A Condition for the Overflow Stability of Second-Order Digital Filters That is Satisfied by All Scaled State-Space Structures Using Saturation

JOHN H. F. RITZERFELD

Abstract —**A set of conditions is derived that ensures overflow stability of second-order digital filters for different classes of overflow arithmetics, involving only the elements of the state-transition matrix. The well-known arithmetic saturation, zeroing, and two's-complement lead to different stability conditions, the condition for saturation being the least restrictive. As a result, all properly scaled second-order state-space structures are zero-input overflow stable if saturation is used for overflow correction. Furthermore, conditions are derived for stable second-order digital filters in a nonzero input situation by introducing a weaker form of stability of the forced response. The presented analysis is based on determining the set of Lyapunov functions for a general second-order state-transition matrix given a certain overflow arithmetic.**

I. Introduction

IN RECENT YEARS digital filter design has been based to an increasing degree on state-space structures. Many design problems have a direct translation in terms of a description in the state space and as such are easily implemented and solved on a computer. For example, the problem of scaling is reduced to solving a set of linear equations, while the nonlinear effects due to quantization and overflow can be handled in a systematic way without any trade-offs. Optimal state-space structures can be computed which achieve the lowest possible quantization noise and yet are overflow stable for any overflow arithmetic as well as free from all zero-input limit cycles when magnitude truncation is used as a quantization characteristic [1].

Whereas all digital filters have a state-space description, not all are state-space structures. The latter are distinguished by the fact that their implementation is a direct translation of the state equation. The elements of the state-transition matrix appear as multipliers, while each state variable is represented by the output signal of a unit-delay element. Signal quantization and overflow correction are applied in principle at the double-precision summation node preceding each delay.

As for an analysis of overflow stability, an important class of state-space structures that was found to be stable for a two's-complement overflow arithmetic is constituted by "minimum-norm" filters [2]. These are characterized by the property that the state-transition matrix has a norm less than unity, denoted $\|A\|<1$, which simply means that the Euclidian norm of the state vector decreases with every state transition, or that the matrix $I-A^tA$ is positive definite (where I stands for the identity matrix and $(\cdot)^t$ denotes transposition). The logical extension was to require positivity for the matrix $D-A^tDA$, where D is any positive diagonal matrix. For second-order digital filters, this led to a condition for overflow stability involving only the elements of the matrix A [3].

In this paper this condition is relaxed by replacing the diagonal matrix D with an arbitrary positive definite matrix P. It will be demonstrated that for saturation arithmetic a relaxed condition for A is found that is satisfied by all scaled state-space structures of second order. From the requirement that the matrix $P-A^tPA$ be positive definite, a set of matrices P is derived that provides candidates for Lyapunov functions $\underline{x}^tP\underline{x}$ of the two state variables, given a state-transition matrix A. The parameters which characterize the matrix P are required to be chosen on the face of an ellipse in a parameter plane. The specific overflow arithmetic also constrains these parameters to a varying degree, since Lyapunov theory demands that the nonlinearity be energy reducing. The most constraining arithmetic in this respect is two's-complement (modulo-2) arithmetic; the least constraining is saturation arithmetic. These constraints will be indicated for various classes of overflow characteristics, yielding a set of allowable Lyapunov functions which is a subset of the above. The requirement that this set be nonempty provides a condition for overflow stability involving only the elements of the state-transition matrix for each class of overflow characteristics.

Initially only zero-input overflow stability is studied. This means that a digital filter "when left to itself" cannot sustain an overflow oscillation of any period, irrespective of the initial state of the filter. A conclusion can be drawn, however, with respect to the "forced response" stability [4] of a digital filter with a certain overflow arithmetic in

Manuscript received January 18, 1988; revised January 27, 1989. This paper was recommended by Associate Editor Y.-C. Jenq.

The author is with the Department of Electrical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands.

IEEE Log Number 89282631.

relation to an otherwise identical filter with another overflow arithmetic. This conclusion will be stated explicitly for the various classes of overflow characteristics. A last preliminary is to be mentioned, one that is commonly adopted in most papers on overflow stability; viz., signal quantization is neglected in the presented analysis. In nearly all practical cases the nonlinear effects of overflow and quantization may be treated as mutually independent.

II. Outline

A well-established condition for the zero-input overflow stability of second-order state-space filters states that the elements of the state-transition matrix $A=[a_{ij}]$ are required to satisfy [3]

$$a_{12}\cdot a_{21} \geqslant 0 \tag{1}$$

or

$$|a_{11}-a_{22}| < 1-\det(A), \qquad \text{if } a_{12}a_{21} < 0. \tag{2}$$

This constraint is valid under the tacit assumption of linear stability, for which we demand

$$|\operatorname{tr}(A)| < 1+\det(A) \tag{3}$$

$$\det(A) < 1 \tag{4}$$

i.e., the well-known stability triangle in a plane with axes $\operatorname{tr}(A)$ (trace of A) and $\det(A)$ (determinant of A). When we write (2) in the alternative form

$$\operatorname{tr}^2(A) < (1+\det(A))^2 + 4\cdot a_{12}\cdot a_{21}$$

we see that this condition is automatically met if (1) and (3) are satisfied. As a result, we need not distinguish between two cases for the product $a_{12}a_{21}$; within the stability triangle, stable overflow behavior is ensured by the condition $|a_{11}-a_{22}| < 1-\det(A)$ alone. Moreover, we note that the case of a nonnegative product $a_{12}a_{21}$ is technically less important, because it precludes the possibility of complex-conjugate poles, in which case we have

$$\operatorname{tr}^2(A) < 4\cdot\det(A) \quad \text{or} \quad (a_{11}-a_{22})^2 < -4\cdot a_{12}\cdot a_{21}. \tag{5}$$

If inequality (2) is satisfied, overflow stability is ensured for all overflow characteristics (saturation, zeroing, modulo-2, slope-inversion [5]). We may expect the constraint on the state-transition matrix to be less restrictive when we require overflow stability for a saturation characteristic only. The second-order direct-form filter, for example, is known to exhibit overflow oscillations for certain pole locations when a modulo-2 overflow nonlinearity is used, whereas it is unconditionally stable using saturation [6]. In Section IV we will prove the following theorem, which provides a condition for overflow stability with saturation.

