On the Stability of Sigma Delta Modulators

Søren Hein, *Member, IEEE,* and Avideh Zakhor, *Member, IEEE*

***Abstract*—In this paper we propose a framework for stability analysis of ΣΔ modulators, and argue that limit cycles for constant inputs are natural objects to investigate in this context. We present a number of analytical and approximate techniques to aid the stability analysis of the double loop and interpolative modulators, and use these techniques to propose ways of improved design that explicitly take stability into account.**

I. Introduction

SIGMA delta (ΣΔ) modulators are playing an increasingly important role in analog-to-digital conversion. They are capable of achieving the same resolution as Nyquist-rate multibit quantizers by employing a one-bit quantizer operating at many times the Nyquist rate. The modulators generally require fewer and simpler components than Nyquist-rate converters, and are more robust against circuit imperfections. As a result they are ideal for on-chip VLSI implementation in relatively low-bandwidth applications such as audio. They have also recently been used in higher bandwidth applications [1], [2].

Historically, single-loop [3] and double-loop [4] ΣΔ modulators were the first to be introduced, analyzed, and implemented. In recent years substantial work has been done on variations of the basic architecture to improve the tradeoff between signal-to-noise ratio (SNR) and oversampling ratio (OSR). These efforts have been focused on complex modulators, as measured by the number of integrators, and as a result two trends have emerged: Higher-order single loop or interpolative modulators [5], and multistage (MASH) or cascaded modulators consisting of cascades of a number of single and double loop modulators [6], [7]. Within these broad categories, a number of designs have been proposed and implemented to meet varying requirements on signal bandwidth, sampling rate, SNR, dynamic range, and other specifications [8], [9]. The main limitation of cascaded structures is their sensitivity to component mismatch between individual stages, while the main limitation of interpolative modulators, especially higher order ones, is their stability problems.

Manuscript received September 3, 1991; revised September 8, 1992. The associate editor coordinating the review of this paper and approving it for publication was Dr. David Nahamoo. This work was supported by the National Science Foundation PYI Award MIP-9057466, ONR Young Investigator Award N00014-92-J-1732, and Analog Devices. It was presented in part at the IEEE DSP Workshop, New York, September 1990, and in part at ISCAS, Singapore, June 1991.

The authors are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720.

IEEE Log Number 9208847.

The purpose of this paper is to discuss stability as an integral part of analysis and design of ΣΔ modulators. It may be argued that the design effort should first concentrate on SNR performance, and that stability problems can be solved by subsequently scaling circuit coefficients. We argue instead that scaling by itself does not entirely solve the instability problem, and that some sacrifice of SNR may be necessary in addition to scaling to stabilize higher order modulators. Specifically, we will demonstrate a tradeoff between SNR and stability for interpolative modulators, and show that stability concepts can be applied even to the double loop modulator which is usually labeled as stable. We do not claim to provide definitive answers to all questions of optimal design, but intend to give a frame of reference for stability considerations, as well as to present a number of analytical techniques to aid design.

Our approach to the stability problem throughout the paper is to examine the large-amplitude limit cycle behavior of the double loop and interpolative modulators for constant inputs. The approach is justified in more detail in Section II, but the main motivation is that constant inputs are a special case of more general inputs, and that limit cycles characterize the long-term behavior of the modulators under constant inputs. Therefore stability under constant inputs is a necessary condition for stability under more general inputs. Furthermore, we will show that results from limit cycle analysis can be used in the design process.

The paper is organized as follows. Section II contains a general discussion of stability issues, and proposes an operational definition of stability. Section III addresses stability issues for the double loop modulator; this modulator is in itself of interest, and is also important because it serves as a building block in cascaded modulators. Section IV considers the class of interpolative modulators, and Section V contains a summary and conclusions. Some variations of architectures proposed in the literature [1], [8] may be accommodated by correspondingly minor modifications in the analysis, while others may require more substantive changes.

II. General Considerations

In this section we first briefly consider the effect of integrator clipping. In Section II-A we suggest a definition of stability which appears to be more operational and better suited to ΣΔ modulators than the definitions of traditional nonlinear systems theory. In Section II-B we argue

that the study of limit cycles for constant inputs provides insight into stability issues.

$\Sigma\Delta$ modulators are generally built around a number of integrators; an example is the double loop modulator shown in Fig. 3 which contains two integrators. Integrator limiting or clipping is an important practical effect which occurs because of voltage saturation in the operational amplifiers of internal integrators. A simple clipping model employs a saturation characteristic of the form

$$\text{sat}\,(x) = \begin{cases} x & \text{for } |x| \leq L \\ L\ \text{sign}\,(x) & \text{otherwise} \end{cases} \tag{1}$$

where L is the clipping level, and sign $(\cdot)$ is the signum function. Modulators are typically designed so that the clipping level is not much larger than the feedback voltage; for instance, the single-loop modulator in Fig. 2 would have L close to b. Clipping entails loss of state variable information and hence performance degradation: All other things being equal, it should be avoided. We will refer to practical modulators suffering from clipping as clipped modulators, and modulators with ideal integrators, that is, modulators with no clipping, as unclipped modulators. Comparing the clipped and unclipped modulators, it is clear that the added nonlinearity shown in (1) in each integrator further complicates analytical attacks on the already nonlinear system. This necessitates a high degree of reliance on computer simulations to assess performance and emphasizes the need for accurate behavioral models of circuit nonidealities [10]. To circumvent these complications, we suggest in Section II-A a way to address the stability problem analytically which avoids introducing the nonlinearity (1).

A. Scaling and Stability

For a given $\Sigma\Delta$ modulator, the only way to avoid clipping is to scale integrator gains, feedback coefficients, and other circuit parameters to keep the signal levels throughout the modulator below saturation most of the time. This description of scaling, however, is vague in two respects:

1) The maximum values of signal levels depend not only upon the modulator, but also upon the class of its input signals.

2) There is an important distinction between scaling which preserves the functionality of the modulator, and one that modifies it. More explicitly, by the functionality of a modulator we mean the transformation it applies to its input to produce the output bit stream. For instance, the functionality of the double loop modulator in Fig. 3 is not affected if we multiply G and b by the same number; we refer to this case as equivalent scaling. On the other hand, changing only G or b affects the functionality; we refer to this case as functional scaling. In Section III we discuss the difference in more detail.

These two points are treated separately here and in the following section. Equivalent scaling is straightforward and may sufficiently reduce signal levels that clipping occurs infrequently under normal operation. There is a simple connection between signal levels at the nodes of the unscaled and scaled, unclipped modulators, since care is taken that the performed scaling at each node only affects the signal level at that node. The main problem with the limited approach of equivalent scaling is that the required scale factors may be excessive: In practice, very large or small loop coefficients are not easily implementable as capacitor ratios. A more attractive option is to first use functional scaling to improve stability, typically at the cost of some SNR performance, and subsequently use equivalent scaling to further reduce the signal levels so that clipping rarely occurs. We take the latter approach in this paper. However, functional scaling also has its problems: First, the effects of scaling on signal levels are more unpredictable than those of equivalent scaling, since scaling at some node is no longer restricted to only have local consequences for that node. Second, the effects of functional scaling on such performance parameters as SNR are not easily predictable, since linearized analyses may be misleading. To use functional scaling appropriately, it is necessary to examine the tradeoffs involved between SNR performance and stability. This tradeoff is a central theme of Sections III and IV.

In order to better address the problems of functional scaling, we propose to define the stability of $\Sigma\Delta$ modulators in terms of the maximum signal levels occurring throughout the unclipped system, i.e., we consider stability to be a matter of degrees. We will call a system K-stable if the signal levels are bounded in absolute value by K for a given class of input signals. We will also call a system very stable or very unstable according to whether maximum signal levels in the corresponding unclipped system are very much smaller or larger than the clipping level, for a given class of inputs. This definition is reasonable because the signal levels dictate whether clipping occurs. The definition is in contrast to traditional stability definitions such as bounded input bounded output (BIBO) and bounded input bounded state (BIBS) stability which are only concerned with categorizing systems as either stable or unstable [11]. Although the input dependence is present in traditional stability theorems [11], these typically assume bounded, square integrable or summable inputs, and allow for no way of distinguishing between, say, two different constant inputs. In contrast we consider the dependence of stability upon DC level or sinusoidal amplitude to be extremely important.

The described viewpoint on stability avoids dealing with the nonlinear clipping operation which would require a nonlinear analysis, and suggests that stability can be assessed by judging the ratio between the maximum signal levels in unclipped systems and the clipping level. The viewpoint appears to be more useful in the present context than traditional definitions of stability: For instance, real $\Sigma\Delta$ modulators are always stable in the BIBO and BIBS sense, in that clipping keeps all voltages bounded. The viewpoint means that stability can be seen as an integral part of the design process for any $\Sigma\Delta$ modulator, even a second-order one. We thus argue that clipping should be

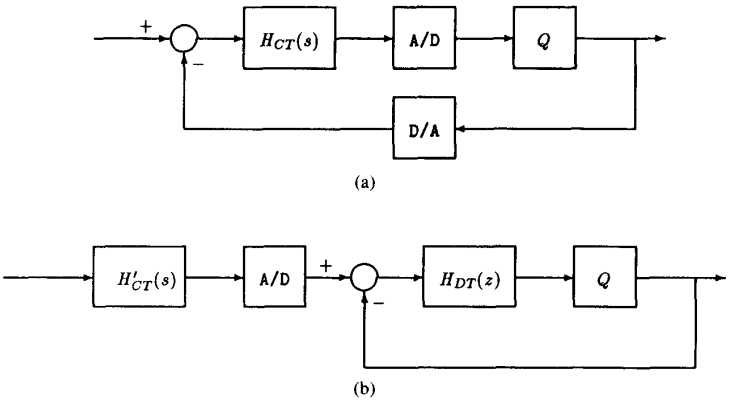


Fig. 1. (a) Continuous-time interpolative modulator with CT filter $H_{CT}(s)$, sample-and-hold A/D converter, 1-b digital quantizer, and ideal D/A converter. (b) Equivalent discrete-time modulator with DT filter $H_{DT}(z)$ and CT prefilter $H'_{CT}(s)$, both depending only upon $H_{CT}(s)$.

precluded by design, and our goal is to guide designs with that objective. This is in contrast with the existing approach [9] where specialized circuitry is used to detect saturation of integrators and reset these.

B. Limit Cycles

This section discusses the applicability of limit cycle analysis to stability investigations. As described above, the purpose of scaling is to keep the maximum values of integrator outputs below the clipping level. Of course, too much downscaling will decrease the ratio between signal and circuit noise levels excessively, and thus will adversely affect performance. In practice a tradeoff must therefore be found between the frequency and effect of clipping on one side, and noise sensitivity on the other. This may be viewed as designing for the right amount of stability.

For a given modulator, the signal levels depend on the input signal class. To eliminate transient phenomena and to focus on long-term behavior, we will consider limit cycles or periodicities for constant inputs. The following arguments justify this.

1) Limit cycles are essential to the operation of ΣΔ modulators, as evidenced by their prominent position in several papers, including [2], [12], and [13]. In [14] and other places is shown that for the single and double loop modulators, limit cycles occur only when the constant input is a rational fraction of the quantizer step size. Limit cycles can thus be seen as a natural result of approximating constant inputs using ΣΔ modulators. Furthermore, a recent paper [15] shows that for a single loop modulator whose integrator has its pole inside the unit circle, almost all constant inputs generate limit cycles.

2) The oversampling of the input in practical situations implies that it appears approximately constant to the modulator.

3) Any modulator designed for dynamic inputs must be able to handle constant inputs as a special case; therefore stability under constant inputs is a necessary condition for stability under more general conditions.

4) The assumption of constant input allows us to make statements which hold for both continuous-time (CT) and discrete-time (DT) ΣΔ modulators. This is because it can be shown that any CT modulator can be converted into a DT modulator [16]; to make the two modulators equivalent, the CT input must be converted into a DT sequence by prefiltering it with a CT filter and then sampling it, as shown in Fig. 1. The required CT filter depends on the specific CT modulator. However, the prefilter can simply be represented by a constant gain for analysis of DC inputs.

5) Constant inputs simplify analytical attacks; indeed, a number of illuminative results have been based on this simplifying assumption [14], [17].

Based on the above discussion, our approach in the following sections is to use limit cycle analysis for constant inputs to find maximum signal levels in ΣΔ modulators. We also examine functional scaling as a way of trading off SNR performance for stability.

C. Comparison with Existing Work

In this section we compare our stability approach to previous results on delta modulation.

Gersho [18] considers single-integration delta modulation with stochastic stationary input processes, and either perfect or leaky integration. The stability concept underlying his approach is consistent with the approach of this paper. For the modulator, he derives upper bounds on the error signal. The corresponding stability result for single-loop ΣΔ modulators with ideal integration is well known, and is stated in Section III below. Gersho's method does not appear to generalize to higher order modulators.

Nielsen [19] considers a special form of double-integration delta modulation with zero input, and examines the specific limit cycle type consisting of a number of positive output bits followed by the same number of negative bits. He states that the limit cycle length is a measure of the stability, and he numerically optimizes a particular

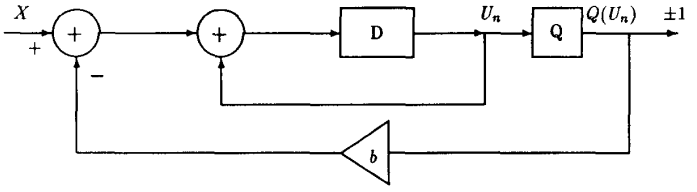


Fig. 2. Discrete-time model of the single loop ΣΔ modulator.

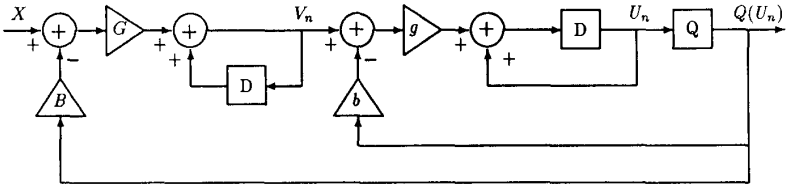


Fig. 3. Discrete-time model of the double loop ΣΔ modulator.

modulator parameter for stability. The approach is less general than the one presented in this paper, because we consider general constant inputs and arbitrary limit cycles, as well as more general modulators.

Finally, Steele [20] considers double-integration delta modulation with ideal integration. For the case of zero input, he derives the peak-to-peak value of the feedback signal. He then introduces prediction around the second integrator to reduce feedback oscillations. Due to the special form of the double-loop ΣΔ modulator with two feedback paths for the quantizer output, Steele's results are not applicable to our problem.

III. Double Loop Modulator

In this section we discuss the stability issues for the double loop ΣΔ modulator within the framework of Section II. Section III-A includes a number of exact analytical methods for investigating the limit cycle behavior of the modulator. In Section III-B we derive exact upper bounds on the largest integrator outputs occurring on limit cycles. This is used to arrive at a design and scaling technique for double loop modulators which results in a more favorable SNR performance than the standard one.

Some of the results of this section may be viewed as extensions of results for the single loop modulator given in [17] and other papers. For completeness we provide a concise overview of these results in the language of the present paper, drawing also on the results in [14]. Fig. 2 shows the block diagram of the single loop modulator with constant discrete-time input X: D represents a unit delay, and Q is a one-bit quantizer or ADC given by

$$Q(U) = \begin{cases} +1 & \text{for } U > 0 \\ -1 & \text{for } U \leq 0. \end{cases}$$

The block labeled b represents a digital-to-analog converter (DAC) whose input is ± 1 and whose output is b times its input. It is shown in [17] that for any constant input $X \in (-b, +b)$, the state variable U_n at any time $n > 0$ lies in the interval $[X - b, X + b]$ provided that the initial state at time 0 lies in the same range. This essentially resolves the stability issue for the single loop modulator: b must be chosen such that $2b$ is less than the clipping level of the integrator. If the constant input X equals some irreducible fraction p/q of the DAC feedback voltage b, there exists exactly one limit cycle. It has period $2q$ if either p or q is even, and period q otherwise; the average of the quantizer outputs $Q(U_n)$ over one period equals the normalized input X/b [14]. The limit cycles or periodicities show up as spikes in the spectrum of the quantization error sequence [21].

A block diagram of the double loop modulator is shown in Fig. 3; it contains four scaling factors, namely, two for the integrators and two for the DAC feedback. The ones corresponding to the outer integrator are denoted by the uppercase letters B and G, and the inner integrator factors are denoted by the lowercase letters b and g. The double loop modulator is superior to the single loop one because it only requires a moderate increase in circuit complexity, and yet it achieves a 15 dB/octave tradeoff between SNR and OSR, whereas the single loop modulator achieves only 9 dB/octave. The double loop modulator is of interest in itself; its analysis and implementation have been described in a number of papers, including [22] and [23]. However, it is also important as a building block in higher order cascaded modulators, as evidenced in [8], [24], and other papers. Linearizing the modulator, it can be viewed as a two-pole digital filter in a feedback loop, and considering measures such as phase margin, it may loosely be characterized as "barely stable."

A. Detection of Specific Limit Cycles

In this section we consider in detail the limit cycles of the canonical unclipped double loop modulator, i.e., the modulator shown in Fig. 3 with $b = B = g = G = 1$ and ideal integrators. The methods readily translate to the more general structure. In Sections III-A1 and A2 we address the following problem: Given a P-bit sequence $\vec{Y} =$

$\{Y_0, \cdots, Y_{P-1}\}$, does there exist a constant input X and a limit cycle with period P such that $\vec{Y}$ corresponds to the modulator output sequence $\{Q(U_0), \cdots, Q(U_{P-1})\}$? If so, what are the largest values of the state variables occurring as the modulator goes through a period of the limit cycle? We define a limit cycle to exist if all internal state variable sequences of the modulator, i.e., U_n and V_n, are periodic. Our technique for solving these problems makes use of the standard Tsypkin method of "opening the loop" and thus circumventing the nonlinearity [25]. A similar approach was used in [26] for the specific case of zero-input symmetric limit cycles with a number of positive bits followed by the same number of negative bits. In Section III-A3 we present a numerical technique to assess the regions in state space that are parts of limit cycles.

