New Properties of Sigma–Delta Modulators with dc Inputs

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

Abstract—**We derive new properties of the single- and double-loop sigma–delta modulators with constant inputs, by exploiting the inherent structure of the output sequences or codewords that the modulators are capable of producing. Specifically, we first derive upper bounds of $O(N^2)$ and $O(N^3)$ on the number of N-bit codewords for the single and double-loop modulators, respectively. We then derive analytical lower bounds on the mean squared error (MSE) obtainable by any decoder, linear or nonlinear, in approximating the constant input; based on N-bit codewords, the bounds are $O(N^{-3})$ and $O(N^{-6})$ for the single and double-loop modulators, respectively. Optimal nonlinear decoders for constant inputs can be based on a table look-up approach which operates directly on the nonuniform quantization intervals. Numerical results show that if the constant input is uniformly distributed, the MSE of such nonlinear decoders are $O(N^{-3})$ and $O(N^{-5})$ for the single- and double-loop modulators, respectively. Using simulations we find that the optimal nonlinear decoders perform better than linear decoders, by about 3 and 20 dB for the single and double-loop modulators, respectively. We also introduce a cascade structure specifically for constant inputs, and derive its corresponding decoding algorithm. The idea behind the cascade structure is to requantize the residue from each stage in order to fully utilize the dynamic range of the next stage. We show that for a fixed latency, the MSE performance of our cascade structure is 12 dB superior, and its throughput is twice the conventional two-stage MASH modulator.**

I. Introduction

SIGMA–DELTA ($\Sigma\Delta$) modulators are becoming increasingly popular for analog-to-digital (A/D) conversion applications due to their insensitivity to circuit imperfections and ease of implementation [1]–[3]. They are based on the principle of using a one-bit quantizer at the expense of operating at sampling rates much higher than the Nyquist rate. The inherent trade-off between sampling rate and resolution in amplitude quantization is well known, and the two-dimensional equivalent has been the subject of a recent investigation [4]: It was shown that reconstruction from multiple level crossings results in sampling schemes whose requirements on position and amplitude quantization can be anywhere between the extremes of Nyquist sampling and a zero-crossing representation.

Paper approved by the Editor for Speech Processing of the IEEE Communications Society. Manuscript received December 7, 1989; revised November 6, 1990 and June 29, 1991. This work was supported in part by the NSF under Grant MIP-9057466 and in part by Analog Devices. This paper was presented in part at the 23rd Annual Asilomar Conference on Signals, Systems, and Computers, October 1989, and at the International Symposium on Circuits and Systems, May 1990.

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

IEEE Log Number 9201162.

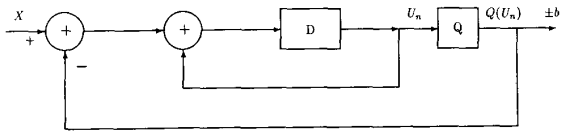


Fig. 1. Discrete-time model of the single-loop $\Sigma\Delta$ encoder.

In this paper we consider the application of the single and double-loop $\Sigma\Delta$ modulators to data acquisition applications, in which the input is approximately constant. The results also provide insight into the more general case of time-varying inputs, since such inputs are typically heavily oversampled. To introduce the problems that we address, let us consider the discrete-time model of the single-loop $\Sigma\Delta$ encoder shown in Fig. 1.[1] The encoder consists of a discrete-time integrator and a one-bit quantizer Q embedded in a negative feedback loop. The quantizer is specified by

$$Q(U_n) = \begin{cases} -b & U_n \le 0 \\ +b & U_n > 0 \end{cases}$$

where b is the quantizer step size. The constant input X is assumed to be in the dynamic range $B = (-b, +b)$, and the encoder output is a binary sequence of $\pm b$'s. The integrator accumulates the error between the input and the quantizer output, and the negative feedback serves to make this error small.

The task of a decoder is to produce an estimate $\hat{X}$ of the constant input X, given a number N of output bits; to be consistent with the terminology in [1] we refer to N as the oversampling ratio (OSR).[2] The collection of estimates $\hat{X}$ that a given decoder is capable of producing for a given OSR and all constant inputs $X \in B$ is referred to as the decoder's reproduction alphabet for that OSR [1]. The mean squared error (MSE) of a decoder is defined as

$$\text{MSE} \triangleq E\left[\left(X - \hat{X}\right)^2\right].$$

A simple decoder is the averaging decoder given by

$$\hat{X} = \frac{1}{N}\sum_{n=0}^{N-1} Q(U_n). \tag{1}$$

The size of its reproduction alphabet is $N + 1$ [1], and if the constant input has a smooth probability distribution, the MSE

[1] We refer to Fig. 1 as an encoder, any linear or nonlinear filter producing estimates of the input as a decoder, and their combination as a modulator.

[2] A different definition is used for dynamic inputs.

of the averaging decoder is upper bounded by [1]

$$\text{MSE} \leq \frac{4b^2}{3N^2}.$$

For the averaging decoder, Gray [5] has also shown that if the input is uniformly distributed on B and N is large, the MSE is given by

$$\text{MSE} \approx \frac{2b^2}{3N^2}.$$

The accuracy of the single-loop modulator as an A/D converter is intuitively linked to the number of output values, that is, the size of the reproduction alphabet at the disposal of its decoder. For instance, the size of the reproduction alphabet is $O(N^2)$ for a decoder consisting of an N-tap FIR filter with triangular impulse response, commonly referred to as a sinc^2 filter. This decoder achieves an asymptotic MSE of $O(N^{-3})$. It is conceivable that there exist a variety of other decoders with larger reproduction alphabets and smaller MSE. The question arises as to the fundamental limits on reproduction alphabet size and MSE performance of the single-loop modulator, as well as more general modulators, with constant inputs.

