THE EFFECT OF INTEGRATOR LEAK IN Σ–Δ MODULATION

by

Orla Feely and Leon O. Chua

Memorandum No. UCB/ERL M90/116

12 December 1990

THE EFFECT OF INTEGRATOR LEAK IN Σ–Δ MODULATION

by

Orla Feely and Leon O. Chua

Memorandum No. UCB/ERL M90/116

12 December 1990

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

THE EFFECT OF INTEGRATOR LEAK IN Σ–Δ MODULATION

by

Orla Feely and Leon O. Chua

Memorandum No. UCB/ERL M90/116

12 December 1990

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

The effect of integrator leak in $\Sigma-\Delta$ modulation *

Orla Feely and Leon O Chua†

December 12, 1990

Abstract

Oversampled sigma-delta modulators are finding widespread use in audio and other signal processing applications, due to their simple structure and robustness against circuit imperfections. Exact analyses of the system are complicated by the presence of a nonlinear element — a one-bit quantizer. The response of most researchers who have studied the system analytically has been to linearize the system and apply standard linear theory, but this approach in general does not yield correct results.

In this paper we apply theory from the field of nonlinear dynamics to provide an analytical description of the behavior of the single-loop modulator with leaky integrators. Integrator leak is inevitable in any practical circuit implementation due to finite op amp gain. The results obtained allow us to discuss in quantitative rather than qualitative terms the robustness of the sigma-delta system to this circuit imperfection.

*This work is supported in part by the Office of Naval Research under Grant N00014-89-J-1402

†The authors are with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720,USA

I Introduction

Oversampled sigma-delta ($\Sigma-\Delta$) modulation [1] has attracted much interest since it was first proposed almost twenty years ago, and the technique is now finding widespread use in audio and other signal processing applications. The basis of this method of analog-to-digital conversion is the exchange of amplitude resolution for time resolution. By operating at sampling rates well above the Nyquist rate, the number of quantization levels used to represent the signal can be reduced. A major advantage of the $\Sigma-\Delta$ modulator lies in its simple structure and its robustness against circuit imperfections and component matching inaccuracy. In the simplest oversampled $\Sigma-\Delta$ modulator — the single-loop $\Sigma-\Delta$ system — a one bit quantizer is used together with a discrete-time integrator inside a feedback loop. This basic structure can be modified by adding more feedback loops or increasing the number of quantization levels. Since such modifications increase the complexity of the system and often give rise to instability, the most commonly used $\Sigma-\Delta$ structures are the single- and double- (feedback) loop modulators.

Despite the simple structure of the modulator, exact analyses of even the single-loop system are highly non-trivial, due to the presence of a nonlinear element — a one-bit quantizer — in the feedback loop. Many researchers have approached the problem by linearizing the nonlinearity, thus allowing standard linear theory to be applied. The results thus obtained, however, in general do not yield correct quantitative or even qualitative results.

One important feature of $\Sigma-\Delta$ modulation is the appearance of periodic behavior, or limit cycles, in the output bit stream. Candy emphasised the importance of this point by proposing the $\Sigma-\Delta$ system in a paper entitled "A use of limit cycle oscillations to obtain robust analog-to-digital converters"[1]. As a result of this oscillatory behavior, the quantization noise of

the single-loop system is not white, but rather contains discrete spikes at frequencies depending on the input. This "pattern noise" can be particularly objectionable in audio applications. Higher order systems suffer from this problem to a lesser extent than does the single-loop system.

Circuit designers have found the performance of the $\Sigma-\Delta$ modulator to be insensitive to the presence of circuit imperfections. To date, however, no attempt has been made to study in a rigorous analytical way the effect of such imperfections. In this paper we focus on the effect of imperfect integration. We confine our study to the single-loop system because, as well as being widely used [2,3], the single-loop system is more amenable to analysis than higher order modulators. Integrator leak is inevitable in any practical realization of the modulator, due to finite op-amp gain. By deriving exact analytical descriptions of the effect of such leak on the behavior of the system, we can discuss in quantitative rather than qualitative terms the robustness of the system.

We approach the problem using techniques from the theory of nonlinear dynamics — in particular the field of symbolic dynamics [4]. With ideal integrators and constant input, the output of the system (when averaged over a "long enough" time period) equals the input. When integrator leak is taken into consideration, however, the input vs. average output plot is no longer linear, but rather has a fractal "staircase" structure. The resolution of the modulator is limited by the width and displacement of these steps. The results of our theoretical analysis are shown to agree exactly with computer simulations.

Section II contains a description of the ideal $\Sigma-\Delta$ system. In Section III the effect of integrator leak is included, and using graphical techniques some qualitative features of the behavior of the system are derived. Section

IV contains the key equation which will be used to test for the existence of a limit cycle, while Section V gives an algorithm, based on the Euclid algorithm from number theory, to generate the structure of the limit cycles. Finally, in Section VI these techniques are used to derive the input-output characteristic of the system, and the implications of these findings are discussed. Proofs of all lemmas and theorems are contained in the Appendix.

II Ideal Single-Loop System

The structure of the single-loop $\Sigma-\Delta$ modulator is as shown in Figure 1.

Figure 1

The only nonlinearity in the modulator is a one-bit quantizer whose output is 1 when its input is ≥ 0; -1 when its input is negative. Since the use of a one-bit quantizer implies minimal amplitude resolution, the system operates at a sampling rate many times higher than the Nyquist rate. In essence, time resolution is traded for amplitude resolution.

We will assume throughout this paper that the input to the modulator is constant. Although this condition is rarely met in practice, the very high sampling rate means that a time-varying input can be approximated by a dc level over relatively large time intervals.

The quantizer, together with a discrete-time integrator, operates in a feedback loop as shown. If the integrator is ideal, its dc gain is infinite and so the feedback connection forces the average value of the output to equal the dc input x. The analog level can be retrieved by averaging the output bit stream.

Assuming an ideal integrator and constant input x, this system is described by the first-order difference equation

$$u_{n+1} = u_n + x - sgn(u_n) \tag{1}$$

The quantizer output is represented symbolically throughout this paper by 1 (+1) and 0 (−1). The case of quantizer output $\pm\Delta$ can be analyzed using a scale change and results in no qualitative change in the behavior of the system.

This ideal single-loop system has been studied by Friedman [5] and Gray [6,7]. They have shown that for rational input the output bit stream is periodic, the average over a complete period being equal to x. With an irrational input to the modulator the output is quasiperiodic. From the perspective of dynamical system theory these results are an immediate consequence of the fact that for states u_n which lie in the interval $[g(x-1), g(x+1))$ the difference equation (1) is topologically conjugate to the well-known translation of the circle

$$\theta_{n+1} = (\theta_n + k)_{mod\ 2\pi} \tag{2}$$

As a model of a real $\Sigma-\Delta$ modulator this system is inadequate, since any slight perturbation will produce a qualitative change in the dynamics. A better model of the circuit is required — one which includes the effect of circuit nonidealities.

III $\Sigma-\Delta$ System with Leaky Integrators

One major approximation made in modeling the single-loop system by (1) is that the integrator is ideal. In any practical implementation of the modulator, of course, circuit nonidealities will result in leaky integration [8], as represented in Figure 2.

Figure 2

Taking integrator leak into account in the ideal single-loop $\Sigma-\Delta$ system, we get the more complete description

$$u_{n+1} = pu_n + g\,(x - sgn(u_n)) \tag{3}$$

Clearly when $p = g = 1$ this reduces to the ideal case (1). In practice finite op-amp gain will mean that p is less than 1, while capacitor mismatch causes g to differ from 1.

As long as $p = 1$, the map (3) on the interval $[g(x-1), g(x+1))$ is still topologically conjugate to the translation of the circle (regardless of g) so the qualitative features described in Section II still hold. If $p > 1$, we have the possibility of chaotic behavior. Since practical considerations cause p to be less than 1, this is the case which will be studied here. The input x will be assumed to lie in the range $(-1, 1)$. [If $x \geq 1$ (resp. ≤ -1) the output bit stream will eventually be fixed at 1 (resp. 0).] The 1-d map $u_n \rightarrow u_{n+1}$ given by (3) takes the form shown in Figure 3.

Figure 3

Since this graph does not intersect the identity line $id(u) = u$, there are no fixed points. Since $p < 1$, any trajectory $\{u_i\}$ will eventually reside in the region $g(x-1) \leq u_i < g(x+1)$. For this reason, we confine our attention to the restricted map $f : [g(x-1), g(x+1)) \rightarrow [g(x-1), g(x+1))$ plotted in Figure 4.

Figure 4

To locate limit cycles of f we must study the graph of the n-fold composition f^n — in particular the intersection of this graph with the identity

line. The following lemmas, proven in the Appendix, describe the structure of f^n.

Lemma 1: $f^n : [g(x-1), g(x+1)) \to [g(x-1), g(x+1))$ is one-to-one, where $f^n(u) = f \circ f \circ f \circ \cdots \circ f(u)$ [n times].

Lemma 2: Let D_n be the set of points in $[g(x-1), g(x+1))$ where f^n is discontinuous. Then $D_n = D_{n-1} \cup \{u > g(x-1) \mid f^{n-1}(u) = 0\}$.

Lemma 3: The graph of f^n consists of at most $n+1$ straight line segments, each of slope p^n.

Lemma 4: For each $n \geq 1$ there exists $u^* \in [g(x-1), g(x+1))$ such that $\hat{f} : [u^*, u^* + 2g) \to [g(x-1), g(x+1))$ given by

$$\hat{f}(u) = \begin{cases} f^n(u) & \text{if } u^* \leq u < g(x+1) \\ f^n(u-2g) & \text{if } g(x+1) \leq u < x^* + 2g \end{cases}$$

is monotone increasing.

Lemma 5: If $f^n(0) < 0$ or $f^n(0^-) \geq 0$ f has no limit cycle of period n.

Using these lemmas, we derive the following theorem, also proven in the Appendix.

Theorem 6: Let N be the least positive integer such that $f^N(0) \geq 0$ and $f^N(0^-) < 0$. [1] f has a globally asymptotically stable limit cycle of least period N.

Several key facts follow immediately from this theorem.

1. Theorem 6 gives two inequalities in x and p which must be satisfied if a period N limit cycle is to exist. Since our interest is in the dependence

[1] $f(x^-) = \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} f(x - \epsilon)$

of the limit cycles on the dc input x for fixed p, we can restate these inequalities in the form $x_{min} \leq x < x_{max}$.

Example: A period-2 limit cycle exists iff $f(g(x-1)) \geq 0$ and $f(g(x+1)) < 0$, i.e. iff

$$\frac{p-1}{p+1} \leq x < \frac{-p+1}{p+1}$$

There are two possibilities for the period-3 limit cycles:

Case 1: $x \geq (-p+1)/(p+1)$. A period-3 limit cycle exists here iff

$$\frac{p^2-p+1}{p^2+p+1} \leq x < \frac{-p^2+p+1}{p^2+p+1}$$

Case 2: $x < (p-1)/(p+1)$. A period-3 limit cycle exists here iff

$$\frac{p^2-p-1}{p^2+p+1} \leq x < \frac{-p^2+p-1}{p^2+p+1}$$

2. One could proceed in this fashion to locate all limit cycles of any period N. An equivalent but more efficient method will be given in Section V.

3. In the ideal single-loop $\Sigma-\Delta$ system each limit cycle could exist for a *fixed* value of x only; the average value of the output over a complete limit cycle being equal to x. In the non-ideal case we see that each limit cycle can exist over a range of x values, introducing error between input and output which depends on the integrator leak factor $(1-p)$.

4. All limit cycles are globally asymptotically stable — i.e. if a limit cycle exists for a particular x, all trajectories will converge to it, regardless of initial condition. As a consequence of this fact we can say that for fixed x and p there can exist at most one limit cycle.

