

Bounding the Integrator Outputs of Second-Order Sigma-Delta Modulators

Ronan Farrell, *Student Member, IEEE*, and Orla Feely, *Member, IEEE*

Abstract—The development to date of the bounds for the outputs of the integrators of second-order sigma-delta modulators has relied on a mixture of theoretical analysis and extensive use of simulation. This paper presents a new approach that analyzes the behavior of such systems, and identifies the external boundaries of the trapping region for the outputs of the integrators. With this approach, it is possible to develop tight results for the standard second-order modulator with constant input, without the need for simulation. This approach is flexible, and can be adapted for other second-order architectures, and gives strong insight into the behavior of these systems.

Index Terms—Chaos, nonlinear analysis, sigma-delta.

I. INTRODUCTION

DUE to advances in very large scale integration (VLSI) technology, sigma-delta modulator-based analog-to-digital converters have become popular, especially in low-frequency, high-precision applications. A sigma-delta modulator produces a coarsely quantized output, requiring only simple analog components. Complex digital circuitry is then used to produce a high-precision digital representation from this coarse quantization. This balance of simple analog components and complex high-speed digital circuitry is particularly suited to VLSI implementation, and facilitates the direct incorporation of these modulators with other digital signal processing systems.

In recent years, a variety of sigma-delta modulator architectures has been suggested for a range of applications [1], [2]. Despite their increasing popularity, accurate analysis of these systems has lagged behind implementation. This weakness in analysis is due to the highly nonlinear nature of the modulator. Recent approaches to analyzing this system have concentrated on determining stability criteria for second- and higher order modulators [3]–[6], on developing upper limits on the voltage span of the integrator outputs [3], [5]–[8], and on spectral analysis of the output [9]. In this paper, a new method for developing tight upper bounds on the output of the integrators of a second-order modulator (Fig. 1) will be presented. These bounds are important as they determine the clipping level of the system, and are a factor in determining the required on-chip

area for the capacitors. It is desirable to minimize any excess in both of these criteria when implementing these systems.

Bounding the outputs of the integrators involves finding a trapping region for the system within the two-dimensional space defined by the integrator outputs. A trapping region has the property that, once within this region, the integrator outputs will never take on values outside this region. Methods for determining the trapping region range from the fully theoretical to approaches which depend on simulations for their final results, but which have a strong theoretical background. The fully theoretical approach is best demonstrated by Hein and Zakhor's work on stability and bounds [3], and in Wang's independent but related analysis of second-order systems [6]. Their approach uses parabolic bounds to produce a smooth approximation to the trapping region. This produces fully analytical results, but these are not exact as the trapping region has been shown to have a complex and jagged shape (as suggested by the trajectory in Fig. 2) [8]. The computational approach to determining the bounds is contained in the work of Wang in his Ph.D. dissertation [8] and of Schreier *et al.* [10]. Wang explained this jagged shape using a computational procedure which identified invariant spaces arising from successive mappings. Schreier *et al.* used a similar computational approach in which a candidate region is iterated and enlarged until an invariant region is found. These approaches produce tighter bounds than those of Hein and Zakhor, and have the advantage that they can be adapted to any second-order system and some higher order systems. The main disadvantage of these computational approaches is the lack of general analytical results.

The aim of the work presented in this paper is to present a new method which produces tighter analytical bounds than those previously obtained, and at the same time yields insight into the behavior of this complex system. Our approach depends upon developing an understanding of the behavior of the system, and then applying this knowledge to determine the jagged outer boundary of the trapping region. Of the work mentioned earlier, this work most resembles that of Wang [8], but differs in that the bounds will be determined analytically. This analytical approach yields general results and insight into the behavior of the second-order system. The analysis considers only systems with constant input, as is the case for most of the work done in this area to date. With high oversampling ratios and slowly changing inputs, bounds for the integrator outputs resulting from a constant input analysis are useful in practical applications.

Manuscript received October 31, 1996. This work was supported in part by Analog Devices and by Forbairt under the Strategic Research Program. This paper was recommended by Associate Editor K. Suyama.

The authors are with the Department of Electronic and Electrical Engineering, University College Dublin, Dublin 4, Ireland (e-mail: ronan.farrell@ucd.ie; orla.feely@ucd.ie).

Publisher Item Identifier S 1057-7130(98)02132-6.

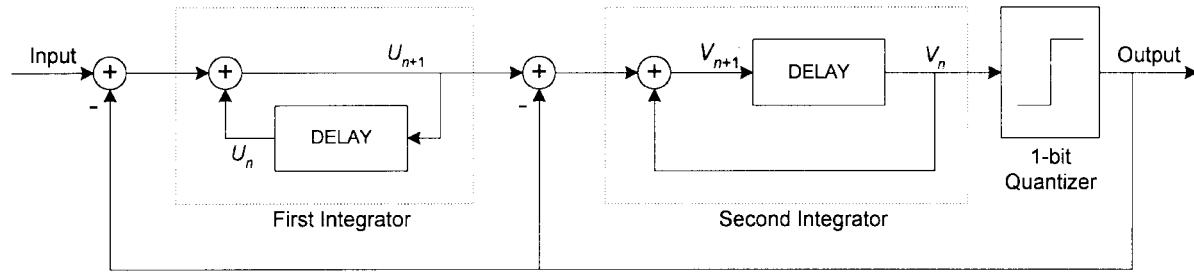


Fig. 1. Standard second-order sigma-delta modulator.

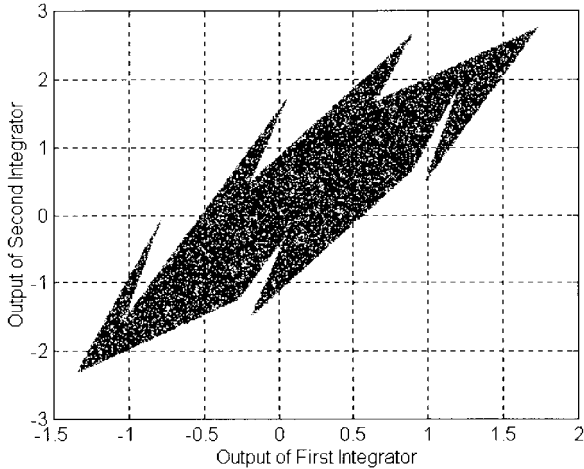


Fig. 2. Trajectory of (1) with input $x = 0.161027$ and initial condition $(0.077, 0.08)$.

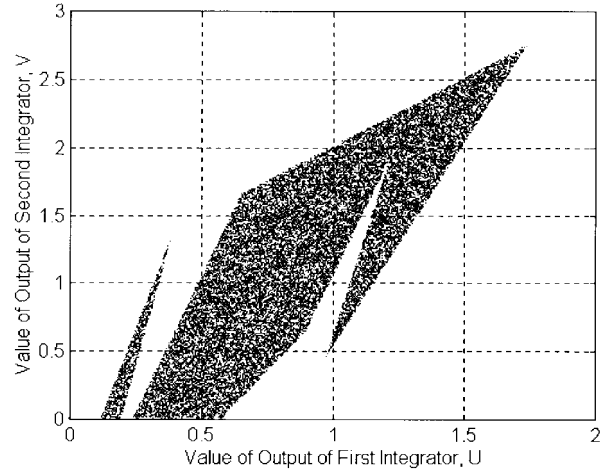


Fig. 3. Subset of the trajectory of Fig. 2 consisting of the points at which the trajectory enters the positive half-plane.