1) Existence of Specific Limit Cycles I: From Fig. 3 with $b = B = g = G = 1$, we obtain the state equations for the variables U_n and V_n:

$$U_n = U_{n-1} + V_{n-1} - Q(U_{n-1})$$
$$V_n = V_{n-1} + X - Q(U_n). \tag{2}$$

Inserting $Y_n = Q(U_n)$ and $Y_{n-1} = Q(U_{n-1})$ in (2), we arrive at the following closed-form formulas in [14]:

$$U_n = U_0 + nV_0 + \tfrac{1}{2}n(n-1)X - Y_0 - \sum_{i=1}^{n-1}(n+1-i)Y_i$$
$$V_n = V_0 + nX - \sum_{i=1}^{n} Y_i. \tag{3}$$

The double loop modulator is defined to have $\vec{Y}$ as the output sequence of a limit cycle if and only if two conditions are satisfied [25]:

C1) The state variable sequences $\{U_n\}$, $\{V_n\}$ obtained by using $\vec{Y} = \{Y_0, \cdots, Y_{P-1}\}$ in (3) are periodic with period P, that is, $U_{P+n} = U_n$, $V_{P+n} = V_n$ for all n. It follows from (2) that this condition holds if and only if $U_P = U_0$ and $V_P = V_0$.

C2) For each $0 \le n \le P-1$, the sign of the quantizer input matches the corresponding bit of the sequence $\vec{Y}$, that is, $Y_n = Q(U_n)$. This corresponds to "closing the loop" and checking the consistency of the resulting system.

To satisfy $V_P = V_0$ in condition C1, (3) implies

$$\frac{1}{P}\sum_{n=0}^{P-1} Y_n = X. \tag{4}$$

There is thus only one constant input which might give rise to the limit cycle under investigation; this constant input is rational and equals the average input. To satisfy $U_P = U_0$ in condition C1, we use the first equation in (3) to obtain

$$V_0 = -\frac{1}{2}(P-1)X + \frac{1}{P}\left(Y_0 + \sum_{n=1}^{P-1}(P+1-n)Y_n\right). \tag{5}$$

To check condition C2 we may proceed in the following way: Let R_0 be the set of all values of U_0 such that the consistency relation at time $n = 0$ is satisfied. For $n \ge 1$, use (3) to recursively compute U_n in terms of U_{n-1}. Determine the set R_n of all values of U_0 such that the consistency relation $Q(U_n) = Y_n$ is satisfied; this amounts to a linear inequality in U_0.[1] At time P, determine the intersection R of all the sets $R_0, \cdots, R_{P-1}$. If R is empty, then the combined constraints on U_0 are impossible to satisfy simultaneously, and the sequence $\vec{Y}$ is not a limit cycle for the double loop modulator. If R is nonempty, the sequence $\vec{Y}$ does exist as a limit cycle for any initial state in R. In this case we can step through (3) to determine the largest values of the state variables over the limit cycle $\vec{Y}$. For brevity, we shall refer to these maxima as the amplitudes of the limit cycles in the state variables U_n and V_n. Without loss of generality we only consider positive constant inputs, and search for maximum absolute values of U_n and V_n.

2) Existence of Specific Limit Cycles II: The method presented in the preceding subsection makes use of the fact that a solution (3) to the difference equation (2) is available; a more general method which does not make use of (3) and which is easily generalized to higher order systems can also be devised. We will use this technique in Section IV-A1 on the interpolative modulator, and in this section we show its application to the double loop modulator.

We can rewrite the two first-order state equations shown in (2) as a single second-order state equation:

$$U_{n+2} - 2U_{n+1} + U_n = -2Y_{n-1} + Y_n + X. \tag{6}$$

We will assume that the state variable sequence $\{U_n\}$ is periodic with period P and enforce the consistency requirement C2. Assuming the periodicity condition C1 holds, (6) can be written as a linear vector equation

$$\begin{bmatrix} -2 & 1 & 0 & \cdots & 0 & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 1 & -2 \\ \vdots & \vdots & & & \vdots & \vdots & \vdots \\ 0 & 1 & -2 & \cdots & 0 & 0 & 0 \\ 1 & -2 & & \cdots & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} U_0 \\ U_1 \\ \vdots \\ \vdots \\ U_{P-1} \end{bmatrix} = \begin{bmatrix} -2 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & -2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & -2 & \cdots & 0 & 0 \end{bmatrix} \begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ \vdots \\ Y_{P-1} \end{bmatrix} + X \begin{bmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix} \tag{7}$$

[1] In the more general case of a multibit quantizer Q, the consistency relation $Q(U_n) = Y_n$ is unchanged, and also gives rise to linear inequalities in U_0.

where the equations for $n = 0$ and $n = P - 1$ are at the bottom and top, respectively. The $P \times P$ matrix on the left-hand side is singular and has rank $P - 1$. By adding together the P scalar equations, we again arrive at the requirement (4) for the constant input X. The right-hand side of (7) is thus known for a given $\vec{Y}$. The equation can be solved by simple forward substitution with, say, U_0 and U_1 as independent variables; as this procedure reaches $n = P - 1$, it produces a linear constraint involving only U_0 and U_1. This means that one of U_0 and U_1 can be used as the sole independent variable; this way the entire sequence $\{U_n\}$ can be specified in terms of only U_0. Enforcing the consistency requirement C2 proceeds in the same manner as before: U_0 is chosen, if possible, such that for $0 \le n \le P - 1$, $Q(U_n) = Y_n$.

The above method can be interpreted geometrically in the following way: Equation (7) specifies a line in the P-dimensional space R^P of P-element real sequences in which $\vec{U} = \{U_0, \cdots, U_{P-1}\}$ lies. The consistency requirement C2 limits the allowable region of this space to a single octant, i.e, all elements U_n of $\vec{U}$ must have signs specified by the corresponding element Y_n of the sequence $\vec{Y}$. The limit cycle $\vec{Y}$ exists if and only if the intersection between the line and the octant is nonempty. The points in the one-dimensional intersection each represent a possible limit cycle in the quantizer input $\{U_n\}$.

We can also explain that there is some latitude in choosing U_0 The left-hand side matrix of (7) has rank $P - 1$, and the eigenvectors corresponding to the eigenvalue 0 are of the form $\vec{K} = (k, k, \cdots, k)^T$. Therefore, if $\vec{U} = \{U_0, \cdots, U_{P-1}\}$ is a P-periodic solution to (7), $\vec{U} - \vec{K} = \{U_0 - k, \cdots, U_{P-1} - k\}$ is another P-periodic solution for any k. As long as the consistency requirement C2 is not violated, both these are limit cycles in the state variable U_n. More precisely, if $\vec{U}$ is a limit cycle, then for all k in the following range, the shifted sequence $\vec{U} - \vec{K}$ is another limit cycle:

$$\max_{U_n \le 0} U_n \le k < \min_{U_n > 0} U_n. \tag{8}$$

We refer to the width of the interval for k in (8) as ΔU. Using the above technique or that of Section III-A1, we can obtain information about short time limit cycles by exhaustively searching over the binary sequences with average value X and different periods P. Table I summarizes such information for periods up to 24 for the cases $X = 0$ and $X = 0.5$: For each period P we list the pair (U_n, V_n) that achieves the largest absolute quantizer input U_n while lying on a limit cycle with period P. We also list the quantity ΔU corresponding to the limit cycle that maximizes U_n. We observe that as X increases, the maximum value of U_n increases comparatively more than that of V_n, so in a practical, clipped modulator, U_n will clip before V_n. At both inputs 0 and 0.5, several limit cycles exist at each period for all but the smallest periods.

The table also suggests that as the period increases, the maximum integrator outputs exhibit an increasing trend,

TABLE I
NUMBER OF LIMIT CYCLES OF THE DOUBLE LOOP MODULATOR WITH CONSTANT INPUTS $X = 0$ AND $X = 0.5$. ALSO SHOWN ARE THE PAIRS (U_n, V_n) MAXIMIZING THE QUANTIZER INPUT, AND THE AMOUNT BY WHICH THE LIMIT CYCLES IN THE STATE PLANE CAN BE SHIFTED IN THE U_n DIRECTION

X	Period	# Limit Cycles	(U_n, V_n) w/max. U_n	ΔU
0	2	1	(1.500, 0.500)	1.500
	4	1	(2.000, 1.000)	1.000
	6	3	(2.500, 1.500)	0.500
	8	2	(2.250, 1.250)	0.500
	10	4	(2.300, 1.300)	0.400
	12	2	(2.333, 1.333)	0.333
	14	8	(2.643, 1.643)	0.071
	16	6	(2.625, 1.625)	0.125
	18	8	(2.611, 1.611)	0.167
	20	4	(2.600, 1.600)	0.200
	22	10	(2.545, 1.591)	0.182
	24	4	(2.500, 1.583)	0.167
0.5	4	1	(2.250, 1.250)	1.250
	8	3	(3.500, 2.000)	0.500
	12	3	(3.333, 1.917)	0.250
	16	4	(3.750, 2.125)	0.125
	20	7	(3.950, 2.150)	0.200
	24	3	(4.000, 2.167)	0.167

and seem to approach limits.[2] In fact, is it conjectured that for $X = 0$, these limits equal $8/3$ for U_n, $5/3$ for V_n; for $X = 0.5$, the limit is about 4.16 for U_n. If our observations are valid for all constant inputs, the results suggest that relatively short limit cycles are good indicators of the maximum signal levels encountered also on longer limit cycles.

3) Graphical State Space Method: The drawback of the methods of Sections III-A1 and A2 is the requirement to examine all binary sequences with average X in order to detect limit cycles for the constant input X. We now present a more graphical approach to obtain an overview of the limit cycle behavior. The approach is based on a state space representation of the double loop modulator where pairs of state variable values (U_n, V_n) are points in a plane with U_n and V_n along the horizontal and vertical axes, respectively. For a given constant input X, each point in state space completely specifies a trajectory that can be found by stepping through the difference equations (2). It is possible in principle to determine for each point whether or not the corresponding trajectory is a limit cycle.[3] The set of states that belong to limit cycles, or equivalently the collection of limit sets [27], can then be shown in a plot such as Fig. 4. In practice the process is implemented numerically by discretizing state space. The number of grid points per unit, referred to as the grid density μ, should be chosen such that X is a grid point, that is, μX is an integer. This is because in (5), the quantity $Y_0 + \Sigma(P + 1 - n) Y_n$ is an integer, so if μV_0 is to be an integer, μX must in general also be an integer.

[2]This is substantiated by considering limit cycles with periods up to several hundred and limit cycles for other constant inputs. These limit cycles were not generated exhaustively, but found with the method described in Section III-A3.

[3]This also holds for the more general case of multibit quantization.

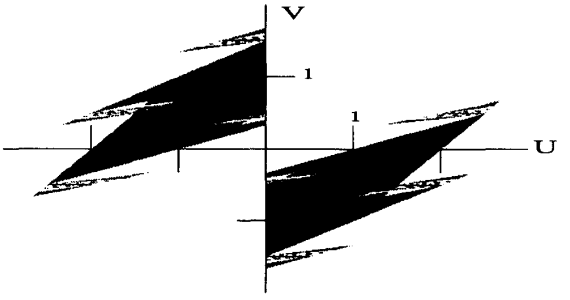


Fig. 4. Plot of the points in state space which lie on a limit cycle for a constant input of $X = 0$. The horizontal and vertical axes represent the state variables U_n and V_n, respectively. The grid density is 80, and each axis tick represents one unit.

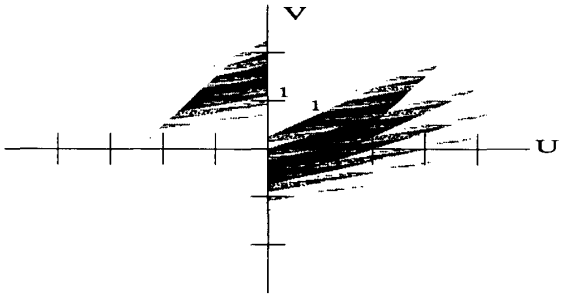


Fig. 5. Plot of the points in state space which lie on a limit cycle for a constant input of $X = 0.5$. The horizontal and vertical axes represent the state variables U_n and V_n, respectively. The grid density is 50, and each axis tick represents one unit.

Figs. 4 and 5 show state space plots obtained for constant inputs $X = 0$ and $X = 0.5$, with grid densities 80 and 50, respectively. The limit cycles which make up the plots have periods that vary from two to many hundred. The maximum state variable values reported for short limit cycles in Table I match the plots well. As the constant input increases, the limit cycles result in larger values of the state variables. It turns out that for a fixed constant input, the state space plots for different, but sufficiently large grid densities look very similar. On the other hand, the periods of the limit cycles which make up the plots can be quite different, and not all limit cycles materialize for an arbitrary grid density. This partly explains that the collection of limit sets appears ragged and irregular. The periods tend to share a number of prime factors with the grid density, even though we found no general rule.

To verify the above results based on limit cycles, Fig. 6 shows simulation results for the maximum value of the quantizer input U_n as a function of the constant input X. The results are obtained by using a large number of random starting points in state space, simulating the modulator for a large number of time steps, and registering the largest quantizer input value. We discard maxima occurring on the first 1000 time steps to get over transients, so that the comparison with limit cycle results is justified; this is not to deny the importance of transients. For inputs $X = 0$ and 0.5 there is good agreement with the limit

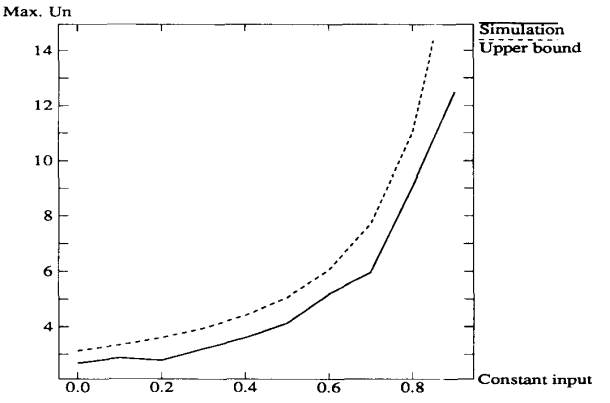


Fig. 6. Maximum value of quantizer input U_n as a function of the constant input X. The lower curve shows simulated values, while the upper curve is the analytical bound (12) derived in Section III-B.

cycle based results in Table I and Figs. 4 and 5, and the agreement is confirmed for other constant inputs. Fig. 6 shows that as X approaches unity, the maximum value of U_n begins to increase rapidly, that is, the modulator becomes less stable and more prone to clipping.

B. Design Implications of Bounds on Limit Cycles

In Section III-B1 we present analytical upper bounds on the limit cycle amplitudes for the general, unclipped double loop modulator with scaling factors B, G, b, G shown in Fig. 3. In Section III-B2 we use the results to propose a modulator with scaling factors that are optimized with respect to both stability and an approximate measure of SNR, and we compare the performance of the clipped modulator to previously suggested scaled modulators.

1) Derivation of Bounds: Our upper bounds on limit cycle amplitudes are derived in Appendix A. To concisely express them, it is convenient to first transform the double loop modulator in Fig. 3 into an equivalent modulator, shown in Fig. 7 and specified in terms of normalized quantities. The equivalence can be confirmed by a series of block diagram manipulations and the observation that the gain g only has the effect of scaling U_n. This is because $Q(U_n)$ only depends on the sign of U_n. To summarize the results in Appendix A, we introduce the normalized quantities

$$\xi = \frac{X}{B}, \quad \phi = GB, \quad \gamma = \frac{2b}{GB} = \frac{2b}{\phi} \tag{9}$$

where X is the constant input, and b, B, and G are defined as in Fig. 3. These three normalized quantities replace X, G, and b as independent variables, so we change our set of independent variables from $(X; G, B, g, b)$ to $(\xi; \phi, B, g, \gamma)$. Some immediate observations on the effects of changing the values of the original variables can be made based on the new set of variables: Multiplying b and G by the same factor only has the effect of increasing ϕ, so the signal levels of U_n and V_n are scaled proportionally. This is equivalent scaling. However, multiplying only G

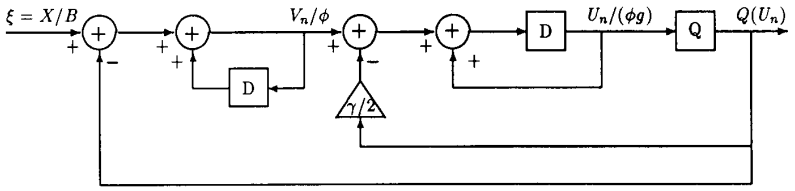


Fig. 7. Discrete-time model of transformed double loop ΣΔ modulator.

by a factor has the effects of scaling ϕ proportionally and scaling γ in inverse proportion. The change in γ affects the signal levels nonlinearly, since the functionality of the modulator is modified. Similarly, changing only B has effects on ξ, ϕ, and γ, and the combined effect on signal levels is nonlinear. These are examples of functional scaling. We show in Appendix A that assuming

$$0 \le \xi < 1 \quad \text{and} \quad \gamma > 1$$

the limit cycle amplitudes of the state variables U_n, V_n are upper bounded by

$$\frac{8|U_n|}{\phi g} \le \begin{cases} \dfrac{\left[(\gamma-1)(1-\xi)+\dfrac{2\gamma}{\gamma-1}\right]^2}{1-\xi} \\ \quad \text{for } 1 < \gamma \le f_1(\xi) \\ \dfrac{16+\gamma^2(1-\xi)^2}{1-\xi} \\ \quad \text{for } f_1(\xi) < \gamma \le f_2(\xi) \\ \dfrac{[\gamma(1+\xi)+2]^2}{1+\xi} \\ \quad \text{for } f_2(\xi) < \gamma \end{cases} \tag{10}$$

$$\frac{2|V_n|}{\phi} \le \begin{cases} \gamma\left(\xi+\dfrac{2}{\gamma-1}\right) \\ \quad \text{for } 1 < \gamma < f_1(\xi) \\ (\gamma+1)\xi+3 \\ \quad \text{for } 1 < \gamma < f_1(\xi) \end{cases} \tag{11}$$

where

$$f_1(\xi) = 1 + \frac{2}{1+\xi},$$

$$f_2(\xi) = \frac{-1+\sqrt{1+2\xi\left(\dfrac{4}{1-\xi}-\dfrac{1}{1+\xi}\right)}}{\xi}.$$

We denote the bounds on $|U_n|$ and $|V_n|$ in (10) and (11) by $U_{\max}$ and $V_{\max}$, respectively. Similar bounds hold for $-1 < \xi \le 0$. For the standard double loop modulator with $b = B = g = G = 1$ or equivalently $\phi = 1$ and $\gamma = 2$, we find in particular that for $0 \le X < 1$,

$$U_{\max} = \frac{(5-X)^2}{8(1-X)}, \quad V_{\max} = X + 2. \tag{12}$$

Fig. 6 shows the bounds in (12) as well the maximum signal levels actually observed in a simulation of the standard double loop modulator. The simulated results are obtained by choosing 500 random pairs of initial states, running the modulator for 1000 samples to get over transients, and observing the largest state variables on the following 1000 samples. The bounds correspond to $\gamma = 2$, $\phi = 1$, and are seen to be relatively tight.