5. Convergence to the limit cycle is asymptotic in the state space, the space in which we have studied the 1-d map. In the $\Sigma-\Delta$ system, however, the output is not the state u_i but $sgn(u_i)$, which takes the values ± 1 (represented symbolically by 1 and 0). From the above results we know that if f has a limit cycle of period N, the graph of f^N consists of N linear segments each of slope p^N. It is proven in Lemma 7 of the Appendix that the kth iterate under f of any of the intervals of continuity of f^N, for any $k \geq 1$, will lie in either $[0, g(x+1))$ or $[g(x-1), 0)$. It follows that any two states lying on one of these N segments will give exactly the same sequence of zeros and ones at the quantizer output. In other words, although the quantizer input converges asymptotically to the limit cycle, the quantizer output begins its limit cycle as soon as the state u_i enters the region $[g(x-1), g(x+1))$.

IV Tsypkin's method

In Section III the 1-d map of (3) was used to derive certain properties of the limit cycles in the leaky $\Sigma-\Delta$ system. Our goal is now to derive the dependence of these limit cycles on the dc input x. This can be done using the inequalities of Section III, but the computation is time consuming. In this section and the next, we derive a more efficient procedure based on techniques from number theory.

If a limit cycle of period N exists in the single-loop $\Sigma-\Delta$ system, we can sum equation (3) over the limit cycle to find the condition

$$u_k = \frac{g}{1-p}x - \frac{g}{1-p^N}\sum_{i=k}^{N+k-1} p^{N+k-1-i}\, sgn\,(u_i) \qquad (4)$$

for $1 \leq k \leq N$. We can use this equation to check for the existence of a given

limit cycle, following the procedure:

1. substitute the assumed bit sequence for the sgn terms in (4);
2. calculate the resulting sequence of states u_i in terms of x, g, and p;
3. impose the conditions that these N u_i [quantizer inputs] must be of such polarities as to give the assumed bit sequence at the quantizer output;
4. calculate the range of possible values for x from the N inequalities in step 3.

This is the basis of Tsypkin's method in relay control theory [9]. As an example, let us derive the range of x values which result in the period-2 limit cycle 10 at the quantizer output.

1. substitute in (4) the conditions $sgn(u_1) = 1$, $sgn(u_2) = -1$;
2. calculate u_1, u_2:

$$\begin{aligned} u_1 &= \frac{g}{1-p}x - \frac{g}{1-p^2}(p-1) \\ u_2 &= \frac{g}{1-p}x - \frac{g}{1-p^2}(-p+1) \end{aligned}$$

3. assign appropriate polarities:

$$\begin{aligned} \frac{g}{1-p}x - \frac{g}{1-p^2}(p-1) &\geq 0 \\ \frac{g}{1-p}x - \frac{g}{1-p^2}(-p+1) &< 0 \end{aligned}$$

4. calculate the range of possible x values from step 3:

$$\begin{aligned} x &\geq \frac{p-1}{1+p} \\ x &< \frac{1-p}{1+p} \end{aligned}$$

The necessary and sufficient condition for the existence of the limit cycle 10 is that the dc input lie in the range $[\frac{p-1}{p+1}, \frac{1-p}{p+1})$. Once again, this method of analysis is exact, but is not useful in any comprehensive study as the amount of labor required to test all possible limit cycles is excessive.

V Euclid algorithm

In [10] Hein and Zakhor show how the limit cycles in the ideal single-loop $\Sigma-\Delta$ system are obtained by stepping through the tree of transition points shown in Figure 5.

Figure 5

To find the limit cycle corresponding to a rational dc input x we start at the top level by comparing x with 0 and setting the initial trajectory to 10. If $x > 0$ we follow the right branch; otherwise we follow the left branch. We continue through the tree in this fashion, adding 1 (resp. 0) to our trajectory each time we take the right (resp. left) branch. If at some level the transition point hit is equal to x, we stop, having found the limit cycle. The limit cycle corresponding to $x = \frac{1}{5}$, for example, is 10101.

In fact, following the method outlined in Section III would yield exactly such a tree for the nonideal case, with the transition points replaced by transition *intervals* whose bounds are functions of p. It can be shown in this manner that the limit cycles which appear at the output of the leaky $\Sigma-\Delta$ modulator are precisely those which appear in the ideal modulator. This statement will not be proven at this point, as it will fall out as a natural consequence of later analysis.

In [5], Friedman shows that the limit cycles which appear at the output of the ideal single-loop $\Sigma-\Delta$ modulator with constant rational input x can

be obtained by applying a particular form of the Euclid algorithm to the continued fraction expansion of x. These limit cycles have the property that the zeros and ones are distributed as uniformly as possible in the output bit stream. The difference between x and the average output as calculated over N successive bits is minimal — when the average is taken over a complete limit cycle this difference is zero. It is reasonable to begin our study of the leaky single-loop structure by determining which of these limit cycles will persist if p is decreased below 1. For reasons which will become clear later we will use the following slightly modified form of the Euclid algorithm.

Algorithm: To find the limit cycle with a ones and $(b-a)$ zeros

(i) Form the continued fraction expansion of a/b

$$\frac{a}{b} = \cfrac{1}{\alpha_1 + \cfrac{1}{\alpha_2 + \cfrac{1}{\alpha_3 + \cfrac{1}{\ddots + \cfrac{1}{\alpha_n}}}}}$$

This fraction will be denoted $[\alpha_1, \alpha_2, \ldots, \alpha_n]$ for convenience. Note that the expansion is not unique: $[\alpha_1, \alpha_2, \ldots, \alpha_n]$ and $[\alpha_1, \alpha_2, \ldots, \alpha_n - 1, 1]$ correspond to the same fraction. By disallowing expansions with final coefficient equal to one we remove this ambiguity.

(ii) Define

$$\begin{aligned} S_0 &= 0 \\ S_1 &= 1(0)^{\alpha_1 - 1} \\ &\vdots \\ S_k &= S_{k-2}(S_{k-1})^{\alpha_k} \end{aligned}$$

$$\vdots$$

$$S_n = S_{n-2}(S_{n-1})^{\alpha_n}$$

where $(S_j)^{\alpha_k}$ consists of the block S_j repeated α_k times, and the α_i are the coefficients of the continued fraction expansion.

Note that at each stage of the iteration the sequence S_k is the shortest possible sequence of zeros and ones where the fraction of ones present is $[\alpha_1, \alpha_2, \ldots, \alpha_k]$. The zeros and ones are distributed as uniformly as possible throughout each S_k.

The sequences derived using this algorithm are identical to those derived by any other of the many possible forms of the Euclid algorithm, modulo a barrel shift. [A barrel shift is a shift with wraparound, so *abcde* is a barrel shifted version of *deabc*. Whenever we use the word shift we shall be referring to a barrel shift.] Applying this algorithm to any of Friedman's examples, for example, gives a shifted version of the sequences obtained by his method.

As in [5], S_n is the limit cycle at the output of the ideal single-loop system with constant input $x = 2a/b - 1$. This value comes from the fact that a 1 (resp. 0) bit output corresponds to an analog value of 1 (resp. -1). The average value over the limit cycle, so, equals $\frac{a-(b-a)}{b}$. Since the ideal system has infinite dc gain in the forward path, this limit cycle can exist only when $x = 2a/b - 1$

For convenience, given the bit sequence $S = (s_1, s_2, \ldots, s_N)$ we define the corresponding sequence $V = (v_1, v_2, \ldots, v_N)$ by

$$v_i = \begin{cases} 1 & \text{if } s_i = 1 \\ -1 & \text{if } s_i = 0 \end{cases}$$

Before presenting an example of this algorithm, we define two more terms which will be of use later. In describing a limit cycle, a shift is clearly of no

consequence. For reasons which will become apparent later, however, the particular version of the limit cycle produced by this algorithm (as opposed to any other version of the Euclid algorithm) is indeed significant. This motivates the following definitions:

For n even we term S_n the R-sequence corresponding to S_n. The first two bits of S_n in this case are 01. Interchanging these two bits gives the L-sequence of S_n. For n odd S_n begins with the bits 10 — we term this the L-sequence of S_n and obtain the R-sequence by interchanging the first two bits. It is proven in Lemma 8 of the Appendix that the L-sequence is a shifted version of the R-sequence.

Example: To find the limit cycle with 24 ones and 31 zeros:

(i) Find the continued fraction expansion of 24/55, which is [2,3,2,3].

(ii) Calculate the S_i:

$$S_0: \quad 0$$

$$S_1: \quad 10$$

$$S_2: \quad 0101010$$

$$S_3: \quad 1001010100101010$$

$$S_4: \quad 010101010010101001010101001010100101010010101001010100101010$$

S_4 is the required limit cycle.

(iii) The R-sequence of 24/55 is S_4. The L-sequence of 24/55 is

$$100101010010101001010101001010100101010010101001010100101010$$

Note that this is identical to the R-sequence if we shift the underlined bit into first place. Lemma 8 through Corollary 13 of the Appendix prove

various properties of these sequences which will be of use in proving our main theorems.

VI Derivation of x Bounds

Using the algorithm of Section V it is possible to determine all limit cycles which occur in the ideal single-loop system with constant input. In order to determine whether or not these limit cycles occur in the leaky system, we apply the Tsypkin-type approach outlined in Section IV to $S = (s_1, s_2, \ldots, s_N)$ or, equivalently, $V = (v_1, v_2, \ldots, v_N)$. (Remember that $s_i \in \{0,1\}$ and $v_i \in \{-1,1\}$.) Restating (4) for this case, we find that the limit cycle S (or V) can occur at the output only when

$$\frac{\sum_{i=k_1}^{N+k_1-1} p^{N+k_1-1-i} v_i}{p^{N-1}+p^{N-2}+\ldots+1} > x \geq \frac{\sum_{i=k_2}^{N+k_2-1} p^{N+k_2-1-i} v_i}{p^{N-1}+p^{N-2}+\ldots+1} \tag{5}$$

where k_1 and k_2 (both $\leq N$) are chosen subject to the constraint $v_{k_1} = -1$ and $v_{k_2} = 1$. The denominators of both bounds in (5) are clearly the same: they are just the $(N-1)$-th order polynomial in p with all coefficients equal to one. We term this $1_N(p)$. To find the greatest lower bound on the range of inputs which gives rise to the limit cycle S (or V) it is necessary to find the shift k_2 of V which maximizes the polynomial

$$\sum_{i=k_2}^{N+k_2-1} p^{N+k_2-1-i} v_i$$

subject to the constraint $v_{k_2} = 1$. Call this maximal polynomial $l(p)$. The least upper bound is produced by finding the shift of V which minimizes the polynomial

$$\sum_{i=k_1}^{N+k_1-1} p^{N+k_1-1-i} v_i$$

subject to the constraint $v_{k_1} = -1$. Call this minimal polynomial $r(p)$.

The bounds on x, so, are derived by finding the appropriate shifts of S which produce $r(p)$ and $l(p)$. In fact it turns out that these shifts are those given by the R- and L-sequences. Theorem 14 of the Appendix proves that the L-sequence gives $l(p)$ — the proof for $r(p)$ is identical.

Theorem 14: Given $S = (s_1, s_2, \ldots, s_N)$ (or, equivalently, $V = (v_1, v_2, \ldots, v_N)$) and $L = (s_k, \ldots, s_{k-1})$ the L-sequence of S. There exists no $\hat{k}$ such that $s_{\hat{k}} = 1$ and

$$\sum_{i=\hat{k}}^{N+\hat{k}-1} p^{N+\hat{k}-1-i} v_i > \sum_{i=k}^{N+k-1} p^{N+k-1-i} v_i$$

for any $p, 0 < p < 1$.

Any limit cycle S obtained by the Euclid algorithm can exist in the leaky system iff $l(p) < r(p)$, where, by Theorem 14, the coefficients of $l(p)$ (resp. $r(p)$) are given by the L-sequence (resp. R-sequence) of S. That this is always true follows from the fact that the L- and R-sequences are identical in all positions except the first two.

$$r(p) - l(p) = \frac{2\,p^{N-2}\,(p-1)}{p^{N-1} + \ldots + p + 1} > 0 \quad \text{for } p \in (0,1) \tag{6}$$

Thus any limit cycle which can exist at the output of the ideal $\Sigma - \Delta$ system can also exist at the output of the leaky system. Figure 6 shows the dependence of the average output over a limit cycle on the dc input x for $p = 0.8$.