II. OVERVIEW

The aim of this paper is to demonstrate a novel approach to the understanding of the behavior of the second-order sigma-delta system, and thus to develop tight bounds on the maximum values that the outputs of the integrators may reach for a constant input. This will be achieved by examining the region that the outputs of the integrators U and V inhabit. We will utilize the fact that the trajectories of these variables are piecewise linear within each half-plane, where the half-planes are defined by the sign of the output of the second integrator (V). The half-planes will be referred to as the positive or negative half-plane according to the sign of V . It is possible to determine closed-form equations to describe the trajectories within a half-plane, given their conditions upon entry into the half-plane. These entry conditions, together with the closed-form equations, determine the maximum integrator output while the trajectory remains within this half-plane. Much of our analysis, therefore, will be concerned with formulating bounds on these entry conditions.

The pair of values (U_α, V_α) will denote the point at which a trajectory enters the positive half-plane after having been in the negative half-plane. The allowable range of these entry conditions is dependent on the number of iterations the trajectory has just spent in the negative half-plane. Thus, knowledge of the maximum number of negative iterations yields bounds on the entry conditions, and hence on the maximum integrator outputs in the positive half-plane. In

Section III, an approach will be presented that will enable the maximum values of U_α and V_α to be determined by considering certain sections of the trapping region. These sections contain all of the possible values that the integrators may take after a certain condition has been satisfied. Two such sections will be of interest: the section within which all possible entry conditions (U_α, V_α) to the positive half-plane must lie (the α sector, Fig. 3), and the section within which all entry conditions (U_β, V_β) to the negative half-plane must lie (the β sector).

To identify the maximum number of negative iterations, an iterative approach will be applied. Specifically, the number of negative iterations results in an outer bound for the α sector, restricting the range of values of U_α and V_α . This bounded region, when mapped forward to the negative half-plane, yields another boundary within the β sector. This boundary can be tested against a set of analytical conditions which determine the number of successive negative iterations that the trajectory may now undergo. This procedure is repeated until the maximum number of negative iterations the system may undergo, for a given input, is determined. Once this is known, the range of U_α and V_α can be ascertained.

In Section IV, functions which describe the rate of change of the individual trajectories will be used to identify the location along the trajectory where the maximum values will occur. With the maximum entry conditions, the maximum possible values for the output of the integrators can be determined.

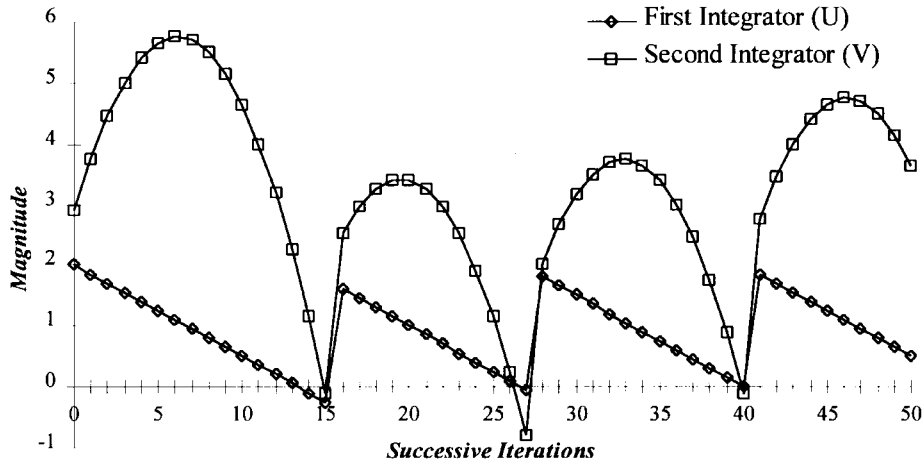


Fig. 4. Plot of the integrator outputs versus time with input $x = 0.85$ and initial conditions $(2.0, 2.9)$.

The fifth section will expand the analyzes of the previous sections to consider the cases of both leaky and chaotic modulators. These are modulators whose integrators are no longer ideal, i.e., whose poles are no longer equal to one. The problem of nonideal integrators does not require any alteration in the methodology of the previous sections—the only difference is due to the much more complex expressions for the trajectories. The final section will then discuss the results and conclusions developed from the analysis, in particular, the strengths of this approach and other avenues where it may be applied.

III. DETERMINING THE TRAPPING REGION

The second-order sigma-delta modulator, shown in Fig. 1 (also known as the standard Candy double-loop modulator [1]), with constant input is modeled by two coupled difference equations:

$$\begin{aligned} U_{n+1} &= U_n + x - \text{sgn}(V_n) \\ V_{n+1} &= V_n + U_{n+1} - \text{sgn}(V_n) \end{aligned} \quad (1)$$

where

- x value of the input;
- U, V output of the first and second integrators, respectively;
- $\text{sgn}(v) = \begin{cases} +1, & \forall v \geq 0 \\ -1, & \forall v < 0. \end{cases}$

It is also possible to have nonunity gains in the feedback path from the output of the quantizer to the input of each integrator. In these cases, the above equations and the following analysis can be easily modified. A positive input will be assumed, but as the system is symmetrical about zero input, the analysis is equally valid for negative inputs. The system can be analyzed in a piecewise-linear manner since the difference equations (1) are linear within each half-plane, as defined by the sign of V . Fig. 4 shows a plot of the way in which the value of each integrator output changes with successive iterations. From (1), closed-form expressions can be developed to express the value of the integrator outputs N iterations after some entry condition, provided that the trajectories stay within a

half-plane:

$$\begin{aligned} U_{k+N} &= U_k + N[x - \text{sgn}(V_k)] \\ V_{k+N} &= V_k + \frac{N[N+1][x - \text{sgn}(V_k)]}{2} + N[U_k - \text{sgn}(V_k)]. \end{aligned} \quad (2)$$

The behavior of the system changes when the output of the second integrator changes sign; therefore, it is important to develop relationships between the two sets of entry conditions to the positive and negative half-planes (U_α, V_α) and (U_β, V_β) , respectively.

A. Determining the Maximum Value of U_α

The boundaries of the β sector, which contains all possible values of U_β and V_β , can be determined by mapping the U axis forward one iteration, and using the axis itself as the other bounding line. This defines a semi-infinite sector, Fig. 5, given by

$$V < 0 \quad (3)$$

and

$$V \geq U - 1. \quad (4)$$

Trajectories enter the negative half-plane at points (U_β, V_β) in the β sector. The number of iterations before the trajectory again enters the positive half-plane depends on the location of (U_β, V_β) within the β sector. It is possible to determine regions within this sector which contain all entry conditions that will result in N^- subsequent iterations in the negative half-plane. The boundaries of these regions within the β sector can be identified by the conditions that, after $N^- - 1$ iterations, the trajectory must be negative and that, after N^- iterations, it must be nonnegative

$$V < -[N^- - 1][U + 1] - \frac{1}{2}[N^-][N^- - 1][x + 1] \quad (5)$$

$$V \geq -[N^-][U + 1] - \frac{1}{2}[N^-][N^- + 1][x + 1]. \quad (6)$$

The β sector is in this manner broken down into a series of subregions bounded by (3)–(6) as shown in Fig. 5. These

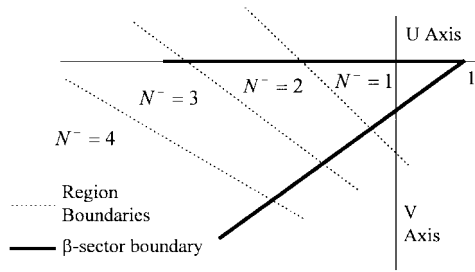


Fig. 5. Subregions within the β sector defined by (5) and (6) according to the number of successive negative iterations.