In this paper we show that the answer to this question can be found by decoupling the encoder and decoder, and studying separately the structure of the output sequences that the encoder is capable of producing. The N-bit output sequences that an encoder can generate for fixed initial integrator states as the constant input ranges over $(-b, +b)$ will be referred to as the codewords for the given initial states, or codewords for short. The size of a modulator's reproduction alphabet is upper-bounded by the number of different N-bit output sequences or codewords that the encoder can produce for constant inputs $X \in B$. The number of codewords is clearly bounded, since there are at most 2^N of them, but we find that for given initial integrator states, many of the 2^N binary N-bit sequences are not codewords. The resolution and hence the MSE performance of a modulator is bounded by the fact that each codeword can be generated by a range of constant inputs, and no decoder can distinguish between constant inputs in such ranges. In principle, an optimal nonlinear decoder could be based on a table look-up approach in which each codeword is mapped to the mean value of the input over the range corresponding to the codeword. These comments are applicable to general $\Sigma\Delta$ modulators, including interpolative ones [6], but we restrict our analysis to the single and double-loop modulators.

Our focus on N-bit codewords and the case of zero or at least known initial integrator states bears some similarity to the approach in [7]. In that paper, N-tap linear decoders are derived which are in some sense optimal. However, the work in [7] is based on an assumption of white, uncorrelated quantization noise—an assumption which is now known not to be valid for low-order encoders with one-bit quantizers [1].

This paper is organized as follows. In Section II, we derive an upper bound of $O(N^2)$ on the number of codewords for the single-loop encoder, and an $O(N^{-3})$ lower bound on the MSE performance of any single-loop decoder with constant inputs. We present simulation results to show that an MSE performance of $O(N^{-3})$ is achievable with the optimal nonlinear decoder, as well as with a number of linear filters. Similar analysis and simulation results for the double-loop encoder are presented in Section III. In Section IV, we first derive the MSE performance of the averaging decoder for the single-loop encoder. Even though the expression for the MSE is known [5], our particular derivation enables us to develop a decoding scheme for a modified cascade structure with dc inputs; the structure may be viewed as a modification of the multistage noise-shaping (MASH) modulator [8]. In Section IV, we discuss this modified cascade structure and its corresponding decoding algorithm. We also compare its performance to that of the MASH modulator with linear decoding. Finally, Section V contains conclusions and directions for future research.

II. The Single- Loop Encoder

In section II-A we derive an upper bound on the number of codewords of the single-loop $\Sigma\Delta$ encoder with constant inputs and a fixed initial integrator state, and compare the bound to simulation results. In Section II-B we use the results of Section II-A to derive a lower bound on the MSE performance of any decoder, and compare the bound to the simulated performance of an optimal nonlinear decoder and two linear decoders.

A. Codewords

In Section II-A1) we derive an upper bound on the number of codewords, and in Section II-A2) we show simulation results on the actual number of codwords.

1) Upper Bound on Number of Codewords: Our approach is to examine the behavior of the encoder state variable U_n in Fig. 1. From the figure, the nonlinear difference equation relating U_n to U_{n-1} is

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

where X is the constant input. It has been shown [1] that if the initial integrator state U_0 is in the range $(X - b, X + b)$, then the integrator state remains in the same range at all future times, that is

$$U_0 \in (X - b, X + b) \Rightarrow U_n \in (X - b, X + b) \quad \text{for all } n > 0. \tag{3}$$

To gain some intuition about the behavior of the state variable, Fig. 2 shows a typical time sequence for U_n. Equation (2) states that if U_{n-1} is positive, the state variable is decremented by $b - X$ in the next time step, whereas if U_{n-1} is negative, the state variable is incremented by $b + X = 2b - (b - X)$. Thus the state variable is always decremented by $b - X$, but if $U_{n-1} < 0$, it is additionally incremented by $2b$, and a negative output bit is generated. To satisfy (3), the number of negative bits j among the first n output bits must therefore satisfy

$$X - b \leq U_0 - n(b - X) + 2jb \leq X + b. \tag{4}$$

In Appendix A.1 we show that equation (4) leads to the following conclusions for a given value of N. The dynamic

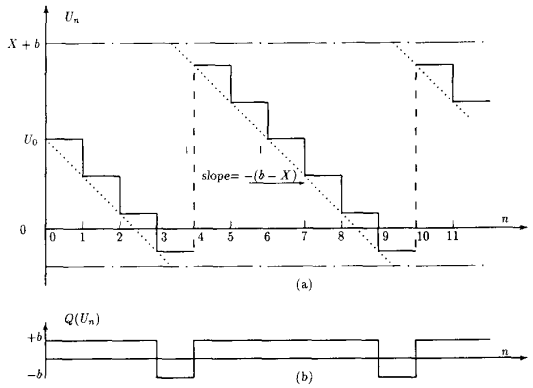


Fig. 2. Typical waveforms for: (a) the state variable U_n, (b) the quantizer output $Q(U_n)$.

TABLE I
TRANSITION POINTS FOR $N = 12$ AND $U_0 = 0$. THE POINTS ARE NORMALIZED BY b

pj	1	2	3	4	5	6	7	8	9	10
2	0									
3	1/3	−1/3								
4	1/2	0	−1/2							
5	3/5	1/5	−1/5	−3/5						
6	2/3	1/3	0	−1/3	−2/3					
7	5/7	3/7	1/7	−1/7	−3/7	−5/7				
8	3/4	1/2	1/4	0	−1/4	−1/2	−3/4			
9	7/9	5/9	1/3	1/9	−1/9	−1/3	−5/9	−7/9		
10	4/5	3/5	2/5	1/5	0	−1/5	−2/5	−3/5	−4/5	
11	9/11	7/11	5/11	3/11	1/11	−1/11	−3/11	−5/11	−7/11	−9/11

range $(-b, +b)$ is divided into quantization intervals whose width and position depend on U_0. Within each interval, all constant inputs generate the same N-bit codeword, but distinct intervals correspond to distinct codewords. The edges of the quantization intervals, referred to as transitions points, are given by

$$X = \frac{pb - (2jb + U_0)}{p} \tag{5}$$

where $1 \leq p \leq N-1$, and j is a positive integer in the appropriate range such that $X \in (-b, +b)$. As an example, the transition points for $N = 12$ and an initial state of $U_0 = 0$ are shown in Table I. By counting the number of parameter combinations (p, j) in (5), we find that the number of quantization intervals, hence the number of codewords, is upper bounded by

$$C_{\max} = \begin{cases} \frac{1}{2}N(N-1)+1 & U_0 \neq 0 \\ \frac{1}{2}N(N-1)(N-2)+1 & U_0 = 0 \end{cases}. \tag{6}$$