Figure 6

The plot was obtained by choosing 20000 dc input values uniformly spaced in the interval $[-1, 1]$. The form of the graph is that of the well known

devil's staircase [11], the qualitative form being replicated at varying levels of resolution. The staircase contains a step at average output q, where q is any rational number in the range $(-1, 1)$. From (6) it is clear that the width of the steps corresponding to limit cycles with period N decreases with N. The widest step is that corresponding to the limit cycle 01 (average output 0) and the next widest are those corresponding to limit cycles 101 (average value $\frac{1}{3}$) and 100 (average value $-\frac{1}{3}$). Figure 7 shows the 27 widest steps predicted by the analysis of Section V for $p = 0.8$.

Figure 7

The correspondence between theory and simulation is clear. (6) also predicts that the width of the steps decreases as p approaches 1. Figure 8 shows the staircase for $p = 0.99$.

Figure 8

Figure 9 shows the locations of the 27 widest steps for varying x and p.

Figure 9

At p = 1, as expected, the widths of all steps shrink to zero, and the "steps" are just the rational numbers. The difference between input and average output is seen to be due to two features — the non-zero step width, and the divergence of the step centers from their ideal values. Figure 10 shows the resultant input-error plot for $p = 0.99$ (the same value as in Figure 8).

Figure 10

One minor point to note is that the steps of the staircase can be taken to be the closed intervals $[x_{min}, x_{max}]$ instead of the half-open intervals $[x_{min}, x_{max})$ defined by our analysis. This is clear from study of the 1-d map. At $x = x_{max}$, the graph of f^N takes the form shown in Figure 11.

Figure 11

There are no fixed points, but the trajectories of f^N converge to the *virtual* fixed points at the rightmost end of each segment, giving at the output the same limit cycle that would be observed for any x inside the half-open interval.

Finally, it was claimed earlier that the limit cycles derived using the Euclid algorithm are the only limit cycles that can appear. This is a consequence of the following theorem, proven in the Appendix.

Theorem 15: The complement C of the projection onto the x-axis of the devil's staircase has measure zero.

Assume another limit-cycle exists — one not given by the Euclid algorithm. Since all limit cycles are globally asymptotically stable, the x values which give rise to this limit cycle must lie in C. But both the 1-d map approach and the Tsypkin approach tell us that any limit cycle that exists in the system will persist over an interval of x values. Since C contains no intervals, this new limit cycle cannot exist.

Theorem 15 indicates another major difference between the limit cycle behavior of the leaky system and that of the ideal system. In the ideal modulator, where $p = 1$, the set of inputs in the range $(-1, 1)$ which give rise to limit cycles at the output has measure zero. If $p > 1$, however, the *complement* of this set has measure zero. In other words, if $p = 1$ almost

no [in the probabilistic sense] inputs give rise to limit cycles, but as soon as p is decreased below 1 almost all inputs give rise to limit cycles. This qualitative change in behavior is a consequence of the fact that f, viewed as a map on the circle, changes from a continuous map to a discontinuous one as p decreases below 1.

VII Conclusions

The method presented in the paper allows us to determine exactly the effect of integrator leak on the performance of the single loop Σ–Δ modulator with dc input. These effects can be summarized as follows:

1. Each limit cycle that can appear at the output of the ideal modulator can appear in the leaky system. The difference is that each limit cycle persists over a range of inputs in the nonideal case.

2. Almost all [in the probabilistic sense] dc inputs give rise to a limit cycle at the quantizer output. That is, if x is chosen from a uniform distribution on $(-1, 1)$, the probability that this input gives rise to a limit cycle is 1.

3. The input versus average output characteristic takes the form of the well-known devil's staircase. Only when the input is 0, 1 or -1 will the average value of the output equal the input. For all other inputs the finite dc gain of the integrators leads to a divergence between average output and input. From Figure 9 we see that this difference is due to two features: non-zero step width and divergence of steps from their ideal ($p = 1$) locations.

4. For p close to 1, a Taylor series truncation can be used to show that the error due to step divergence is approximately $\pm(1-p)x$ for $|\ x\ | \ll 1-2(1-p)$. This explains the underlying trend of the graph of Figure 10. This error could easily be removed by introducing a gain in the decoder.

5. No decoder could remove the error due to non-zero step width. The consequent loss of resolution is a highly nonlinear function of the input. The greatest loss of resolution occurs in the neighborhood of the rational numbers with lowest denominators (after the affine transformation described earlier). Around the dc input level 0, for example, the range of inputs $[-(1-p)/(1+p), (1-p)/(1+p)]$ will all give rise to the same limit cycle at the output. Near dc input $\frac{1}{3}$, all inputs in the range $[(p^2-p+1)/(p^2+p+1), (-p^2+p+1)/(p^2+p+1)]$ produce the same limit cycle. The width of each of these uncertainty intervals is, for p close to 1, approximately $2p^{N-2}(1-p)/N$, where N is the period of the limit cycle.

6. The capacitor mismatch factor g has no significant effect on the behavior described in this paper. The integrator leak is the crucial quantity.

We have not discussed in this paper the effect of the oversampling ratio, but our results could easily be modified to take this into account. Assuming a simple averaging decoder, one can derive bounds on the input versus average output plot where the average is now taken over M bits instead of a complete limit cycle. In this way, circuit designers can quantify the trade-off between oversampling ratio and op amp gain.

VIII Acknowledgements

The authors would like to thank Cormac Conroy for first bringing $\Sigma - \Delta$ modulation to our attention, and for many helpful suggestions. Thanks also to Professor Avideh Zakhor and Soren Hein for interesting discussions on the subject.

A Appendix

Lemma 1: $f^n : [g(x-1), g(x+1)) \rightarrow [g(x-1), g(x+1))$ is one-to-one, where $f^n(u) = f \circ f \circ f \circ \cdots \circ f(u)$ [n times].

Proof: Follows immediately from the fact that f is one-to-one.

Lemma 2: Let D_n be the set of points in $(g(x-1), g(x+1))$ where f^n is discontinuous. Then $D_n = D_{n-1} \cup \{u > g(x-1) \mid f^{n-1}(u) = 0\}$.

Proof: If f^{n-1} is discontinuous at u, f^n is also discontinuous at u. Thus $D_{n-1} \subset D_n$. If $f^{n-1}(u)$ equals zero and $u > g(x-1)$, f^n is discontinuous at u. Thus $D_{n-1} \cup \{u > g(x-1) \mid f^{n-1}(u) = 0\} \subset D_n$.

If f^{n-1} is continuous at u and f is continuous at $f^{n-1}(u)$, f^n is continuous at u. Thus $D_n \subset D_{n-1} \cup \{u > g(x-1) \mid f^{n-1}(u) = 0\}$, and the lemma is proven.

Lemma 3: The graph of f^n consists of at most $n+1$ straight line segments, each of slope p^n.

Proof: By induction.

The statement clearly holds for $n = 1$.

Assume the statement is true for $n = k$, so D_k has at most k points. Then D_{k+1} ($= D_k \cup \{u > g(x-1) \mid f^k(u) = 0\}$) has at most $k+1$ points, by Lemma 1. These points divide the domain of f^{k+1} into at most $k+2$ intervals. On each of these intervals f^k is a continuous affine linear function of slope p^k whose image does not include the origin. It follows that on each of these intervals $f^{k+1} = f \circ f^k$ is a continuous affine linear function of slope p^{k+1}.

Lemma 4: For each $n \geq 1$ there exists $u^* \in [g(x-1), g(x+1))$ such that $\hat{f} : [u^*, u^* + 2g) \to [g(x-1), g(x+1))$ given by

$$\hat{f}(u) = \begin{cases} f^n(u) & \text{if } u^* \leq u < g(x+1) \\ f^n(u-2g) & \text{if } g(x+1) \leq u < u^* + 2g \end{cases}$$

is monotone increasing.

Proof: By induction.

The statement is clearly true for $n = 1$, with $u^* = 0$. Assume it holds for $n = k$.

CASE 1: $Im(f^k) \subset \overline{\mathbb{R}^+}$ (resp. $\mathbb{R}^-$). Since f is monotone increasing on $\overline{\mathbb{R}^+}$ (resp. $\mathbb{R}^-$), the property will hold for f^{k+1} with $u^*_{k+1} = u^*_k$.

CASE 2: $Im(f^k)$ includes both positive and negative values. Say

$$f^k(u) \begin{cases} < 0 & \text{for } g(x-1) \leq u < \hat{u} \\ \geq 0 & \text{for } \hat{u} \leq u < u^* \\ < 0 & \text{for } u^* \leq u < g(x+1) \end{cases}$$

(The only other possibility, where the zero crossing of f^k lies to the right of u^*, is similar.)

Clearly f^{k+1} is increasing on each of $[g(x-1), \hat{u})$, $[\hat{u}, u^*)$ and $[u^*, g(x+1))$. f^{k+1} is increasing at u^* since $f^k((u^*)^-) > 0$ and $f^k(u^*) < 0$. Also, since

$f^k(g(x-1)) > f^k(g(x+1)^-)$ and both are negative we have $f^{k+1}(g(x-1)) > f^{k+1}(g(x+1)^-)$. Thus the property holds in this case also.

Lemma 5: If $f^n(0) < 0$ or $f^n(0^-) \geq 0$ f has no limit cycle of period n.

Proof: CASE 1: $0 \notin Im(f^{n-1})$. In this case (by the proof of Lemma 3) the graph of f contains at most n straight line segments, each of slope $p^n < 1$. For f to have a limit cycle of minimum period n, it is necessary that the identity line $id(u) = u$ intersect the graph of f^n n times. However, if $f^n(0) < 0$ (respectively $f^n(0^-) \geq 0$), the segment of this graph immediately to the right (resp. left) of the origin cannot intersect the identity line, so there can be at most $n-1$ intersections and therefore no period n limit cycle.

CASE 2: There exists $\hat{u}$ such that $f^{n-1}(\hat{u}) = 0$.

If $\hat{u} = g(x-1)$ the graph of f^n has at most n segments so, as in case 1, there is no limit cycle of period n.

If $\hat{u} \neq g(x-1)$ then $f^n(\hat{u}) = g(x-1)$, $f^n(\hat{u}^-) = g(x+1)$. In this situation clearly neither the segment of the graph immediately to the right nor that immediately to the left of $\hat{u}$ can intersect the identity line. Since two of the possible $n+1$ segments are now known not to intersect the identity line, there can be no limit cycle of period n.

Theorem 6: Let N be the least positive integer such that $f^N(0) \geq 0$ and $f^N(0^-) < 0$. f has a globally asymptotically stable limit cycle of least period N.

Proof: Since N is the least positive integer satisfying the conditions, f has

no limit cycle of period less than N.

$$f^N(0) \geq 0 \Rightarrow f^{N-1}(g(x-1)) \geq 0$$

$$f^N(0^-) < 0 \Rightarrow f^{N-1}(g(x+1)) < 0$$

It follows from the proof of Lemma 4 that the graph of f^N is monotone increasing on $[g(x-1), g(x+1))$. Since

$$g(x-1) \leq f^N(g(x-1)) < f^N(0^-) < 0$$

and f is monotone increasing on $[g(x-1), 0)$, there must be at least one intersection point in $[g(x-1), 0)$ where $u = f^N(u)$. Since f has no limit cycle of period less than N, there must be N such points and so f has a limit cycle of period N. Global asymptotic stability follows immediately from the form of the graph of f^N.

Lemma 7: Given that f has a limit cycle of minimum period N, take the kth iterate under f (where k is any positive integer) of any of the intervals of continuity of f^N. The result is an interval in either $[0, g(x+1))$ or $[g(x-1), 0)$.