subregions consist of all entry conditions that will experience N^- iterations before reentering the positive half-plane. The location of this reentry is obtained by mapping these boundaries forward through N^- iterations. After N^- iterations, boundaries (5) and (6) give boundaries

$$V \geq 0 \quad (7)$$

$$V < U + 1. \quad (8)$$

These form the boundaries of the α sector, containing the location of all entry points into the positive half-plane. [They could also have been obtained in the manner used to obtain the boundaries (3) and (4) of the β sector.] After N^- iterations, boundaries (3) and (4) give the boundaries

$$V < UN^- + N^- - \frac{N^-[N^- - 1][x + 1]}{2} \quad (9)$$

$$V \geq U[N^- + 1] + [N^- - 1] - \frac{N^-[N^- + 1][x + 1]}{2}. \quad (10)$$

In this way, the α sector is divided into subregions, bounded by (9) and (10), based on the number of iterations that the trajectory spent in the negative half-plane before entering the positive half-plane at point (U_α, V_α) . The upper bound on U_α in any of these subregions occurs at the intersection of the boundaries (8) and (10), and the upper bound on V_α follows from (8). Letting U_{mN^-} denote the upper bound on U_α in the subregion corresponding to N^- preceding negative iterations, we have

$$U_\alpha < U_{mN^-} = \frac{-[N^- - 2] + \frac{1}{2}[N^-][N^- + 1][x + 1]}{N^-} \quad (11)$$

and, in particular,

$$\begin{aligned} U_{m2} &= \frac{3}{2}[x + 1] \\ U_{m3} &= 2x + \frac{5}{3}. \end{aligned} \quad (12)$$

Thus, after a given number of negative iterations N^- , there is an outer bound (10) on the region of the α sector in which the entry conditions (U_α, V_α) must reside. In fact, (10) also forms the outer bound on the entry conditions after any number of iterations less than or equal to N^- . (The case of only one negative iteration needs special attention since, from (11), U_{m1} is greater than either U_{m2} or U_{m3} . Appendix A shows that, due to other considerations, the bound resulting from only one negative iteration does not exceed those due to more than one negative iteration.)

B. Determining the Maximum Number of Negative Iterations

For a given number of negative iterations N^- , there is a maximum number of iterations that the trajectory may spend in the positive half-plane N^+ . As long as the trajectory remains within the positive half-plane, we have from (2)

$$\begin{aligned} U_{\alpha+N} &= U_\alpha + N[x - 1] \\ V_{\alpha+N} &= V_\alpha + \frac{N[N + 1][x - 1]}{2} + N[U_\alpha - 1]. \end{aligned} \quad (13)$$

The maximum number of iterations in the positive half-plane N^+ can be obtained by finding the smallest integer value of N that satisfies the following inequality, which has been obtained from (13) with $U_\alpha = U_{mN^-}$ and $V_\alpha = U_{mN^-} + 1$ and the condition $V_\beta < 0$:

$$U_{mN^-}[N + 1] - [N - 1] + \frac{N[N + 1][x - 1]}{2} < 0. \quad (14)$$

If the region bounded by (8) and (10) is mapped forward, an outer boundary for the entire trapping region is formed. This is demonstrated in Fig. 6. The outer boundary of the trapping region in the β sector is found by mapping the upper boundary (8) of the α sector N^+ times.¹ Once this outer boundary is known, then it is possible to identify which areas of the β sector lie within the trapping region. This will enable us to determine which regions within the sector the trajectories may enter, and hence the maximum number of negative iterations possible. If the outer boundary (8) is mapped forward N^+ iterations, it will produce a boundary line inside the β sector:

$$V < U[N^+ + 1] - [N^+ - 1] - \frac{N^+[N^+ + 1][x - 1]}{2}. \quad (15)$$

It is possible to develop a measure of the distance between this line and the line defining the edge of the regions representing a specific number of iterations in the negative half-plane by comparing the intersection points of each with the boundary line (4), as shown in Fig. 7. For the system to have the possibility of entering the region corresponding to N^- successive negative iterations

$$I_T < I_R \quad (16)$$

where

- I_T intersection of the outer boundary of the trapping region boundary (15) and the lower boundary of the β sector (4), and
- I_R intersection of the N^- region boundary (5) and the lower boundary of the β sector (4).

¹The case of a single positive iteration would seem to produce an outer boundary in the β sector which is outside that resulting from the maximum number of positive iterations. In fact this is not the case as there is a restriction on the left-hand side of the α sector which excludes any area which extends beyond the boundary due to the maximum number of negative iterations. This boundary in the α sector arises from the outer boundary of the trapping region in the β sector, even if the boundary resulting from $N^+ = 1$ is used.

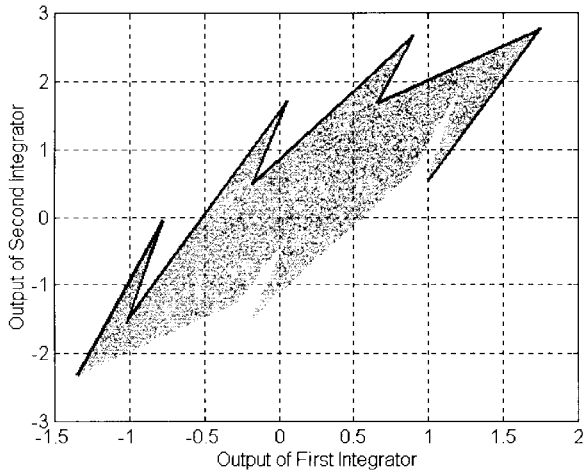


Fig. 6. Boundaries of the α sector mapped forward to bound the trapping region.

This can be expanded to produce a condition on the input value in terms of the maximum number of positive iterations and the number of negative iterations being tested:

$$I_T = \frac{[N^+ - 2]}{N^+} + \frac{[N^+ + 1][x - 1]}{2} \quad (17)$$

$$I_R = \frac{[2 - N^-]}{N^-} - \frac{[N^- - 1][x + 1]}{2} \quad (18)$$

which produces a general expression for (16):

$$\left[2 - \frac{2}{N^+} - \frac{2}{N^-}\right] + x \left[\frac{N^+ + N^-}{2}\right] + \left[\frac{N^- - N^+ - 2}{2}\right] < 0. \quad (19)$$

In summary, a candidate number of successive negative iterations gives bounds on U_α and V_α which are used to find N^+ from (14). The allowable number of successive negative iterations N^- on the next pass through the negative half-plane is then constrained by (19). Therefore, for any given number of negative iterations, it is possible to develop a set of conditions that can be solved to determine whether that number will increase or decrease with time. In this way, it is possible to show that for all values of input less than one, the system will collapse down to three negative iterations, and often at most two. Fig. 8 shows the transient behavior of the system. The trajectories collapse very quickly from a high number of negative iterations down to at most three, and often only two, negative iterations in fewer than ten cycles. In this manner, it is possible to prove the global stability of the second-order modulator for inputs of magnitude less than one. It would be possible, using the same technique, to identify any regions of instability that may arise (as happens in the chaotic modulators discussed in Section V). The same method can also be used to prove that one negative iteration can be followed by two or three, but that one, two, or three consecutive negative iterations cannot be followed by four or more. Thus, this method can identify an invariant region even if unstable regions exist outside it.

It is now important to identify the input ranges in which the system will experience at most two successive negative iterations in steady state and those in which it can experience

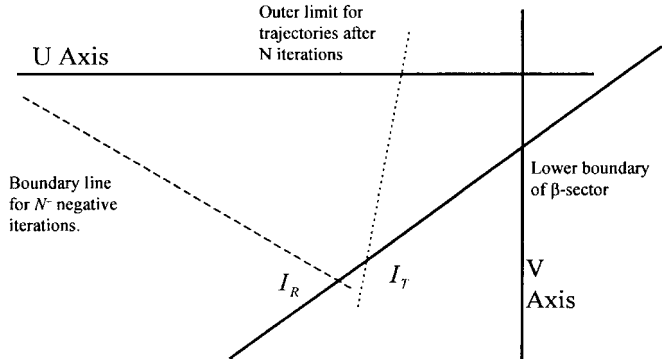


Fig. 7. Intersection points of the two boundaries (5) and (15) with the lower boundary of the β sector (4).