We find in general that for $1 < \gamma < 2$, the derived bounds on state variables are valid, but not extremely tight. For $\gamma > 2$ the bounds are tighter, especially for moderate and large constant inputs ξ, and are thus suitable for design. Interestingly, the analytical bounds that are valid for $1 < \gamma \le f_1(\xi)$ are very good general approximations to simulated maximum state variable values, even when $\gamma > f_1(\xi)$. When $\gamma < 1$, the technique in Appendix A does not yield upper bounds on limit cycle amplitudes. This of course does not imply that the modulator is unstable. However, it is interesting to compare with [16], where it is claimed that under a number of approximations the double loop modulator with $b = B = 1$ is stable provided $G < 2$ or, equivalently, $\gamma > 1$.

2) Design Implications: In this section we use the results of Section III-B1 to design a double loop modulator that takes into account both stability and SNR performance. The design is based on the constant input assumption, but simulation results for sinusoidal inputs are shown to verify the design. We assume that our circuit technology dictates a given clipping level L defined in (1).

The design problem has five degrees of freedom, namely ($X_{\max}$; G, B, g, b) where $X_{\max}$ is the largest constant input for which the design guarantees absence of clipping. Equivalently, we can use the parameters ($\xi_{\max}$; ϕ, B, g, γ) where $\xi_{\max} = X_{\max}/B$. We reduce the number of degrees of freedom to two by the following three equality constraints:

$$U_{\max} = L, \quad V_{\max} = L, \quad B(1+\xi_{\max}) = L. \tag{13}$$

The first two constraints are stability constraints to avoid saturation, and the last constraint states that the maximum signal level at the output of the input summer, $B + X_{\max} = B(1+\xi_{\max})$, should also equal the clipping level. The constraints ensure that we take maximum advantage of the dynamic ranges of the circuit elements while maintaining stability.

We use the remaining two degrees of freedom, $\xi_{\max}$ and γ, to find a tradeoff between two goals, namely, maximizing an approximate measure of SNR performance and

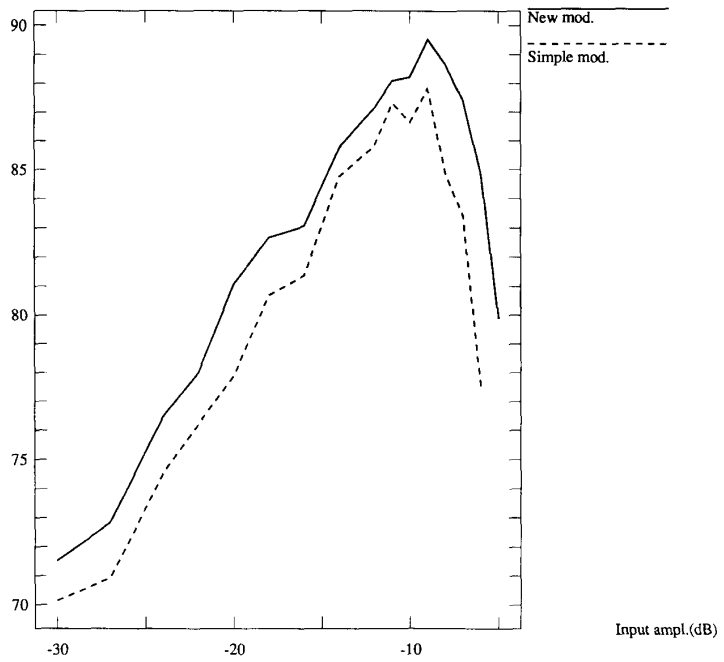


Fig. 8. Comparison of two differently scaled double loop modulators with sinusoidal input. The curve labeled "Simple mod." represents scaling factors $(G, B, g, b) = (0.5, 1, 0.5, 1)$, while the other curve represents scaling factors $(G, B, g, b) = (0.64, 1.12, 0.54, 0.85)$.

maximizing the largest constant input $X_{\max}$ which does not saturate the integrators. The SNR measure that we use is the product gG which is shown in Appendix B.1 to approximately control the baseband noise suppression. We show in Appendix B.2 that for a given $\xi_{\max} \in (\sqrt{5} - 2 = 0.2361, 1)$, the product gG is maximized by choosing $\gamma = 1 + 2/(1 + \xi_{\max})$. We are thus left with a single degree of freedom, $\xi_{\max}$, on which both the product gG and the largest allowed constant input given by

$$X_{\max} = L \cdot \xi_{\max}/(1 + \xi_{\max})$$

depend. Fig. 17 shows gG and $X_{\max}/L$ as functions of the normalized maximum input $\xi_{\max}$. We see that the design goals of maximizing both quantities are conflicting. As a compromise, we choose $\xi_{\max} = 0.5$; this choice is commented on in section III-B3. Using (10), (11), and (13), we then find $\gamma = 2.33$, $\phi = 0.429L$, and the scaling factors

$$B = \frac{L}{1 + \xi} = 0.667L,$$
$$g = \frac{L}{\phi} \cdot \frac{8(1 - \xi)}{16 + \gamma^2(1 - \xi)^2} = 0.538$$
$$G = \frac{\phi}{B} = 0.643, \quad b = \frac{\gamma\phi}{2} = 0.500L \qquad (14)$$

which results in a maximum permissible constant input of $X = B\xi = 0.333L$, and a performance product of $gG = 0.346$. For a given clipping level L, (14) presents the designer with a choice of scaling factors that take into account both stability and SNR performance.

3) Design Comparisons: For comparison we mention a few scaling schemes that have appeared in the literature. In [8] and [22] scaling factors of $G = g = 1/2$, $B = b = 1$ are found to make the double loop modulator sufficiently stable with a clipping level of $L = 1.7$. In [24] scaling factors of $G = 1/4$, $g = 1$, $B = 1$, $b = 1/2$ are chosen, although G is split into two gains of $1/2$ each: One before and one after the outer integrator. The clipping level is not reported. These two designs both correspond to $\gamma = 4$ and $gG = 0.25$. In the setup of the present chapter the value of γ would appear to be somewhat large, and the product gG correspondingly small; however, considerations such as ease and regularity of implementation may also have played a part in the described choices.

Fig. 8 shows the simulated SNR performance of a modulator designed with our technique, and a modulator with $B = b = 1$, $G = g = 0.5$, for a clipping level of $L = 1.7$ [8], [22]. The vertical axis shows the SNR, while the horizontal axis represents the amplitude of a sinusoid with a fixed frequency of 1020 Hz relative to a sampling frequency of 1.024 MHz. The amplitude is measured in decibels relative to the level 1. A sinc^3 decimation filter is used for both modulators, and the oversampling ratio is 128. The plot shows that the peak SNR for our modulator is 1.5–2 dB above that of the simpler modulator, and the dynamic range is 2–3 dB larger. The increase in dynamic range is due to our modulator's ability to operate on larger inputs, and indicates a more stable design. The increase in peak SNR may reflect the explicit design with respect to an SNR performance measure, even though the measure is approximate.

Fig. 8 demonstrates that although our modulator design is based on a limit cycle analysis for constant inputs, the results are also useful for dynamic inputs. For instance, reducing ξ_{max} to 0.3 results in a reduction in dynamic range. However, we should bear in mind that the analysis is a worst case one, and that it is not strictly valid for time-varying inputs. For instance, we find numerically that choosing ξ_{max} between 0.5 and 0.7 has little effect on the dynamic range.

IV. Interpolative Modulator

This section discusses stability issues for the interpolative $\Sigma\Delta$ modulator. Section IV-A presents an exact analytical method to determine the existence and amplitude of given limit cycles. Section IV-B addresses the problem of finding the maximum amplitudes of limit cycles without requiring knowledge of their specific form: In Section IV-B1 we derive an approximate result, based on the describing-function approach and aimed at systems with open-loop poles close to the unit circle, and in Section IV-B2 we present a numerical method for finding upper bounds on limit cycle amplitudes. Throughout results are exemplified using the fourth-order interpolative modulator introduced in [5], often referred to below as "the fourth-order modulator." The section is similar in form to Section III, but significant differences in the methods and results will be pointed out.

Fig. 9 shows a discrete-time model of the interpolative $\Sigma\Delta$ modulator, consisting of an arbitrary discrete-time filter $H(z)$ embedded in a nonlinear negative feedback loop including also a one-bit quantizer Q. The poles of $H(z)$ may be inside, on or outside the unit circle. To avoid race-around the filter must contain at least one delay. For simplicity no scaling is performed in the feedback path. The quantizer can be viewed as adding a noise sequence $\{E_n\}$ with z-transform $E(z)$ to its input sequence; assuming for a moment that the input sequence $\{X_n\}$ and the noise sequence are independent, the signal and noise transfer functions are

$$H_X(z) = \left.\frac{Y(z)}{X(z)}\right|_{E(z)=0} = \frac{H(z)}{1 + H(z)}$$

$$H_E(z) = \left.\frac{Y(z)}{E(z)}\right|_{X(z)=0} = \frac{1}{1 + H(z)}. \qquad (15)$$

$H(z)$ is chosen to have large gains over a passband corresponding to the frequency range in which the input signal is concentrated, and to have small gains outside of passband. As a result, the signal and noise transfer functions are low pass and high pass, respectively. For the special case

$$H(z) = \frac{z^{-1}}{1 - z^{-1}}$$

the interpolative modulator reduces to the single loop modulator shown in Fig. 2.

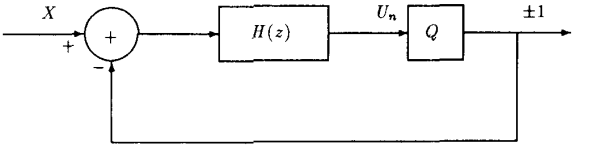


Fig. 9. Discrete-time model of the interpolative $\Sigma\Delta$ modulator.

The main advantage of choosing a higher order $H(z)$ is an improved tradeoff between OSR and SNR, as measured in decibels/octave. In addition, a higher order $H(z)$ is usually designed for a specific OSR and input bandwidth: Equation (15) shows that a desirable $H(z)$ is a very sharp low-pass filter which cuts off immediately above the signal bandwidth. However, higher order loops have problems that are not shared by lower order ones. One such problem is that they require specially designed decimation filters which may require significant chip area and power consumption [9], [28]. Another problem is that higher order designs appear to be inherently plagued with the potential for large-amplitude low-frequency oscillations.[4] These may be detrimental to performance because they can drive the modulators into sustained modes of integrator saturation.

The occurrence of large oscillations is not predictable from the simple linearized equations (15) which indicate that the modulator specified by $H(z)$ is stable if and only if the zeros of $1 + H(z)$ are inside the unit circle, and that stability is independent of the initial states of the integrators or the level of dc inputs. In contrast we find empirically that both these factors profoundly affect the behavior of the modulator. In addition, we show in Section IV-B1 that the proximity of the poles of the open-loop filter $H(z)$ to the unit circle can be important to stability. This is despite the fact that these poles do not manifest themselves in $H_X(z)$, and only appear as zeros in the error transfer function $H_E(z)$. An immediate observation demonstrating the importance of the poles of $H(z)$ is that a modulator is guaranteed to be BIBS stable if the poles of $H(z)$ are inside the unit circle. This is because both the quantizer output and the modulator inputs are bounded, implying that the open-loop filter input is bounded. Another such result emphasizing the natural role of these poles is presented in Section IV-A. The shift of focus from the zeros of $1 + H(z)$ to the poles of $H(z)$ is important, especially in view of the fact that interpolative modulators are frequently designed with open-loop poles on the unit circle itself [5], [9].

The limitations of linearized analysis suggest that the phenomenon of large oscillations should be considered from a state space point of view, and that the limit cycle framework set forth in Section II may provide insight. As the setup is more general than that of Section III, we find it convenient to focus on the quantizer input U_n as representative of limit cycle amplitudes, and to not consider amplitudes of oscillations in other internal state variables.

[4]As before, we are referring to oscillations in the unclipped modulators.

This is because the filter $H(z)$ may be realized in various ways leading to different natural choices of state variables, and we wish to separate the problems of realization and transfer function design.

The flow of the proposed design process is as follows: We assume that we are given an unclipped modulator with satisfactory SNR performance, and that there exists a realization of the modulator filter such that its filter coefficients are all of the same order of magnitude. The modulator could therefore potentially be implemented in switched-capacitor technology. We assume that the modulator has stability problems at one or more internal nodes. Finally, we assume that equivalent scaling is insufficient to solve the stability problem, because the resulting scaling factors would result in capacitor ratios that were too large to be implemented in practice. Our goal is to perform functional scaling to make the modulator more stable, so that subsequent equivalent scaling will not result in excessive scaling factors.

A. Detection of Specific Limit Cycles

By way of motivation, consider the fourth-order interpolative modulator with transfer function [5]

$$H(z) = \frac{0.8653 - 2.2692z^{-1} + 2.0064z^{-2} - 0.59714z^{-3} + 0.000035z^{-4}}{1 - 3.99646z^{-1} + 5.992922z^{-2} - 3.996460866z^{-3} + 1.000000433z^{-4}} \cdot z^{-1}. \quad (16)$$

This open-loop filter $H(z)$ has both its two pole pairs on the unit circle in signal baseband, and the filter is realized with a cascade of integrators from which outputs are fed forward and backward. For constant inputs and zero initial integrator outputs, the modulator is reported in [5] to be unstable when $|X| > 0.65 - 0.7$, in the sense that the SNR decreases dramatically. Whether or not the oscillations are in fact bounded, this behavior is undesirable as it limits the dynamic range. Expanding on this we find that if the initial integrator states are all chosen to be 1, large oscillations occur when the constant input exceeds approximately 0.2803. This underscores the influence of the initial states on the system trajectory, and shows that even small inputs may excite large oscillations. It appears difficult to describe exactly the relationship between initial states and open-loop filter that gives rise to large oscillations for various inputs. Therefore, rather than avoiding initial states that might result in large oscillations, it is desirable to design modulators that do not exhibit large oscillations for any initial states. As seen in Section IV-B such designs effectively sacrifice SNR to improve stability.

In this section we describe in more detail the instability problems of interpolative modulators. In Section IV-A1 we present a method which can be used to answer the following questions: Given a P-bit pattern $\vec{Y} = \{Y_0, \cdots, Y_{P-1}\}$, does there exist a constant input X and an initial state for the modulator such that the corresponding output sequence is a periodically repeated version of $\{Y_0, \cdots, Y_{P-1}\}$? If so, what is the amplitude of the limit cycle, as measured at the quantizer input? As in Section III, our method is based on the standard technique of "opening the loop" frequently associated with Tsypkin's name [29]. In Section IV-A2 we present results of applying the method to the fourth order modulator. In Section IV-A3 we discuss the conditions under which the limit cycles of an interpolative modulator are attracting.

1) Existence of Specific Limit Cycles: Consider the general interpolative modulator of order N. We may write its open-loop transfer function as

$$H(z) = \frac{\sum_{n=0}^{N} A_n(z-1)^{N-n}}{(z-1)^N - \sum_{n=1}^{N} B_n(z-1)^{N-n}} \cdot z^{-1} \quad (17)$$

where $A_0, \cdots, A_N$ and $B_1, \cdots, B_N$ are filter coefficients.[5] Let us define an N-dimensional state vector $\vec{S}_n$ for the open-loop filter, where n is the time step. Further, define an $N \times N$ matrix $\boldsymbol{B}$ in terms of the filter coefficients, and an N-dimensional input vector $\vec{L}_n$ depending on the constant input X as well as quantizer outputs, such that the state space representation of the system is of the form

$$\vec{S}_{n+1} = \boldsymbol{B}\vec{S}_n + \vec{L}_n. \quad (18)$$

It is well known from systems theory that there are many representations of this form. For the fourth-order modulator, one such representation is obtained by defining the four-dimensional state vector $\vec{S}_n$ such that

$$\boldsymbol{B} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ B_4 & B_3 & B_2 & 1 + B_1 \end{bmatrix}, \quad \vec{L}_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ X - Q(U_n) \end{bmatrix}.$$

In general, $\vec{L}_n$ depends linearly on the constant input X. Finally, define two N-dimensional vectors $\vec{C}$ and $\vec{D}$, depending on the filter coefficients, such that the filter output satisfies

$$\vec{U}_{n+1} = \vec{C}^T\vec{S}_n + \vec{D}^T\vec{L}_n \quad (19)$$

where T denotes transposition. For instance, the fourth-order modulator with transfer function (17) has

$$\vec{C} = \begin{bmatrix} A_4 + A_0B_4 \\ A_3 + A_0B_3 \\ A_2 + A_0B_2 \\ A_1 + A_0B_1 \end{bmatrix}, \quad \vec{D} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ A_0 \end{bmatrix}.$$

[5]For the fourth-order modulator with transfer function (16), $(A_0, \cdots, A_4) = (0.8653, 1.1920, 0.3906, 0.06926, 0.005395)$ and $(B_1, \cdots, B_4) = (-3.540 \cdot 10^{-3}, -3.542 \cdot 10^{-3}, -3.134 \cdot 10^{-6}, -1.567 \cdot 10^{-6})$.

We now address the question stated above, i.e., does there exist a limit cycle of period P corresponding to the binary output sequence $\vec{Y} = \{Y_0, \cdots, Y_{P-1}\}$? It is convenient to associate with each sequence $\vec{Y}$ a vector $\vec{L}'_n$ which is obtained from $\vec{L}_n$ by replacing $Q(U_n)$ with Y_n. Specifically, $\vec{L}'_n = [0, 0, 0, X - Y_n]^T$ for the fourth-order modulator. To answer the question we must determine whether or not there exists an X such that two conditions hold:

C1) Periodicity, i.e., $\vec{S}_{P+n} = \vec{S}_n$ for all n. From (18), this condition is equivalent to $S_p = S_0$.

C2) Consistency, i.e., as we step through the difference equation (18) using $\{\vec{L}'_n\}$ as the input vectors, we have $Y_n = Q(U_n)$ for all $0 \le n \le P - 1$.

Condition C1 can be used to find the initial state vector $\vec{S}_0$ from (18) as a linear function of X,

$$\vec{S}_P = \boldsymbol{B}^P \vec{S}_0 + \sum_{n=0}^{P-1} \boldsymbol{B}^{P-1-n} \vec{L}'_n.$$

We set $\vec{S}_P = \vec{S}_0$ to enforce condition C1:

$$\vec{S}_0 = (\boldsymbol{I} - \boldsymbol{B}^P)^{-1} \sum_{i=0}^{P-1} \boldsymbol{B}^{P-1-n} \vec{L}'_n \qquad (20)$$

where $\boldsymbol{I}$ is the $N \times N$ identity matrix, and $\boldsymbol{I} - \boldsymbol{B}^P$ is assumed invertible. If $\boldsymbol{I} - \boldsymbol{B}^P$ is singular, a generalization of the technique in Section III-A1 must be used instead.[6] The right-hand side of (20) depends linearly on X. Using $\vec{S}_0$ from (20), we can step through the difference equation (18) of the system, and at each time step use (19) to find U_n as a linear function of X. The sequence $\vec{Y}$ then is a limit cycle output sequence if and only if condition C2 holds, that is, there exists an X such that all the linear inequalities $Q(U_n) = Y_n$ in X can simultaneously be satisfied.[7] The amplitude of the limit cycle, if it exists, is also found by stepping through equations (18) and (19).