Theorem 1: Second-order state-space filters using saturation for overflow correction are zero-input overflow stable if the elements of the state matrix $A=[a_{ij}]$ are restricted by the condition

$$|a_{11}-a_{22}| \leqslant 2\cdot\min(|a_{12}|,|a_{21}|)+1-\det(A). \tag{6}$$

As it should be, (6) is satisfied for both direct-form realizations, $A=\begin{bmatrix} a & 1 \\ b & 0 \end{bmatrix}$ (DF1) and $A=\begin{bmatrix} a & b \\ 1 & 0 \end{bmatrix}$ (DF2), for any value of the multipliers $a=\operatorname{tr}(A)$ and $b=-\det(A)$ within the stability triangle. As with the original condition for overflow stability given by (1) and (2), condition (6) may be accompanied with an optional condition for the product $a_{12}a_{21}$. To be specific, overflow stability with saturation is achieved with the condition

$$a_{12}a_{21} \geqslant -m(m+1-\det(A)) \tag{7}$$

or

$$|a_{11}-a_{22}| \leqslant 2m+1-\det(A)$$
$$\text{if} \quad a_{12}a_{21} < -m(m+1-\det(A))$$

where

$$m=\min(|a_{12}|,|a_{21}|).$$

The distinction between two cases for $a_{12}a_{21}$ is optional, because if (7) holds then (6) is implied by the premise (3) of linear stability.

We note that the result of this test for overflow stability, as opposed to the original one, is affected by the proper scaling of the filter, since this will change a_{12} and a_{21} (without changing their product). The following surprising result will be shown to hold (Section V).

Theorem 2: All properly scaled second-order state-space filters using saturation for overflow correction are zero-input overflow stable.

As it turns out, Theorem 1 may be stated in more general terms to allow for alternative overflow arithmetics: a second-order state-space filter is zero-input overflow stable if[1]

$$|a_{11}-a_{22}| \leqslant (1+s)m+1-\det(A) \tag{8}$$

where $s=1$ for saturation and slope-inversion arithmetic, $s=0$ for zeroing, and $s=-1$ for modulo-2 arithmetic. In Section VI the set of classes O^s $(-1\leqslant s\leqslant 1)$ of overflow characteristics is introduced, such that, e.g., O^1 contains saturation arithmetic. The validity of the generalized condition (8) will then be demonstrated. Also, a digital filter that is zero-input overflow stable with respect to O^{s_1} will be shown to be forced response stable with respect to O^{s_2} for a class of input signals whose amplitudes do not exceed $(s_2-s_1)/(3+s_1)$.

III. Stability Analysis for Saturation

In order to decide if a second-order digital filter is zero-input overflow stable, we seek some *norm* $\|\underline{x}\|$ of the state $\underline{x}=(x_1,x_2)^t$ which is

- nonincreasing with respect to the linear state transition,
- decreasing with respect to the proposed overflow correction.

[1] Under the condition

$$a_{12}\cdot a_{21} < -m\cdot\left[m+\frac{1+s}{2}(1-\det(A))\right].$$

Together with the fact that a norm of $\underline{x}$ is a nonnegative function which is zero only if $\underline{x}$ is zero, these two requirements imply that the state will become asymptotically zero in a zero-input situation starting from any initial state inside the unit square $|x_i| \leqslant 1$, $i = 1,2$. Incidentally, we have assumed the overflow level to be unity.

We define

$$\|\underline{x}\| = \left(\underline{x}'P\underline{x}\right)^{(1/2)} \tag{9}$$

where

$$P = \begin{bmatrix} \alpha & \gamma \\ \gamma & \beta \end{bmatrix}$$

is a positive definite matrix, so

$$\alpha > 0 \qquad \beta > 0 \tag{10}$$

and

$$\det(P) > 0 \qquad \text{or} \qquad \alpha \cdot \beta > \gamma^2. \tag{11}$$

First we demand $\|A\underline{x}\| \leqslant \|\underline{x}\|$ or $\underline{x}'(A'PA)\underline{x} \leqslant \underline{x}'P\underline{x}$. This implies that the matrix $P - A'PA$ is positive (semi-) definite, which will be the case if

$$\operatorname{tr}(P - A'PA) > 0 \tag{12}$$

and

$$\det(P - A'PA) \geqslant 0. \tag{13}$$

Second, we demand $\|f(\underline{x})\| < \|\underline{x}\|$ for any $\underline{x}$ outside the unit square, where $f(\cdot)$ stands for the overflow arithmetic, which for saturation reads

$$f(x_i) = \begin{cases} -1, & \text{if } x_i < -1 \\ x_i, & \text{if } |x_i| \leqslant 1 \\ 1, & \text{if } x_i > 1 \end{cases} \qquad (i = 1,2). \tag{14}$$

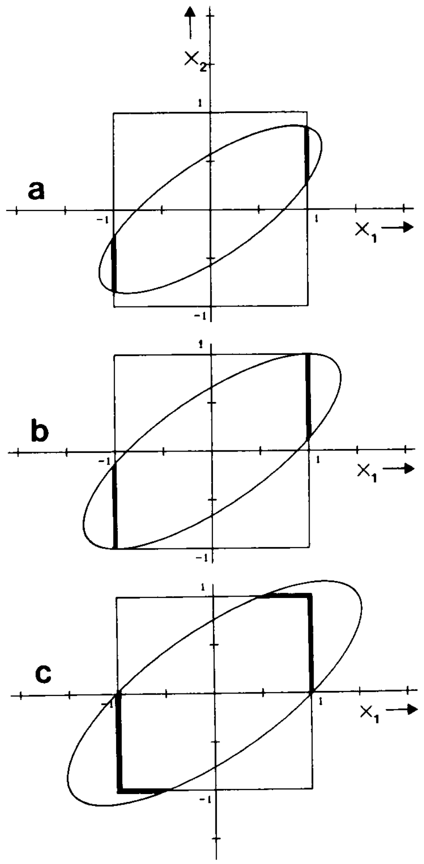

Fig. 1. Borderline case for the geometry of the set $\underline{x}'P\underline{x} = \text{constant}$ if saturation is used for overflow correction.