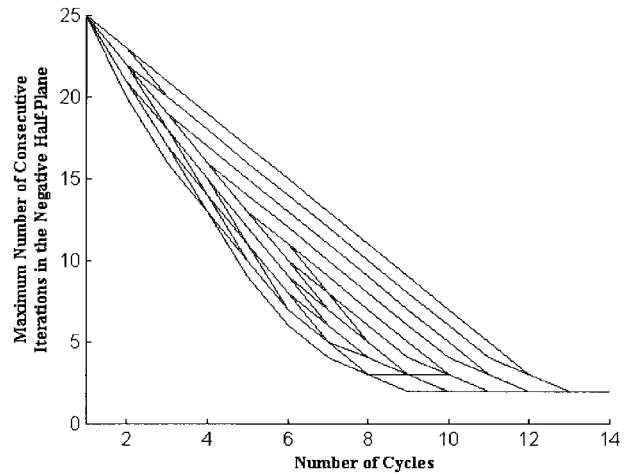


Fig. 8. Decay in the maximum number of possible consecutive negative iterations upon successive cycles shown for various values of input, from 0.014 to 0.95 in steps of 0.0094 (increasing to right).

three. This will be tackled in two steps: first, by identifying the input ranges where two successive negative iterations can be followed by three, and then, by seeing if three negative iterations can be sustained outside these ranges. From (19), three negative iterations can follow N^+ positive iterations if

$$x < x_{\max} = \frac{3N^+[N^+ + 1] - [14N^+ - 12]}{3N^+[N^+ + 3]}. \quad (20)$$

This must be compared with the minimum input value that is required to obtain N^+ iterations. This can be found from the fact that the inequality in (14) must fail for $N = N^+ - 1$:

$$U_{mN^-} \geq \frac{[N^+ - 2]}{N^+} - \frac{[N^+ - 1][x - 1]}{2}. \quad (21)$$

With $N^- = 2$ and $U_{m2} = \frac{3}{2}[x + 1]$, (21) reduces to

$$x > x_{\min} = \frac{2}{N^+} \left[\frac{N^+ - 2}{N^+ + 2}\right] + \left[\frac{N^+ - 4}{N^+ + 2}\right]. \quad (22)$$

Therefore, two negative iterations may be followed by three negative iterations if the input lies between the bounds (20) and (22) for some value N^+ of the intervening number of

TABLE I
BOUNDS DEFINING THE INPUT RANGES WHERE THREE NEGATIVE
ITERATIONS MAY FOLLOW TWO NEGATIVE ITERATIONS

N^+	x_{\min}	x_{\max}
3	0.0000	0.1111
4	0.1667	0.1905
5	0.3143	0.2667

TABLE II
BOUNDS DEFINING THE INPUT RANGES WHERE THREE NEGATIVE
ITERATIONS MAY FOLLOW THREE NEGATIVE ITERATIONS

N^+	x_{\min}	x_{\max}
3	-0.1111	0.1111
4	0.0952	0.1905
5	0.2333	0.2667
6	0.3333	0.3333
7	0.4095	0.3905

positive iterations, i.e.,

$$x_{\min} < x < x_{\max}. \quad (23)$$

The results from this inequality are expressed in Table I.

For $N^+ \geq 5$, $x_{\min} > x_{\max}$, and therefore, three successive negative iterations may follow two only if the intervening number of positive iterations is fewer than five and if the input is in the range (0.000, 0.1111) or (0.1667, 0.1905). Outside these input ranges, two successive negative iterations can never lead to three consecutive negative iterations.

Similarly, we can check to see if three successive negative iterations can follow three successive negative iterations. If not, then three successive negative iterations will arise again only in the input ranges given in Table I. The analysis is the same as before, except this time we substitute $U_{mN^-} = U_{m3}$ in (21). This yields a similar table of results (Table II).

For $N^+ \geq 6$, $x_{\min} \geq x_{\max}$, and therefore, three successive negative iterations may follow three only if the intervening number of positive iterations is fewer than six. Three negative iterations may follow three for inputs less than 0.1111, but this is not important as it overlaps with the results from Table I. For inputs in the range (0.1111, 0.1905), three negative iterations may follow three if the intervening number of positive iterations is four, and for inputs in the range (0.2333, 0.2667) if the intervening number of positive iterations is five. If, within these input ranges, the number of negative iterations drops to two, it can never again reach three except in the range (0.1667, 0.1905) from Table I. Thus for the input ranges (0.1111, 0.1667) and (0.2333, 0.2667), three successive negative iterations will arise in steady state only within an infinitely repeating sequence of three negative iterations interspersed with exactly N^+ positive iterations where N^+ can be either four or five. Appendix B uses this condition to identify the input values at which this may occur and the initial conditions required to obtain them. It will also be shown that these repeating sequences are very fragile, and can be ignored in the general analysis. Thus, excluding these sequences, three successive negative iterations can only occur for inputs (0.000, 0.1111) and (0.1667, 0.1905), and outside these ranges, the maximum number of successive negative iterations during stable operation is two.

IV. BOUNDS ON THE INTEGRATOR OUTPUTS

Given the range of entry conditions to the α sector, it is possible to determine the maximum values for both U and V using (2). To determine the location of the maximum value for the output of each integrator along the trajectory, it is necessary to be able to evaluate the rate of change of the trajectory at all points along the trajectory. The maximum obtained by the trajectory within a half-plane is a local maximum, but if the maximum values for the entry conditions are used for that trajectory, it is clear from (2) that the resultant maximum is a global maximum for that half-plane. It can also be shown by comparing the magnitudes of the local maxima in both half-planes that the local maximum in the half-plane of the same sign as the input is the greater. In a continuous system, a local maximum occurs when the rate of change of a function equals zero. The equivalent property of a discrete-time system is that the maximum will occur on the iteration before the rate of change goes negative.

A. Determining the Maximum Values

The rate-of-change function of a discrete-time system may be defined as the difference between two successive iterations. These functions in the positive half-plane can be obtained directly from (1):

$$\Delta U_n = U_{n+1} - U_n = x - 1 \quad (24)$$

and

$$\Delta V_n = V_{n+1} - V_n = U_{n+1} - 1. \quad (25)$$

From inspection, it can be seen that ΔU_n is a constant value, independent of the iteration number. If ΔU_n is positive, then the trajectory of the output of the first integrator will climb linearly to infinity or to the clipping voltage. If ΔU_n is negative, which is the case for stable operation, then the maximum value of the first integrator must occur on the first iteration in the positive half-plane, the entry condition U_α . This maximum value has been determined in the previous section to be (12)

$$U_{m2} = \frac{3}{2}[x + 1], \quad \text{after two negative iterations}$$

$$U_{m3} = 2x + \frac{5}{3}, \quad \text{after three negative iterations.}$$

The rate of change function for V shows that it will become negative when U no longer exceeds one. This occurs after a number of iterations given by

$$N_{\max} = \text{int} \left[\frac{U_\alpha - 1}{1 - x} \right]. \quad (26)$$

Thus, the maximum value will occur N_{\max} iterations after entering the α sector, when the maximum entry conditions are chosen:

$$V_{\max} = U_{mN^-} [N_{\max} + 1] - [N_{\max} - 1] + \frac{N_{\max} [N_{\max} + 1] [x - 1]}{2}. \quad (27)$$

These equations can be used to determine tight bounds on the maximum values that either integrator will reach for any