Our procedure is similar in spirit to the derivation in Section III-A2 for the double loop modulator. If we were to apply the above technique to this modulator,[8] however, we find that with the state vector $\vec{S}_n = (U_n, V_n)^T$,

$$\boldsymbol{B} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}; \quad \boldsymbol{B}^n = \begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix}, \quad n \ge 1.$$

Therefore $\boldsymbol{I} - \boldsymbol{B}^n$ is singular. This is an indication that for a given X there either exists infinitely many initial states satisfying $\vec{S}_P = \vec{S}_0$ or none at all, confirming the result of Section III-A.

[6]Assuming that the realization of $H(z)$ is minimal, the characteristic polynomial of the matrix $\boldsymbol{B}$ is the denominator of the open-loop filter $H(z)$, so $\boldsymbol{B}$ has the eigenvalue 1 if and only if $H(z)$ has a pole at DC. Since $\det[\boldsymbol{I} - \boldsymbol{B}] = 0 \Rightarrow \det[\boldsymbol{I} - \boldsymbol{B}^P] = 0$, a dc pole of $H(z)$ implies noninvertibility. The complete result is that $\det[\boldsymbol{I} - \boldsymbol{B}^P] = 0 \Leftrightarrow H(e^{2\pi n/P}) = \infty$ for some integer n. Therefore the outlined technique will not work if and only if the open-loop filter has a pole on the unit circle with an argument which is a multiple of $2\pi/P$, where P is the period in question.

[7]The method generalizes easily to multibit quantization.

[8]The argument is somewhat imprecise because the structure of the double loop modulator is inherently different from that of an interpolative modulator, but the argument can be made rigorous.

Another difference between the double loop and the general interpolative modulator is that in the former there is a range of initial states supporting a given limit cycle, but there is only one specific constant input supporting it. In the latter there is also a range of initial states supporting a given limit cycle, but there is only one possible initial state $\vec{S}_0$, specified by (20), for each constant input, and there is a range of possible constant inputs. This difference is due to the finite dc gain $H(1)$ of the open-loop filter for the interpolative modulator: Consider the average input to the open-loop filter over one period,

$$Z = X - \frac{1}{P} \sum_{n=0}^{P-1} Y_n. \qquad (21)$$

If the open-loop filter has infinite dc gain, Z must be zero to maintain the limit cycle, and the situation is analogous to that of the double loop modulator. However, if $H(1)$ is finite, the dc level of the quantizer input sequence is $ZH(1)$. From a time-domain point of view the constant input X can be varied around a nominal value without affecting the output sequence, as long as no quantizer input U_n is shifted so much that it changes sign. All other things being equal, it is undesirable that several values of X give rise to indistinguishable output sequences, since it implies that any decimation filter is inherently limited in resolution when the modulator is on such limit cycles.

2) Numerical Results on Specific Limit Cycles: To illustrate the method of Section IV-A1 we consider the fourth-order modulator with transfer function (16) [5], as well as a variation of this modulator in which the poles of the open-loop transfer function have been scaled by a factor of 0.98 to move them inside the unit circle.[9] We find empirically that the limit cycles with the largest amplitudes are the ones with relatively large periods, that is, low frequencies, and that the output sequences on these limit cycles tend to take on the special form of a number of positive bits q followed by a number of negative bits r. We focus on these limit cycles, and characterize them for brevity by pairs of the form (q, r).

For the fourth-order modulator with transfer function (16) we find that many limit cycles with fairly short periods exist, but that limit cycles of the form (q, r) with the periods around 100–140 fail to materialize. As we will see in Section IV-B1, this is the range in which we would expect to find large-amplitude limit cycles, because of the pole frequencies of the transfer function (16). We attribute the absence of these limit cycles to the fact that the poles of $H(z)$ are on the unit circle, so if the input to the open-loop filter contains a sinusoidal component at a pole frequency, the filter output will contain an unbounded oscillation at the pole frequency with linearly increasing amplitude. Although unbounded oscillations do not nec-

[9]In terms of the filter coefficients of (17), $(A_0, \cdots, A_4)$ are unchanged and $(B_1, \cdots, B_4) = -8.347 \cdot 10^{-2}, -6.010 \cdot 10^{-3}, -1.752 \cdot 10^{-4}, -3.053 \cdot 10^{6})$.

essarily occur when $H(z)$ has poles on the unit circle, they are a possibility which manifests itself in the present case.

We next turn to the modulator whose open-loop pole moduli have been reduced by two percent. For this modulator we find a large number of limit cycles of the form (q, r) with period $P = q + r = 117$; Table II summarizes the characteristics of all the ones with more than 50% positive bits, that is, $P/2 \le q \le P$. We find similar results for other periods close to 117. The first two columns of the table show the center and width of the X-interval supporting the limit cycle, while the next two columns show the maximum and minimum values of the quantizer input U_n on the limit cycle. The table shows that as q moves from its smallest to its largest value, the width of the input variable supporting the limit cycle first increases from close to zero, then reaches a maximum and finally returns to zero. The maximum value of the quantizer input U_n follows the same pattern, while the most negative value of U_n is an increasing function of q. As shown in Section IV-B1, the limit cycles in question are close to sinusoidal, so the average of the extremes of U_n is a good estimate of the dc level of the quantizer input. We therefore expect the following quantity to be small:

$$\Delta = \frac{\frac{1}{2}(U_{\max} + U_{\min}) - ZH(1)}{ZH(1)}$$

where Z is defined in (21). The last column of Table II showing Δ confirms that the quantity is small, namely, on the order of 2%. The amplitude of limit cycles of the form (q, r) with period 117 are upper bounded by approximately 1435, which is a disturbingly large number.

3) Attracting Limit Cycles: In this section we show that if the open-loop filter has all its poles inside the unit circle, almost all limit cycles are attracting. More precisely, if we take almost any limit cycle in state space and consider a sufficiently small region around any point on the limit cycle, then for all initial states in the region, the system trajectory will converge to the limit cycle. This follows from the fact that if all points on a limit cycle satisfy $U_n \neq 0$, the collection of Lyapunov exponents for the limit cycle equals the set of eigenvalues of the matrix $\boldsymbol{B}$, or equivalently the poles of $H(z)$ [27]. Therefore, if the poles of $H(z)$ are all inside the unit circle, the limit cycles is attracting or stable. Note that the concept of stability of limit cycles is different from the concept of stability of $\Sigma\Delta$ modulators [27].

The result of the previous paragraph calls for further comparison between the double loop and the interpolative modulator. The double loop modulator only has limit cycles for rational constant inputs, and since its open-loop transfer functions has poles on the unit circle, its limit cycles are not attracting. An interpolative modulator with a stable open-loop transfer function, on the other hand, has limit cycles in many intervals of constant inputs, and its limit cycles are attracting. The intervals include both rational and irrational inputs. These facts may imply that limit cycles play an even greater role for interpolative modulators than for single and double loop modulators.

TABLE II
CHARACTERISTICS OF LIMIT CYCLES OF THE FORM (q, r) WITH PERIOD $P = q + r = 117$ FOR A FOURTH-ORDER INTERPOLATIVE MODULATOR. ALL SUCH LIMIT CYCLES SHOWN WITH MORE THAN 50% POSTIIVE BITS, I.E., $P/2 \le q \le P$. FIRST TWO COLUMNS SHOW CENTER AND WIDTH OF X-INTERVAL SUPPORTING THE LIMIT CYCLE, THE NEXT TWO COLUMNS SHOW MAXIMUM AND MINIMUM VALUES OF QUANTIZER INPUT U_n ON LIMIT CYCLE. THE LAST COLUMN MEASURES DIFFERENCE IN PERCENT BETWEEN ACTUAL DC INPUT TO THE QUANTIZER AND AN APPROXIMATION

q	Average X	$10^3 \times$ Width	max. U_n	min. U_n	Δ(%)
73	0.4645	1.68	1400	−647	−1.5
74	0.4932	3.88	1410	−618	−1.8
75	0.5312	6.10	1419	−588	−1.6
76	0.5486	8.33	1426	−560	−1.8
77	0.5753	10.6	1431	−532	−1.7
78	0.6012	12.7	1434	−504	−1.8
79	0.6265	14.9	1435	−477	−1.8
80	0.6509	17.0	1434	−451	−1.8
81	0.6745	19.0	1431	−425	−1.8
82	0.6973	20.8	1425	−400	−1.8
83	0.7187	21.3	1417	−376	−1.7
84	0.7393	19.1	1401	−350	−1.9
85	0.7596	15.9	1385	−323	−2.0
86	0.7791	12.9	1367	−296	−1.8
87	0.7977	10.0	1347	−271	−2.0
88	0.8154	7.31	1325	−248	−2.0
89	0.8323	4.81	1301	−225	−2.1
90	0.8482	2.52	1275	−203	−2.1
91	0.8633	0.45	1248	−183	−2.0

B. Amplitudes of Limit Cycles

In this section we take a different view on stability issues, and relax the requirement of known limit cycles. In Section IV-B1 we apply the describing-function approach to obtain approximate relationships between characteristics of an open-loop filter $H(z)$ and the corresponding large-amplitude limit cycles. We will see that the approximation is useful for filters with poles inside and close to the unit circle, as is the case for many practical interpolative modulators. In Section IV-B2 we present a numerical method for deriving upper bounds on limit cycle amplitudes.

1) Describing Function Approximation: In this section we use the describing function method to obtain approximate relationships between an open-loop filter $H(z)$ and the corresponding large-amplitude limit cycles. The approximation is valid for filters with their poles close to the unit circle, and gets better as the poles move closer to the unit circle. We will use the analysis to demonstrate design tradeoffs between SNR and stability.

Our approximate describing function approach used in Section IV-B1 bears some resemblance to the work of Ardalan and Paulos [16], who use a frequency-domain approach and model the quantizer as two linearized gains: One for either a dc or a sinusoidal component, and one for the residual which is assumed Gaussian. In contrast, we are specifically interested in situations where instability can occur, and can thus use a relatively simple approach without a Gaussian assumption. In addition, we do not assume that the open-loop transfer function has a pole at dc.

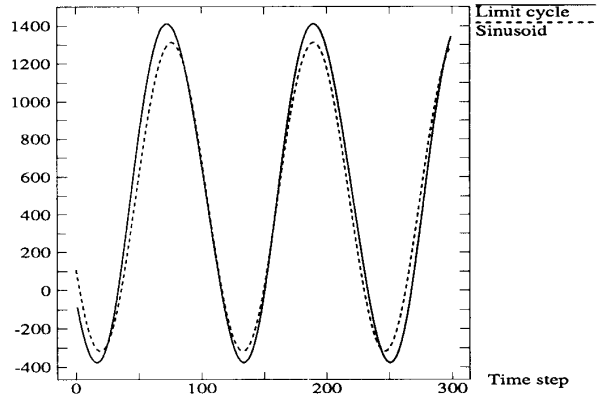


Fig. 10. Limit cycle in the quantizer input sequence U_n for the fourth-order modulator of [5] with pole moduli scaled by 0.98. The constant input is $X = 0.7$. Also shown is the describing function approximation.

TABLE III
SIMULATED AND ANALYTICAL RESULTS FOR LIMIT CYCLE PERIODS AND EXTREME VALUES FOR THE FOURTH-ORDER MODULATOR OF [5] WITH ITS POLE MODULI SCALED BY r. THE CONSTANT INPUT IS CHOSEN TO BE THE SMALLEST INPUT RESULTING IN A LARGE-AMPLITUDE LIMIT CYCLE IF THE INITIAL STATES OF $H(z)$ ARE ALL ZERO. ALSO SHOWN ARE THE DC GAIN AND GAIN MARGIN OF $H(z)$

		LC Period				LC Extremes	
r	X	Sim.	Theor.	k_{DC}	k_{GM}	Sim.	Theor.
0.99	0.7004	127	125	2845	2194	+3700, −1400	+3600, −1400
0.98	0.7839	117	113	1767	808	+1370, −300	+1250, −220
0.97	0.8292	101	96	1004	293	+350, −25	+450, −50

A motivation for using the describing-function method is the observation that large-amplitude limit cycles are often close to sinusoidal. For example, Fig. 10 shows a limit cycle in the fourth-order modulator (16) with its pole moduli scaled by 0.98. Also shown is the result of using the describing function approximation derived below. The maximum value on the limit cycle is predicted to within about 10%.

In applying the describing function method, we assume that the quantizer input sequence is of the form

$$U_n = C - A \sin \omega_0 n, \qquad 0 \le C < A \tag{22}$$

where ω_0 is the frequency of oscillation, and A, C, ω_0 are unknowns which are to be determined for a given open-loop filter $H(z)$ and a given constant input $X \ge 0$. Equation (22) is clearly an approximation, although bounds on its quality may be derived using the methods of [30]; in general terms it is best if higher harmonics of the fundamental frequency ω_0 are highly suppressed by $H(z)$. As a further approximation we consider the problem in continuous time rather than discrete time,

$$U(t) = C - A \sin \omega_0 t, \qquad 0 \le C < A.$$

In Appendix C we derive a method to find the unknown constants A, C, ω_0: The frequency ω_0 is given by $\angle H(e^{j\omega_0}) = -180°$, where the phase of H is reduced to $[-180°, +180°)$; therefore the expected limit cycle period is $P = 2\pi/\omega_0$. The constants A and C are found by solving the following set of nonlinear equations depending only on X, the dc gain $k_{DC} = H(1)$ and the gain margin $k_{GM} = |H(e^{j\omega_0})|$:

$$C = k_{DC}\left(X - \frac{2}{\pi}\arcsin\frac{C}{A}\right)$$

$$A = k_{GM} \cdot \frac{4}{\pi}\sqrt{1 - \left(\frac{C}{A}\right)^2}. \tag{23}$$

It is shown in Appendix C that as a first-order approximation, the nonlinear equations (23) are also valid for determining A and C when the quantizer is two bit or infinite bit, rather than one bit[10]; these quantizers are defined in the Appendix. This indicates that for interpolative modulators with large-amplitude limit cycles, the quantizer resolution does not affect stability; the stability problem is intrinsically linked to the open-loop transfer function.

Returning to one-bit quantization, Table III and IV illustrate the results for variations on the fourth order-modulator (16) [5]: In Table III we scale the pole moduli by a factor r, and in Table IV we instead scale the pole angles by a factor a. The tables show simulated and analyt-

[10]It is found by simulation that using a two-bit or infinite-bit quantizer does not change the amplitude or the period of the resulting large-amplitude limit cycles.

TABLE IV
SIMULATED AND ANALYTICAL RESULTS FOR LIMIT CYCLE PERIODS AND EXTREME VALUES FOR THE FOURTH-ORDER MODULATOR OF [5] WITH ITS POLE ARGUMENTS SCALED BY a, AND ITS POLE MODULI FIXED AT $r = 0.99$. THE CONSTANT INPUT IS CHOSEN TO BE THE SMALLEST INPUT RESULTING IN A LARGE-AMPLITUDE LIMIT CYCLE IF THE INITIAL STATES OF $H(z)$ ARE ALL ZERO. ALSO SHOWN ARE THE DC GAIN AND GAIN MARGIN OF $H(z)$

		LC Period				LC Extremes	
a	X	Sim.	Theor.	k_{DC}	k_{GM}	Sim.	Theor.
1.2	0.7253	104	102	1459	1245	+2100, −800	+2050, −800
1.0	0.7004	127	125	2845	2194	+3700, −1400	+3600, −1400
0.8	0.6928	161	159	6248	4324	+7400, −2600	+7100, −2500

ical results for limit cycle periods and extreme values; the constant input is in each case chosen to be the smallest input resulting in a large-amplitude limit cycle if the initial states of $H(z)$ are all zero. The predictions of the describing function approximation are generally close to the observed amplitudes, and the approximation is better for poles close to the unit circle.

We also show in Appendix C that for a given open-loop filter, the largest value of $U(t)$ for any constant input, that is, the largest limit cycle amplitude equals

$$U_{\max} = \max_X (A + C) = \frac{3\sqrt{3}}{\pi} k_{GM} \doteq 1.65 k_{GM}. \quad (24)$$

This quantity is seen to only depend on the gain margin, not on the dc gain. Although the analysis is approximate, it does suggest the existence of a design conflict: On one hand the gain margin should be kept small to minimize limit cycle amplitudes and thus maximize system stability. On the other hand, a linearized system model suggests that the magnitude of the open-loop transfer function $H(z)$ should be large over all of baseband, including the frequency ω_0, so that the baseband noise suppression and thus the SNR are maximized [5]. In fact, we can use the dc gain as a rough indicator of the SNR, because the dc gain sets the level of the transfer function magnitude in baseband for interpolative modulators such as the one in [5]. Within this setup, the tradeoff is between maximizing the dc gain and minimizing the gain margin.

The tradeoff between stability and SNR performance can be explored in various ways. As an illustration we consider the following problem: For the fourth-order modulator with transfer function (16), how should the pole moduli of the open-loop transfer function be modified to maximize stability, given that only a certain degradation of SNR is acceptable? We consider k_{DC} and k_{GM} to be indicators of the SNR and maximum limit cycle amplitudes, because SNR depends approximately linearly on 20 $\log_{10} k_{DC}$, and $U_{\max} \approx 1.65\, k_{GM}$. Let us call the modulus scaling factors for the lower frequency (LF) and higher frequency (HF) poles r_{LF} and r_{HF}, respectively. For a given dc gain we do the optimization by choosing the pair (r_{LF}, r_{HF}) resulting in the smallest k_{GM} subject to $H(1) = k_{DC}$ and $r_{LF} < 1$.

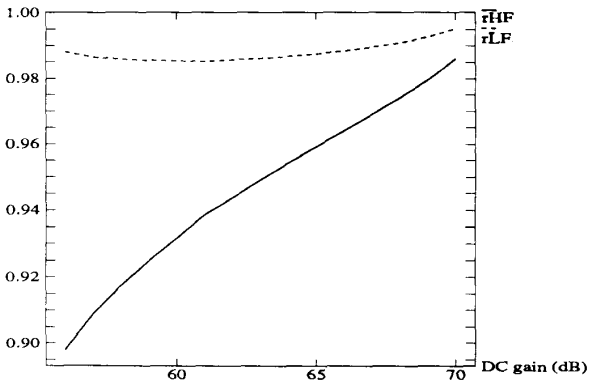


Fig. 11. Figure showing the optimum pole modulus scaling factors for the fourth-order interpolative modulator [5]. For a given DC gain, these scaling factors achieve the smallest gain margin.