In the following, we take a positive definite matrix P for which (12) and (13) hold and a constant K which may assume any positive real value. The set of curves $\|\underline{x}\|^2 = \underline{x}'P\underline{x} = K$ represents "concentric" ellipses in the x_1, x_2 plane, which can be interpreted as curves of constant energy, since the quadratic form $\underline{x}'P\underline{x}$ is nonnegative. With our choice of P this energy cannot increase with the linear state transition, so successive state points are located on ellipses with nonincreasing value of the constant K, as long as no overflow occurs. In the case of overflow, some point outside the unit square will be mapped to a point inside the unit square (or on its perimeter) by the overflow correction. For overflow stability we demand that the overflow nonlinearity reduce energy, or equivalently, be norm-decreasing. This will be the case if, for any K, the two arcs of the ellipse $\underline{x}'P\underline{x} = K$ which are outside the unit square are mapped into the ellipse's interior by the overflow characteristic $f(\cdot)$. Now suppose, to guide our thoughts, that the set $\underline{x}'P\underline{x} = K$ is oriented such that, with an increasing value of K, $|x_1|$ is first to exceed unity. Then, for saturation, the set of ellipses shows the desired property if the points of extreme x_2 value (where the derivative dx_2/dx_1 is zero) lie inside the unit square as long as $\max(|x_2|) < 1$ and on its perimeter for $\max(|x_2|) = 1$.[2] More specifically, the x_1 coordinate corresponding to $\max(|x_2|) = 1$ should satisfy $|x_1| \leqslant 1$. Fig. 1 depicts the case in which this condition is met marginally, i.e., $|x_1| = 1$ when $\max(|x_2|) = 1$. As a function of K, the points of extreme x_2 value lie on a straight line through the origin, given by

$$\alpha x_1 + \gamma x_2 = 0. \tag{15}$$

Likewise, $|x_1|$ has its maximum on the line

$$\gamma x_1 + \beta x_2 = 0. \tag{16}$$

So finally, when we distinguish between the cases "$\alpha < \beta$" (i.e., x_1 is first to "overflow") and "$\alpha > \beta$" (i.e., x_2 is first to "overflow"), we arrive at the following condition for overflow stability with saturation:

$$|\gamma| \leqslant \min(\alpha, \beta). \tag{17}$$

Note that if $\alpha = \beta$ then the main axis of the ellipses has an inclination of 45° for $\gamma < 0$ and of 135° for $\gamma > 0$. In that case (17) is satisfied automatically, since we have $\gamma^2 < \alpha\beta$ on account of (11).

[2] Of course this condition must be formulated with respect to x_1 if the geometry of the set $\underline{x}'P\underline{x} = \text{constant}$ is such that x_2 is first to "overflow."

Summarizing, given a matrix A which satisfies (3) and (4), we seek a matrix P which satisfies (11)–(13) and (17), or more accurately, we want to determine the set of matrices A for which such a matrix P exists. The next section presents the proof that for all matrices A that satisfy constraint (6) or (7), at least one Lyapunov function $\underline{x}^tP\underline{x}$ can be found.

Note that when we let

$$\gamma = 0 \tag{18}$$

the set $\underline{x}^tP\underline{x} = \text{constant}$ has a horizontal or vertical orientation. It is easily recognized that any overflow correction will be norm-decreasing. Indeed, as we will see, (18) leads to the original condition (2) for stability.

IV. The Set of Lyapunov Functions and Proof of Theorem 1

Lemma 1: Let A be a 2×2 matrix which satisfies conditions (3) and (4) for linear stability. Let P be positive definite; then (12) is implied by (13). The proof of this lemma is given in Appendix I.

In the following we assume that the product $a_{12}a_{21}$ is negative.[3] With that premise we will further examine condition (13), $\det(P - A^tPA) \geqslant 0$, which is equivalent to

$$\alpha\beta(a_{11}-a_{22})^2 - 2\gamma(\alpha a_{12}-\beta a_{21})(a_{11}-a_{22}) - 4\gamma^2 a_{12}a_{21} + (\alpha a_{12}+\beta a_{21})^2 \leqslant \det(P)(1-\det(A))^2. \tag{19}$$

Note that by inserting $\gamma = 0$ and letting $\det(P) = \alpha\beta = 1$,[4] we see that (19) yields

$$(a_{11}-a_{22})^2 + (\alpha a_{12}+\beta a_{21})^2 \leqslant (1-\det(A))^2 \tag{20}$$

which is equivalent to (2) with the optimal choice of the matrix P as

$$P = \begin{bmatrix} \sqrt{(-a_{21}/a_{12})} & 0 \\ 0 & \sqrt{(-a_{12}/a_{21})} \end{bmatrix}. \tag{21}$$

We are free to choose $\alpha = 1$ in (19). Furthermore, we take the sign of γ as

$$\operatorname{sign}(\gamma) = \operatorname{sign}[a_{12}(a_{11}-a_{22})] = \operatorname{sign}[-a_{21}(a_{11}-a_{22})]. \tag{22}$$

This choice of $\operatorname{sign}(\gamma)$ is dictated by the following argument. The left-hand side of inequality (19) is a concave quadratic function of the absolute difference $|a_{11}-a_{22}|$, for which we want to find an upper bound for overflow stability. This bound will be largest if the second term on the left-hand side of (19) (the linear term) is strictly negative, which will be the case with the choice of $\operatorname{sign}(\gamma)$ as in (22). Finally, with the definitions $\xi = \beta$, $\eta = |\gamma|$, $m = \min(|a_{12}|, |a_{21}|)$, and $M = \max(|a_{12}|, |a_{21}|)$, we can rewrite (19) as

$$\xi|a_{11}-a_{22}|^2 - 2\eta(M+\xi m)|a_{11}-a_{22}| + 4\eta^2 mM + (M-\xi m)^2 \leqslant (\xi-\eta^2)\cdot(1-\det(A))^2 \tag{23}$$

where we have let $|a_{12}| > |a_{21}|$.[5] This inequality represents an ellipse (including its interior) in the ξ, η plane centered on the point (ξ_0, η_0), where

$$\xi_0 = \frac{(1-\det(A))^2 + 2mM}{2m^2} \quad \text{and} \quad \eta_0 = \frac{|a_{11}-a_{22}|}{2m}. \tag{24}$$

This fact may be verified if we write (23) in shorthand as $\underline{v}^tQ\underline{v} \leqslant \kappa$, where $\underline{v} = (\xi-\xi_0, \eta-\eta_0)^t$, $\kappa = \frac{(1-\det(A))^2}{4m^2}\cdot[(1+\det(A))^2 - \operatorname{tr}^2(A)] > 0$, and Q is a positive definite matrix,[6] given by

$$Q = \begin{bmatrix} m^2 & -m|a_{11}-a_{22}| \\ -m|a_{11}-a_{22}| & (1-\det(A))^2 + 4mM \end{bmatrix}.$$

We conclude that the matrix P satisfies (11), (12), and (13) if (ξ, η) is a point on the face of an ellipse, described by (23), with $\xi > \eta^2$. If we want P to meet (17) as well, we demand $\eta \leqslant \min(\xi, 1)$.