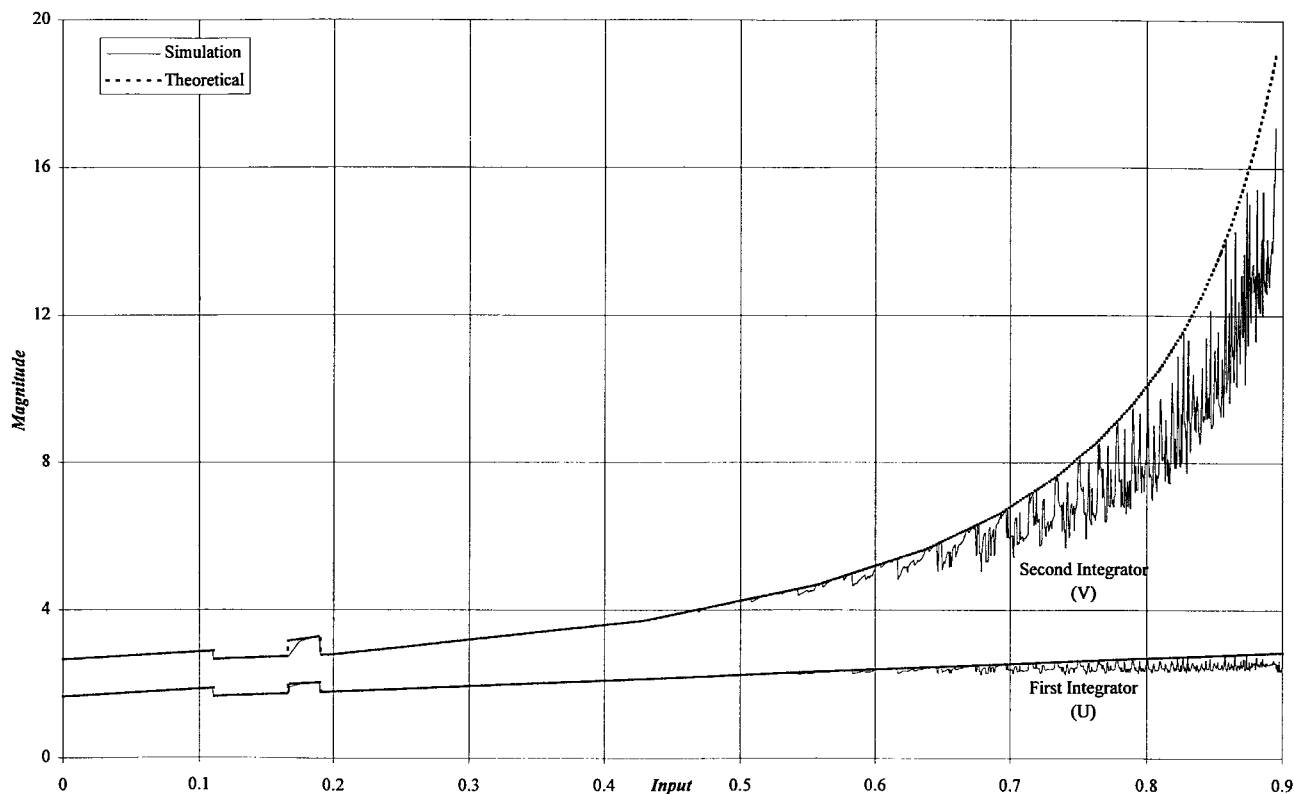


Fig. 9. Graph of the theoretical results (12), (27) compared with the maxima obtained from simulations for the ideal modulator.

combination of input and feedback gains. The appropriate value of U_{mN^-} can be selected using Table I.

B. Comments Upon the Results

Fig. 9 shows a comparison of the theoretical bounds, obtained from (11) and (27), with the maximum values obtained through simulation. The method used in the simulations was to randomly select initial conditions for the integrators. The values of the outputs of the integrators for the first 100 cycles were ignored to avoid transient effects, although this is unnecessary if the initial conditions are within the trapping region. The system was then iterated another 5000 times, and the largest value obtained by each integrator was retained. This process was repeated 500 times, with a different set of initial conditions each time, and the greatest maximum was kept. This method of repeatedly restarting the simulation is in order to prevent limit-cycles or limit-cycle-like behavior from developing, which may prevent the largest possible values from occurring.

There is a good match between the simulated maxima and the theoretical bounds, with the percentage difference varying between parts of 1% for the lower input values, and up to 5% for the higher values of input. The reason for the relatively poor results toward the higher end of the input spectrum is due to the internal structure of the α and β sectors. From Fig. 3, it is possible to see distinct regions containing the entry conditions that the trajectory may take, and the gaps between them representing values which the entry conditions do not take. This structure is due to the discrete nature of the trajectories. The entry conditions (U_α, V_α) are mapped across

the positive half-plane. If the entry conditions are mapped N times, they will form a certain region, and if they are mapped $N + 1$ times, then another region will be described. These regions will not always overlap, and are the cause of the internal structure that can be seen in the α sector (Fig. 3). If the maximum value of (U_α, V_α) lies in one of the gaps between the regions, then it cannot be achieved. The true maximum value of (U_α, V_α) could be found by determining this internal structure, but the equations predicting these bounds become significantly more complex with little benefit. This problem is particularly noticeable when N^+ increases by one, for example, at an input value of 0.1667 in Fig. 9. The problem is more visible for larger input values due to the increasingly close changes in the maximum number of positive iterations. These rapid changes in the value of N^+ increase the difficulty for simulations in obtaining the maximum possible values.

Another feature discernible from the graph is the step-like nature of the bounds. This is due to the variation in the number of iterations in the negative half-plane. Also visible is the noticeable change in slope as the value of the input increases, particularly at an input value of 0.42. This is one of the input values at which the value of N_{\max} changes by one. This results in a change of slope in (27).

In summary, the equations developed in Sections III and IV produce tight bounds for the maximum values that the output of the integrators can reach. The analysis also gives good insight into how the trajectories spend different numbers of iterations in each half-plane depending upon the input value, and how this affects the maximum values that can be reached by the outputs of the integrators.

V. CHAOTIC AND LEAKY SYSTEMS

Discrete-time integrators can be modeled by a unit delay and a feedback loop, where the gain of the feedback loop represents the pole of the integrator. In an ideal system, the gain of the feedback loop is unity, but in real systems, the gain is slightly less than one, and such an integrator is said to be leaky [11]. If the feedback gain is greater than one, the system is chaotic. A leaky system is stable, but is prone to limit cycle behavior, which reduces the accuracy of the output. A chaotic system has no stable limit cycles, which is beneficial in terms of tone removal, but suffers from reduced noise performance and dynamic range [12]. Most real modulators have poles which are slightly less than one. When the poles deviate from unity, the analysis of these systems becomes much more complex. It is possible to use the same approach as was used for the ideal case to determine the maximum values that the outputs of the integrators may reach for any combination of pole positions. Once the closed-form equations for the trajectories within a half-plane are determined, the rest of the analysis is as before. In order to get these closed-form equations, it is necessary to consider four cases: 1) only the first pole is not equal to one, 2) only the second pole is not equal to one, 3) both poles are not equal to one, but are equal in value, and finally, 4) both poles are not equal to one, and are not equal to each other.