This is an $O(N^2)$ upper bound. Clearly, the number of codewords is a diminishingly small fraction of 2^N for large oversampling ratios.

The actual number of codewords may be less than indicated by (6), since some parameter combinations may correspond to the same transition point. Such "degeneration" occurs only if U_0/b is rational, which happens with probability zero if U_0 has a smooth probability distribution on $(X-b, X+b)$. If U_0/b is irrational, the exact number of codewords is $\frac{1}{2}N(N-1)+1$. In the special case $U_0 = 0$, the transition points of equation (5) are known as the Farey series of number theory [9], and an asymptotic expansion of their number is available [9],

$$C \to \frac{3}{\pi^2}N^2 + O(N \log N) \approx 0.304N^2 \quad \text{as } N \to \infty. \tag{7}$$

The upper bound (6) is thus asymptotically 65% too high for the case $U_0 = 0$, but it has the correct dependence on N.

2) Numerical Results: For the special case of zero initial integrator state, $U_0 = 0$, we have used computer simulations to find the actual number of codewords for the single-loop encoder as a function of oversampling ratio. Fig. 3 shows the results, as well as the upper bound derived in Section II-A1). The upper bound is seen to be rather tight. The analytical approximation (7) is not plotted, since it is indistinguishable from the simulated results.

B. Lower Bounds on MSE

In Section II-B1) we use the results of Section II-A to derive a number of lower bounds on the MSE obtainable with any decoder operating on a single-loop encoder. In Section II-B2) we show numerical results on the actual MSE of a number of decoders.

1) Lower Bounds: Our first lower bound on the MSE follows directly from the upper bound on the number of codewords

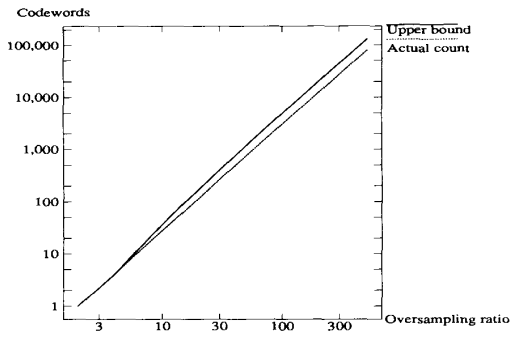


Fig. 3. Single-loop encoder: Actual number of codewords and the upper bound (6) as functions of the oversampling ratio.

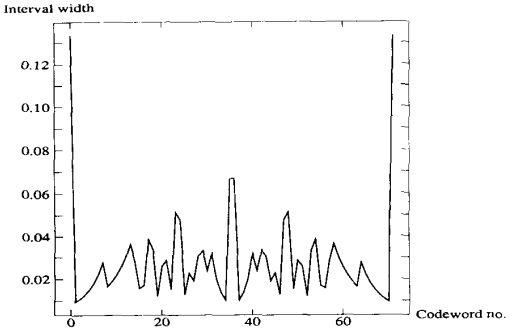


Fig. 4. Single-loop encoder: Width of the quantization interval associated with the ith codeword versus i for an oversampling ratio of $N = 16$. Widths are normalized by b.

derived in Section II-A1). To state the bound, we quote a well-known result from quantization theory [10]: If the constant input X is restricted to $(-b, +b)$ and has probability density function $p(x)$, then the minimum MSE obtainable with a large number C of quantization levels is

$$\mathrm{MSE}_{\min} = \frac{1}{12C^2}\left[\int_{-b}^{+b} p(x)^{1/3}\,dx\right]^3. \tag{8}$$

If in particular X is uniformly distributed on $(-b, +b)$, the minimum MSE is $b^2/(3C^2)$. Since the single-loop $\Sigma\Delta$ encoder has a maximum of $O(N^2)$ codewords, (8) implies that a lower bound on its MSE for any given $p(x)$ is $O(N^{-4})$. If the input is uniformly distributed on $(-b, +b)$, the following $O(N^{-4})$ lower bound holds on the MSE:

$$\mathrm{MSE}_{\min} = \frac{4b^2}{3N^4}.$$

We now derive an $O(N^{-3})$ lower bound on the MSE by considering in more detail the quantization intervals of the single-loop encoder. If the number of codewords C is large, and the density function $p(x)$ is smooth, then the constant input X is approximately uniformly distributed on each quantization interval I_i, $1 \le i \le C$. An optimal decoder will therefore decode any codeword into the midpoint of the corresponding quantization interval. Denoting the width of the ith interval by d_i, a lower bound on the MSE is thus

$$\mathrm{MSE}_{\mathrm{optimal}} = \sum_{i=1}^{C} P[X \in I_i] \cdot \frac{d_i^2}{12} = \sum_{i=1}^{C} \frac{d_i^3}{24b} \tag{9}$$