Lemma 2: Condition (23), together with $\xi > \eta^2$ and $\eta \leqslant \min(\xi, 1)$, is met for at least one pair (ξ, η) if the matrix A is constrained by (6) or (7). The proof of this lemma is given in Appendix II.

Having completed the proof of Theorem 1, let us study inequality (23) more closely. If the two parameters ξ and η, which determine the symmetric 2×2 matrix P, satisfy (23) with inequality, then the square norm $\underline{x}^tP\underline{x}$ is a Lyapunov function of the state variables if, in addition, the overflow nonlinearity causes this quadratic form to decrease. Incidentally, positivity of the matrix P demands that $\xi > \eta^2$. Hence, the set of potential Lyapunov functions is given by the section of an ellipse in the ξ, η plane that is to the right of the parabola $\xi = \eta^2$. The area of this ellipse shrinks with decreasing "distance from the stability triangle" of the parameters determining linear stability, i.e., $\operatorname{tr}(A)$ and $\det(A)$. For $\det(A) \to 1$ the ellipse shrinks to a single point given by (cf. (24))

$$(\hat{\xi}, \hat{\eta}) = (\xi_0, \eta_0)\big|_{\det(A)\to 1} = \left[\frac{M}{m}, \frac{|a_{11}-a_{22}|}{2m}\right]. \tag{25}$$

For $|\operatorname{tr}(A)| \to 1 + \det(A)$ the ellipse degenerates into a line described by

$$m\cdot\xi + M = |a_{11}-a_{22}|\cdot\eta. \tag{26}$$

Note also that all points on the face of the ellipse satisfy $\xi > \eta^2$ in the case of complex conjugate poles (cf. (5)); the ellipse and the parabola touch at (η_0^2, η_0) for coinciding real poles.

When we let $\underline{x}^t\hat{P}\underline{x}$ denote the Lyapunov function corresponding to $(\hat{\xi}, \hat{\eta})$, then the following unique property

[3] This assumption poses no real restriction, since overflow stability has already been established for $a_{12}a_{21} \geqslant 0$ (cf. (1)).

[4] The matrix P may always be multiplied by some positive scaling factor.

[5] In case $|a_{12}| < |a_{21}|$, we let $\xi = \alpha$ and $\beta = 1$ to arrive at (23).

[6] $\operatorname{tr}(Q) > 0$ and $\det(Q) = m^2[(1+\det(A))^2 - \operatorname{tr}^2(A)] > 0$ on account of (3).

holds, irrespective of the condition $\det(A) \to 1$:

$$\underline{x}'A'\hat{P}A\underline{x} = \det(A) \cdot \underline{x}'\hat{P}\underline{x}. \tag{27}$$

Note that $(\hat{\xi}, \hat{\eta})$ satisfies both (23) with inequality and $\hat{\xi} > \hat{\eta}^2$, as long as condition (5) for complex poles is met.[7] The interpretation of (27) is appealing: in a zero-input situation without overflow, the state $\underline{x}(k)$ moves on the spiral $\underline{x}'\hat{P}\underline{x} = \det^k(A) \cdot \underline{x}_0'\hat{P}\underline{x}_0$, where $\underline{x}_0$ is some initial state at time instant $k = 0$. The state would remain on the ellipse $\underline{x}'\hat{P}\underline{x} = \underline{x}_0'\hat{P}\underline{x}_0$ if $\det(A) = 1$ held.

V. The Effect of Scaling on Overflow Stability

In order to scale a second-order state-space filter, we need to determine the ℓ_2 norms of the impulse responses $f_i(k)$ $(i = 1, 2)$ from filter input to both state variables. The squares of these norms are on the main diagonal of the matrix K, defined by [1]

$$K = AKA' + \underline{b}\underline{b}' \tag{28}$$

where $\underline{b} = (b_1, b_2)'$ describes how the input signal $u(k)$ is linked to the state variables by the state equation

$$\underline{x}(k+1) = A\underline{x}(k) + \underline{b}u(k). \tag{29}$$

When we define the properly scaled filter as the filter for which the elements on the main diagonal of K are equal,[8] we can prove Theorem 2, which for convenience is repeated.

Theorem 2: All properly scaled second-order state-space filters using saturation for overflow correction are zero-input overflow stable.

Proof: The state-transition matrix A' of the scaled filter is related to the matrix A of the unscaled filter as [1]

$$A' = \begin{bmatrix} a_{11} & a_{12}\sqrt{(K_{22}/K_{11})} \\ a_{21}\sqrt{(K_{11}/K_{22})} & a_{22} \end{bmatrix} \tag{30}$$

where K_{11} and K_{22} are the elements on the main diagonal of the matrix K, to be calculated using (28). We have

$$K_{11} = (1 - \det(A))^{-1}\left[(1 + \det(A))^2 - \operatorname{tr}^2(A)\right]^{-1} \cdot \det \begin{bmatrix} b_1^2 & -a_{12}^2 & -2a_{11}a_{12} \\ b_2^2 & 1 - a_{22}^2 & -2a_{21}a_{22} \\ b_1 b_2 & -a_{12}a_{22} & 1 - a_{11}a_{22} - a_{12}a_{21} \end{bmatrix} \tag{31}$$

$$K_{22} = (1 - \det(A))^{-1}\left[(1 + \det(A))^2 - \operatorname{tr}^2(A)\right]^{-1} \cdot \det \begin{bmatrix} 1 - a_{11}^2 & b_1^2 & -2a_{11}a_{12} \\ -a_{21}^2 & b_2^2 & -2a_{21}a_{22} \\ -a_{11}a_{21} & b_1 b_2 & 1 - a_{11}a_{22} - a_{12}a_{21} \end{bmatrix}. \tag{32}$$

If A' satisfies condition (7), then the scaled filter is zero-input overflow stable with saturation for overflow correction. This is the case if[9]

$$(1 - \det(A))^2 \geqslant \left[a_{12}\sqrt{(K_{22}/K_{11})} + a_{21}\sqrt{(K_{11}/K_{22})}\right]^2. \tag{33}$$

A moderate amount of algebraic manipulation is needed to show that

$$K_{11}K_{22}(1 - \det(A))^2 - (a_{12}K_{22} + a_{21}K_{11})^2 = \frac{\left[a_{21}a_{22}b_1^2 + a_{11}a_{12}b_2^2 + (1 - a_{11}a_{22} - a_{12}a_{21})b_1 b_2\right]^2}{(1 + \det(A))^2 - \operatorname{tr}^2(A)}. \tag{34}$$

This expression is nonnegative on account of (3).