Proof: By our previous results, (i) the graph of f^N consists of N affine linear segments, each of slope p^N; (ii) $f^N(0) \geq 0$ and $f^N(0^-) < 0$; and (iii) f^N is monotone increasing on $[g(x-1), g(x+1))$. Thus the graph of f^k for $k \geq N$ consists of N affine linear segments, none of which can intersect the u-axis. This proves the lemma for $k \geq N$.

Suppose that for some $\hat{k} < N$ the image under $f^{\hat{k}}$ of one of our N intervals contains the origin at a non-boundary point. f^N will then be discontinuous at some non-boundary point of this interval, which is not possible. This proves the lemma.

Lemma 8: For $k \geq 1$, the sequence $\hat{S}_k$ obtained by interchanging the first two bits of S_k is a shifted version of S_k.

Proof: $S_k = S_{k-2}\,(S_{k-1})^{\alpha_k}$, so if we can show that the sequences $S_{k-2}\,S_{k-1}$ and $S_{k-1}\,S_{k-2}$ are identical but for the reversal of the first two bits then the lemma is proven. We will proceed by induction.

Assume sequence $S_{i-2}\,S_{i-1}$ is obtained from $S_{i-1}\,S_{i-2}$ by reversal of the first two bits. Then $S_{i-1}\,S_i = S_{i-1}\,S_{i-2}\,(S_{i-1})^{\alpha_i}$, which is (by our assumption) identical to $S_{i-2}\,S_{i-1}\,(S_{i-1})^{\alpha_i}$ in all positions except the first two. But $S_{i-2}\,S_{i-1}\,(S_{i-1})^{\alpha_i}$ is $S_i\,S_{i-1}$. Thus, since it is clear that $S_1\,S_2$ and $S_2\,S_1$ are identical in all positions except the first two, it has been shown that sequences $S_{k-2}\,S_{k-1}$ and $S_{k-1}\,S_{k-2}$, for $k \geq 2$, are obtained from each other by reversal of the first two bits.

The lemma as stated follows from this fact. $S_k = S_{k-2}\,(S_{k-1})^{\alpha_k}$ for $k \geq 2$, so $S_{k-1}\,S_{k-2}\,(S_{k-1})^{\alpha_k - 1}$ is a shifted version of S_k. But $S_{k-1}\,S_{k-2}\,(S_{k-1})^{\alpha_k - 1}$ is obtained from $S_{k-2}\,S_{k-1}\,(S_{k-1})^{\alpha_k - 1}$ $(= S_k)$ by changing the first two bits, proving the lemma. The case where $k = 1$ is trivial.

Lemma 9: Given a subsequence r of S_{k+1}, of length less than l_k, the length of S_k. r is a subsequence of S_k. [Note: All inclusions are modulo a shift.]

$S_{k+1} = S_{k-1}\,(S_k)^{\alpha_{k+1}}$. Clearly if r is a subsequence of $(S_k)^{\alpha_{k+1}}$ it is a subsequence of S_k. Therefore we need only consider those r contained in

$$S_k\,S_{k-1}\,S_k = S_{k-2}\,(S_{k-1})^{\alpha_k}\,S_{k-1}\,S_{k-2}\,(S_{k-1})^{\alpha_k}.$$

We want to show that such an r is a subsequence of $S_{k-2}\,(S_{k-1})^{\alpha_k}$. This is trivial for all r which do not intersect all $(\alpha_k + 1)$ of the S_{k-1} in the longest

S_{k-1} block. There are three cases to be considered:

$$\text{(i)} \qquad \underbrace{S_{k-2}\,(S_{k-1})^{\alpha_k}\,S_{k-1}}_{r}$$

(This notation signifies that the subsequence r is the concatenation of a block from the end of S_{k-2}, α_k S_{k-1} blocks and a block from the start of S_{k-1}.) Since the length of r is less than l_k, and $l_k = \alpha_k l_{k-1} + l_{k-2}$, r does not contain the first two elements of S_{k-2}. Thus r is also given by

$$\underbrace{S_{k-1}\,S_{k-2}\,(S_{k-1})^{\alpha_k-1}\,S_{k-1}}_{r},$$

by Lemma 8. This r is clearly a subsequence of S_k.

$$\text{(ii)} \qquad \underbrace{S_{k-1}\,(S_{k-1})^{\alpha_k-1}\,S_{k-1}}_{r}$$

Again by Lemma 8, r is also given by

$$\underbrace{S_{k-1}\,S_{k-2}\,(S_{k-1})^{\alpha_k-1}\,S_{k-1}}_{r},$$

so it is a subsequence of S_k.

$$\text{(iii)} \qquad \underbrace{S_{k-1}\,(S_{k-1})^{\alpha_k}\,S_{k-2}}_{r}.$$

Since the length of r is less than l_k, r is given by

$$\underbrace{S_{k-2}\,(S_{k-1})^{\alpha_k}\,S_{k-2}}_{r},$$

which is contained in S_k.

Corollary 10: Given a subsequence r of S_{k+1} of length l_k, where the first bit of r is the same as that of S_k. r is a shifted version of S_k.

Proof: It is easy to see, by induction, that for $k \geq 1$, the first two bits of S_k are of opposite sign. With this information, the proof of the corollary is identical to that of Lemma 9, except for case (i), where we say "Since r has length l_k, and r has the same first bit as S_k (and therefore S_{k-2}), r does not contain the first two elements of S_{k-2} ....." , and case (iii), where we say "Since r has length l_k.....".

Lemma 11: Given r, r^* subsequences of S_k of the same length.
Let Z_r denote the total number of zeros in r, and Z_{r^*} denote the total number of zeros in r^*.
Then $\mid Z_r - Z_{r^*} \mid \leq 1$.

Proof: By induction. The statement is clearly true for $k = 1, 2$. Assume it holds for $k \leq i-1$. Recall $S_i = S_{i-2}\,(S_{i-1})^{a_i}$

If r has length less than l_{i-1}, Lemma 9 and the inductive assumption imply $\mid Z_r - Z_{r^*} \mid \leq 1$.

Otherwise, if r and r^* are both subsets of $(S_{i-1})^{a_i}$, the inductive assumption again implies $\mid Z_r - Z_{r^*} \mid \leq 1$.

Finally, suppose r is of the form

$$\underbrace{S_{i-1}}_{a \text{ bits}} \underbrace{\ldots S_{i-1}\, S_{i-2}\, S_{i-1} \ldots}_{r} \underbrace{S_{i-1}}_{b \text{ bits}}$$

(Once again, this notation signifies that r is the concatenation of the last $(l_{i-1} - a)$ bits of S_{i-1}, a number of S_{i-1} blocks, S_{i-2}, a number of S_{i-1} blocks and the first $(l_{i-1} - b)$ bits of S_{i-1}.) Since we have proven the lemma for the case where r and r^* are subsets of $(S_{i-1})^{a_i}$, we will use Lemma 8 to remove

the S_{i-2} term from consideration. If $a \neq 1$ and $b \neq l_{i-2} - 1$, $Z_r = Z_{\hat{r}}$, where $\hat{r}$ is given by

$$\underbrace{S_{i-2}}_{a}\underbrace{S_{i-1}\ldots S_{i-1}}_{\hat{r}}\underbrace{}_{b} = \underbrace{S_{i-1}\ldots S_{i-1}}_{\hat{r}} \quad \text{if } a \neq 1$$

$$\text{or} \quad \underbrace{S_{i-1}}_{a}\underbrace{S_{i-1}\ldots S_{i-2}}_{\hat{r}}\underbrace{}_{b} \rightarrow \underbrace{S_{i-1}\ldots S_{i-1}}_{\hat{r}} \quad \text{if } b \neq l_{i-2} - 1$$

It follows that the lemma holds if neither r nor r^* has $a = 1$ and $b = l_{i-2} - 1$. The only remaining case is that where r or r^* is of the form

$$\underbrace{S_{i-1}}_{1}\underbrace{\ldots S_{i-1}\, S_{i-2}\, S_{i-1}\ldots S}_{r}\underbrace{}_{l_{i-2}-1}\underbrace{S_{i-1}}$$

In this case the length of r is $M\, l_{i-1}$. If S_{i-2} begins with 0, $Z_r = M\, Z_{S_{i-1}} + 1$; if S_{i-2} begins with 1, $Z_r = M\, Z_{S_{i-1}} - 1$. But if S_{i-2} begins with 0 Z_{r^*} is either $M\, Z_{S_{i-1}} + 1$ or $M\, Z_{S_{i-1}}$. Similarly, if S_{i-2} begins with 1 Z_{r^*} is either $M\, Z_{S_{i-1}} - 1$ or $M\, Z_{S_{i-1}}$. The lemma holds in this case also.

Lemma 12: Given S_k with first element 1. No shift of S_k which keeps a 1 in the first position can move the zeros to a higher position. That is, if

$$Z_{S_{k1,\ldots r}}$$

denotes the number of zeros in positions 1 through r of S_k, there exists no shift S_k^* of S_k beginning with 1 such that

$$Z_{S_{k1,\ldots r}^*} > Z_{S_{k1,\ldots r}}$$

for some r.

Proof: By induction. The statement holds for $k = 1, 2$. Assume it holds for $k \leq i - 2$. By Corollary 10, the first l_{i-2} bits of S_i are a shifted version

of S_{i-2}. Thus to maximize the cumulative zero count over the first l_{i-2} bits we must shift S_i so that the first l_{i-2} bits are precisely S_{i-2} (unshifted). [This follows from the inductive assumption.] Lemma 8, together with the inductive assumption, imply that our opening S_{i-2} block must be followed by an S_{i-3} block. Since $S_i = S_{i-2}\,(S_{i-3}\,(S_{i-2})^{\alpha_{i-1}})^{\alpha_i}$, the optimal shift is of the form

$$S_{i-2}\,(S_{i-3}\,(S_{i-2})^{\alpha_{i-1}})^{\beta}\,S_{i-2}\,(S_{i-3}\,(S_{i-2})^{\alpha_{i-1}})^{\alpha_i-\beta-1}\,(S_{i-3}\,(S_{i-2})^{\alpha_{i-1}-1}).$$

[Here we have used the fact that $S_{i-2}\,(S_{i-3}\,(S_{i-2})^{\alpha_{i-1}})^{\alpha_i}$ contains no "hidden" $S_{i-2}\,S_{i-3}$ blocks — this is a consequence of Lemma 8.] It follows immediately, since S_{i-2} begins with a 1 and S_{i-3} with a 0, that our optimal shift is given by $\beta = \alpha_i$, i.e. the optimal shift is just S_i.

One point remains to be mentioned, concerning the existence of an optimal shift. Suppose

$$Z_{S_{k_1,\ldots,r}} < Z_{S^*_{k_1,\ldots,r}}$$

but

$$Z_{S_{k_1,\ldots,t}} > Z_{S^*_{k_1,\ldots,t}}$$

— if this situation is possible there may not be an optimal shift. Lemma 11 guarantees that this cannot happen, since in the above case

$$Z_{S_{k_{r+1},\ldots,t}} - Z_{S^*_{k_{r+1},\ldots,t}} = Z_{S_{k_1,\ldots,t}} - Z_{S^*_{k_1,\ldots,t}} - Z_{S_{k_1,\ldots,r}} + Z_{S^*_{k_1,\ldots,r}}$$

$$> 1 + 1,$$

contradicting Lemma 11.

Corollary 13: Given S_k with first element 0. By Lemma 8, the sequence L_k obtained from S_k by changing the first two bits is a shifted version of S_k. L_k

is the shift of S_k which maximizes the cumulative zero total while keeping a 1 in the first position. That is, if

$$Z_{L_{k1,\ldots r}}$$

denotes the number of zeros in positions 1 through r of L_k, there exists no shift L_k^* of L_k beginning with 1 such that

$$Z_{L_{k1,\ldots r}^*} > Z_{L_{k1,\ldots r}}$$

for some r.