where the last equality holds if X is uniformly distributed on $(-b, +b)$. Clearly, the optimal MSE is lower bounded by an expression similar to (9) in which only some of the C terms are included. Our strategy for lower-bounding the MSE is therefore to find a few large interval widths d_i of order $O(N^{-1})$. As long as there is nonzero probability of inputs in these intervals, we arrive at an $O(N^{-3})$ lower bound on the MSE, but for simplicity we state the MSE bounds for the case of uniformly distributed input. Appendix A.2 shows that if the initial state is zero, the two intervals closest to the extreme inputs $-b$ and $+b$ are the largest intervals, and they have widths

$$d_1 = d_C = \frac{2b}{N-1} \tag{10}$$

while the two intervals immediately above and below $X = 0$ have widths

$$d_{\frac{C}{2}-1} = d_{\frac{C}{2}} \ge \frac{b}{N-1}. \tag{11}$$

Including only the four terms corresponding to (10) and (11) in (9), we thus find the $O(N^{-3})$ lower bound

$$\mathrm{MSE}_{\mathrm{optimal}} \ge \frac{d_1^3 + d_C^3 + d_{\frac{C}{2}}^3 + d_{\frac{C}{2}+1}^3}{24b} \ge \frac{3b^2}{4(N-1)^3} \approx \frac{0.75b^2}{N^3} \tag{12}$$

As an aside, it is also shown in Appendix A.2 that the smallest quantization intervals are located next to the two largest ones, near $+b$ and $-b$, and have widths

$$d_2 = d_{C-1} = \frac{2b}{(N-1)(N-2)}. \tag{13}$$

This shows that the ratio between the largest and smallest interval widths is $N-2$, and gives a bound on the nonuniformity of the quantization intervals of the single-loop encoder.

The $O(N^{-3})$ lower bound (12) on the MSE was based on the assumption $U_0 = 0$. For $U_0 \neq 0$ a similar bound can be obtained, although the intervals of width $O(N^{-1})$ are harder to find. To demonstrate their existence, Appendix A.2 shows that for $U_0 > 0$, the interval near $+b$ has width

$$d_1 = \frac{b}{N-1}U_0. \tag{14}$$

A similar result holds for d_C if $U_0 < 0$.

2) Numerical Results: To illustrate the results in (10), (11), and (13), Fig. 4 shows a plot of quantization interval widths for an oversampling ratio of 16 under the assumption of zero initial state. The results seem to be in excellent agreement with the experimental results in [11]. The figure indicates that the single-loop encoder is a highly nonuniform quantizer.

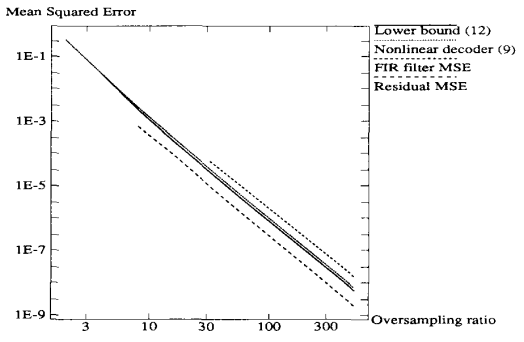


Fig. 5. Single-loop encoder: MSE performance (9) of the optimal nonlinear decoder as a function of oversampling ratio. Also shown is the lower bound (12), and the residual MSE obtained by discarding the two large edge intervals. The last curve shows the performance achievable with the asymptotically optimal linear decoder derived by Gray [5]. The MSE is normalized by b^2.

Fig. 5 shows several curves of MSE performance versus oversampling ratio, assuming the input to be uniformly distributed on $(-b, +b)$. One curve shows the actual performance of the optimal nonlinear decoder whose MSE is given by (9); numerically, we find that

$$\mathrm{MSE}_{\mathrm{optimal}} \approx \frac{0.91b^2}{N^3}. \tag{15}$$

Another curve shows the lower bound (12), which is rather tight. This indicates that the four terms used to derive (12) contribute most of the error to the MSE summation (9). A third curve shows the residual MSE obtained by ignoring the contributions from the edge intervals d_1 and d_C, corresponding to a slight decrease in dynamic range. The reduction in dynamic range improves the MSE by 5.8 dB, but the performance remains at $O(N^{-3})$, in part because of the intervals of width $O(N^{-1})$ in (11).

The last curve in Fig. 5 shows the MSE performance achievable with a specific linear N-tap finite impulse response (FIR) filter as the decoder; the filter was derived by Gray [5] as the asymptotically optimal linear decoder for constant inputs. The MSE of the FIR filter is asymptotically $2b^2/N^3$ for large N. At a fixed oversampling ratio, the optimal nonlinear decoder is 3.4 dB superior to the optimal linear decoder.

The optimal linear and nonlinear decoders are not the only decoders with $O(N^{-3})$ MSE characteristic. Gray showed in [1] that for uniform input distribution and large oversampling ratio, the MSE associated with the ideal low-pass filter with cutoff frequency $1/N$ is $2\pi^2 b^2/(9N^3) \approx 2.2b^2/N^3$. The particular proportionality constant should not be compared directly to that of our nonlinear decoder, since our decoder only has access to N output bits at a time.

The simulation results presented in this section assume that $U_0 = 0$. If we had chosen a different, but known initial state, the codewords and transition points would have changed, but similar conclusions would have been reached. Specifically, the bounds of $O(N^2)$ on the number of codewords and $O(N^{-3})$ on the MSE of any decoder would still hold. Choosing $U_0 = 0$ is merely a convenient way to ensure that $U_0 \in (X-b, X+b)$ for any input $X \in (-b, +b)$.

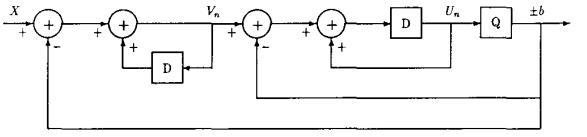


Fig. 6. Discrete-time model of the double-loop $\Sigma\Delta$ encoder.