Q.E.D.

VI. Generalization

Definition: The class O^s of overflow characteristics is given by those characteristics $f(\cdot)$ that are bounded, so $|f(x)| \leqslant 1$, and that satisfy

$$\begin{aligned} f(x) &= x, && \text{if } |x| \leqslant 1 \\ f(x) &\geqslant -x + (1+s), && \text{if } x > 1 \\ f(x) &\leqslant -x - (1+s), && \text{if } x < -1 \end{aligned} \tag{35}$$

where s is a fixed parameter in the range $-1 \leqslant s \leqslant 1$.

Hence, O^s contains all overflow characteristics that do not leave the hatched region in Fig. 2. Note that O^1 includes slope-inversion arithmetic and saturation, O^0 includes zeroing, and O^{-1} features all characteristics for which $|f(x)| \leqslant 1(|x| > 1)$, including two's-complement.

What would happen if we were to use the overflow characteristic that is on the edge of the class O^1, i.e., $f(x) = -x + 2 \cdot \operatorname{sgn}(x)$ for $1 < |x| \leqslant 3$, instead of saturation? The answer to this question can be given simply by taking another look at Fig. 1. Instead of being mapped onto the perimeter of the unit square, the arcs of the ellipses that are outside the unit square are mirrored with respect to its perimeter. These mirrored parts are on the face of the corresponding ellipse (as can be checked easily) for any matrix P satisfying (17). As a result, the conditions for overflow stability formulated in Theorem 1 are valid for the whole class O^1, which contains saturation only as a special case. Also, Theorem 2 may be restated as follows.

Theorem 2': All properly scaled second-order state-space filters are zero-input overflow stable with respect to the class O^1 of overflow characteristics.

Finally, we may generalize Theorem 1 even further for a general class O^s.

Theorem 1': A sufficient condition for the zero-input overflow stability of second-order state-space filters with respect to the class O^s of overflow characteristics is given

[7] This implies strict positivity of the matrices $\hat{P}$ and $\hat{P} - A'\hat{P}A$.

[8] $K_{11} = K_{22} = \delta^{-2}$. A large value of δ means that scaling is conservative [1].

[9] Again, we consider only the nontrivial case $a_{12}a_{21} < 0$, so (7) is equivalent to $1 - \det(A) \geqslant M' - m'$.

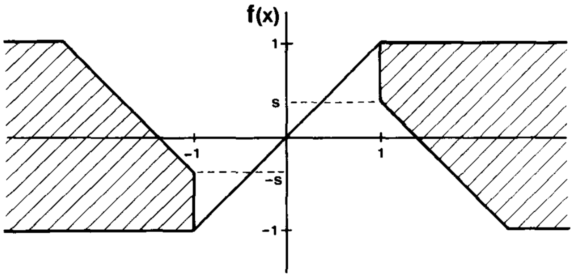

Fig. 2. Region in which overflow characteristics must be located in order to belong to the class O^s.

by

$$t(1-\det(A)) \geqslant \left[(M-m)^2+4mM(1-t^2)\right]^{1/2} \tag{36}$$

or, if

$$M-m \leqslant t(1-\det(A)) < \left[(M-m)^2+4mM(1-t^2)\right]^{1/2} \tag{i}$$

by

$$|a_{11}-a_{22}| \leqslant t\cdot(M+m) + \left[(1-t^2)\cdot\left((1-\det(A))^2-(M-m)^2\right)\right]^{1/2} \tag{37}$$

or, if

$$t(1-\det(A)) < M-m \tag{ii}$$

by

$$|a_{11}-a_{22}| \leqslant 2t\cdot m+1-\det(A) \tag{8}$$

where $t=(1+s)/2$ and $a_{12}\cdot a_{21}=-m\cdot M$. Note that if $t=1$ $(s=1)$ then (i) is never fulfilled and only two regions remain, yielding the original theorem. If $t=0$ $(s=-1)$ then (36) is never fulfilled and (i) is met only for $M=m$, but in that case both (37) and (8) yield the same condition $|a_{11}-a_{22}| \leqslant 1-\det(A)$, in accordance with (2).

The proof of this primed theorem follows the same line of reasoning as the proof given in Section IV. Again, we consider the ellipse face (23) in combination with the restriction on the matrix P that is dictated by the overflow arithmetic. We take the overflow characteristic that is on the edge of O^s, i.e.,

$$f(x) = -x+(1+s)\cdot\operatorname{sign}(x), \qquad \text{for } 1<|x|\leqslant 2+s. \tag{38}$$

The borderline case for the geometry of the set of "concentric" ellipses $\underline{x}^tP\underline{x} = \text{constant}^{10}$ is depicted in Fig. 3 for the characteristic given by (38). Only the most critical ellipse of this set is drawn (the middle ellipse in the corresponding figure for saturation, cf. Fig. 1), where any increase of the constant would cause both states to overflow. The arcs outside the unit square are mirrored with respect to the lines $|x_i|=(1+s)/2=t$ by the overflow arithmetic. The images are on the face of the ellipse as is shown in Fig. 3. From the requirement that the ellipse in Fig. 3, i.e.,