A. Detailing the Approach to Determining the Bounds

The state equations for a system with nonideal integrators can be expressed in the following manner, where the poles of the integrators are represented by p and q :

$$\begin{aligned} U_{n+1} &= qU_n + x - \text{sgn}(V_n) \\ V_{n+1} &= pV_n + U_{n+1} - \text{sgn}(V_n). \end{aligned} \quad (28)$$

From this, it is possible to develop closed-form equations for the trajectories of the two integrators in the positive half-plane. For the first integrator, a single equation will cover all combinations of the two poles:

$$U_{\alpha+N} = q^N U_\alpha + [x-1]f_q(N-1) \quad (29)$$

whereas for the second integrator, four separate equations are required:

$$\begin{aligned} V_{\alpha+N} &= V_\alpha - N + U_\alpha f_q^*(N) + \left[\frac{x-1}{1-q} \right] [N - f_q^*(N)], \\ &\text{where } q \neq 1, p = 1 \end{aligned} \quad (30)$$

$$\begin{aligned} V_{\alpha+N} &= p^N V_\alpha - f_p(N-1) + U_\alpha p^N f_{1/p}^*(N) \\ &+ p^N (x-1)g_p^*(N), \quad \text{where } q = 1, p \neq 1 \end{aligned} \quad (31)$$

$$\begin{aligned} V_{\alpha+N} &= p^N V_\alpha - f_p(N-1) + Np^N U_\alpha \\ &+ p^N \left[\frac{x-1}{1-p} \right] [f_{1/p}^*(N) - N], \\ &\text{where } q \neq 1, p \neq 1, p = q \end{aligned} \quad (32)$$

$$\begin{aligned} V_{\alpha+N} &= p^N V_\alpha - f_p(N-1) + U_\alpha p^N f_{q/p}^*(N) \\ &+ p^N \left[\frac{x-1}{1-q} \right] [f_{1/p}^*(N) - f_{q/p}^*(N)], \\ &\text{where } q \neq 1, p \neq 1, p \neq q \end{aligned} \quad (33)$$

where

$$f_{q/p}(j) = \sum_{i=0}^j \left[\frac{q}{p} \right]^i = \frac{1 - \left[\frac{q}{p} \right]^{j+1}}{1 - \frac{q}{p}} \quad (34)$$

$$f_{q/p}^*(j) = \sum_{i=1}^j \left[\frac{q}{p} \right]^i = f_{q/p}(j) - 1 \quad (35)$$

$$g_p^*(j) = \sum_{i=1}^j ip^{-i} = p^{-j} \left[\frac{j-p-jp+p^{1+j}}{(p-1)^2} \right]. \quad (36)$$

From the above equations and (28), it is possible to repeat every step used in the analysis of the ideal system (Section III). As before, the maximum number of consecutive iterations the trajectories can undergo in the negative half-plane is a major factor in the maximum value that the outputs of the integrators may reach. It is possible to use this approach to determine exact results, but it is cumbersome due to the presence of implicit equations. For practical purposes, the approach used can be converted into a simple test to check for the maximum number of iterations in the negative half-plane.

It is possible to develop a closed-form expression bound on the maximum value that the output of the first integrator may reach, as given in (37) and (38), shown at the bottom of the page. These bounds were developed in the same manner as those in Section III, although the nonunity pole values complicate the expression. These bounds are valid for any combination of pole positions. With each pole at unity, the expressions will simplify to those for the ideal case (11).

Due to the complexity of the equations for the second integrator output (30)–(36), it has not been possible to determine a closed-form expression for the bound on the maximum value of this quantity. Only for the case when the second pole is equal to one can a closed-form expression be developed for N_{\max} . For the other cases, numerical solutions will be necessary, but as N_{\max} must be an integer, this is relatively simple. The correct value for U_{mN-} may be obtained by using (37) and (38) to determine the maximum value of U_α . The same bound on V_α may be used as in the ideal case (8).

With the relevant expression (30)–(33) and N_{\max} , the maximum output of the second integrator may be determined. Unlike the ideal case, the largest absolute value of the integrator outputs may arise in the negative half-plane for a positive input. This may be checked by comparing U_{mN-} and the modified form of I_T . The maximum value of V can then be easily obtained [13]. An example of the results that can be obtained

$$U_{m2} = q^3 \left[\frac{(2p+q-2) - x(q+p+1)}{(q^2+qp)} \right] + x[q^2+q+1] + [1+q-q^2] \quad (37)$$

$$U_{m3} = q^4 \left[\frac{(q^2-q-2-2p+qp+2p^2) - x(q^2+qp+p^2+q+p+1)}{q^3+q^2p+p^2q} \right] + x[q^3+q^2+q+1] + [1+q+q^2-q^3] \quad (38)$$

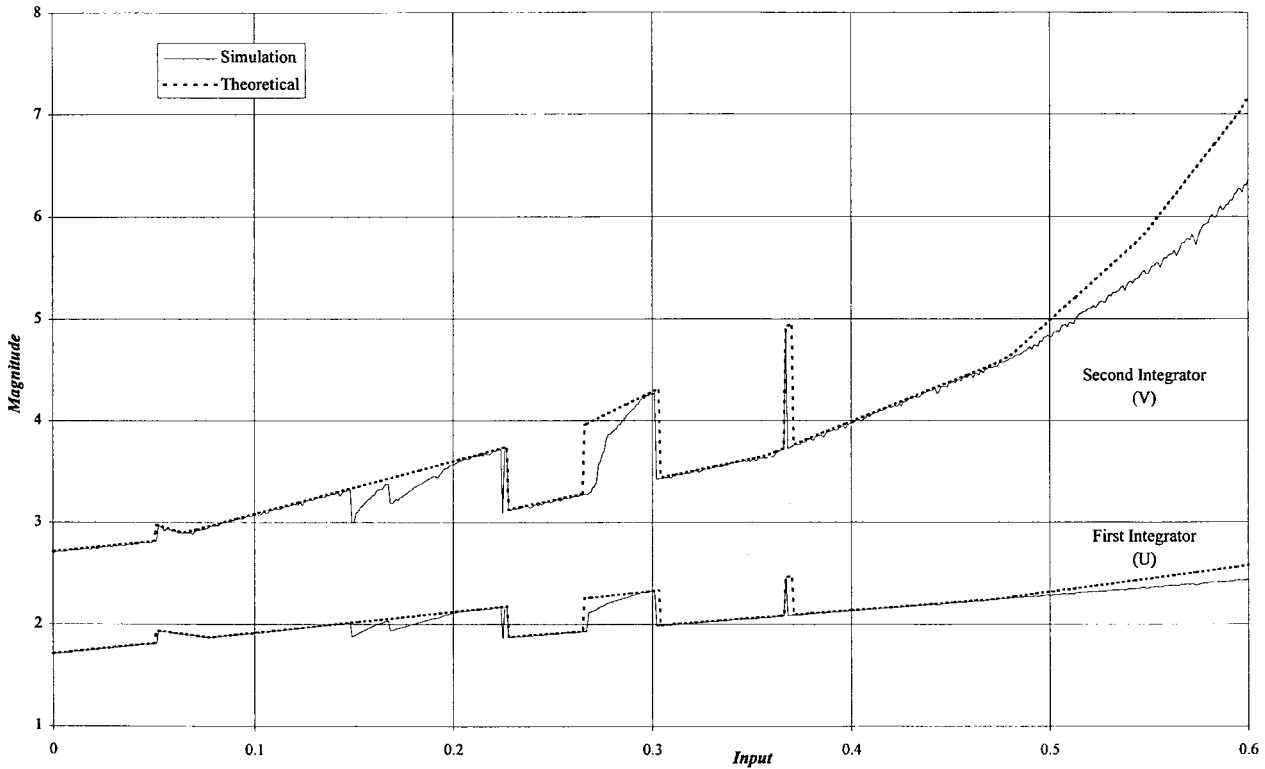


Fig. 10. Graph of the theoretical results from Section V compared with the maxima obtained from simulations for the chaotic modulator with the first pole at 1.023 and the second pole at 1.035.

from these equations is shown in Fig. 10 for the case where both poles are greater than one and different from each other.