III. The Double Loop Encoder

In this section we consider the double-loop $\Sigma\Delta$ encoder whose discrete-time model is shown in Fig. 6. In Section III-A we derive an upper bound on the number of codewords of the double-loop encoder with constant inputs and a fixed initial integrator state, and compare the bound to simulation results. In Section III-B we use the results of Section III-A to derive lower bounds on the MSE performance of any decoder.

A. Codewords

Section III-A1) contains an upper bound on the number of codewords, and Section III-A2) contains simulation results on the actual number of codewords.

1) Upper Bound on Number of Codewords: Our approach is to examine the state variables U_n and V_n in Fig. 6. We obtain the state equations

$$\begin{aligned} U_n &= U_{n-1} + V_{n-1} - Q(U_{n-1}) \\ V_n &= X + V_{n-1} - Q(U_n). \end{aligned} \tag{16}$$

Appendix B contains a derivation of an expression for the transition points, analogous to (5) for the single-loop encoder. For all transition points there exist integers (j, n) such that

$$X = \frac{\frac{1}{2}n(n+1)b - (2jb + U_0 + nV_0)}{\frac{1}{2}n(n-1)}, \qquad 2 \le n \le N-1 \tag{17}$$

where U_0 and V_0 are initial integrator states. Unlike for the single-loop encoder, (17) only gives the values of the constant input X that *might be* transition points, but any transition point corresponds to some parameter set (j, n). The absence of a better formula is due to the fact that no result analogous to (3) exists for the double-loop encoder.

From (17) we can find the range of integers j such that $X \in B$. Counting the number of possible parameter sets (p, j), we obtain the following upper bound on the number of codewords, derived in more detail in Appendix 3.1:

$$C_{\max} = \frac{1}{6}N(N-1)(N-2) + 1, \quad N \ge 3. \tag{18}$$

This is an $O(N^3)$ upper bound, in contrast with the upper bound of $O(N^2)$ for the single-loop encoder. For large N, the upper bound is a diminishingly small fraction of 2^N.

2) Numerical Results: We have used computer simulations to find the actual number of codewords for the case $U_0 = V_0 = 0$. Fig. 7 shows the results, as well as the $O(N^3)$ upper bound derived in Section III-A1). The actual number of codewords is seen to also be $O(N^3)$; the upper bound is about three times larger than the actual count.

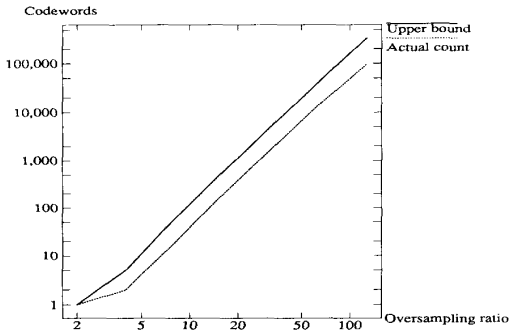


Fig. 7. Double-loop encoder: Actual number of codewords and the upper bound (18) as functions of the oversampling ratio.

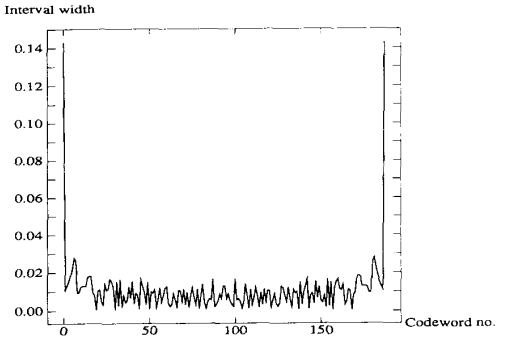


Fig. 8. Double-loop encoder: Width of the quantization interval associated with the ith codeword versus i for an oversampling ratio of $N = 16$. Widths are normalized by b.

B. Lower Bounds on MSE

In Section III-B1) we derive lower bounds on the MSE obtainable with any decoder operating on a double-loop encoder. In Section III-B2) we show numerical results on the actual MSE of the optimal nonlinear decoder.

1) Lower Bounds: Our first lower bound is based on (8) as for the single-loop encoder. Since the number of codewords is upper bounded by $O(N^3)$, the minimum MSE is lower-bounded by $O(N^{-6})$. If in particular the constant input X is uniformly distributed on $(-b, +b)$, then

$$\mathrm{MSE}_{\min} \geq \frac{b^2}{3\left[\frac{1}{6}N(N-1)(N-2)+1\right]^2} \approx \frac{12b^2}{N^6}. \quad (19)$$

We next derive a tighter lower bound on the MSE by finding the widths of some actual quantization intervals. Appendix B shows that for initial integrator states $U_0 = V_0 = 0$, the intervals closest to the extreme inputs are the largest intervals, and they have widths

$$d_1 = d_C = \frac{2b}{N-2}.$$

We can use these interval widths in (9) to derive the following $O(N^{-3})$ lower bound on the MSE, assuming uniformly distributed input:

$$\mathrm{MSE} \geq \frac{1}{24b}\left[2\left(\frac{2b}{N-2}\right)^3\right] = \frac{2b^2}{3(N-2)^3} \approx \frac{0.67b^2}{N^3}, \quad N \geq 3. \quad (20)$$

At first sight, this bound is disappointing since it does not improve on the $O(N^{-3})$ lower bound for the single-loop encoder, except for a proportionality constant. However, if we restrict the dynamic range of the input so that the two largest quantization intervals in the neighborhood of $+b$ and $-b$ are avoided, the performance can be improved drastically, as shown in the next section.

2) Numerical Results: To demonstrate the nonuniformity of the double-loop encoder, Fig. 8 shows a plot of the quantization interval widths for an oversampling ratio of 16 under the assumption of zero initial states. Comparing with Fig. 4 for the single-loop encoder, we see that the double-loop encoder is a more uniform quantizer, the only significant non-uniformities being located at the edges of the dynamic range.