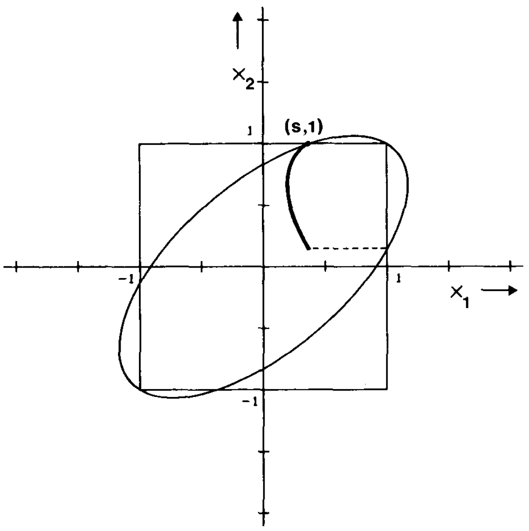

Fig. 3. Borderline case for the geometry of the set $\underline{x}^tP\underline{x}$ = constant if overflow is corrected by a characteristic in the class O^s.

$$\alpha x_1^2 - 2|\gamma|x_1x_2 + \beta x_2^2 = \alpha - 2|\gamma| + \beta$$

contains the point $(s,1)$, we conclude that $2|\gamma| = (1+s)\cdot\alpha$. Hence, the new restriction on the matrix P to replace (17) for a general class O^s of overflow characteristics is simply

$$|\gamma| \leqslant \frac{1+s}{2}\cdot\min(\alpha,\beta) \qquad \text{or} \qquad \eta \leqslant t\cdot\min(\xi,1). \tag{39}$$

Condition (8) is found by demanding that the minimum value of η in the set of potential Lyapunov functions (see Appendix II) satisfies $\min(\eta) \leqslant t$. Of course this solution is valid only if the corresponding value of ξ is greater than 1. The distinction between the cases $\xi \leqslant 1$ and $\xi > 1$ leads to the above regions (i) and (ii). The condition (37) is found by demanding that the point $(\xi,\eta)=(1,t)$ represent a Lyapunov function. The region (36) covers the case where this condition ceases to pose a real restriction on $|a_{11}-a_{22}|$ in view of linear stability.

Example: The direct-form filter with state-transition matrix $A=\begin{bmatrix} a & b \\ 1 & 0 \end{bmatrix}$ is linearly stable if $|a|+b<1\Lambda b>-1$. It is zero-input overflow stable under the additional condition $|a|-b\leqslant 1$ for modulo-2 arithmetic and $|a|\leqslant 1$ for zeroing. For saturation and slope-inversion arithmetic no additional condition is required.

VII. Stability of the Forced Response

A digital filter is said to be forced response stable if it recovers from overflow in the presence of a nonzero input signal; i.e., after an overflow the forced response is re-

[10] Again, we assume that this set is oriented such that $|x_1|$ is first to exceed unity with increasing value of the constant, i.e., $\alpha<\beta$.

gained asymptotically in time [4].[11] More specifically, this recovery may be described as follows.

The forced response $\underline{x}_p(k)$ is defined as the particular solution of the linear state equation (29). Similarly, the response of the actual filter (with overflow correction f) is given by

$$\underline{x}(k+1)=f\left[A\underline{x}(k)+\underline{b}u(k)\right]. \tag{40}$$

The deviation from the forced response $\underline{d}(k)=\underline{x}(k)-\underline{x}_p(k)$ satisfies the equation

$$\begin{aligned}\underline{d}(k+1)&=f\left[A\underline{d}(k)+\underline{x}_p(k+1)\right]-\underline{x}_p(k+1)\\&=\underline{F}\left[A\underline{d}(k)\right]\end{aligned} \tag{41}$$

where $\underline{F}(\cdot)$ is a time-varying overflow arithmetic [4]. The norm of the difference vector $\underline{d}$ decreases asymptotically to zero, or equivalently the actual filter is forced response stable, if the filter with state vector $\underline{d}$ and state-transition matrix A is zero-input overflow stable using the overflow arithmetic $\underline{F}$. Note that $\underline{F}$ is a vector function, since the two states d_1 and d_2 are corrected differently, according to (41). The two components of $\underline{F}$ are shifted versions of f along the line $f(x)=x$ in the f plot. We can picture these shifts in Fig. 3[12] as a movement of the unit square. The unit square moves about in the state plane with a time-varying center $-\underline{x}_p(k+1)$. The amplitude of the horizontal and vertical motion is less than unity, since the forced response is assumed not to cause overflow. Likewise, this amplitude will be less than c if the input signal $u(k)$ is scaled down by a factor c, i.e., if $u(k)$ is replaced with $c\cdot u(k)$. With the help of such a scaling factor $(0\leqslant c\leqslant 1)$ we can introduce a limited forced response stability, where $c=0$ corresponds to zero-input stability.

Theorem 3: A second-order state-space filter that is zero-input overflow stable with respect to the class O^{s_1} of overflow characteristics is forced response stable with respect to O^{s_2} for a class of input signals whose amplitudes do not exceed $(s_2-s_1)/(3+s_1)$.

The proof of this theorem is based on the notion that the worst case for the movement of the unit square in Fig. 3 is a shift along the line $d_1+d_2=0$. With the upper right corner of the unit square at the point $(1-c,1+c)$, the set $\underline{d}'P\underline{d}=$ constant represents the borderline case (for an overflow characteristic in the class O^{s_2}) if it contains an ellipse through this upper right corner and the point $(s_2-c,1+c)$. On the other hand the set should contain an ellipse through the points $(1,1)$ and $(s_1,1)$ for zero-input stability with respect to O^{s_1}. With these requirements it follows that $c=(s_2-s_1)/(3+s_1)$.

Example: The direct-form filter with state-transition matrix $A=\begin{bmatrix} a & b\\ 1 & 0\end{bmatrix}$ is forced response stable for input signals with amplitudes up to $1/3$ if $|a|\leqslant 1$ and saturation is used for overflow correction. The wave digital filter with state-matrix

$$A=\tfrac{1}{2}\begin{bmatrix}1+a+b & -1+a+b\\ 1+a-b & -1+a-b\end{bmatrix}$$

is forced response stable for input signals with amplitudes up to $1/2$ if zeroing is used for overflow correction. With respect to O^1, wave digital filters are altogether forced response stable $(c=1)$. This latter result was stated in [4].

VIII. Conclusion and Summary

A general analysis is presented that establishes a set of conditions for the overflow stability of second-order state-space structures. The well-known overflow arithmetics saturation, zeroing, and two's-complement represent three classes of overflow characteristics, for each of which a stability condition is derived involving only the elements of the state-transition matrix. The zero-input stability condition for the class containing saturation is such that it is satisfied by all properly scaled second-order state-space filters. This result need not be restricted to state-space structures. It also holds more generally for all second-order digital filters if overflow at non-state-variable nodes is ruled out through the use of an extra bit for those nodes whose l_2 norms exceed that of the scaled state variables.

