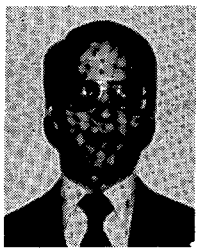
[24a] S. R. Parker and S. F. Hess, "Limit-cycle oscillations in digital filters," *IEEE Trans. Circuit Theory*, vol. CT-16, pp. 302–311, 1969.

[25] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[26] C. M. Rader and B. Gold, "Effects of parameter quantization on the poles of a digital filter," *Proc. IEEE*, vol. 55, pp 688–689, 1967.

[27] I. W. Sandberg, "A theorem concerning limit cycles in digital filters," in *Proc. Seventh Allerton Conf. Circuit and System Theory*, Univ. Illinois, Urbana, IL, pp. 63–68, Oct. 1969.
[28] A. N. Willson, "Limit cycles due to adder overflow in digital filters," *IEEE Trans. Circuit Theory*, vol. CT-19. pp. 342–346, 1972.

✤

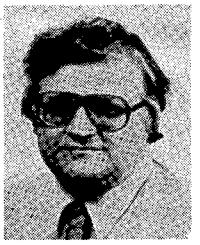


Kelvin T. Erickson (S'77–M'84) was born in Ridgeway, PA, on August 9, 1957. He received the B.S. and M.S. degrees in electrical engineering from the University of Missouri, Rolla, in 1978 and 1979, respectively, and the Ph.D. degree in electrical engineering from Iowa State University, Ames, in 1983. He has held a National Science Foundation Graduate Fellowship.

Since 1979, he has bee employed by Fisher Controls International, Inc., Marshalltown, IA. His present research interests are in the areas of nonlinear digital control systems and distributed control systems.

Dr. Erickson is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

✤



Anthony N. Michel (S'55–M'59–SM'79–F'82) was born in Rekasch, Romania, on November 17, 1935. He received the B.S.E.E., the M.S. degree in mathematics and the Ph.D. degree in electrical engineering from Marquette University, Milwaukee, WI, and the D.Sc. degree in applied mathematics from the Technical University of Graz, Austria.

He has seven years of industrial experience and has held positions with Stearns Magnetic Products, Milwaukee, WI, the U.S. Army Corps of Engineers, and A. C. Electronics, a division of G. M., Oak Creek, WI. In 1968, he joined the faculty of Iowa State University, Ames, where he was a professor in the Department of Electrical Engineering until 1984. Currently, he is a Professor and Chairman of the Department of Electrical Engineering at the University of Notre Dame, Notre Dame, IN. He is co-author of the books *Qualitative Analysis of Large Scale Dynamical Systems* (with R. K. Miller) (New York: Academic Press, 1977), *Mathematical Foundations in Engineering and Science: Algebra and Analysis* (with C. J. Herget) (Englewood Cliffs, NJ: Prentice-Hall, 1981), and *Ordinary Differential Equations* (with R. K. Miller) (New York: Academic Press, 1982). His recent research is in the areas of nonlinear systems and large-scale dynamical systems. He received the Best Transactions Paper Award from the IEEE Control Systems Society in 1978 (with R. D. Rasmussen) and the 1984 Guillemin–Cauer Award from the IEEE Circuits and Systems Society (with R. K. Miller and B. H. Nam) for the best paper published in an IEEE Circuits and Systems Society Transactions during the previous two calendar years. He has received an IEEE Centennial Medal. He is a past Associate Editor of the *IEEE Transactions on Circuits and Systems* (1977–1979), a past Associate Editor of the *IEEE Transactions on Automatic Control* (1981), the past Editor of the *IEEE Transactions on Circuits and Systems* (1981–1983), an ADCOM member of the IEEE Circuits and Systems Society, and an IEEE *Ad Hoc* Visitor for ABET (formerly ECPD). He was Co-chairman (with H. W. Hale) of the Organizing Committee of the 1978 Midwest Symposium on Circuits and Systems, he was the Program Chairman of the 1982 IEEE International Large Scale Systems Symposium, and he is the Program Chairman of the 1985 IEEE Conference on Décision and Control.

Dr. Michel is a member of Pi Mu Epsilon, Eta Kappa Nu, Phi Kappa Phi, and Sigma Xi, and a Registered Professional Engineer (in the State of Wisconsin).

Stability Analysis of Fixed-Point Digital Filters Using Computer Generated Lyapunov Functions—Part II: Wave Digital Filters and Lattice Digital Filters

KELVIN T. ERICKSON, MEMBER, IEEE, AND ANTHONY N. MICHEL, FELLOW, IEEE

Abstract — **In a companion paper [4], we utilize the *constructive stability algorithm* of Brayton and Tong in the stability analysis of fixed-point digital filters which are in the direct form and in the coupled form. We continue this work in the present paper by considering wave digital filters and lattice digital filters. We believe that the results of the present paper and its companion paper demonstrate that the *constructive algorithm* constitutes an *effective* and *general* approach in the qualitative analysis of fixed-pointed digital filters.**

Manuscript received August 19, 1983. This work was supported in part by the National Science Foundation under Grant ECS-8100690 and by the Engineering Research Institute, Iowa State University, Ames, IA 50011.

K. T. Erickson is with Fisher Controls International, Inc., Marshalltown, IA 50158.

A. N. Michel was with Iowa State University, Ames, IA 50011. He is now with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556.

I. INTRODUCTION

IN THE companion paper [4], we first showed how the constructive stability algorithm of Brayton and Tong [2], [3] may be applied in the stability analysis of rather broad classes of fixed-point digital filters which may be endowed with various types of quantization and overflow nonlinearities. We then considered, in particular, direct form digital filters and coupled form digital filters. Our objective was to determine a region in the parameter plane of a given digital filter for which the zero-input digital filter is globally asymptotically stable, and consequently, does not possess any zero-input limit cycles. The results in [4], which use only *one* approach of stability analysis, seem rather encouraging when compared to many of the existing

0098-4094/85/0200-0132$01.00