Fig. 11 shows the optimum scaling factors as functions of the dc gain; the main observation is that r_{LF} and r_{HF} remain close to 1 for moderate decreases in dc gain from the nominal 70.7 dB. In general the HF pole pair is scaled more than the LF one for a given dc gain. This is intuitively understandable, since the oscillation frequency is close to the frequency of the HF pole. Therefore the most efficient way of reducing k_{GM} is to increase the distance between the HF pole pair and $e^{\pm j\omega_0}$, where ω_0 is the oscillation frequency. Fig. 12 shows the tradeoff between k_{DC} and $U_{\max} = 1.65\, k_{GM}$ obtained by using the optimum scaling factors r_{LF} and r_{HF} from Fig. 11. As an example, if the designer is willing to sacrifice 6 dB of dc gain, the limit cycle amplitudes can be reduced from essentially infinite to approximately 800. This is still a large number, but it can be remedied by equivalent scaling. This way the modulator will not be as heavily clipped as the original modulator, and due to the more stable design, it will be able to leave such saturation modes more gracefully.

2) Bounding the Amplitude: In Appendix A we have derived analytical upper bounds on limit cycle amplitudes for the double loop modulator. The idea is to divide trajectories in state space into two types of segments in time: Positive and negative ones as specified by the sign of the quantizer input. For each segment type we define a po-

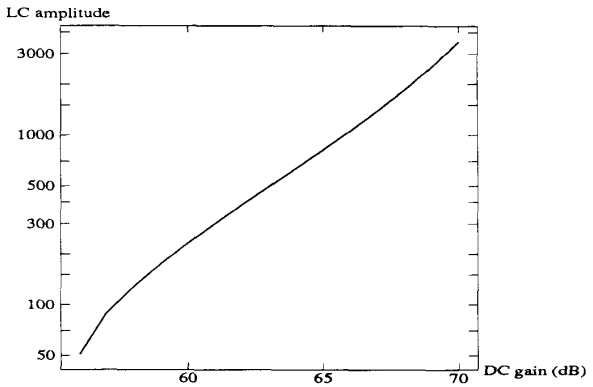


Fig. 12. Figure showing the best achievable tradeoff between DC gain and limit cycle amplitudes. Note the logarithmic axes.

tential related to the maximum values that the state variables can assume on the segment. We derive an upper bound on the potential obtainable on a positive segment as a function of the potential on the previous positive segment; the bound is a monotonic function for all but very small values of the potential. We argue that an upper bound on the positive potential on any limit cycle could be found by considering all initial potentials for which the next potential can be as least as large: The largest such initial potential is the desired upper bound, because potentials exceeding the bound cannot recur and so cannot correspond to limit cycles.

This technique does not immediately generalize to interpolative modulators. The main reason is that solutions to the difference equations are more involved for higher order systems. The idea of positive and negative segments can, however, still be put to use. In this section, we describe a numerical way of finding upper bounds on limit cycle amplitudes, and present numerical results in agreement with the limit cycle results in Section IV-B1. We first describe our method conceptually, then describe its numerical implementation and show resulting bounds on limit cycle amplitudes.

Our approach is to focus directly on the quantizer input U_n, rather than a potential, for a given constant input X. Fig. 13 shows some possible trajectories $\{U_n\}$ as functions of time. We divide all trajectories into positive and negative segments characterized by the sign of U_n as before, and consider the peak values assumed by the quantizer input in each positive segment.[11] The basic idea is to derive an upper bound on the positive peak value as a function of the previous positive peak value. The fact that this function is not in general monotonic is a complicating factor compared to the double loop case, but as detailed below, it is still possible to use the function to find an upper bound on the positive peak value on any limit cycle.

[11]The method can be generalized to multibit quantization, if a positive segment is understood to be a segment for which all $U_n > M$, where M is the smallest number for which $Q(M) = 1$. Negative segments are defined similarly.

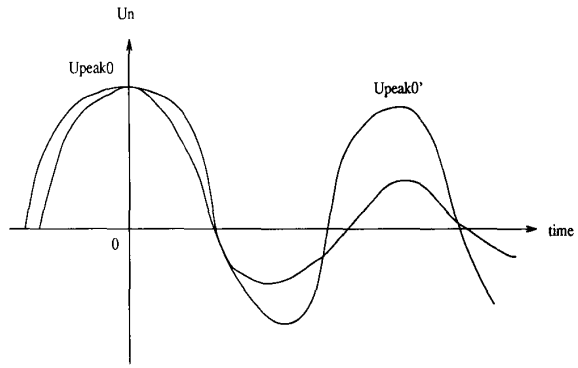


Fig. 13. Some time sequences of the quantizer input $\{U_n\}$ for a given constant input X. Both peak at the value U_{peak0} within the positive segment around time 0. U'_{peak0} is the largest peak value of all trajectories on the next positive segment.

To be more specific, we define a parameter U_{peak} which physically denotes a peak value in the positive segment around time 0. In Fig. 13 a particular value, U_{peak0}, is used. For each value of U_{peak}, we consider all trajectories that have the following two properties:

P1) At some arbitrary time $n = 0$, $U_0 = U_{peak} > 0$.

P2) Within the positive segment to which $n = 0$ belongs, the trajectory peaks at $n = 0$.

Both trajectories in Fig. 13 have these properties with the particular value $U_{peak} = U_{peak0}$. Let us follow all such trajectories forward in time until they reach their next positive segment, and register the largest value of U_n, denoted by U'_{peak}, attained by any such trajectory. In Fig. 13, a particular value $U'_{peak} = U'_{peak0}$ is shown, assuming that none of the other trajectories with properties P1 and P2 have a value of U'_{peak} exceeding U'_{peak0}. We can plot U'_{peak} as a function of U_{peak}, and an example of this is shown schematically in Fig. 14. As follows shortly, it is not necessary for the plot to be monotonically increasing. Our approach only allows us to draw conclusions if there exists some value U_{max} such that for all U_{peak} greater than U_{max}, U'_{peak} is less than U_{peak}, and for all U_{peak} less than U_{max}, U'_{peak} is less than U_{max}; if in particular the plot in Fig. 14 is monotonically increasing, U_{max} is always the intersection of the plot with the 45° line. Geometrically, the two requirements on the plot in Fig. 14 are that there exists some U_{max} such that

R1) For all $U_{peak} > U_{max}$, the plot is below the 45° line.

R2) For $0 \le U_{peak} \le U_{max}$, the plot is completely contained in the square $(U_{peak}, U'_{peak}) \in [0, U_{max}]^2$.

If both requirements on the plot holds, we can conclude that U_{max} is an upper bound on the amplitude of any limit cycle. This follows by contradiction: Assume that some limit cycle has a peak value $U_{peak} > U_{max}$. Then the series of peak values in the subsequent segments must be strictly decreasing as long as the peak values are above U_{max} because of requirement R1, and once the peak value is below U_{max}, it can never again exceed U_{max} because of requirement R2. In particular, the peak value can never

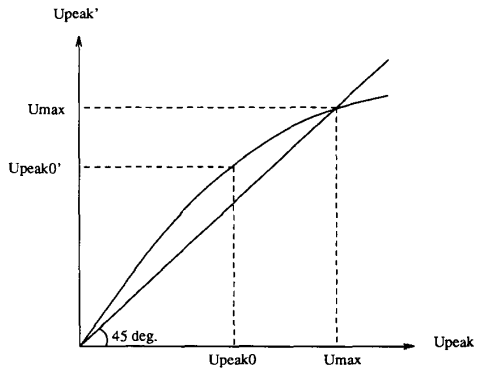


Fig. 14. Stylized plot of the largest possible peak value U'_{peak} as a function of the peak value U_{peak} at the previous positive segment. Also shown is the 45° line. U_{max} is an upper bound on limit cycle amplitudes.

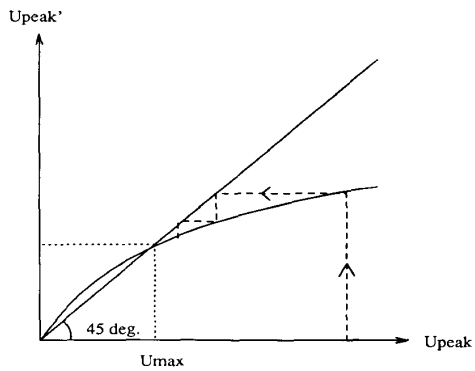


Fig. 15. Demonstration that U_{max} is an upper bound on limit cycle amplitudes. The trajectory marked with arrows is a worst case scenario, because the peak values decrease as slowly as possible. In general, the plot of U'_{peak} versus U_{peak} is an upper bound, obtained only by one or a few trajectories.

again reach U_{peak}. Therefore U_{peak} could not correspond to a limit cycle. The process is illustrated in Fig. 15.

This argument can be further refined. Consider the same approach as before, except that instead of observing the sequence $\{U_n\}$ on the next positive segment, we wait a fixed number s of positive segments. Specifically, U'_{peak} now denotes the largest positive peak on the sth positive segment, and the $s - 1$ intermediate positive peak values are ignored. The exact same graphical technique of plotting U'_{peak} against U_{peak} as in Fig. 14 will then yield an upper bound on U_{peak} on any limit cycle, as shown in Appendix D-1. Using a fixed number $s > 1$ of positive segments in general produces better bounds than using $s = 1$, as shown rigorously in Appendix D-2. Fig. 16 shows some actual upper bound curves for different values of s for a specific fourth order modulator, as described in more detail below. As seen, the curves tend to become flatter as s increases.

We now outline a numerical implementation of the method for bounding limit cycle amplitudes. For a given constant input and each value of U_{peak}, we need to determine U'_{peak} so we can make a plot similar to Fig. 14. We must therefore find all the trajectories with properties P1 and P2. For convenience, we express the system equation (18) as a scalar difference equation of the form

$$\sum_{m=0}^{N} a_m U_{n-m} = cX + \sum_{m=0}^{N} b_m Q(U_{n-1-m}). \quad (25)$$

The coefficients $\{a_m\}$ and $\{b_m\}$ can be obtained directly from the denominator and numerator, respectively, of the open-loop transfer function $H(z)$. Restating our goal, we want to find all trajectories of (25) subject to P1 and P2.

Equation (25) is nonlinear and of order $N + 1$. If we consider the special case where all the $N + 1$ output bits used on the right-hand side of (25) are identical over some period of time, the equation reduces to an Nth order linear difference equation,

$$\sum_{m=0}^{N} a_m U_{n-m} = K \quad (26)$$

where the constant K depends on whether the segment is positive or negative.[12] We will limit our numerical technique to trajectories with at least $N + 1$ positive bits directly before each value of U_{peak} considered.[13] This limitation appears reasonable for the modulators in question: For example, the fourth-order modulators investigated in Section IV-B1 have on the order of 30 positive bits before the positive peaks on their large-amplitude limit cycles.

It is shown in Appendix D-3 that an explicit solution can be found to the simplified equation (26). The solution is valid over the positive segment to which U_{peak} belongs, and depends linearly on N arbitrary constants. Our goal is to select those trajectories that have properties P1 and P2, for a given constant input and a given peak value U_{peak}. It is shown in Appendix D-3 that enforcing constraints P1 and P2 reduces the number of degrees of freedom from N to $N - 2$. The constraints manifest themselves as linear equations in the arbitrary constants. To express the constraints conveniently, trajectories are approximated by continuous-time curves.

Conceptually, we can now generate a figure similar to Fig. 14 by maximizing U'_{peak} over a sufficiently fine grid in the $(N - 2)$-dimensional space of arbitrary constants, for each value of U_{peak} and given constant input X. For $N = 3$ this task is relatively simple, because the maximization is over a one-dimensional space. For $N = 4$ an example is given below. For larger N, more sophisticated search methods in the $(N - 2)$-dimensional space may be necessary.

Since the figure analogous to Fig. 14 is generated entirely numerically, we have no information about the behavior of the curve for values of U_{peak} which are not explicitly investigated. However, if the curve satisfies requirements R1 and R2 up to some large value $U_{big} >$

[12]Incidentally, on such segments the system behavior is determined almost exclusively the $\{a_m\}$ coefficients, that is, by the poles of $H(z)$. The only effect of the $\{b_m\}$ coefficients, that is, the zeros of $H(z)$, is to set the constant value K.

[13]The same argument can be made for $N + 1$ negative bits.

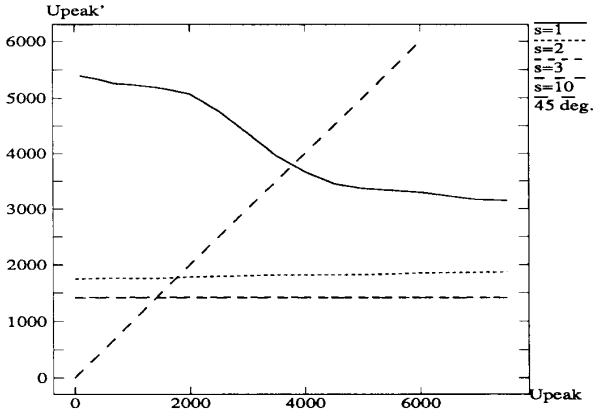


Fig. 16. Actual plot of the U'_{peak} as a function of U_{peak} for a specific fourth-order modulator. Also shown is the 45° line.

U_{max}, we can at least make the following statement: No limit cycles with peak values in the range between U_{max} and U_{big} can exist, that is, if limit cycles with amplitudes above U_{big} exist, they cannot be excited from trajectories with any peak value $U_{peak} \leq U_{big}$. As U_{big} tends to infinity, U_{max} becomes a guaranteed upper bound on amplitudes.

To illustrate our technique, we consider the fourth-order modulator (16) with its pole moduli reduced by 2%. The modulator has pole arguments $\omega_1 = 0.02277$, $\omega_2 = 0.05498$, and pole moduli $r_1 = r_2 = 0.98$. We consider a constant input of $X = 0.70$, so by Table II we know that a limit cycle with amplitude 1425 exists. If we choose the number of positive segments s in our method to be 10, we obtain the results shown as one of the curves in Fig. 16. We can satisfy requirements R1 and R2, at least up to U_{big} which was 10^6 in our case, by choosing $U'_{peak} = 1413$. We conclude that no limit cycles with peak values between 1413 and 10^6 can exist; the upper bound 1413 is about 1% below the observed amplitude 1425 due to the continuous-time approximation made in Appendix D-3, and is thus in good agreement with the observed limit cycle amplitude. The flatness of the plot indicates that the modulator tends to a limit cycle with amplitude 1413 regardless of the initial peak value. The results suggest that the limit cycles shown in Table II are the ones with largest amplitudes, and are thus in agreement with the derivations in Section IV-B1.

V. Summary and Conclusions

We have suggested a framework for stability analysis of ΣΔ modulators, and we have argued that limit cycles for constant inputs are natural objects to investigate in this context. We have presented the following analytical and approximate techniques to aid analysis and design of ΣΔ modulators:

1) For the double loop modulator, we have presented a variation on a standard technique to determine the existence and amplitudes of specific limit cycles. We have introduced a graphical state space approach which provides intuition in itself, and used it to derive analytical upper bounds on limit cycle amplitudes in a general, scaled double loop modulator with constant input. This led to a suggestion for improved design of scaling factors so as to maximize stability as well as a measure of SNR performance. Simulations indicated that the peak SNR of the resulting modulator is about 1–1.5 dB above that of another often used double loop modulator, and the dynamic range is about 2 dB greater.

2) For the interpolative modulator, we presented another variation on a Tsypkin-type method to determine existence and amplitudes of specific limit cycles. We used the describing function method to approximately quantify an inherent tradeoff between SNR performance and stability, and used this to suggest a way of functional scaling which leads to as stable modulators as possible for a given acceptable degradation in SNR. We argued that remaining stability problems could be fixed with subsequent equivalent scaling. We also presented a numerical counterpart to the analytical derivation of upper bounds on limit cycle amplitudes for the double loop modulator.

In concluding this work, we emphasize that there is more to be discovered about the important stability aspect. Although our limitations of focus appear to produce useful results, further research and experiments are needed to strictly verify their validity; transient as well as possibly chaotic behavior may require attention.

Appendix A
Upper Bounds on Double Loop Limit Cycles

In this Appendix we derive the upper bounds stated without proof in Section III-B. An outline of the proof is as follows: We divide the state space of pairs (U_n, V_n) into two half planes according to the value of $Q(U_n)$. Any trajectory, that is, any sequence of state variable pairs $\{(U_n, V_n)\}$, can be decomposed into a number of trajectory segments over which the quantizer outputs are identical. A segment is referred to as positive or negative depending on whether the quantizer outputs all equal $+1$ or -1 on that segment, that is, on whether the segment lies wholly in the positive or negative half plane. In each half plane, the difference equations describing the modulator can be solved in closed form. Each segment defines a number, called the potential, which is preserved for all state variable pairs on that segment. The potential is referred to as the positive or negative potential depending on the segment type; the potential itself can be defined so that it is always a nonnegative number. We derive an analytical expression for the minimum and maximum potential which can occur in one half plane as a function of the potential in the other. Using this information, we derive bounds on the positive potential as a function of the positive potential on the previous positive segment, and similarly, we derive bounds on the negative potential as a function of the previous negative potential.