Proof: The existence of an optimal shift is proven exactly as in Lemma 12. By Corollary 10 and Lemma 12 the optimal shift of L_k must begin with an S_{k-1} block. By Lemmas 8 and 12, this S_{k-1} block must be followed by an S_{k-2} block. Since $L_k = S_{k-1}\,S_{k-2}\,(S_{k-1})^{\alpha_k - 1}$ the optimal shift is just L_k.

Theorem 14: Given $S = (s_1, s_2, \ldots, s_N)$ (or, equivalently, $V = (v_1, v_2, \ldots, v_N)$) and $L = (s_k, \ldots, s_{k-1})$ the L-sequence of S. There exists no $\hat{k}$ such that $s_{\hat{k}} = 1$ and

$$\sum_{i=\hat{k}}^{N+\hat{k}-1} p^{N+\hat{k}-1-i} v_i > \sum_{i=k}^{N+k-1} p^{N+k-1-i} v_i$$

for any $p, 0 < p < 1$.

Proof: Case 1: $s_1 = 1$, so $L = S$.
Let S^* be a shifted version of S with $s_1^* = 1$. Let μ_1 be the first element where S and S^* differ, i.e.

$$s_i = s_i^* \quad \text{if } i < \mu_1$$

$$s_i \neq s_i^* \quad \text{if } i = \mu_1$$

By Lemma 12 $s_{\mu_1} = 0$ and $s^*_{\mu_1} = 1$. Let μ_2 be the next element where S and S^* differ. By Lemma 11 $s_{\mu_2} = 1$ and $s^*_{\mu_2} = 0$.

$$\sum_{i=1}^{\mu_2} p^{N-i} v_i - \sum_{i=1}^{\mu_2} p^{N-i} v_i^* = 2\, p^{N-\mu_2} \left(1 - p^{\mu_2 - \mu_1}\right) > 0 \quad \text{for } 0 < p < 1$$

Continue in this manner to enumerate all elements where S and S^* differ. By Lemma 12

$$s_{\mu_i} = 0 \qquad s^*_{\mu_i} = 1 \qquad \text{for } i \text{ odd}$$

By Lemma 11

$$s_{\mu_i} = 1 \qquad s^*_{\mu_i} = 0 \qquad \text{for } i \text{ even}$$

Since S^* is a shifted version of S, the total number of such μ_i is even, so we can treat them in pairs. Considering the μ_i, μ_{i+1} pair, for i odd, we find

$$\sum_{i=\mu_{i-1}+1}^{\mu_{i+1}} p^{N-i} v_i - \sum_{i=\mu_{i-1}+1}^{\mu_{i+1}} p^{N-i} v_i^* = 2\, p^{N-\mu_{i+1}} \left(1 - p^{\mu_{i+1} - \mu_i}\right) > 0 \quad \text{for } 0 < p < 1$$

Summing over all such pairs yields a positive quantity, so the theorem holds in this case.

Case 2: $s_1 = 0$, so $L = S$ with the first two bits reversed. By Corollary 13, L is the shift of S which maximizes the cumulative zero count. The proof is from then on identical to that of Case 1.

Theorem 15: The complement C of the projection onto the x-axis of the devil's staircase described in Section VI has measure zero.

Proof: The proof will consist of three stages. First we will derive an upper bound on the length of the complementary intervals at each stage of the formation of the devil's staircase. Then we will derive an upper bound on the number of such intervals. Finally we will show that the sum of the lengths

of these intervals tends to zero as the staircase is filled in. One minor point to note is that we will consider the y-axis to represent not average output but the fraction of ones in the limit-cycle. This is permissible as the two quantities are related by an affine linear transformation. When we refer to the step "corresponding to rational number q" we mean the step representing the limit cycle where the fraction of ones present is q.

Step 1: Define B_N to be the projection onto the x–axis of the steps of the staircase corresponding to rational numbers with denominator less than or equal to N. Let C_N be the complement of this set in $(-1,1)$ — clearly C_N consists of a number of intervals and $C_N \subset C_{N-1}$. We will prove that the length of each of these intervals is bounded above by the quantity $2p^{N-1}/(p^{N-1}+\cdots+p+1)$.

Using the theory of Farey fractions [4, Ch 4], we know that if a and b are rational numbers corresponding to neighboring steps of B_N then their continued fraction expansions are of one of the following two forms:

$$a = \frac{n_a}{d_a} = [\alpha_1, \ldots, \alpha_n] \quad \text{and} \quad b = \frac{n_b}{d_b} = [\alpha_1, \ldots, \alpha_n + 1] \quad \text{or}$$

$$a = \frac{n_a}{d_a} = [\alpha_1, \ldots, \alpha_n] \quad \text{and} \quad b = \frac{n_b}{d_b} = [\alpha_1, \ldots, \alpha_n - 1, 2]$$

with $d_a + d_b > N$.

The length of the interval between steps a and b (with $b > a$) is

$$\frac{l_b(p)}{1_{d_b}(p)} - \frac{r_a(p)}{1_{d_a}(p)} = \frac{l_b(p).1_{d_a}(p) - r_a(p).1_{d_b}(p)}{1_{d_a}(p).1_{d_b}(p)}$$

$l_b(p).1_{d_a}(p)$ is a polynomial in p of order $d_a + d_b - 2$; the coefficient of p^k being

for $d_a \geq d_b$:

the sum of the coefficients of p^0 through p^k in $l_b(p)$ for $0 \leq k \leq d_b - 1$

the sum of the coefficients of p^{k-d_a+1} through p^{d_b-1} in $l_b(p)$

for $d_a \le k \le d_a + d_b - 2$

the sum of all coefficients in $l_b(p)$ for $d_b \le k \le d_a - 1$

for $d_a < d_b$:
the sum of the coefficients of p^0 through p^k in $l_b(p)$ for $0 \le k \le d_a - 1$

the sum of the coefficients of p^{k-d_a+1} through p^{d_b-1} in $l_b(p)$
for $d_b - 1 \le k \le d_a + d_b - 2$

the sum of the coefficients of p^{k-d_a+1} through p^k in $l_b(p)$ for $d_a \le k \le d_b - 2$

There are five cases to be considered:

(i) $a = [\alpha_1, \ldots, \alpha_n] < b = [\alpha_1, \ldots, \alpha_n + 1]$ (n even; S_{n-2} begins with 01)

$r_a(p)$ has coefficients $S_{n-2}(S_{n-1})^{\alpha_n}$
$l_b(p)$ has coefficients $S_{n-2}(S_{n-1})^{\alpha_n}S_{n-1}$ with the first two bits reversed

Since $S_{n-1}S_{n-2}$ is obtained from $S_{n-2}S_{n-1}$ by reversing the first two bits, the coefficients of $l_b(p).1_{d_a}(p)$ and $r_a(p).1_{d_b}(p)$ are identical in all positions except the first. It follows that

$$\frac{l_b(p)}{1_{d_b}(p)} - \frac{r_a(p)}{1_{d_a}(p)} = \frac{2p^{d_a+d_b-2}}{1_{d_b}(p).1_{d_a}(p)} \le \frac{2p^{d_a+d_b-2}}{1_{d_a+d_b-1}(p)} \le \frac{2p^{N-1}}{1_N(p)}$$

(ii) $a = [\alpha_1, \ldots, \alpha_n + 1] < b = [\alpha_1, \ldots, \alpha_n]$ (n odd; S_{n-2} begins with 10)

$r_a(p)$ has coefficients $S_{n-2}(S_{n-1})^{\alpha_n}S_{n-1}$ with the first two bits reversed
$l_b(p)$ has coefficients $S_{n-2}(S_{n-1})^{\alpha_n}$

By the argument of case (i), we find

$$\frac{l_b(p)}{1_{d_b}(p)} - \frac{r_a(p)}{1_{d_a}(p)} = \cdots \leq \frac{2p^{N-1}}{1_N(p)}$$

(iii) $a = [\alpha_1, \ldots, \alpha_n] < b = [\alpha_1, \ldots, \alpha_n - 1, 2]$ (n odd; S_{n-2} begins with 10)

$r_a(p)$ has coefficients $S_{n-2}(S_{n-1})^{\alpha_n}$ with the first two bits reversed
$l_b(p)$ has coefficients $S_{n-1}S_{n-2}(S_{n-1})^{\alpha_n-1}S_{n-2}(S_{n-1})^{\alpha_n-1}$ with the first two bits reversed

By the argument of case (i), we find once again

$$\frac{l_b(p)}{1_{d_b}(p)} - \frac{r_a(p)}{1_{d_a}(p)} = \cdots \leq \frac{2p^{N-1}}{1_N(p)}$$

(iv) $a = [\alpha_1, \ldots, \alpha_n - 1, 2] < b = [\alpha_1, \ldots, \alpha_n]$ (n even; S_{n-2} begins with 01)

$r_a(p)$ has coefficients $S_{n-1}S_{n-2}(S_{n-1})^{\alpha_n-1}S_{n-2}(S_{n-1})^{\alpha_n-1}$ with the first two bits reversed
$l_b(p)$ has coefficients $S_{n-2}(S_{n-1})^{\alpha_n}$ with the first two bits reversed

By the argument of case (i), we find once again

$$\frac{l_b(p)}{1_{d_b}(p)} - \frac{r_a(p)}{1_{d_a}(p)} = \cdots \leq \frac{2p^{N-1}}{1_N(p)}$$

(v) Finally, it is easily seen that the outermost intervals of C_N have length $(2p^{N-1})/(p^{N-1} + \cdots + 1)$.

Step 2: At stage m of the formation of the staircase the number of steps added is at most $m-1$ (i.e. those steps corresponding to the rational numbers with denominators $\leq m$). The total number of steps at stage N is, therefore, less

than or equal to $N(N-1)/2$. If follows that the total number of intervals of C_N is at most $N(N-1)/2+1$.

Step 3: The total length of C_N is less than or equal to

$$\left(\frac{N(N-1)}{2}+1\right)\left(\frac{2p^{N-1}}{p^{N-1}+\cdots+1}\right)$$

which tends to zero as N tends to infinity, since $p<1$. For any $\epsilon>0$, in other words, there exists a covering of C by intervals with total length less than ϵ. C, therefore, has measure zero.

B References

[1] J. C. Candy, "A use of Limit Cycle Oscillations to Obtain Robust Analog-to-Digital Converters". *IEEE Trans. Comm*, vol. COM-22, pp. 298–305, Mar 1974.

[2] Fathy F. Yassa and Steven L. Garverick, "A Multichannel Digital Demodulator for LVDT/RVDT Position Sensors". *IEEE J. Solid-State Circuits*, vol. SC-25, pp. 441–450, Apr 1990.

[3] Bosco Hok-Chung Leung, "Multichannel PCM A/D interfaces using oversampling techniques". PhD thesis, University of California, Berkeley, Dec 1987.

[4] Hao Bai-Lin, *Elementary Symbolic Dynamics.* Singapore: World Scientific, 1989.

[5] Vladimir Friedman, "The Structure of the Limit Cycles in Sigma Delta Modulation". *IEEE Trans Comm.*, vol. COM-36, pp. 972-979, August 1988

[6] R. M. Gray, "Oversampled Sigma-Delta Modulation". *IEEE Trans. Comm.*, vol. COM-35, pp. 481–489, May 1987.

[7] R. M. Gray, "Spectral Analysis of Quantization Noise in a Single Loop Sigma Delta Modulator with dc Input". *IEEE Trans. Comm*, vol. COM-37, pp. 588–599, June 1989.

[8] Roubik Gregorian and Gabor C. Temes, *Analog MOS Integrated Circuits for Signal Processing.* New York: Wiley 1986, p. 485.

[9] E. I. Jury, *Theory and Application of the z-transform Method.* New York: Wiley, 1964

[10] Soren Hein and Avideh Zakhor, "Lower Bounds on the MSE of the Single and Double Loop Sigma Delta Modulators". *Proc. Int. Conf. Circuits and Systems*, pp. 1751–1755, May 1990.