The analysis is based on determining the set of Lyapunov functions for a general second-order state-transition matrix. The bounds within which the elements of the state matrix are to be restricted are derived from the requirement that this set contain only one element. These bounds represent sufficient conditions for zero-input overflow stability. In Appendix III it is shown that these conditions are not strictly necessary.

As for the stability of the forced response, a digital filter that is zero-input overflow stable for one class of overflow characteristics is shown to be forced response stable (possibly in a limited sense) for a related class of characteristics.

Appendix I
Proof of Lemma 1

Since P is a positive definite and symmetric matrix, we may write $P=T^{-\prime}T^{-1}$, where $T^{-1}=(\underline{t}_1\ \underline{t}_2)$ is a nonsingular matrix, so $\underline{t}_1\neq\lambda\underline{t}_2$. Note that T is not unique. Substitution yields

$$P-A'PA=T^{-\prime}(I-S'S)T^{-1}$$

where I is the identity matrix and $S=T^{-1}AT$; i.e., S and A are linked by a similarity transformation. Note that $\det(S'S)=\det^2(A)<1$.

(i) If $\det(P-A'PA)\geqslant 0$ then $I-S'S$ is positive (semi) definite.

Proof: Let $\det(P-A'PA)\geqslant 0$ then

$$\det(I-S'S)=\det{}^2(T)\det(P-A'PA)\geqslant 0$$

and

$$\begin{aligned}\operatorname{tr}(I-S'S)=2-\operatorname{tr}(S'S)&>1-\operatorname{tr}(S'S)+\det(S'S)\\&=\det(I-S'S).\end{aligned}$$

[11] In this definition it is assumed that the state motion associated with the forced response, $\underline{x}_p(k)$, ultimately does not leave the unit square, or less formally, that overflows are sufficiently far apart in time. This will be the case for any properly scaled filter.

[12] If the state plane in Fig. 3 is understood to be the d_1, d_2 plane.

From $\mathrm{tr}(I - S'S) > \det(I - S'S) \geqslant 0$ we conclude that $I - S'S$ is positive (semi)definite.

(ii) If $I - S'S$ is positive (semi)definite, then $\mathrm{tr}(P - A'PA) > 0$.

Proof: $\mathrm{tr}(P - A'PA) = \underline{t}_1'(I - S'S)\underline{t}_1 + \underline{t}_2'(I - S'S)\underline{t}_2$. Let $I - S'S$ be positive (semi)definite; then both terms to the right of the equal sign are nonnegative. But since $\underline{t}_1 \neq \lambda \underline{t}_2$ and $I \neq S'S$, they are never simultaneously zero, so $\mathrm{tr}(P - A'PA) > 0$.

From (i) and (ii) we conclude that (12) is implied by (13).

Q.E.D.

Appendix II
Proof of Lemma 2

Distinguish between two regions:

(i) $1 - \det(A) \geqslant M - m$. The line $\xi = 1$ intersects the ellipse in the ξ, η plane (given by (23) satisfied with equality) at two, possibly coinciding, points, say $(1, \eta_1)$ and $(1, \eta_2)$. The point $(1, \eta_3)$, where $\eta_3 = (\eta_1 + \eta_2)/2$, lies on the face of the ellipse; for the matrix P to exist, we need only check whether $\eta_3 < 1$. We find

$$\eta_3 = \frac{(M+m)|a_{11} - a_{22}|}{(1 - \det(A))^2 + 4mM} < 1$$

because $|a_{12} - a_{21}|^2 < (1 - \det(A))^2 + 4mM$ on account of (3), and $(M+m)^2 \leqslant (1 - \det(A))^2 + 4mM$ in the region under consideration. So, with $(\xi, \eta) = (1, \eta_3)$, P exists without any further constraint on A.

(ii) $1 - \det(A) < M - m$. All points on the above ellipse satisfy $\xi > 1$; for the matrix P to exist we need at least one point on the face of the ellipse for which $\eta \leqslant 1$.[13] This will be the case if the minimum of η on the ellipse satisfies $\min(\eta) \leqslant 1$. This minimum if found to be

$$\min(\eta) = \frac{|a_{11} - a_{22}| - (1 - \det(A))}{2m}.$$

From $\min(\eta) \leqslant 1$ we conclude that

$$|a_{11} - a_{22}| \leqslant 2m + 1 - \det(A). \tag{6}$$

The choice

$$(\xi, \eta) = \left(\frac{M}{m} - \frac{1 - \det(A)}{m}, 1 \right)$$

leads to a Lyapunov function $\underline{x}'P\underline{x}$.

Since the condition "$1 - \det(A) \geqslant M - m$" is equivalent to (7) for $a_{12}a_{21} < 0$, the proof is complete.

Q.E.D.

Appendix III

The conditions for overflow stability (laid down in the general Theorem 1′) are sufficient but not necessary, owing to the following observations.

(i) The adopted norm need not be required to be decreasing with respect to overflow correction, as long as a possible nonlinear increase of $\|A\underline{x}\|$ due to overflow correction is compensated by a foregoing, larger decrease of $\|\underline{x}\|$ as a result of the linear state transition.

(ii) Even if the combined operation $f(A\underline{x})$ shows an occasional increase of the chosen norm, this need not necessarily lead to instability. For a periodic overflow oscillation to exist, the norm $\|\underline{x}\|$ should come "full circle"; i.e., after one period it should assume its original value.

(iii) If the overflow characteristic $f(\cdot)$ does not map the two arcs outside the unit square of some ellipse $\|\underline{x}\| =$ constant entirely into its interior, this does not necessarily mean that overflow correction will ever increase the norm of $\underline{x}$. Since a digital filter has a bounded state space, only a limited region of the state plane outside the unit square can actually be reached, i.e., be an image under the mapping of the unit square by the state-transition matrix A. Those parts of the ellipse $\|\underline{x}\| =$ constant that are mapped beyond its perimeter (by the overflow characteristic) might never be reached at all. A good example is given by the direct-form filter with $x_2(k+1) = x_1(k)$: any point (x_1, x_2) with $|x_2| > 1$ cannot be an image of an original inside the unit square.