1. Potential Bounds

We assume without loss of generality that the normalized constant input ξ satisfies $0 \le \xi < 1$. We also assume $\gamma > 1$, that is, the internal stabilizing feedback is not too weak. The standard double loop modulator corresponds to $\gamma = 2$. The difference equations describing the normalized double loop modulator are

$$U_n = U_{n-1} + gV_{n-1} - \frac{g\gamma\phi}{2} Q(U_{n-1})$$

$$V_n = V_{n-1} + \phi\xi - \phi Q(U_n). \tag{27}$$

Consider a segment on which a number of quantizer outputs are constant, $Q(U_0) = Q(U_1) = \cdots = Q(U_n) = a$, where $a = \pm 1$. Then the last equation in (27) yields

$$V_n = V_0 - n\phi(a - \xi) \Rightarrow n = \frac{V_0 - V_n}{\phi(a - \xi)}.$$

The first equation in (27) yields

$$\begin{aligned} U_n &= U_0 - n\frac{g\gamma\phi}{2}a + g\sum_{i=0}^{n-1} V_i \\ &= U_0 - n\frac{g\gamma\phi}{2}a + g\left(nV_0 - \frac{1}{2}n(n-1)\phi(a-\xi)\right) \\ &= U_0 + g\frac{V_0 - V_n}{\phi(a-\xi)}\left[-\frac{1}{2}\gamma\phi a + V_0 \right. \\ &\quad \left. - \frac{1}{2}\phi(a-\xi)\left(\frac{V_0 - V_n}{\phi(a-\xi)} - 1\right)\right]. \end{aligned}$$

This can be written

$$\begin{aligned} &\phi(a-\xi)U_n + \frac{g}{2}\left[V_n - \frac{\phi}{2}(\xi + a(\gamma - 1))\right]^2 \\ &\quad = \phi(a-\xi)U_0 + \frac{g}{2}\left[V_0 - \frac{\phi}{2}(\xi + a(\gamma-1))\right]^2. \end{aligned}$$

Therefore the following nonnegative quantity, referred to as the potential, is preserved on a segment passing through the point (U_n, V_n) in state space:

$$P_a = \phi(a-\xi)U_n + \frac{g}{2}\left[V_n - \frac{\phi}{2}(\xi + a(\gamma-1))\right]^2. \tag{28}$$

We define the quantity

$$Q_a = \sqrt{\frac{2P_a}{g}}.$$

When there is no possibility of confusion, we also refer to Q_a as the potential. We now consider the transition from a segment characterized by the constant quantizer output a to the other possible quantizer output $c = -a$. If $Q(U_n) = a$ and $Q(U_{n+1}) = c$, we have from (27) that

$$\begin{aligned} U_{n+1} &= U_n + gV_n - \frac{g\gamma\phi}{2}a \\ V_{n+1} &= V_n - \phi(c - \xi). \end{aligned} \tag{29}$$

The potential P_c on the segment characterized by the constant quantizer output c is thus

$$\begin{aligned} P_c &= \phi(a-\xi)U_{n+1} + \frac{g}{2}\left[V_{n+1} - \frac{\phi}{2}(\xi + a(\gamma-1))\right]^2 \\ &= \phi(a-\xi)\left(U_n + gV_n - \frac{g\gamma\phi}{2}a\right) \\ &\quad + \frac{g}{2}\left[V_n - \phi(c-\xi) - \frac{\phi}{2}(\xi + a(\gamma-1))\right]^2. \end{aligned} \tag{30}$$

We can isolate U_n in (28) and substitute it in (30). Rearranging the result, we find that the potential transition from quantizer output a to c obeys the law

$$\begin{aligned} &\frac{(a-\xi)P_c - (c-\xi)P_a}{g} \\ &\quad = \frac{a-c}{2}\left[\left(V_n - \frac{\gamma\phi\xi}{2}\right)^2 \right. \\ &\qquad \left. - \frac{\phi^2}{2}(\gamma+1)^2(a-\xi)(c-\xi)\right]. \end{aligned} \tag{31}$$

Consider the transition from $a = +1$ to $c = -1$. The constraint $a = Q(U_n) = +1$ implies that $U_n > 0$, and (28) therefore implies that V_n must lie in the interval

$$V_n \in (M_1 - Q_1, M_1 + Q_1) \tag{32}$$

where $M_1 = (\xi + \gamma - 1)\phi/2$. The constraint $c = Q(U_{n+1}) = -1$ implies $U_{n+1} \le 0$, which by the first equation in (29) implies a bound on $U_n + gV_n$. The state variable U_n can be isolated in (29), so the bound on $U_n + gV_n$ can be expressed in terms of V_n and P_1 or equivalently Q_1. This bound is

$$V_n \notin (M_2 - Q_1, M_2 + Q_1) \tag{33}$$

where $M_2 = (-\xi + \gamma + 1)\phi/2$. For a given P_1 or Q_1, we will derive upper and lower bounds on the achievable P_{-1} on the negative segment following the positive one. We must therefore consider all values of V_n that satisfy both (32) and (33), and find the value which results in the minimum and maximum values of $(V_n - \gamma\phi\xi/2)^2$ in (31). As $M_2 - M_1 = (1 - \xi)\phi > 0$ and the intervals in (32) and (33) have the same length, there is no value satisfying (32) for which $V_n \ge M_2 + Q_1 + 1$. Therefore (33) reduces to the bound $V_n \le M_2 - Q_1$. This upper bound is weaker than the upper bound $V_n < M_1 + Q_1$ in (32) if and only if $0 \le Q_1 \le (M_2 - M_1)/2 = (1 - \xi)\phi/2$. The lower bound $V_n > M_1 - Q_1$ in (32) is always below both upper bounds $M_1 + Q_1$ and $M_2 - Q_1$. Thus

$$\begin{aligned} &M_1 - Q_1 < V_n < M_1 + Q_1 \\ &\qquad \text{for } 0 \le Q_1 \le (1-\xi)\phi/2 \\ &M_1 - Q_1 < V_n \le M_2 - Q_1 \\ &\qquad \text{for } (1-\xi)\phi/2 < Q_1 \end{aligned}$$

and so

$$N_1 - Q_1 < V_n - \frac{\gamma\phi\xi}{2} < N_1 + Q_1 \quad \text{for } 0 \le Q_1 \le (1-\xi)\phi/2$$

$$N_1 - Q_1 < V_n - \frac{\gamma\phi\xi}{2} \le N_2 - Q_1 \quad \text{for } (1-\xi)\phi/2 < Q_1$$

where $N_1 = M_1 - \gamma\phi\xi/2 = (\gamma - 1)(1-\xi)\phi/2 > 0$, $N_2 = M_2 - \gamma\phi\xi/2 = (\gamma+1)(1-\xi)\phi/2 > N_1$. When both the lower and upper bounds on $V_n - \gamma\phi\xi/2$ are positive, the square of these bounds are the lower and upper bounds on $(V_n - \gamma\phi\xi/2)^2$, respectively. When both the lower and upper bounds on $V_n - \gamma\phi\xi/2$ are negative, the square of these bounds are the upper and lower bounds on $(V_n - \gamma\phi\xi/2)^2$, respectively; note the reverse order. When the lower bound on $V_n - \gamma\phi\xi/2$ is negative and the upper bound is positive, as is the case for $N_1 < Q_1 < N_2$, the square of the bound with largest absolute value is the upper bound on $(V_n - \gamma\phi\xi/2)^2$, and the square of the other bound on $V_n - \gamma\phi\xi/2$ is the lower bound on $(V_n - \gamma\phi\xi/2)^2$. To summarize,

$$\begin{aligned}
&(N_1 - Q_1)^2 < (V_n - \gamma\phi\xi/2)^2 < (N_1 + Q_1)^2 && \text{for } 0 \le Q_1 \le (1-\xi)\phi/2\\
&(N_1 - Q_1)^2 < (V_n - \gamma\phi\xi/2)^2 \le (N_2 - Q_1)^2 && \text{for } (1-\xi)\phi/2 < Q_1 \le N_1\\
&0 \le (V_n - \gamma\phi\xi/2)^2 \le (N_2 - Q_1)^2 && \text{for } N_1 < Q_1 \le (N_1+N_2)/2\\
&0 \le (V_n - \gamma\phi\xi/2)^2 < (N_1 - Q_1)^2 && \text{for } N_1 + N_2)/2 < Q_1 \le N_2\\
&(N_2 - Q_1) \le (V_n - \gamma\phi\xi/2)^2 < (N_1 + Q_1)^2 && \text{for } N_2 < Q_1. \qquad (34)
\end{aligned}$$

Inserting each of the lower bounds on $(V_n - \gamma\phi\xi/2)^2 < (N_1 + Q_1)^2$ in (31), we arrive at the following lower bounds on Q_{-1} as a function of Q_1:

$$\begin{aligned}
&Q_{-1} > \sqrt{[Q_1 - \phi(\gamma-1)]^2 + 2\phi^2\gamma(1+\xi)} && \text{for } 0 \le Q_1 \le N_1\\
&Q_{-1} \ge \sqrt{-\frac{1+\xi}{1-\xi}Q_1 + \frac{\phi^2}{2}(1+\xi)(\gamma+1)^2} && \text{for } N_1 < Q_1 \le N_2\\
&Q_{-1} \ge |Q_1 - \phi(\gamma+1)| && \text{for } N_2 < Q_1.
\end{aligned}$$

The lower bound on Q_{-1} is a decreasing function of Q_1 for $0 \le Q_1 \le \phi(\gamma+1)$. For our purposes, it is sufficient to use the weaker bound

$$\begin{aligned}
&Q_{-1} \ge N_{-2} && \text{for } 0 \le Q_1 \le N_2\\
&Q_{-1} \ge \phi(\gamma+1) - Q_1 && \text{for } N_2 < Q_1 \le \phi(\gamma+1)\\
&Q_{-1} \ge 0 && \text{for } \phi(\gamma+1) < Q_1 \qquad (35)
\end{aligned}$$

where we have defined $N_{-2} = (\gamma+1)(1+\xi)\phi/2$, analogously to the definition of N_2. Note that $N_2 + N_{-2} = \phi(\gamma+1)$. For future use, we also define $N_{-1} = (\gamma - 1)(1-\xi)\phi/2 < N_{-2}$. Inserting each of the upper bounds on $(V_n - \gamma\phi\xi/2)^2 < (N_1 + Q_1)^2$ from (34) in (31), we arrive at the following upper bounds on Q_{-1} as a function of Q_1:

$$\begin{aligned}
&Q_{-1} < \sqrt{[Q_1 + \phi(\gamma-1)]^2 + 2\phi^2\gamma(1+\xi)} && \text{for } 0 \le Q_1 \le (1-\xi)\phi/2\\
&Q_{-1} < \phi(\gamma+1) - Q_1 && \text{for } (1-\xi)\phi/2 < Q_1 \le (N_1+N_2)/2\\
&Q_{-1} < \sqrt{[Q_1 + \phi(\gamma-1)]^2 + 2\phi^2\gamma(1+\xi)} && \text{for } (N_1+N_2)/2 < Q_1.
\end{aligned}$$

For our purposes, it is sufficient to use the weaker bound

$$\begin{aligned}
&Q_{-1} < \phi(\gamma+1) - Q_1 && \text{for } 0 \le Q_1 \le (N_1+N_2)/2\\
&Q_{-1} < \sqrt{[Q_1 - \phi(\gamma-1)]^2 + 2\phi^2\gamma(1+\xi)} && \text{for } (N_1+N_2)/2 < Q_1. \qquad (36)
\end{aligned}$$

Note that the second bound on Q_{-1} in (36) is always above $\phi(\gamma+1) - Q_1$ in its interval of validity. For the potential transition from $a = -1$ to $c = +1$, we can similarly derive upper and lower bounds on the positive potential as a function of the negative potential on the previous negative segment. The results analogous to (35) and (36) are

$$\begin{aligned}
&Q_1 \ge N_2 && \text{for } 0 \le Q_{-1} \le N_{-2}\\
&Q_1 \ge \phi(\gamma+1) - Q_{-1} && \text{for } N_{-2} < Q_{-1} \le \phi(\gamma+1)\\
&Q_1 \ge 0 && \text{for } \phi(\gamma+1) < Q_{-1} \qquad (37)
\end{aligned}$$

and

$$\begin{aligned}
&Q_1 < \phi(\gamma+1) - Q_{-1} && \text{for } 0 \le Q_1 \le (N_{-1}+N_{-2})/2\\
&Q_1 < \sqrt{[Q_{-1} - \phi(\gamma-1)]^2 + 2\phi^2\gamma(1+\xi)} && \text{for } (N_{-1}+N_{-2})/2 < Q_{-1}. \qquad (38)
\end{aligned}$$

Note that the second bound on Q_1 in (36) is always above $\phi(\gamma+1) - Q_{-1}$ in its interval of validity. We now consider a trajectory as it goes from a positive segment over a negative segment back to a positive segment. We call the potentials on the two positive segments Q_1^{old} and Q_1^{new}, respectively, and the potential on the intervening negative segment Q_{-1}^{mid}. We can find upper and lower bounds on Q_1^{new} as a function of Q_1^{old} by using the bounds

(35)–(38). The lower bound on Q_1 in (37) is nonincreasing, so the lower bound on Q_1^{new} corresponds to choosing Q_{-1}^{mid} as large as possible as a function of Q_1^{old} in (36). For $0 \le Q_1^{\text{old}} \le (N_1 + N_2)/2$, the upper bound $\phi(\gamma + 1)$ on Q_{-1}^{mid} is between N_{-2} and $\phi(\gamma + 1)$, so Q_1^{new} is upper bounded by Q_1^{old}. For $Q_1^{\text{old}} > (N_1 + N_2)/2$, we must use a different upper bound on Q_1^{mid} in (36). This upper bound is below N_{-2} if and only if $\xi \ge -1 + 8\gamma/(1 + \gamma)^2$; note that $0 \le -1 + 8\gamma/(1 + \gamma)^2 < 1$ for $\gamma > 1$. Similarly, the upper bound on Q_{-1}^{mid} is above $\phi(\gamma + 1)$ for

$$Q_1 \ge \phi(\gamma - 1) + \phi\sqrt{\gamma^2 - 2\gamma\xi + 1}$$

which always exists for $0 \le \xi < 1$. To summarize, we find that

$$\begin{aligned} Q_1^{\text{new}} &> Q_1^{\text{old}} \quad \text{for } 0 \le Q_1^{\text{old}} \le (N_1 + N_2)/2 \\ Q_1^{\text{new}} &> \min\{N_2, \phi(\gamma + 1) - K_1\} \\ &\qquad \text{for } (N_1 + N_2)/2 \le Q_1^{\text{old}} \le N_3 \\ Q_1^{\text{new}} &\ge 0 \quad \text{for } N_3 < Q_1^{\text{old}} \end{aligned} \tag{39}$$

where we have defined

$$N_3 = \phi(\gamma - 1) + \phi\sqrt{\gamma^2 - 2\gamma\xi + 1}$$

and

$$K_1 = \sqrt{[Q_1^{\text{old}} - \phi(\gamma - 1)]^2 + 2\phi^2\gamma(1 + \xi)}.$$

The upper bound on Q_1^{new} as a function of Q_1^{old} is found in a similar way. The upper bound (38) on Q_1^{new} as a function of Q_{-1}^{mid} is decreasing for $0 \le Q_{-1}^{\text{mid}} \le \max\{\gamma(1 + \xi)\phi/2, (\gamma - 1)\phi\}$, and increasing for $\max\{\gamma(1 + \xi)\phi/2, (\gamma - 1)\phi\} \le Q_{-1}^{\text{mid}}$, so for a given interval of possible values of Q_{-1}^{mid}, the largest value of Q_1^{new} always occurs at one of the endpoints of the interval. We first consider the ways in which we can get $Q_1^{\text{new}} \ge Q_1^{\text{old}}$ for $Q_1^{\text{old}} > N_2$. For $Q_1^{\text{old}} > \phi(\gamma + 1)$, choosing the lower bound on Q_{-1}^{mid} in (35) will not work, because the largest possible value of Q_1^{new} for $Q_{-1}^{\text{mid}} = 0$ is $\phi(\gamma + 1)$. For $(\gamma - \gamma\xi + 2)\phi/2 < Q_1^{\text{mid}} \le \phi(\gamma + 1)$, the lower bound $\phi(\gamma + 1) - Q_1^{\text{old}}$ on Q_{-1}^{mid} is below $(N_{-1} + N_{-2})/2$, so the resulting largest Q_1^{new} is less than Q_1^{old}. For $N_2 < Q_1^{\text{old}} \le (\gamma - \gamma\xi + 2)\phi/2$, however, the lower bound on Q_{-1}^{mid} is above $(N_{-1} + N_{-2})/2$, and so by (38), it is possible to get a value of Q_1^{new} above Q_1^{old}. We can also attempt to get $Q_1^{\text{new}} \ge Q_1^{\text{old}}$ by choosing the upper bound on Q_{-1}^{mid} in (35), and use that value of Q_{-1}^{mid} in (38) to find the largest value of Q_1^{new} achievable in this way. It can be shown that in this way, we can achieve $Q_1^{\text{new}} \ge Q_1^{\text{old}}$ if and only if

$$Q_1^{\text{old}} \le \frac{\phi}{2}\left((\gamma - 1)(1 - \xi) + \frac{2\gamma}{\gamma - 1}\right) \triangleq B_1.$$

We then consider the interval $(N_1 + N_2)/2 \le Q_1^{\text{old}} \le \max\{B_1, B_2\}$ where we have defined $B_2 = (\gamma - \gamma\phi + 2)\phi/2$. The lower bound on the corresponding interval for Q_{-1}^{mid} is achieved at $Q_1^{\text{old}} = \max\{B_1, B_2\}$, because the lower bound on Q_{-1}^{mid} is a nonincreasing function of Q_1^{old}. The upper bound on the corresponding interval for Q_{-1}^{mid} is achieved at either $Q_1^{\text{mid}} = (N_1 + N_2)/2$ or at $Q_1^{\text{mid}} = \max\{B_1, B_2\}$. In the former case, inserting $Q_{-1}^{\text{mid}} = (\gamma + \gamma\xi + 2)\phi/2 \ge (N_{-1} + N_{-2})/2$ in (38) yields the upper bound

$$Q_1^{\text{new}} \le \sqrt{16 + \gamma^2(1 - \xi)^2} \cdot \frac{\phi}{2} \triangleq B_3$$

for $(N_1 + N_2)/2 \le Q_1^{\text{old}} \le \max\{B_1, B_2\}$, and in the latter case, we know from above that $Q_{-1}^{\text{mid}} \le \max\{B_1, B_2\}$.