B. Comments on Results

The simulated results were obtained in the same manner as that used for the ideal system in Section IV. The predicted bounds closely match the maxima obtained from the simulations, and compare favorably with the bounds developed by Hein [14] for chaotic modulators. By examining the equations and comparing them to the graph, it is possible to explain all of the major features, especially the step-like nature of the curve, in the same manner as for the ideal system. Other features noted in the ideal system are also valid here, in particular, the comments on the effect of assuming that the maximum value is always obtained. One new feature is the sudden increase in bounds at about 0.05. This is due to the bounds arising in the negative half-plane.

There are some limitations to the use of this method in determining the bounds of the system. For a leaky system, strong limit cycle behavior will develop. Once these limit cycles have developed, they will restrict the outputs of the integrators to a set of values, and the maximum values of this set may be substantially lower than that predicted by the bounds. This becomes more pronounced as the two poles become smaller. There is also the problem of instability as the values of the poles increase. The dynamic range of the system is reduced as the poles increase in value. The approach presented will fail once instability develops, as it will be impossible to determine a value of N^- that does not increase after every cycle. This failure can be used as a test for instability [15],

and provides some insight into the development of instability. Once these two restrictions on the usability of this approach are considered, the results for the range of acceptable pole positions are tight enough for design purposes.

VI. CONCLUSIONS

A new method has been presented for developing upper bounds on the maximum values that the outputs of the integrators of the standard second-order sigma-delta modulator may obtain. This has been achieved through an analysis of the behavior of the system, and by identifying the external boundaries of the trapping region of the trajectories. This approach to analyzing the second-order modulator has been shown to be flexible, and can be used to determine bounds for not only the standard second-order modulator, but also for chaotic and leaky systems. This approach can also be used to develop bounds for other modulator architectures once closed-form equations for the trajectories of the modulators can be obtained. Further developments of this work include the extension of this analysis to consider time-varying inputs and the analysis of higher order systems.

APPENDIX A
THE MAXIMUM VALUE OF U_{m1}

According to the analysis in Section III, the maximum value of the first integrator after only one negative iteration, U_{m1} , can exceed the maximum possible value of U_α after either two or three negative iterations. In practice, this is not the case as there are additional constraints on the maximum value of U_β that were not considered in Section III. In Fig. 5, the maximum

value that U_β can obtain within the β sector is one. It was not considered if this maximum value of U_β is permissible.

The maximum value of U_{m1} is dependent on the maximum possible number of iterations that the trajectory may undergo on its previous passage through the negative half-plane, before entering the positive half-plane and then experiencing only one negative iteration. To show this, it will be necessary to determine the maximum value that U_β may reach given a certain number of consecutive iterations in the negative half-plane on the previous cycle. This can be determined by examining the range of values of (U_α, V_α) which forms the α sector. The α sector is bounded by (7), (8), and (10). The third boundary (10) is dependent on the maximum number of consecutive iterations that the system may undergo in the negative half-plane during steady-state behavior, and which defines the upper limit on the maximum value of U_α . Given the maximum number of consecutive negative iterations, it is possible to identify the range of values of U_α and V_α .

Given this set of permissible entry conditions, it is possible to subdivide the α sector into regions representing the number of iterations that the trajectory will spend in the positive half-plane. As we are in search of the maximum value of U_β , it is only necessary to consider the maximum value of U_α within each of these regions. This value can then be mapped forward into the negative half-plane, resulting in U_β . The regions in the α sector representing a given number of positive iterations can be bounded by the same technique as was used for the β sector in Section III, resulting in the boundaries

$$V \geq [N^+ - 1][1 - U] - \frac{N^+[N^+ - 1][x - 1]}{2} \quad (\text{A1})$$

$$V < N^+[1 - U] - \frac{N^+[N^+ + 1][x - 1]}{2}. \quad (\text{A2})$$

The maximum value of U_α for a given number of positive and negative iterations N^+ and N^- can be obtained from the solution of (10) and (A2). If this point is now mapped forward N^+ iterations, the maximum value for U_β will be obtained. The relationship between U_α and U_β is given by

$$U_\beta = U_\alpha + N^+(x - 1). \quad (\text{A3})$$

Using this, an expression for the maximum value of U_β , after a given number of iterations in both the positive and negative half-planes, can be obtained, as shown in (A4) at the bottom of the page. If this value of U_β is mapped forward one iteration, the maximum value of U_{m1} can be obtained in terms of the previous number of negative (N^-) and positive (N^+) iterations, as shown in (A5) at the bottom of the page.

This result suggests that, in certain cases, where more than one consecutive iteration in the negative half-plane may occur, the maximum value of U_α that can occur after only one

negative iteration may still exceed the maxima that can occur after a higher number of negative iterations if there is only one intervening positive iteration, e.g., $U_{m1,3,1}$. This will henceforth be referred to as U_{m1,N^-} . This does not arise as there is a further constraint on the minimum number of positive iterations that may occur, N_{\min}^+ , after undergoing N^- negative iterations. If this minimum number is greater than one, then from (A5), the maximum value of U_{m1} will never exceed the maxima that may occur after a higher number of negative iterations. This constraint on the minimum value of N_{\min}^+ means that, for example, a single positive iteration may not follow three negative iterations during stable operation. It is possible to check if such a combination is possible by reverse mapping N^- times the upper boundary of the region representing N_{\min}^+ positive iterations (A2) from the α sector to the β sector. This will produce a boundary in the β sector which, if it is inside the trapping region defined in Section III, will mean that N_{\min}^+ positive iterations may follow N^- negative iterations. To show that U_{m1,N^-} will never exceed U_{mN^-} for all values of N^- , we must prove that for the larger values of input, for example, where $U_{m1,2}$ exceeds U_{m2} (12), the trajectories must spend more than one iteration in the positive half-plane. If this can be shown, then by (A5), it can be easily shown that the resulting values of U_{m1,N^-} will not exceed U_{mN^-} for $N^- > 1$.

To check if one iteration in the positive half-plane may follow N^- negative iterations, it is necessary to reverse map N^- times the upper boundary (A2) of the region representing only one positive iteration. This will result in the following boundary in the β sector:

$$V < -[N^- + 1]U - 2[N^- - 1] - x[N^- + 1] - \frac{N^-[N^- + 1][x + 1]}{2}. \quad (\text{A6})$$

It is possible to test this boundary in the same manner as was used to test if a region in the β sector, representing N^- negative iterations, was within the trapping region. This will produce several inequalities which can be obtained by examining the intersection point of (A6) with (4), and that of the trapping region (17) with (4). This will give an inequality defined in terms of N^+ and the input.

$$x < \frac{N^{+2} - 3N^+ + 4}{N^+(N^+ + 4)}, \quad \text{for one positive iteration to follow two negative iterations} \quad (\text{A7})$$

$$x < \frac{N^{+2} - \frac{23}{5}N^+ + 4}{N^+(N^+ + 5)}, \quad \text{for one positive iteration to follow three negative iterations} \quad (\text{A8})$$

$$U_\beta < \frac{[N^+ - N^- + 1] + \frac{1}{2}[N^{+2} + N^+ + 2N^+N^-][x - 1] + \frac{1}{2}[N^{-2} + N^-][x + 1]}{N^+ + N^- + 1} \quad (\text{A4})$$

$$U_{m1,N^-,N^+} < \frac{[N^+ - N^- + 1] + \frac{1}{2}[N^{+2} + N^+ + 2N^+N^-][x - 1] + \frac{1}{2}[N^{-2} + 3N^- + 2N^+ + 2][x + 1]}{N^+ + N^- + 1} \quad (\text{A5})$$

where N^+ is the maximum number of iterations in the positive half-plane. The value of N^+ can be determined from (21) given the maximum value of U_α , as given by the greatest of $U_{m1,2}$, $U_{m1,3}$, U_{m2} , or U_{m3} .