In view of these observations a necessary condition for overflow stability can only be derived from a comprehensive analysis of the overflow behavior. Such an analysis would show a strong dependence on the choice of the matrix A and would probably not lead to a closed range of stability for the absolute difference $|a_{11} - a_{22}|$. We can use the first of the above concepts, however, to relax the stability condition (8) even further.

In order to do so, we recall that there is a unique Lyapunov function $\underline{x}'\hat{P}\underline{x}$ for which property (27) holds. With the choice $P = \hat{P}$, the norm $\|\underline{x}\|$ decreases by a factor $\sqrt{(\det(A))}$ with every state transition. As a result, we may allow an increase of $\|\underline{x}\|$ due to overflow correction by a factor less than $\sqrt{(\det(A))}$. In other words, $\hat{\eta}$ may exceed the bound set by (38), i.e., $t \cdot \min(\hat{\xi}, 1) = t$. If we can establish just how far $\hat{\eta}$ may exceed t, we will have found a new set of conditions for overflow stability, since $|a_{11} - a_{22}| = 2m \cdot \hat{\eta}$. Following the reasoning of Section III, we consider the set of ellipses $\underline{x}'\hat{P}\underline{x} =$ constant (with the usual assumption $\hat{\alpha} < \hat{\beta}$, so $\hat{\alpha} = 1$ and $\hat{\beta} = M/m$). The borderline case for the geometry of this set is shown in Fig. 4 for the overflow characteristic given by (39). This figure should be read as follows. The inner ellipse $\underline{x}'\hat{P}\underline{x} = K_1$ contains the point $Q = (1,1)$ which has a mirror image $R = (s, 1)$ with respect to the line $x_1 = (1+s)/2$. The constant K_1 is equal to $1 - 2\hat{\eta} + M/m$. The outer ellipse $\underline{x}'\hat{P}\underline{x} = K_2$ contains the original, denoted P, of Q under the mapping A, so $A(P) = Q$ and $K_2 = K_1/(\det(A))$.

[13] Note that for this point $\xi > \eta^2$ is met as well, since $\xi > 1$.

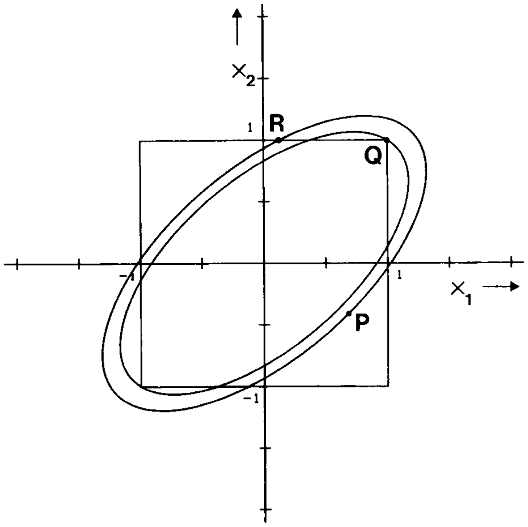

Fig. 4. Borderline case for the geometry of the set $\underline{x}'\hat{P}\underline{x}$ = constant.

The combined operation $f(A\underline{x})$ decreases the norm of $\underline{x}$ if $R=(s,1)$ is on the face of the outer ellipse, or on its perimeter in the borderline case. This condition is met if $s^2 - 2\hat{\eta}\cdot s + M/m < (1-2\hat{\eta} + M/m)/(\det(A))$, or

$$|a_{11} - a_{22}| < \frac{M\cdot(1-\det(A)) + m\cdot(1-s^2\det(A))}{1-s\cdot\det(A)}. \quad \text{(A1)}$$

This is an alternative condition for zero-input overflow stability to replace (8). Note that the bound on $|a_{11}-a_{22}|$ given by (A1) exceeds the bound given by (8) if $1-s\cdot\det(A) < M - m\cdot s$, so (A1) is preferred over (8) only if $\det(A)$ is sufficiently close to unity. This must be attributed to the fact that the choice $P=\hat{P}$ is not optimal. Note also that both bounds are equal to $(1+s)\cdot m$ in the limit $\det(A)\to 1$.

Note that the new condition (A1) applies only if the inner ellipse in Fig. 4 actually intersects the line $x_2 = 1$ in a (second) point (next to Q) with the x_1 value smaller than or equal to 1. This is the case only if $\hat{\eta} \leqslant 1$ or $|a_{11}-a_{22}| \leqslant 2m$. Hence, (A1) is valid under the condition

$$s \leqslant 1 - \left[\frac{(M-m)\cdot(1-\det(A))}{m\cdot\det(A)}\right]^{1/2}. \quad \text{(A2)}$$

Note that in view of this restriction the bound of (A1) does not exceed $2\sqrt{(mM)}$, as indeed is necessary due to (5) (recall that $\underline{x}'\hat{P}\underline{x}$ is a Lyapunov function only in the case of complex poles).

Acknowledgment

The author wishes to thank Prof. H. J. Butterweck for triggering the original idea that led to this paper and M. J. Werter for stimulating discussions along the way.

References

[1] R. A. Roberts and C. T. Mullis, *Digital Signal Processing.* Reading, MA: Addison-Wesley, 1987, ch. 9.

[2] 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, p. 569–574, Oct. 1977.

[3] W. L. Mills, C. T. Mullis, and R. A. Roberts, "Digital filter realizations without overflow oscillations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 334–338, Aug. 1978.

[4] T. A. C. M. Claasen, W. F. G. Mecklenbräuker, and J. B. H. Peek, "On the stability of the forced response of digital filters with overflow nonlinearities," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 692–696, Aug. 1975.

[5] T. A. C. M. Claasen, W. F. G. Mecklenbräuker, and J. B. H. Peek, "Effects of quantization and overflow in recursive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 517–529, Dec. 1976.

[6] P. M. Ebert, J. E. Mazo, and M. G. Taylor, "Overflow oscillations in recursive digital filters," *Bell Syst. Tech. J.*, vol. 48, pp. 2999–3020, Nov. 1969.

John H. F. Ritzerfeld received the master's degree in electrical engineering from the Eindhoven University of Technology, Eindhoven, The Netherlands, where he is currently employed as a research fellow with the digital signal processing group.

His main research interests are in digital signal processing, nonlinear effects in digital filters, and discrete-time nonlinear systems in general.