Finally, consider the interval $0 \le Q_1^{\text{old}} < (N_1 + N_2)/2$. For a given Q_1^{old} in this interval, the possible interval for Q_{-1}^{mid} according to (35) and (36) is a superset of the possible interval for Q_{-1}^{mid} at $Q_1^{\text{old}} = (N_1 + N_2)/2$. Therefore, the largest achievable value of Q_1^{new} for a given Q_1^{old} is the maximum of $C_1 \triangleq \max\{B_1, B_2, B_3\}$ and the value we get by choosing the upper bound for Q_{-1}^{mid} in (36) and the corresponding upper bound for Q_1^{new} in (38). To summarize,

$$\begin{aligned} Q_1^{\text{new}} &\le \max\{C_1, \sqrt{Q_1^{\text{old}} - 2\phi)^2 + 2\phi^2\gamma(1 - \xi)}\} \\ &\qquad \text{for } 0 \le Q_1^{\text{old}} \le \gamma(1 - \xi)\phi/2 \\ Q_1^{\text{new}} &\le C_1 \quad \text{for } \gamma(1 - \xi)\phi/2 < Q_1^{\text{old}} \le C_1 \\ Q_1^{\text{new}} &\le Q_1^{\text{old}} \quad \text{for } C_1 < Q_1^{\text{old}}. \end{aligned} \tag{40}$$

For the potential transition from a quantizer output of -1 over $+1$ back to -1, we can similarly derive upper and lower bounds on the potential on a negative segment, Q_{-1}^{new}, as a function of the potential on the previous negative segment, Q_{-1}^{old}. The results analogous to (39) and (40) are

$$\begin{aligned} Q_1^{\text{new}} &> Q_{-1}^{\text{old}} \quad \text{for } 0 \le Q_{-1}^{\text{old}} \le (N_{-1} + N_{-2})/2 \\ Q_{-1}^{\text{new}} &> \min\{N_{-2}, \phi(\gamma + 1) - K{-1}\} \\ &\qquad \text{for } (N_{-1} + N_{-2})/2 \le Q_{-1}^{\text{old}} \le N_3 \\ Q_{-1}^{\text{new}} &\ge 0 \quad \text{for } N_{-3} < Q_{-1}^{\text{old}} \end{aligned} \tag{41}$$

where we have defined $N_{-3} = \phi(\gamma - 1) + \phi\sqrt{\gamma^2 + 2\gamma\xi + 1}$ and

$$K_{-1} = \sqrt{[Q_{-1}^{\text{old}} - \phi(\gamma - 1)]^2 + 2\phi^2\gamma(1 + \xi)}$$

and

$$\begin{aligned} Q_{-1}^{\text{new}} &\le \max\{C_{-1}, \sqrt{Q_{-1}^{\text{old}} - 2\phi)^2 + 2\phi^2\gamma(1 - \xi)}\} \\ &\qquad \text{for } 0 \le Q_{-1}^{\text{old}} \le \gamma(1 + \xi)\phi/2 \\ Q_{-1}^{\text{new}} &\le C_{-1} \quad \text{for } \gamma(1 - \xi)\phi/2 < Q_{-1}^{\text{old}} \le C_{-1} \\ Q_{-1}^{\text{new}} &\le Q_{-1}^{\text{old}} \quad \text{for } C_{-1} < Q_{-1}^{\text{old}}. \end{aligned}$$

where we have defined

$$B_{-1} = \frac{\phi}{2}\left((\gamma - 1)(1 + \xi) + \frac{2\gamma}{\gamma - 1}\right)$$

$$B_{-2} = \frac{\phi}{2}(\gamma + \gamma\xi + 2), \quad B_{-3} = \frac{\phi}{2}\sqrt{16 + \gamma^2(1 - \xi)^2}$$

and $C_{-1} = \max\{B_{-1}, B_{-2}, B_{-3}\}$. We now show that on any limit cycle, the positive potential is bounded by $Q_1 \le C_1$ and the negative potential is bounded by $Q_{-1} \le C_{-1}$. Consider first the positive potential, and suppose that a positive potential $Q_1 = A_1 > C_1$ is the largest positive

potential occurring on a limit cycle. The limit cycle must also contain a positive potential A_2 for which the upper bound on Q_1^{new} in (40) equals A_1 when $Q_1^{old} = A_2$. Equation (40) shows that A_2 cannot be between $\gamma(1 - \xi)\phi/2$ and C_1, because $Q_1^{new} \leq C_1$ in this interval of Q_1^{old}. Nor can A_2 be above C_1, because $Q_1^{new} < Q_1^{old}$ when $Q_1^{old} > C_1$, so A_2 would have to be above A_1, and we have assumed that A_1 is the largest positive potential occurring on a limit cycle. Therefore A_2 must be between 0 and $\gamma(1 - \xi)\phi/2$.

To get $A_1 > C_1$, the negative potential Q_{-1}^{mid} generating $Q_1^{new} = A_1$ must be above B_{-1}. We define D_{-1} to be the smallest negative potential Q_{-1}^{mid} exceeding B_{-1} such that the next positive potential Q_1^{new} can reach A_1. Since we must have $Q_1^{old} \leq \gamma(1 - \xi)\phi/2$, the largest value of Q_1^{old} for which $Q_1^{new} = A_1$ is achievable, must satisfy $Q_1^{old} \leq \phi(\gamma + 1) - D_{-1}$, so in particular, $A_2 \leq \phi(\gamma + 1) - D_{-1}$. We will now show that such values of Q_1^{old} could in turn only be generated by values of Q_1 that exceed A_1, thus violating the assumption that A_1 is the largest occurring positive potential.

To see this, consider (41). To generate a positive potential $Q_1^{new} \leq \gamma(1 - \xi)\phi/2$, the previous negative potential must satisfy $Q_{-1}^{mid} > \phi(\gamma + 1) - Q_1^{new} \geq (\gamma + \gamma\xi + 2)\phi/2$. We are trying to generate a positive potential $Q_1^{new} \geq \phi(\gamma + 1) - D_{-1}$, so the previous negative potential must satisfy $Q_{-1}^{mid} > D_{-1}$. If such a large negative potential is generated by a positive potential exceeding $\gamma(1 - \xi)\phi/2$, the negative potential must also exceed B_1; we define D_1 to be the smallest positive potential Q_1^{old} for which the next negative potential Q_{-1}^{mid} equals D_{-1}. Now $D_1 > B_1$, and $Q_1^{old} = D_1$ maps into $Q_{-1}^{mid} = D_{-1}$ using the upper bound in (36), and Q_{-1}^{mid} maps into $Q_1^{new} = A_1$ using the upper bound in (38). Therefore $D_1 > A_1$, proving that if the negative potential D_{-1} is generated by a positive potential above $\gamma(1 - \xi)\phi/2$, the same positive potential must also exceed A_1. This violates the assumption defining A_1. Alternatively, a negative potential exceeding $\phi(\gamma + 1) - Q_1^{new}$ may be generated for a positive potential Q_1^{old} below $\gamma(1 - \xi)\phi/2$; however, we must then also have $Q_1^{old} < Q_1^{new}$. Using the same argument on Q_1^{old} and going backwards in time by another positive segment, the previous positive potential must either exceed A_1 or be below Q_1^{old}. On the other hand, A_1 is assumed to be a positive potential on a limit cycle, so the positive potential A_1 must recur with a periodicity that corresponds to a finite number of positive segments. Therefore, going backwards in time by a sufficiently large number of positive segments must result in the positive potential $A_1 > \gamma(1 - \xi)\phi/2$ by assumption. This establishes a contradiction, showing that $A_1 > C_1$ cannot be a positive potential on a limit cycle, and that $C_1 = \max\{B_1, B_2, B_3\}$ is an upper bound on positive potentials on limit cycles. A similar argument shows that $C_{-1} = \max\{B_{-1}, B_{-2}, B_{-3}\}$ is an upper bound on negative potentials on limit cycles.

The potential bounds can be written in more detail as follows:

$$\frac{2Q_1}{\phi} \leq \begin{cases} (\gamma - 1)(1 - \xi) + \dfrac{2\gamma}{\gamma - 1} & \text{for } 1 < \gamma \leq 1 + \dfrac{2}{1 + \xi} \\ \sqrt{16 + \gamma^2(1 - \xi)^2} & \text{for } 1 + \dfrac{2}{1 + \xi} < \gamma \leq \dfrac{3}{1 - \xi} \\ \gamma(1 - \xi) + 2 & \text{for } \dfrac{3}{1 - \xi} < \gamma \end{cases} \tag{42}$$

$$\frac{2Q_{-1}}{\phi} \leq \begin{cases} (\gamma - 1)(1 + \xi) + \dfrac{2\gamma}{\gamma - 1} & \text{for } 1 < \gamma \leq 1 + \dfrac{2}{1 + \xi} \\ \gamma(1 + \xi) + 2 & \text{for } 1 + \dfrac{2}{1 + \xi} < \gamma \end{cases}. \tag{43}$$

2. State Variable Bounds

In this section we use the potential bounds (42) and (43) to derive bounds on the largest absolute values of the state variables U_n and V_n on limit cycles for the double loop modulator.

The definition of the potential (28) shows that for a given potential P_a, the largest absolute value of the state variable U_n occurs when $V_n = (\xi + a(\gamma - 1))\phi/2$ and equals $P_a/(\phi|a - \xi|)$. We must consider both the negative and the positive potential in order to arrive at an upper bound on the absolute value of U_n. We find the following result:

$$\frac{8|U_n|}{\phi g} \leq \begin{cases} \dfrac{\left[(\gamma - 1)(1 - \xi) + \dfrac{2\gamma}{\gamma - 1}\right]^2}{1 - \xi} & \text{for } 1 < \gamma \leq f_1(\xi) \\ \dfrac{16 + \gamma^2(1 - \xi)^2}{1 - \xi} & \text{for } f_1(\xi) < \gamma \leq f_2(\xi) \\ \dfrac{[\gamma(1 + \xi) + 2]^2}{1 + \xi} & \text{for } f_2(\xi) < \gamma \end{cases} \tag{44}$$

where

$$f_2(\xi) = \frac{-1 + \sqrt{1 + 2\xi\left(\frac{4}{1-\xi} - \frac{1}{1+\xi}\right)}}{\xi}.$$

An upper bound on the absolute value of V_n on limit cycles can be found as follows. The state equation (27) for V_n shows that on negative segments V_n increases, while on positive segments V_n decreases. Furthermore, V_n decreases when making a transition from a negative to a positive segment, and increases when making a transition from a positive to a negative segment. Therefore the largest positive value of V_n is bounded by the largest positive value of V_n on a negative segment. The definition of the potential (28) shows that this value equals $(\xi - \gamma + 1)\phi/2 + Q_{-1}$. Similarly, the negative value of V_n with the largest absolute value occurs on a positive segment, and this absolute value equals $-(\xi + \gamma - 1)\phi/2 + Q_1$. Choosing the largest of these two absolute values for V_n yields

$$\frac{2|V_n|}{\phi} \le \begin{cases} \gamma\left(\xi + \dfrac{2}{\gamma - 1}\right) & \text{for } 1 < \gamma \le 1 + \dfrac{2}{1+\xi} \\ (\gamma + 1)\xi + 3 & \text{for } 1 + \dfrac{2}{1+\xi} < \gamma. \end{cases} \quad (45)$$

The derived bounds are only upper bounds for two reasons, both related to the discrete nature of the modulator: First, the trajectories in state space may not go through the parabola extremes specified by the potentials. Second, it is in general not possible for a trajectory to attain the maximum negative potential P_{-1} for a given positive potential P_1, and subsequently to obtain the maximum P_1 for that P_{-1}.

Appendix B
Design of Double Loop Modulator

1. Approximate SNR Measure

We will show that the product gG is an approximate measure of the SNR performance of the double loop modulator. Consider a linearized model of the double loop modulator in which the quantizer is viewed as an independent noise source.[14] The transfer function between the z transform of the input and output sequences, $X(z)$ and $Y(z)$, is then

$$H_X(z) = \left.\frac{Y(z)}{X(z)}\right|_{E(z)=0} = \frac{gGz^{-1}}{1 + (gb + gGB - 2)z^{-1} + (1 - gb)z^{-2}};$$

$$H_X(1) = \frac{1}{B} \quad (46)$$

and the transfer function between the z transforms of the noise and output sequences is

$$H_E(z) = \left.\frac{Y(z)}{E(z)}\right|_{x(z)=0} = \frac{(1 - z^{-1})^2}{1 + (gb + gGB - 2)z^{-1} + (1 - gb)z^{-2}}. \quad (47)$$

If the scaling factors are chosen such that the poles of the transfer functions are safely out of baseband, the baseband transfer functions are approximately given by

$$H_X(z) \approx \frac{z^{-1}}{B} \quad (48)$$

$$H_E(z) \approx \frac{(1 - z^{-1})^2}{gGB}. \quad (49)$$

From (48), the noise shaping function of the modulator is seen to be preserved if the approximation leading to (48) is valid. From (49), the product gG adjusts the amount of baseband noise suppression, and hence we will use this product as an approximate measure of SNR to be maximized

2. Design Tradeoff

To perform the design optimization presented in Section III-B2, we will consider ξ and γ to be the two independent variables. Our optimization must take into account the validity regions of the different bounds in (44) and (45), but we first perform some general manipulations. Let us denote the upper bound on $8U_n/(\phi g)$ in (44) by $h_1(\xi, \gamma)$, and the upper bound on $2V_n/\phi$ in (45) by $h_2(\xi, \gamma)$. The constraint $V_{\max}$ implies that

$$\phi = \frac{2L}{h_2(\xi, \gamma)}, \quad G = \frac{\phi}{B} = \frac{2(1 + \xi)}{h_2(\xi, \gamma)}$$

where we have used the constraint $B = L/(1 - \xi)$. The constraint $U_{\max} = L$ implies that

$$g = \frac{8L}{\phi h_1(\xi, \gamma)} = \frac{4h_2(\xi, \gamma)}{h_1(\xi, \gamma)}.$$

Therefore the product gG equals

$$gG = \frac{8(1 + \xi)}{h_1(\xi, \gamma)}.$$

Maximizing the product gG for a fixed value of ξ with respect to γ is equivalent to minimizing $h_1(\xi, \gamma)$. Equation (44) shows that for $\gamma > 1 + 2/(1 + \xi)$, $h_1(\xi, \gamma)$ is

[14] For the purposes of the present discussion, we need not make the assumption of white noise, but an assumption of independence is implicit. As is well known, these assumptions are in many respects inadequate [17].

an increasing function of γ. Therefore it is advantageous to choose γ as small as possible, that is, $\gamma = 1 + 2/(1 + \xi)$, given that γ cannot be below that value. But we still need to consider the range $1 < \gamma \leq 1 + 2/(1 + \xi)$. We can show that $h_1(\xi, \gamma)$ has a minimum at

$$\gamma = 1 + \sqrt{2/(1 - \xi)}$$

which is below $1 + 2/(1 + \xi)$ for $0 \leq \xi < \sqrt{5} - 2 \approx 0.23$. Therefore the optimum choice of γ is

$$\gamma = \begin{cases} 1 + \sqrt{\dfrac{2}{1 - \xi}} & \text{for } 0 \leq \xi \leq \sqrt{5} - 2 \\ 1 + \dfrac{2}{1 + \xi} & \text{for } \sqrt{5} - 2 < \xi < 1 \end{cases} \tag{50}$$

where ξ is understood to be the largest constant normalized input for which we design. We want to maximize the largest unnormalized constant input $X = B\xi$, that is,

$$X = L \cdot \frac{\xi}{1 + \xi} \tag{51}$$

which is equivalent to maximizing ξ. However, we also want to optimize the performance product gG which depends on ξ directly as well as through γ and (50). It can be shown that setting

$$\gamma = 1 + \sqrt{2/(1 - \xi)}$$

results in the product

$$gG = \frac{2(1 - \xi^2)}{[1 + \sqrt{2(1 - \xi)}]^2} \tag{52}$$

whose only extremum between 0 and 1 is a maximum at $\xi \approx 0.3611 > \sqrt{5} - 2$, so for $0 \leq \gamma \leq 1 + \sqrt{2/(1 - \xi)}$, gG is an increasing function of ξ, and within the range $0 \leq \xi \leq \sqrt{5} - 2$, the upper limit on ξ is optimal. For $\xi > \sqrt{5} - 2$, we find that

$$gG = \frac{8(1 + \xi)^3(1 - \xi)}{[4 + (1 + \xi)^2]^2} \tag{53}$$

whose only extremum between 0 and 1 is a rather broad maximum at $\xi = \sqrt{28} - 5 \approx 0.2915$. Fig. 17 shows the product gG as a function of ξ, taking into account the ranges of validity of (52) and (53). The figure also shows X/L from (51) which measures the designed dynamic range. As a tradeoff between the conflicting requirements of maximizing the dynamic range and maximizing the approximate performance measure, we choose $\xi = 0.5$ in Section III-B2, which leads to

$$\gamma = 1 + \frac{2}{1 + \xi} = 2.33,$$

$$\phi = \frac{2L}{(\gamma + 1)\xi + 3} = 0.429L.$$

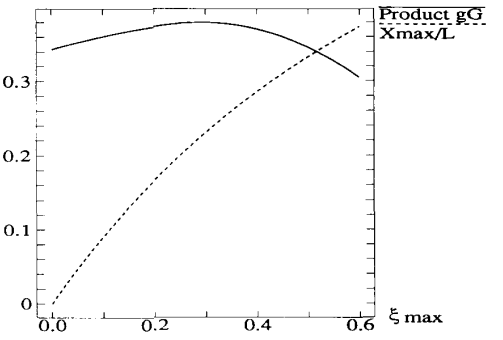


Fig. 17. The product gG is an approximate measure of the SNR performance. The dynamic range is the range of constant inputs permissible according to the design. These are shown as a function of the constant normalized input ξ_{max}.

The corresponding physical scaling factors are

$$B = \frac{L}{1 + \xi} = 0.667L,$$

$$g = \frac{L}{\phi} \cdot \frac{8(1 - \xi)}{16 + \gamma^2(1 - \xi)^2} = 0.538,$$

$$G = \frac{\phi}{B} = 0.643, \quad b = \frac{\gamma\phi}{2} = 0.500L \tag{54}$$

which results in a maximum permissible constant input of $X = B\xi = 0.333L$, and a performance product of $gG = 0.346$. The main result of the appendix is thus that (54) dictates the choice of the scaling factors for a given clipping level L.

Appendix C
Describing Function Method

Consider the setup of Section IV-B1. It can be shown that the dc component and first harmonic of $Q(U(t))$ equal

$$Q(U(t)) = \frac{2}{\pi}\arcsin\frac{C}{A} - \frac{4}{\pi}\sqrt{1 - \left(\frac{C}{A}\right)^2}\sin\omega_0 t + \cdots . \tag{55}$$

If the frequency ω_0 is chosen such that $\angle H(e^{j\omega_0}) = -180°$, and if furthermore all higher harmonics of ω_0 are neglected, we find that the assumption $U(t) = C - A\sin(\omega_0 t)$ is consistent if and only if

$$C = k_{DC}\left(X - \frac{2}{\pi}\arcsin\frac{C}{A}\right)$$

$$A = k_{GM} \cdot \frac{4}{\pi}\sqrt{1 - \left(\frac{C}{A}\right)^2} \tag{56}$$

where we have introduced the dc gain $k_{DC} = H(1)$ and the gain margin $k_{GM} = |H(e^{j\omega_0})|$. It is convenient to introduce the new variable

$$D = \sqrt{1 - \left(\frac{C}{A}\right)^2}, \quad 0 < D \leq 1$$

implying

$$\frac{C}{A} = \sqrt{1 - D^2}, \qquad C^2 = A^2(1 - D^2). \tag{57}$$

Using this and the fact that $\arcsin t = \arccos \sqrt{1 - t^2}$, (56) becomes

$$A = \frac{4}{\pi} k_{GM} D$$

$$D\sqrt{1 - D^2} = g\left(X - \frac{2}{\pi} \arccos D\right), \quad g = \frac{\pi}{4} \frac{k_{DC}}{k_{GM}}. \tag{58}$$

The latter equation can be solved numerically for D, but insight can be gained from a graphical representation. Fig. 18 shows the left- and right-hand sides of (58), termed LHS and RHS, as functions of D; the specific value used for g corresponds to the fourth-order interpolative modulator of [5] with its pole moduli scaled by 0.98, while the constant input is chosen to be $X = 0.7$. The constant C is proportional to the LHS, with proportionality constant $4k_{GM}/\pi$. It can be shown that the LHS reaches its maximum value $1/2$ at $D = \sqrt{2}/2$, and that for any $0 \le X < 1$, $g > 0$, the RHS and LHS have a unique intersection point between 0 and 1. The latter fact follows from the observation that the RHS is only positive for $\cos(\pi X/2) < D \le 1$, and in this interval the derivative of the RHS with respect to D always exceeds the derivative of the LHS.