[11] M.P.Kennedy, K.R.Krieg and L.O.Chua, "The Devil's Staircase: The Electrical Engineer's Fractal". *IEEE Trans. CAS*, vol. CAS-36, pp.1133–1139, August 1989.

List of Footnotes

. Manuscript received: _____

* This work is supported in part by the Office of Naval Research under Grant N00014-89-J-1402

† The authors are with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720, USA

1 $f(x^-) = \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} f(x - \epsilon)$

List of Figures

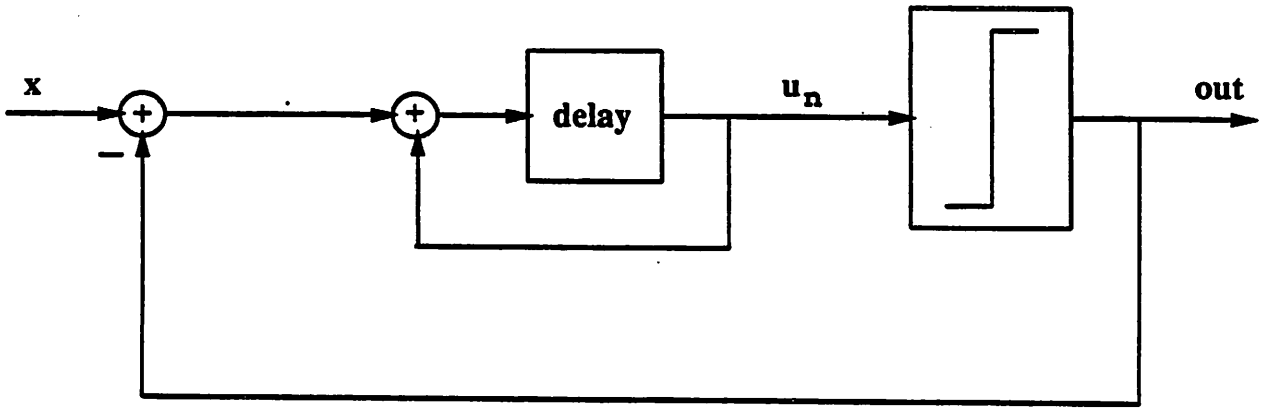

Figure 1

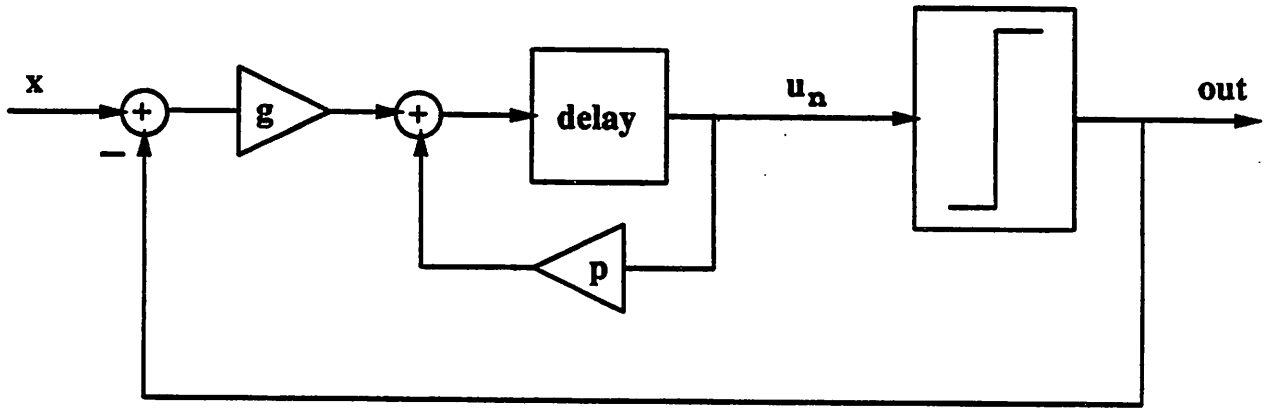

Figure 2

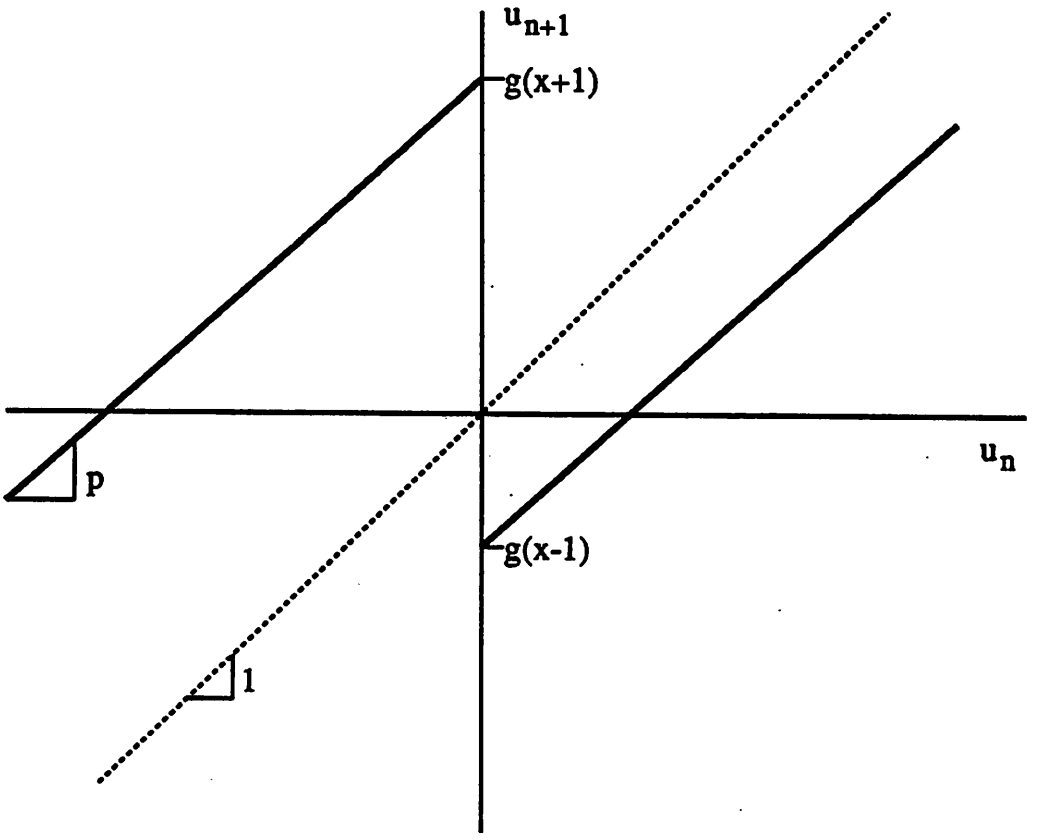

Figure 3

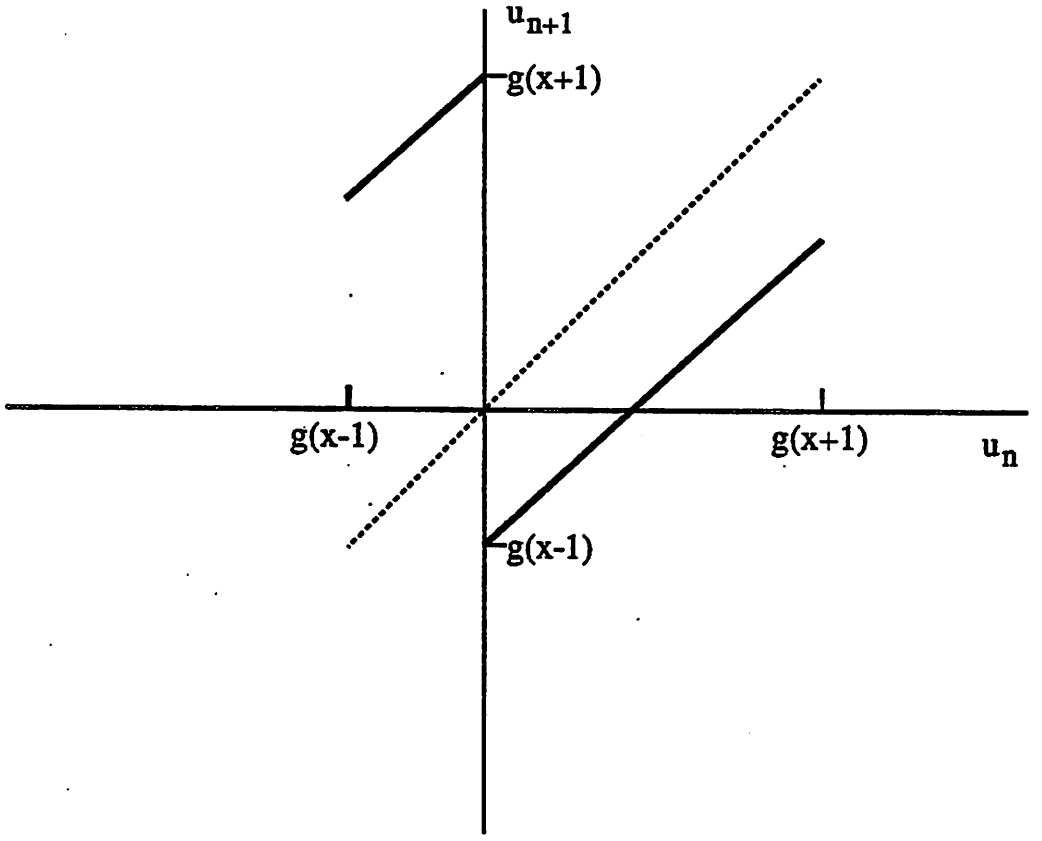

Figure 4

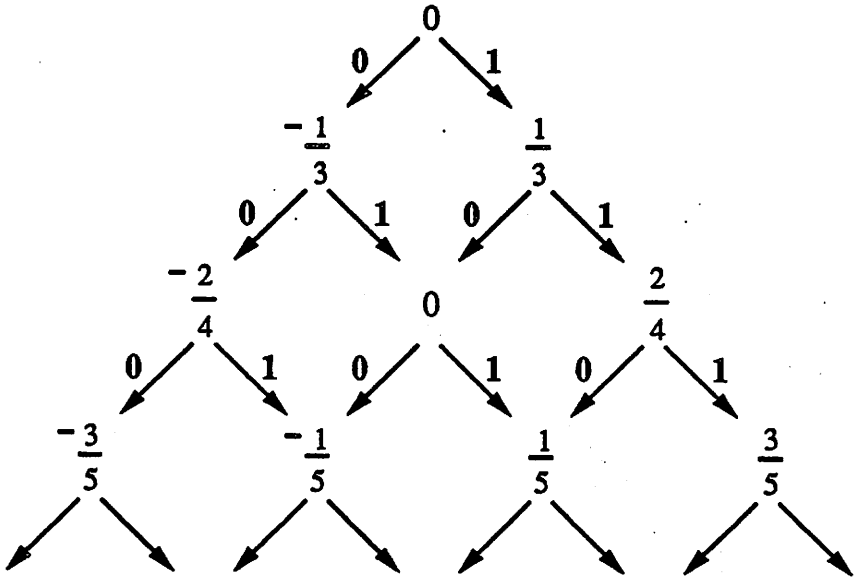
0
0
1
$-\frac{1}{3}$
$\frac{1}{3}$
0
1
0
1
$-\frac{2}{4}$
0
$\frac{2}{4}$
0
1
0
1
0
1
$-\frac{3}{5}$
$-\frac{1}{5}$
$\frac{1}{5}$
$\frac{3}{5}$