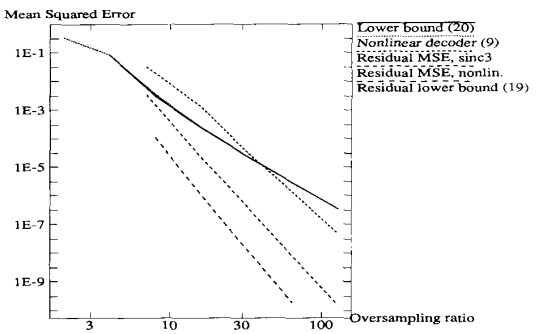


Fig. 9. Double-loop encoder: MSE performance of the optimal nonlinear decoder as a function of oversampling ratio, and $O(N^{-3})$ lower bound (20). Also shown is the $O(N^{-5})$ residual MSE obtained by limiting the dynamic range by 10%. For comparison, a curve shows the residual MSE of sinc^3 decoder. The last curve shows the $O(N^{-6})$ lower bound (19). The MSE is normalized by b^2.

Fig. 9 shows a number of theoretical and computed MSE curves for the double-loop encoder, assuming uniformly distributed input. One curve shows the performance of an optimal nonlinear decoder with an MSE computed from (9) using the actual interval widths. Another curve represents the $O(N^{-3})$ lower bound of (20). The lower bound and the simulated curve are extremely close, indicating that almost all the MSE comes from the two edge intervals with widths $O(N^{-1})$. The MSE can be significantly reduced by excluding these two edge intervals. A third curve in Fig. 9 shows simulated values of the residual MSE obtained by limiting the dynamic range to $(-0.9b, +0.9b)$; this curve is denoted by "residual MSE, nonlin" in Fig. 9. The corresponding MSE performance is approximately $O(N^{-5})$. The improvement provides a quantitative reason for not fully utilizing the dynamic range, since avoiding the edge intervals is an advantageous sacrifice of dynamic range for resolution. For comparison, another curve shows the MSE performance of the sinc^3 decoder with N taps. At a given oversampling ratio, the optimum nonlinear decoder is about 20 dB better than the sinc^3 decoder. Finally, the fifth curve in the figure corresponds to the $O(N^{-6})$ lower bound of (19).

IV. A Modified Cascade Structure for dc Inputs

In Section IV-A we present a derivation of the MSE performance of the single-loop encoder with an averaging decoder. Although the result is known for the general case of arbitrary FIR filtering [5], our derivation is necessary to understand the cascade structure introduced in the following sections. In Sections IV-B and C we describe a modified cascade encoder structure suitable for dc inputs, and a decoding algorithm for it. The encoder structure can have arbitrarily many stages, but for convenience we only consider two stages. Section IV-D contains numerical results for the structure. We compare our structure to the MASH modulator [8] which is applicable to more general band-limited inputs, and not designed specifically for dc inputs. Similar to the MASH modulator, our structure does not suffer from instability problems since it consists of stable single-loop encoders.

A. MSE Analysis of the Averaging Decoder

We begin by decomposing the constant input $X \in (-b, +b)$ into an integer multiple P of the step size

$$\Delta = \frac{2b}{N}$$

and a residue R in the range $[0, \Delta)$. Specifically,

$$X = -b + P\Delta - R. \tag{21}$$

From (21),

$$1 \le P = \left\lceil \frac{N(X+b)}{2b} \right\rceil \le N - 1.$$

For large values of N, the random variable R is uniformly distributed on $[0, \Delta)$. The residue can be considered the quantization error in representing the input X by the integer P.

Consider now the single-loop encoder with state variable U_n. We assume that the initial state U_0 is uniformly distributed on the range $(X - b, X + b)$. Equation (3) then shows that $U_N \in (X - b, X + b)$. From (4), the number of negative output bits A in an N-bit codeword is specified by

$$U_N = U_0 - N(b - X) + 2Ab. \tag{22}$$

Using the decomposition (21), (22) can be rewritten

$$RN = U_0 - U_N - 2b(N - P - A). \tag{23}$$

The product RN is in the range $[0, 2b)$, and since both U_0 and U_N are in the range $(X - b, X + b)$, their difference is in the range $(-2b, +2b)$. The last term on the right-hand side of (23) is an integer multiple of $2b$. We therefore find that

$$RN = \begin{cases} U_0 - U_N & \text{if } U_0 \le U_N \\ U_0 - U_N + 2b & \text{if } U_0 > U_N \end{cases}$$

or more concisely,

$$RN - b = (U_0 - U_N) - Q(U_0 - U_N). \tag{24}$$

Inserting (24) in (23),

$$N - P - A = \frac{b - Q(U_0 - U_N)}{2b} \triangleq \Theta \tag{25}$$

where the random variable Θ takes on the values 0 and 1. The following three theorems investigate the statistical properties of Θ; their proofs can be found in Appendix C.

Theorem 1: If U_0 is uniformly distributed on $(X - b, X + b)$, then for a given input X,

$$P[\Theta = 1|X] = \frac{R}{\Delta}.$$

Theorem 2: If U_0 is uniformly distributed on $(X - b, X + b)$, then for large oversampling ratios, the MSE of the averaging decoder (1) for a given X is given by

$$E\left[\left(X - \hat{X}\right)^2 \,\middle|\, X\right] = \Delta^2 \cdot P[\Theta = 1|X] \cdot P[\Theta = 0|X] = R(\Delta - R).$$

Theorem 3: If the constant input X has a smooth probability distribution of $(-b, +b)$, and the initial state U_0 is uniformly distributed on $(X - b, X + b)$, then for large values of N, the MSE of the averaging decoder is

$$\text{MSE} = E\left[\left(X - \hat{X}\right)^2\right] = \frac{\Delta^2}{6}.$$

The result of Theorem 3 agrees with that of [5]. However, in [5] the input is assumed to be uniformly distributed on $(-b, +b)$, whereas we assume smoothly distributed input; furthermore the result in [5] is independent of the distribution of U_0, while we assume U_0 to be uniformly distributed. There are also fundamental differences between our MSE derivation and those in [2], [3], [11]. The latter derivations assume that the error sequence $\{U_n - Q(U_n)\}$ is white. Gray showed in [5] that in fact this assumption does not hold true in situations where Q quantizes to as few bits as one.