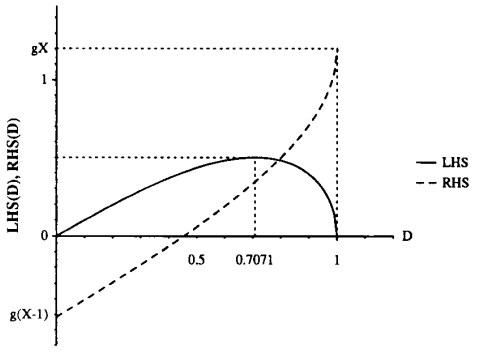


Fig. 18. Figure showing the left and right-hand sides of (58).

From the figure and (57), (58) we observe that for $X = 0$, the constants are given by $D = 1$, $C = 0$, and $A = 4k_{GM}/\pi \doteq 1.27k_{GM}$. As X increases, C increases while D and A decrease; as D passes $\sqrt{2}/2$, the LHS and C begin to decrease along with A. In terms of limit cycle amplitudes, this has the following interpretation: There is a value of X between 0 and 1 which maximizes the amplitude. This is confirmed by Table II. Changing the open-loop transfer function such that k_{GM} decreases will increase g, which decreases D and A; the effect on C depends on the particular value of D.

We next find the largest limit cycle amplitude that can occur for a given open-loop transfer function and a constant input $0 \le X < 1$. We observe that A and C are proportional to the abscissa and ordinate of the LHS curve, respectively, and that in both cases, the proportionality constant is $4k_{GM}/\pi$. The maximum sum of D and $D\sqrt{1 - D^2}$ subject to $0 \le D \le 1$ can easily be shown to occur at $D = \sqrt{3}/2$, and the resulting maximum value of U_n is

$$U_{max} = \max_X (A + C) = \frac{3\sqrt{3}}{\pi} k_{GM} \doteq 1.65k_{GM}.$$

Finally, we consider the case of multibit quantization. Specifically, let a two-bit quantizer be defined by

$$Q_2(U) = \begin{cases} +1 & \text{for } \frac{2}{3} < U \\ +\frac{1}{3} & \text{for } 0 < U \le \frac{2}{3} \\ -\frac{1}{3} & \text{for } -\frac{2}{3} < U \le 0 \\ -1 & \text{for } U \le -\frac{2}{3}. \end{cases}$$

It can be shown that

$$Q_2(C - A\sin(\omega_0 t)) = \frac{2}{3\pi}\left[\arcsin\left(\frac{C - \frac{2}{3}}{A}\right) + \arcsin\frac{C}{A} + \arcsin\left(\frac{C + \frac{2}{3}}{A}\right)\right] - \frac{4}{3\pi}\left[\sqrt{1 - \left(\frac{C - \frac{2}{3}}{A}\right)^2} + \sqrt{1 - \left(\frac{C}{A}\right)^2} + \sqrt{1 - \left(\frac{C + \frac{2}{3}}{A}\right)^2}\right]\sin(\omega_0 t) + \cdots. \tag{59}$$

To first order in $1/A$ and $1/C$, it can be shown that (59) reduces to

$$Q_2(C - A\sin(\omega_0 t)) \approx \frac{2}{\pi}\arcsin\frac{C}{A} - \frac{4}{\pi}\sqrt{1 - \left(\frac{C}{A}\right)^2}\sin\omega_0 t + \cdots$$

which is identical to (55). This implies that if the solution to (56) for the one-bit quantizer has large values of A and C, then approximately the same solution is valid for the two-bit quantizer.

In the limit as the number of bits in the quantizer goes to infinity, the quantizer approaches the saturation characteristic

$$Q_\infty(U) = \begin{cases} +1 & \text{for } 1 < U \\ U & \text{for } -1 \le U \le 1 \\ -1 & \text{for } U < -1. \end{cases}$$

It can be shown that

$$Q_\infty(C - A\sin(\omega_0 t)) = \frac{A}{\pi}\left[\frac{C+1}{A}\arcsin\left(\frac{C+1}{A}\right) - \frac{C-1}{A}\arcsin\left(\frac{C-1}{A}\right) + \sqrt{1-\left(\frac{C+1}{A}\right)^2} - \sqrt{1-\left(\frac{C-1}{A}\right)^2}\right] - \frac{A}{\pi}\left[\arcsin\left(\frac{C+1}{A}\right) - \arcsin\left(\frac{C-1}{A}\right) + \frac{C+1}{A}\sqrt{1-\left(\frac{C+1}{A}\right)^2} - \frac{C-1}{A}\sqrt{1-\left(\frac{C-1}{A}\right)^2}\right]\sin(\omega_0 t) + \cdots. \quad (60)$$

To first order in $1/A$ and $1/C$, it can be shown that (60) reduces to (55). This implies that if the solution to (56) for the one-bit quantizer has large values of A and C, the approximately the same solution is valid for a quantizer with infinitely many bits.

The conclusions for the two-bit and infinite-bit quantizers show that if a modulator with a one-bit quantizer has large-amplitude sinusoidal limit cycles, then the modulator will have the same stability problems even with multibit quantizers.

Appendix D
Proof of Statements in Section IV-B2

1. The General Case $s > 1$

We want to show that using the outlined technique with $s > 1$ positive segments produces an upper bound on limit cycle amplitudes. This can be seen by considering a limit cycle with amplitude U_{peak0}, and dividing it into positive segments. Say that the peak value U_{peak0} occurs every t positive segments. It we start at time zero with a peak value of U_{peak0} and go through any multiple of t positive segments, the peak value U'_{peak0} must again be U_{peak0}. This is true in particular for traversing st positive segments. Consider on the other hand a plot of the form in Fig. 14, only modified by waiting s positive segments rather than only one, before registering U'_{peak} versus U_{peak}. If U_{peak0} is greater than $U_{\max}$, then all subsequent peak values after multiples of s positive segments must be below U_{peak0}. This holds in particular after st positive segments. Putting together the two statements on the peak value after st positive segments, it follows that $U_{\max}$ in the modified figure is an upper bound on limit cycle amplitudes as asserted.

2. Bounds with $s > 1$ Are Tighter

In general, if one value of $U_{\max}$ satisfies requirements R1 and R2, then all values $U > U_{\max}$ also satisfy the requirements. Let $U_{\max}(s)$ denote the infimum of all the possible values of $U_{\max}$ in the two requirements, for a given value of s. We want to show that if $s > 1$, $U_{\max}(s) \le U_{\max}(1)$.

Let us denote the peak values in Fig. 14 by $U_{\text{peak}}(s)$ and $U'_{\text{peak}}(s)$ for a given value of s. Let us further introduce the function f_s for the curve in the figure, $U'_{\text{peak}}(s) = f_s(U_{\text{peak}})$. The maximum value $U_{\max}(s)$ is uniquely determined by f_s. Consider first any $U > U_{\max}(1)$; clearly, $f_1(U) < U$. If $f_1(U) > U_{\max}(1)$, any trajectory with initial positive peak value will have all subsequent peak values below $f_1(U)$. In particular, the trajectory that corresponds to the point $(U, f_s(U))$ in the plot of f_s, must have $f_s(U) < U$. If instead $f_1(U) \le U_{\max}(1)$, then any trajectory with initial positive peak value U will have all subsequent positive peak values $\le U_{\max}(1)$. In both cases, $f_s(U) \le U$, so requirement R1 for f_s is satisfied with $U_{\max} = U_{\max}(1)$. Consider next any $0 \le U \le U_{\max}$. No trajectory with this initial peak value can ever exceed $U_{\max}(1)$. In particular, the trajectory corresponding to the point $(U, f_s(U))$ must have $f_s(U) \le U_{\max}(1)$, so requirement R2 for f_s is again satisfied with $U_{\max} = U_{\max}(1)$. Since $U_{\max}(s)$ is defined as the infimum of all possible values of $U_{\max}$, we have shown that $U_{\max}(s) \le U_{\max}(1)$ for $s > 1$.

3. Solution to Simplified System Equation

If $H(z)$ has a pole of order l at dc, a particular solution to (26) is of the form $U_{n,\text{part}} = k_0 n^l$ where we solve for k_0 by inserting $U_{n,\text{part}}$ in (26).[15] If $H(z)$ has no poles at dc, the particular solution reduces to a constant. The solution to the homogeneous equation obtained by setting $K = 0$ is determined by the poles of $H(z)$. Assume for simplicity that N is even, that all poles are simple with nonzero imaginary part, and denote them by $\{p_m, \bar{p}_m = r_m e^{\pm j\omega_m}: 1 \le m \le N/2\}$. The homogeneous solution can then be written

$$U_{u,\text{hom}} = \sum_{m=1}^{N/2} r_m^n\{k_m \cos(\omega_m n) + l_m \sin(\omega_m n)\} \quad (61)$$

where $\{k_m, l_m: 1 \le m \le N/2\}$ are N arbitrary constants. The complete solution to (26) is then

$$U_n = U_{n,\text{part}} + U_{n,\text{hom}}. \quad (62)$$

We can find the trajectories with property P1 by enforcing the linear constraint $U_0 = U_{\text{peak}}$ on the arbitrary constants. To find a necessary condition for property P2 to approximately hold, we consider the continuous-time function

$$U(t) = k_0 + \sum_{m=1}^{N/2} r_m^n\{k_m \cos(\omega_m t) + l_m \sin(\omega_m t)\}. \quad (63)$$

[15] In fact, $k_0 = (X-1)H(1) < 0$ for positive segments, $k_0 = (X+1)H(1) > 0$ for negative segments.

We can obtain another linear constraint on the arbitrary constants by setting the derivative of $U(t)$ at $t = 0$, $U'(0)$ to zero.[16] To check whether property P2 holds for a trajectory with arbitrary constants satisfying both linear constraints, we must then check whether the trajectory actually attains its segment maximum at $n = 0$, rather than having a minimum, a local extremum or an inflection point. This check is done by running the difference equation backwards and forwards in time form $n = 0$ until $U_n \leq 0$.

ACKNOWLEDGMENT

The authors wish to thank Prof. C. G. Sodini and O. Feely for interesting discussions on the stability of interpolative modulators, and Prof. S. Sanders for suggesting helpful references. They also thank M. Hauser and M. Mar, whose simulator was used to generate Fig. 8.

REFERENCES

[1] F. O. Eynde, G. M. Yin, and W. Sansen, "A CMOS fourth-order 14b 500k-sample/s sigma-delta ADC converter," in *Proc. Int. Solid-State Circuits Conf.*, 1991, pp. 62–63.

[2] E. F. Stikvoort, "Some remarks on the stability and performance of the noise shaper or sigma-delta modulator," *IEEE Trans. Commun.*, vol. 36, pp. 1157–1162, Oct. 1988.

[3] H. Inose and Y. Yasuda, "A unity feedback coding method by negative feedback," *Proc. IEEE*, vol. 51, pp. 1524–1535, Nov. 1963.

[4] J. C. Candy, "A use of double integration in sigma-delta modulation," *IEEE Trans. Commun.*, vol. COM-33, pp. 249–258, Mar. 1985.

[5] K. C. Chao, S. Nadeem, W. L. Lee, and C. G. Sodini, "A higher order topology for interpolative modulators for oversampling A/D converters," *IEEE Trans. Circuits Syst.*, vol. 37, pp. 309–318, Mar. 1990.

[6] K. Uchimura, T. Hayashi, T. Kimura, and A. Iwata, "Oversampling A-to-D and D-to-A converters with multistage noise shaping modulators," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1899–1905, Dec. 1988.

[7] W. Chou, P. W. Wong, and R. M. Gray, "Multistage sigma-delta modulation," *IEEE Trans. Inform. Theory*, vol. 35, pp. 784–796, July 1989.

[8] B. P. Brandt and B. A. Wooley, "A CMOS oversampling A/D converter with 12b resolution at conversion rates above 1 MHz," in *Proc. Solid-State Circuits Conf.*, 1991, pp. 64–65.

[9] P. Ferguson, Jr., A. Ganesan, R. Adams, S. Vincelette, R. Libert, A. Volpe, D. Andreas, A. Charpentier and J. Dattorro, "An 18b 20 kHz dual ΣΔ A/D converter," in *Proc. Int. Solid-State Circuits Conf.*, 1991, pp. 68–69.

[10] G. T. Brauns, R. J. Bishop, M. B. Steer, and J. J. Paulos, "Table-based modeling of delta-sigma modulators using ZSIM," *IEEE Trans. Computer-Aided Design*, vol. 9, pp. 142–150, Feb. 1990.

[11] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*. New York: Academic, 1975.

[12] V. Friedman, "The structure of limit cycles in sigma delta modulation," *IEEE Trans. Commun.*, vol. COM-36, pp. 972–979, Aug. 1988.

[13] J. C. Candy, "A use of limit cycle oscillations to obtain robust analog-to-digital converters," *IEEE Trans. Commun.*, vol. COM-22, pp. 298–305, Mar. 1974.

[14] S. Hein, K. Ibraham, and A. Zakhor, "New properties of sigma delta modulators with DC inputs," *IEEE Trans. Commun.*, vol. 40, pp. 1375–1387, Aug. 1992.

[15] S. J. Park and R. M. Gray, "Sigma delta modulation with leaky integration and constant input." May 8, 1991, preprint.

[16] Enforcing $U'(0) = 0$ is an approximation since the discrete-time sequence $\{U_k\}$ may not peak exactly at the points where $U'(t) = 0$, even if t is an integer.

[16] S. H. Ardalan and J. J. Paulos, "An analysis of nonlinear behavior in delta-sigma modulators," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 593–603, June 1987.

[17] R. M. Gray, "Oversampled sigma-delta modulation," *IEEE Trans. Commun.*, vol. COM-35, pp. 481–488, May 1987.

[18] A. Gersho, "Stochastic stability of delta modulation," *Bell Syst. Tech. J.*, vol. 51, pp. 821–841, Apr. 1972.

[19] P. T. Nielsen, "On the stability of a double integration delta modulator," *IEEE Trans. Commun.*, vol. COM-19, pp. 364–366, June 1971.

[20] R. Steele, *Delta Modulation Systems*. New York: Wiley, 1975, ch. 2.

[21] R. M. Gray, "Spectral analysis of quantization noise in a single-loop sigma-delta modulator with dc input," *IEEE Trans. Commun.*, vol. 37, pp. 588–599, June 1989.

[22] B. E. Boser, "Design and implementation of oversampled analog-to-digital converters," Ph.D. dissertation, Stanford Univ., Oct. 1988.

[23] N. He, F. Kuhlmann, and A. Buzo, "Double-loop sigma-delta modulation with dc input," *IEEE Trans. Commun.*, vol. 38, pp. 487–495, Apr. 1990.

[24] D. B. Ribner, R. D. Baertsch, S. L. Garverick, D. T. McGrath, J. E. Krisciunas, and T. Fuji, "16b third-order sigma delta modulator with reduced sensitivity to nonidealities," in *Proc. Int. Solid-State Circuits Conf.*, 1991, pp. 66–67.

[25] E. I. Jury, *Theory and Applications of the z-Transform Method*. Robert E. Krieger, 1964.

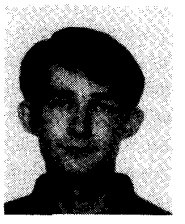
[26] M. H. H. Höfelt, "On the stability of a 1-bit-quantized feedback system," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1979, pp. 844–848.

[27] T. S. Parker and L. O. Chua, "Chaos: A tutorial for engineers," *Proc. IEEE*, vol. 75, pp. 982–1008, Aug. 1987.

[28] A. Sherstinsky and C. G. Sodini, "A programmable demodulator for oversampled analog-to-digital converters," *IEEE Trans. Circuits Syst.*, vol. 37, pp. 1092–1103, Sept. 1990.

[29] M. E. Pai, "Oscillations in nonlinear sampled-data systems," *IRE Trans. Automat. Contr.*, pp. 350–355, Jan. 1963.

[30] A. R. Bergen and R. L. Franks, "Justification of the describing function method," *SIAM J. Cont.*, vol. 9, pp. 568–589, Nov. 1971.



Søren Hein (S'88-M'89) was born in May 1968 in Copenhagen, Denmark. He received the M.Sc. degree in electrical engineering from the Technical University of Denmark in January 1989, and the Ph.D. degree in electrical engineering from the University of California at Berkeley in May 1992.

Since October 1992 he has worked for Siemens ZFE in Germany. His research interests include algorithmic aspects of oversampled A/D conversion and signal and image processing. He has also worked on error-correction coding for satellite communications.

Avideh Zakhor (S'87-M'87) received the B.S. degree from the California Institute of Technology, Pasadena, and the S.M. and Ph.D. degrees from Massachusetts Institute of Technology, Cambridge, all in electrical engineering, in 1983, 1985, and 1987, respectively.

In 1988, she joined the faculty at the University of California, Berkeley, where she is currently Assistant Professor in the Department of Electrical Engineering and Computer Sciences. Her research interests are in the general area of signal processing and its applications to images and video, and biomedical data. She has been a consultant to a number of industrial organizations and holds four U.S. patents.

Dr. Zakhor was a General Motors Scholar from 1982 to 1983, received the Henry Ford Engineering Award and Caltech Prize in 1983, was a Hertz fellow from 1984 to 1988, received the Presidential Young Investigators (PYI) Award, IBM Junior Faculty Development Award, and Analog Devices Junior Faculty Development Award in 1990, and Office of Naval Research (ONR) Young Investigator Award in 1992. She is currently an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, and a member of the Technical Committee for Multidimensional Digital Signal Processing.