Figure 5

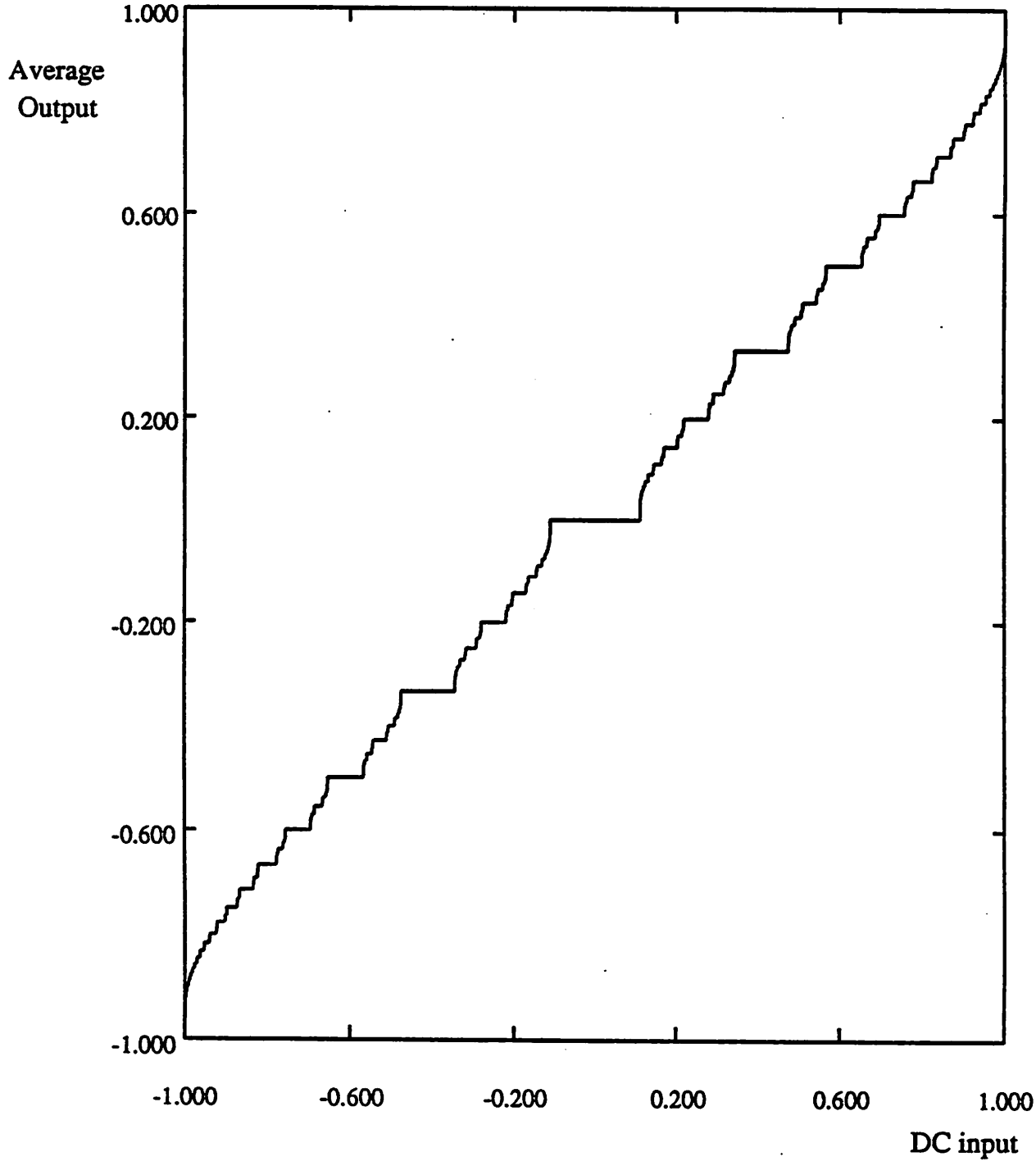
1.000
Average
Output
0.600
0.200
-0.200
-0.600
-1.000
-1.000
-0.600
-0.200
0.200
0.600
1.000
DC input

Figure 6

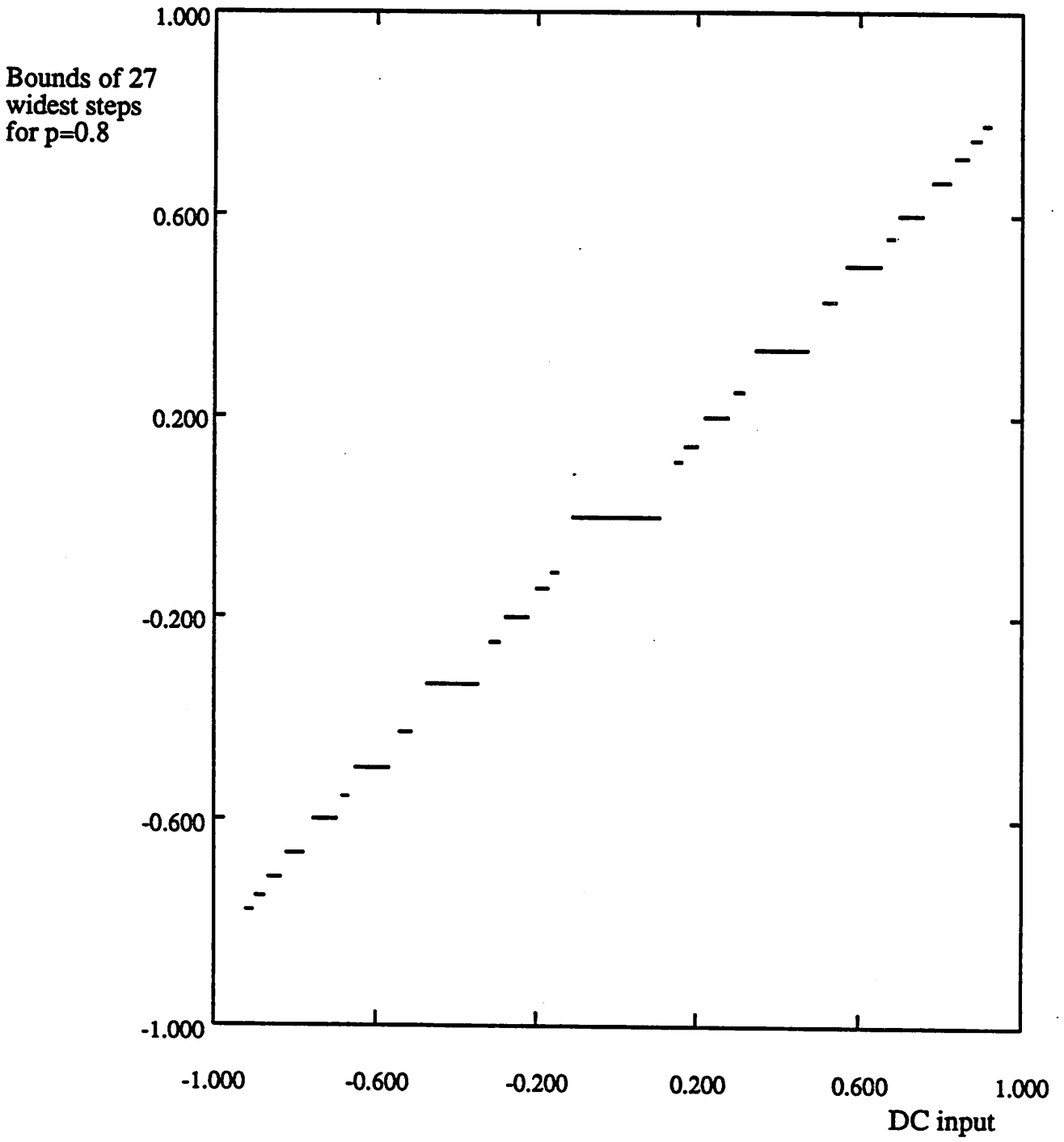
Bounds of 27
widest steps
for p=0.8
1.000
0.600
0.200
-0.200
-0.600
-1.000
-1.000
-0.600
-0.200
0.200
0.600
1.000
DC input

Figure 7

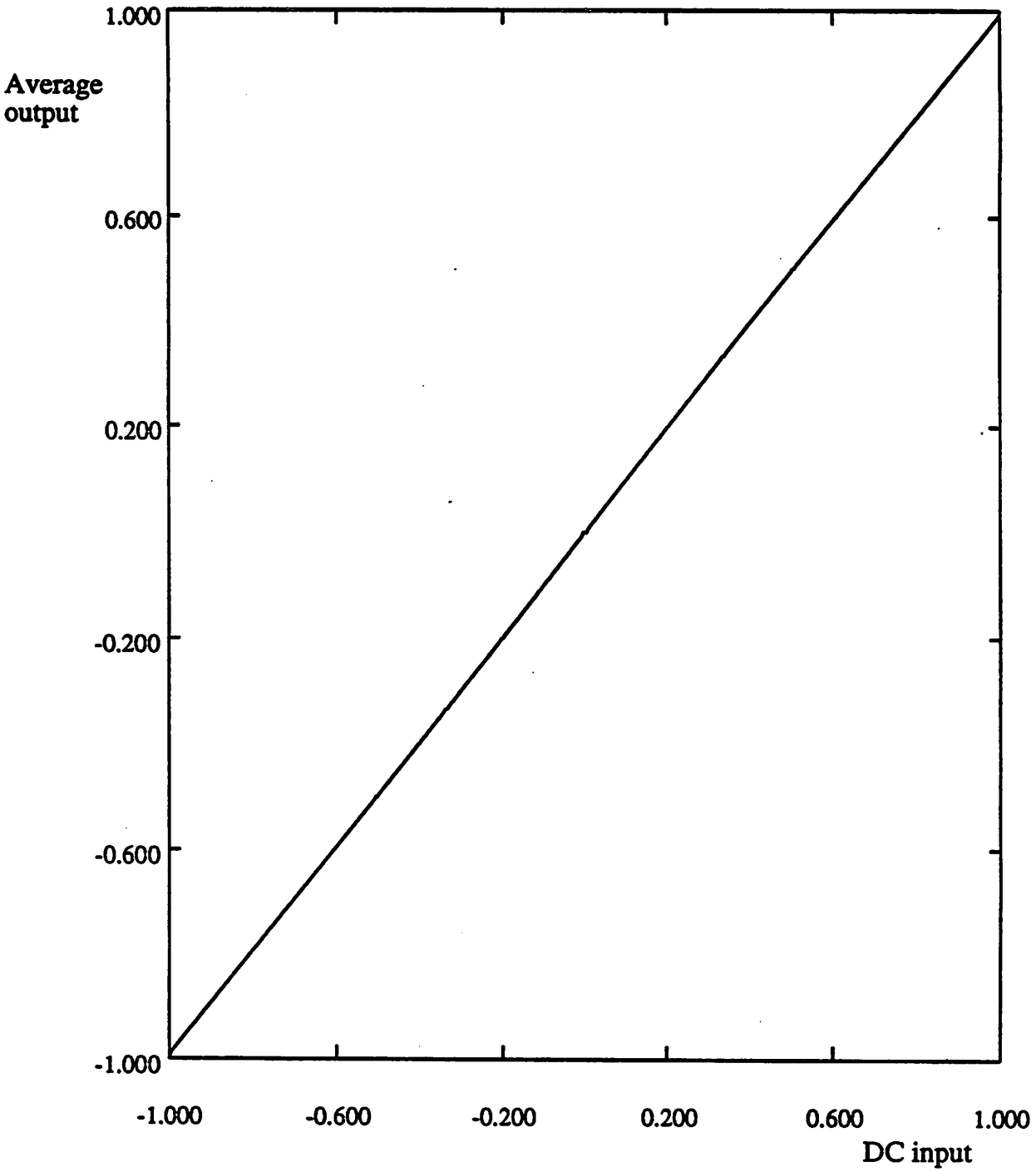
1.000
Average output
0.600
0.200
-0.200
-0.600
-1.000
-1.000
-0.600
-0.200
0.200
0.600
1.000
DC input