B. Cascade Encoder Structure

In this section we describe a cascade structure specifically for constant inputs, based on the derivations of Section IV-A. The idea is to feed a constant input X_1 into a single-loop encoder with an oversampling ratio of N_1, and subsequently use a residual error of this first stage as the constant input X_2 to another single-loop encoder with oversampling ratio N_2. Using the output sequences of both encoders, the input X_1 can be estimated with greater accuracy than if only the output of the first stage is used. The idea generalizes easily to more than two stages, but we restrict attention to two stages for simplicity. The cascade structure may be viewed as a variation on the MASH encoder proposed by Uchimura *et al.* [8].

We recall that by (21), any constant input X_1 to a single-loop encoder with oversampling ratio N_1 can be decomposed as

$$X_1 = -b + P_1\Delta_1 = R_1 \quad \text{where } \Delta \triangleq \frac{2b}{N_1}. \tag{26}$$

For large oversampling ratios and a smoothly distributed constant input X_1, the residue R_1 is uniformly distributed on $[0, \Delta_1)$, so the random variable

$$X_2 \triangleq R_1 N_1 - b \tag{27}$$

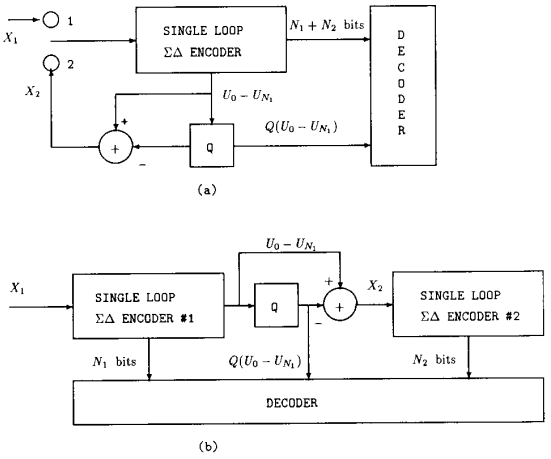


Fig. 10. Cascade structure: (a) One single-loop encoder performs the functions of both stages, controlled by a switch; (b) Two single-loop encoders are cascaded, and the first one passes a requantized residue as the constant input to the second one.

is uniformly distributed on $(-b, +b)$. It can therefore be used as the constant input to a second single-loop encoder. We refer to X_2 as the requantized residue. Denoting the state variable of the single-loop encoder by U_n, the requantized residue can be written

$$X_2 = (U_0 - U_{N_1}) - Q(U_0 - U_{N_1}). \tag{28}$$

To calculate X_2 we therefore need access to U_{N_1}, which is immediately available, and U_0, which can be stored in a sample-and-hold circuit. Alternatively, U_0 can be initialized to zero at the beginning at each conversion cycle; in that case, the residue X_2 is simply given by $Q(U_{N_1}) - U_{N_1}$. In either case, the requantizer residue is used as the constant input to the second stage over N_2 cycles. The set-up can be contrasted with that of the two-stage MASH encoder, in which the input to the second stage is the quantization error of the first stage at each time step, that is

$$X_{2,n} = Q(U_n) - U_n.$$

To summarize, the operation of the modified cascade encoder structure is as follows.

1) The constant input X_1 to the first stage gives rise to an N_1-bit codeword.

2) The requantized residue X_2 is found using (28) and used as the constant input to the second stage over N_2 samples.

Decoding for this encoder is described in the following section. Our described cascade structure is shown in Fig. 10. In Fig. 10(a) the same single-loop encoder is used to perform the tasks of both stages, controlled by the position of the switch. In Fig. 10(b) two separate single-loop encoders are used, and the first encoder can start encoding the next input while the second encoder is encoding the requantized version of the first input.

C. Decoding

In this section we derive a decoding scheme for the cascade structure described in Section IV-B. By (26) we need to determine the integer number of steps P_1 and estimate the residue R_1 in order to achieve our goal of estimating the constant input X_1. Using (25), P_1 is given by

$$P_1 = N_1 - A_1 + \frac{Q(U_0 - U_{N_1}) - b}{2b} \tag{29}$$

where the number of negative bits A_1 is easily counted, and $Q(U_0 - U_{N1})$ is already used to find the requantized residue X_2 in (28). Since

$$A_1 = \frac{1}{2}\left(N_1 - \frac{1}{b}\sum_{i=0}^{N_1-1} Q(U_i)\right)$$

(29) can also be written

$$P_1 = \frac{N_1+1}{2} + \frac{1}{2b}\left(Q(U_0 - U_{N_1}) + \sum_{i=0}^{N_1-1} Q(U_i)\right).$$

This shows that the algorithm for finding P_1 is linear in the output bits $\{Q(U_0), \cdots, Q(U_{N_1-1})\}$, and also linear in $Q(U_{N1})$ if U_0 is initialized to zero.