If the condition given in (A8) and the condition required to obtain N^+ iterations (21) are compared, one can see that three negative iterations can never be followed by only one positive iteration during stable operation. Examination of $U_{m1,2}$ shows that for the range of values where it can exist, it will never exceed U_{m2} , and that for higher values of input, two negative iterations will not be followed by only one positive iteration. Thus, at no point does the maximum value of U_{m1} exceed the value of U_{mN^-} , for $N^- > 1$.

APPENDIX B DETERMINING THE MAXIMUM NUMBER OF NEGATIVE ITERATIONS

In Section III, it was shown that there were some input ranges where three consecutive negative iterations could be sustained indefinitely within an infinite sequence consisting of three negative iterations and either four or five positive iterations, depending on the input value. After one cycle, i.e., one pass through the positive half-plane (N^+ iterations) and one pass through the negative half-plane (N^- iterations), U_α is incremented by the amount

$$[N^+ + N^-]x + [N^- - N^+]. \quad (\text{B1})$$

Since U_α can never leave the region of the α sector corresponding to N^+ positive iterations, this increment must be zero; hence

$$x = \frac{N^+ - N^-}{N^+ + N^-}. \quad (\text{B2})$$

For $N^- = 3$ and $N^+ = 4$ or 5, this allows two options:

$$x = \frac{1}{7}, \frac{1}{4}. \quad (\text{B3})$$

As before, over one cycle, V_α is incremented by the amount

$$\frac{N^+[N^+ + 1][x - 1]}{2} + \frac{N^-[N^- + 1][x + 1]}{2} + U_\alpha[N^+ + N^-] + N^-N^+[x - 1] + N^- - N^+. \quad (\text{B4})$$

Again, this must be zero if V_α is not to leave the constrained region; thus,

$$U_\alpha = \begin{cases} \frac{13}{7}, & \text{for } x = \frac{1}{7} \\ \frac{17}{8}, & \text{for } x = \frac{1}{4}. \end{cases} \quad (\text{B5})$$

The only restriction on the value of V_α is that it must be in the range necessary to obtain the required number of iterations in the positive and negative half-planes. These limits can be found by considering the boundaries of the α sector (7), (8), and those of the region representing N^+ iterations in the α

sector (A1), (A2), giving the ranges below:

$$\begin{aligned} \frac{18}{7} \leq V_\alpha < \frac{20}{7}, & \quad \text{for } x = \frac{1}{7} \\ 3 \leq V_\alpha < \frac{25}{8}, & \quad \text{for } x = \frac{1}{4}. \end{aligned} \quad (\text{B6})$$

Since both U and V must return to their starting points after one cycle, the trajectory is a periodic orbit. Equations (B5) and (B6) define very strict conditions on the set of entry conditions (U_α, V_α) that are necessary for these periodic sequences to form, and any slight deviation from them will result in an eventual collapse of the periodic sequences into the trapping region defined in Section III. The periodic orbits described in this section are the only such cycles that can exist for the given conditions and, due to their fragile nature and very specific entry conditions, may be ignored in the general analysis of the system. In a leaky modulator where there is significant limit cycle behavior, increased ranges of input values could satisfy the above requirements. These ranges can be determined, and they may be tested to see if any limit cycles, outside the trapping region, will persist. In the chaotic modulator, stable periodic orbits do not form, although when close to unstable orbits, it may require a larger number of iterations than normal before the trajectories collapse to the trapping region.

ACKNOWLEDGMENT

The authors wish to thank D. McCartney for his advice and comments.

REFERENCES

- [1] J. Candy and G. Temes, Eds., *Oversampling Delta-Sigma Data Converters*. New York: IEEE Press, 1992.
- [2] S. Norsworthy, R. Schreier, and G. Temes, *Delta-Sigma Data Converters*. New York: IEEE Press, 1997.
- [3] S. Hein and A. Zakhor, "On the stability of sigma delta modulators," *IEEE Trans. Signal Processing*, vol. 41, pp. 2322–2348, 1993.
- [4] M. Motamed, A. Zakhor, and S. Sanders, "The double loop sigma delta modulator with unstable filter dynamics: Stability analysis and tone behavior," *IEEE Trans. Circuits Syst. II*, vol. 43, pp. 549–559, 1996.
- [5] S. Pinault and P. Lopresti, "On the behavior of the double-loop sigma delta modulator," *IEEE Trans. Circuits Syst. II*, vol. 40, pp. 467–479, 1993.
- [6] H. Wang, "A geometric view of sigma delta modulations," *IEEE Trans. Circuits Syst. II*, vol. 39, pp. 402–405, 1992.
- [7] R. Farrell and O. Feely, "A novel approach to bounding the integrator outputs of second order sigma delta converters," in *IEEE Int. Symp. Circuits Syst.*, 1996, vol. I, pp. 521–524.
- [8] H. Wang, "A study of sigma delta modulations as dynamical systems," Ph.D. dissertation, Columbia Univ., New York, NY, 1993.
- [9] M. Motamed, A. Zakhor, S. Sanders, and T. Lee, "Spectral characteristics of the double loop $\Sigma\Delta$ modulator," in *IEEE Int. Symp. Circuits Syst.*, 1996, vol. I, pp. 457–460.
- [10] R. Schreier, M. Goodson, and B. Zhang, "On the behavior of the double-loop sigma delta modulator," *IEEE Trans. Circuits Syst. I*, vol. 44, pp. 38–44, 1997.
- [11] O. Feely and L. O. Chua, "The effect of integrator leak in $\Sigma\Delta$ modulation," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 1293–1305, 1991.
- [12] R. Schreier, "On the use of chaos to reduce idle channel tones in delta sigma modulators," *IEEE Trans. Circuits Syst. I*, vol. 41, pp. 539–547, 1994.
- [13] R. Farrell, "Non-linear analysis of sigma delta modulators," Ph.D. dissertation, unpublished.

- [14] S. Hein, "Exploiting chaos to suppress spurious tones in general double loop $\Sigma\Delta$ modulator," *IEEE Trans. Circuits Syst. II*, vol. 40, pp. 651–659, 1993.
- [15] R. Farrell and O. Feely, "Determining regions of stable behavior for chaotic sigma delta modulators," in *Irish Digital Signal Processing Contr. Conf.*, 1996, pp. 117–124.



Ronan Farrell (S'95) received the B.E. degree in electronic engineering from the National University of Ireland in 1993. Upon graduation he commenced employment as a Control Systems and Instrumentation Engineer in the chemical industry in England. In December 1994, he returned to University College Dublin to study for the Ph.D. degree in the area of nonlinear analysis techniques and sigma–delta modulators. His areas of interest include nonlinear analysis techniques, system modeling, control systems, and instrumentation.



Orla Feely (S'87–M'92) received the B.E. degree in electronic engineering from University College Dublin, in 1986, and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1990 and 1992, respectively. Since 1992, she has been a Lecturer in the Department of Electronic and Electrical Engineering at University College Dublin, teaching courses in the areas of circuit theory, electronic circuits, and nonlinear dynamics. Her research interests lie in the area of nonlinear dynamics of electronic circuits.

Dr. Feely's Ph.D. thesis on sigma–delta modulation won the D. J. Sakrison Memorial Prize for outstanding and innovative research, awarded annually by the Department of Electrical Engineering and Computer Sciences at UC Berkeley. She also received the UC Berkeley Outstanding Graduate Student Instructor Award in Electrical Engineering for 1990, and the Best Paper Award presented at the European Conference on Circuit Theory and Design, 1997. She has served as Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS PART I: FUNDAMENTAL THEORY AND APPLICATIONS, and as Guest Editor of a special issue of the *International Journal of Circuit Theory and Applications* on "Delta–Sigma Modulators and Noise Shaping."