Figure 8

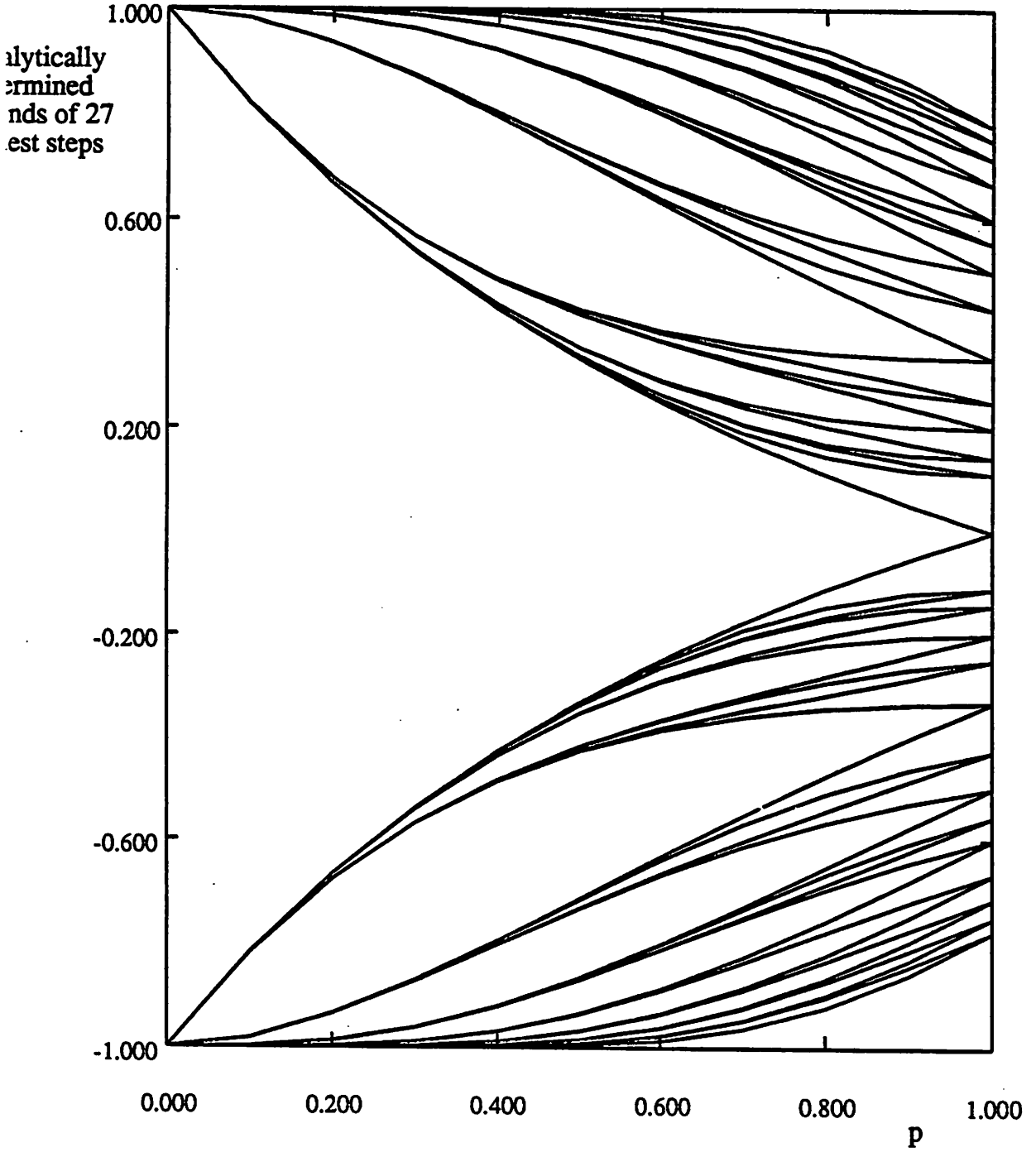
ılytically
ɛrmined
nds of 27
est steps
1.000
0.600
0.200
-0.200
-0.600
-1.000
0.000
0.200
0.400
0.600
0.800
1.000
p

Figure 9

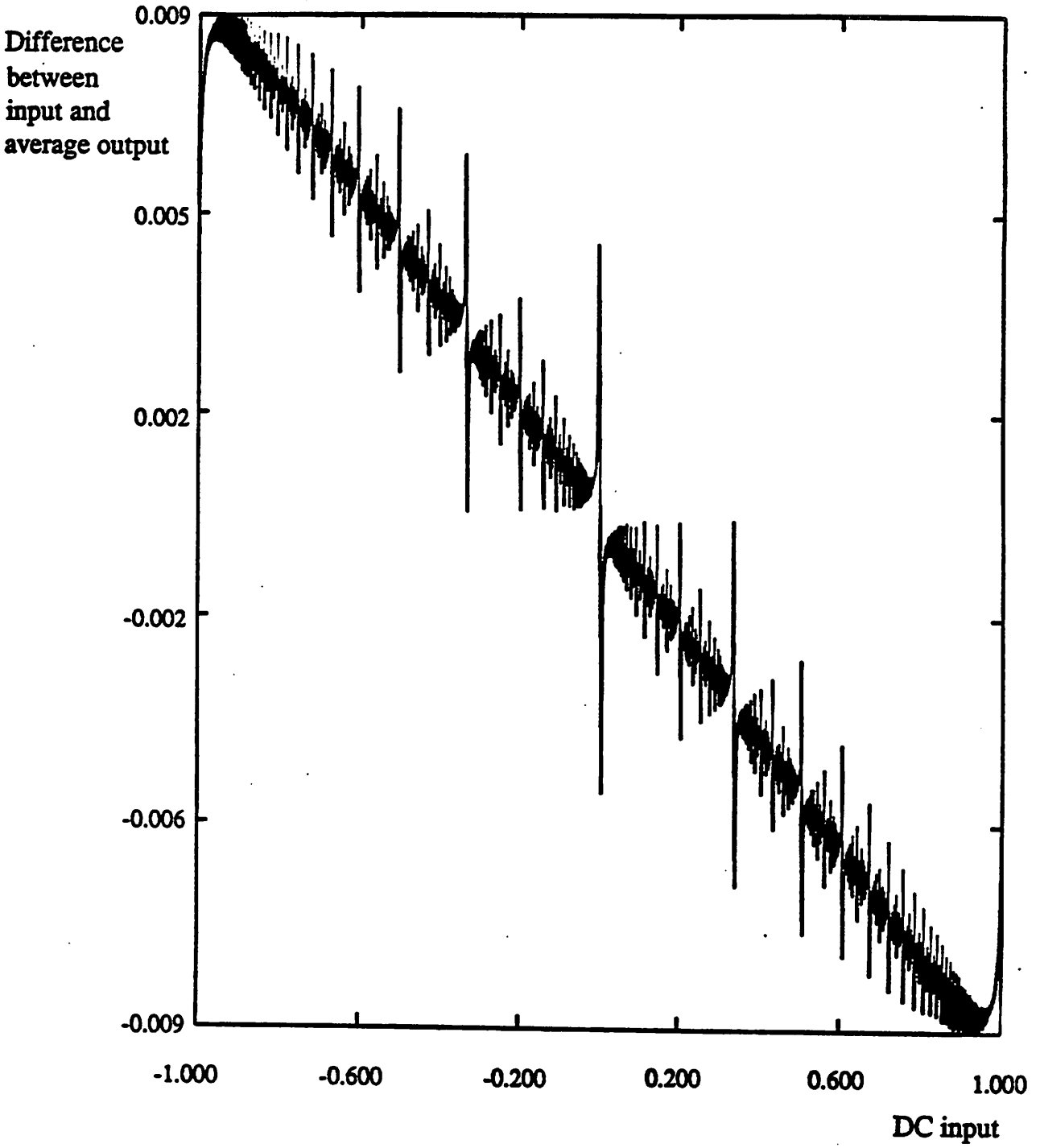
Difference between input and average output
0.009
0.005
0.002
-0.002
-0.006
-0.009
-1.000
-0.600
-0.200
0.200
0.600
1.000
DC input

Figure 10

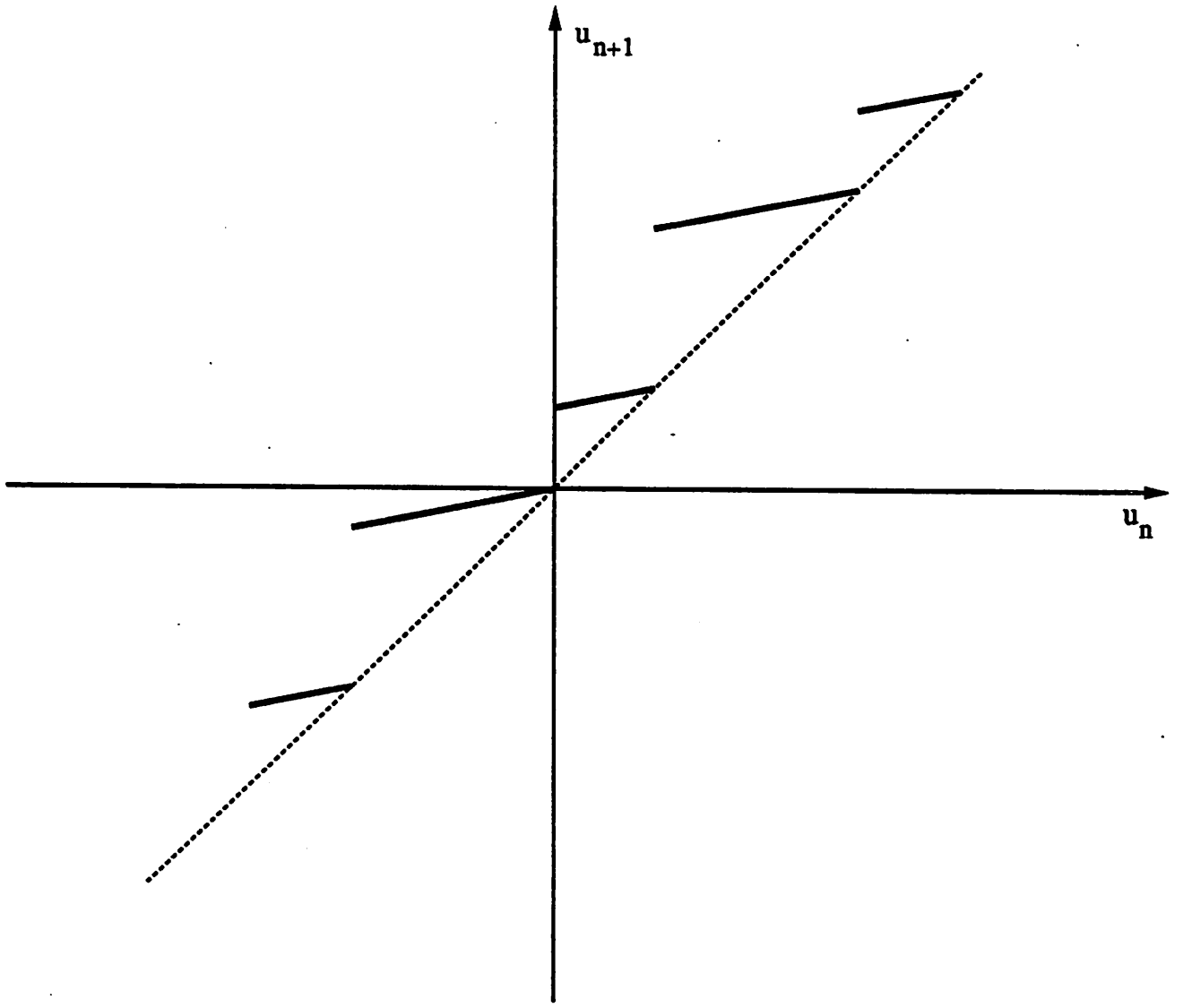

Figure 11

List of Footnotes

Manuscript received: _____

* This work is supported in part by the Office of Naval Research under Grant N00014-89-J-1402

† The authors are with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720, USA

1 $f(x^-) = \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} f(x - \epsilon)$

List of Figures