An estimate $\hat{X}_2$ of X_2 can be found from the N_2-bit codeword of the second stage using any decoder, linear or nonlinear. From (26) and (27) we then obtain the following estimate $\hat{X}_1$ of the input X_1:

$$\hat{X}_1 = P_1\Delta_1 = \frac{\hat{X}_2 + b}{N_1} = \Delta_1\left(P_1 - \frac{N_1+1}{2} - \frac{\hat{X}_2}{2b}\right).$$

D. MSE Performance

In this section we consider the MSE performance of our proposed modified cascade structure shown in Fig. 10. From (27) the MSE in estimating the constant input X_1 is $1/N_1^2$ times the MSE in estimating the residue R_1. Any decoder can be used for the second stage, but for specificity we consider the optimal nonlinear decoder with MSE given by (15). The overall MSE is then

$$\mathrm{MSE}_{2\mathrm{stage}} \approx \frac{0.91b^2}{N_1^2 N_1^3} \approx \frac{0.91b^2}{N^5} \tag{30}$$

where the last equality follows if $N_1 = N_2 = N$. In this case, the cascade decoder can potentially process a different constant input every N cycles, but it takes $2N$ cycles to obtain an estimate of any given input. For a given number of total samples $N_1 + N_2$, the optimal ratio of N_1 to N_2 is $2:3$.

The decoder for the first stage of our cascade encoder is fixed and described in the previous section, however, we can use any decoder for the second stage. If we use the optimal N_2-tap linear filter derived by Gray [5] instead of the optimal nonlinear decoder, we lose about 3 dB compared to (30), as shown in Section II-B2).

The MSE performance (30) can be compared to that of the two-stage MASH modulator. We consider the case where a two-stage MASH encoder is operated for N cycles, and a decoder must base its estimate on a total of $2N$ output bits, namely, N bits from each stage. Such a modulator produces an input estimate every N cycles and takes N cycles to obtain an input estimate. It thus has the same throughput as our cascade structure, but only half the latency. Fig. 11 shows the MSE

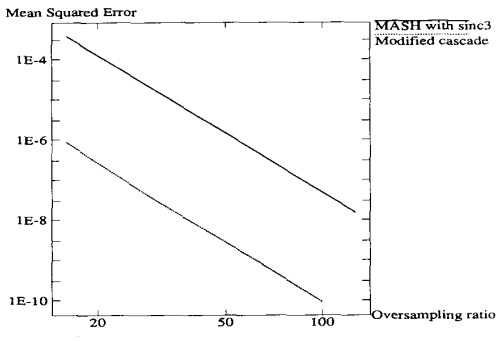


Fig. 11. MSE curves for the cascade structure proposed in Section IV with optimal decoding of second stage, and for the MASH two stage modulator using a sinc^3 filter as its decoder. The MSE is normalized by b^2.

performance of the two cascade modulators.[3] For the MASH modulator we use a standard N-tap sinc^3 filter operating on a linearly filtered combination of the output sequences of the two stages [12]. Our modified cascade structure performs about 27 dB better than the MASH modulator for dc inputs. If we reduce the total number of samples used in our cascade structure such that it achieves the same latency as the MASH modulator, the modified cascade structure still performs 12 dB better than the MASH modulator, and has twice its throughput.

V. Conclusion

We analyzed the single and double-loop $\Sigma\Delta$ modulators as quantizers of constant inputs, in a set-up where N encoder output bits are available at a time to a decoder. By considering in detail the structure of the encoders, we derived upper bounds of $O(N^2)$ and $O(N^3)$ on the number of codewords for the single and double-loop encoders, respectively. Numerical results indicated that the actual numbers of codewords are also $O(N^2)$ and $O(N^3)$. These results bound the sizes of reproductions alphabets of all decoders operating on single and double-loop encoder outputs under the defined set-up. Other simulations showed that the single-loop encoder is a highly nonuniform quantizer. In fact, the ratio between the largest and smallest quantization interval widths is approximately equal to the oversampling ratio. On the other hand, the double-loop encoder is a more uniform quantizer than the single-loop one.

We derived analytical lower bounds on the MSE of all decoders, linear or nonlinear with constant input. The bounds are $O(N^{-3})$ for both the single and double-loop modulators, but if the dynamic range of the double-loop modulator is decreased slightly, our best lower MSE bound is $O(N^{-6})$. We simulated the performance of an optimal nonlinear decoder, based on a table look-up approach in which each possible codeword is mapped to the midpoint of the corresponding quantization interval. Such a decoder, although impractical, represents a bound on the dc performance obtainable with any other decoder. Simulations indicated that the actual MSE performance of the optimal nonlinear decoder is $O(N^{-3})$ for the single-loop modulator, and $O(N^{-5})$ for the double-loop one when the dynamic range is slightly decreased. The same dependence on oversampling ratio can be obtained with linear decoders, but there is a constant gain in dB associated with the optimal decoder. The gain is independent of the OSR and is about 3 dB for the single-loop modulator, 20 dB for the double-loop encoder.

We described a cascade structure designed specifically for constant inputs, consisting of two single-loop encoders. The idea is to coarsely estimate the constant input using essentially an averaging decoder, and subsequently estimate the error of the first stage using another modulator. For a fixed latency, the MSE performance of the cascade structure for constant inputs is about 12 dB better than that of the two-stage MASH modulator, and it achieves twice the throughput rate of the two-stage modulator. On the other hand, for the same throughput the modified cascade structure shows a 27 dB improvement over the two-stage MASH modulator, but twice the latency.

The above conclusions are valid under the assumption of ideal circuit components and operating conditions. Our conclusions are therefore tentative, and experimental results are needed to demonstrate the practical applicability of the results in the presence of various circuit imperfections. Other directions for future research include nonlinear decoder design and analysis of the nonconstant input case.

Appendix A
Derivations for Single Loop Encoder

Appendixes A and B contain derivations to justify the assertions of Sections II-A1) and II-B1), respectively.

A. Transition Points of Single-Loop Encoder

For any given number of negative bits j, there are in general several consecutive time steps n satisfying (4). The largest of these time steps is the position of the $(j+1)$st negative bit, denoted by n_{j+1}. Thus, n_{j+1} is determined by the following inequalities:

$$n_{j+1}(b-X) \geq U_0 + 2jb$$
$$(n_{j+1}-1)(b-X) < U_0 + 2jb.$$

Taken together, these two inequalities imply that the position of the $(j+1)$st negative bit is

$$n_{j+1} = \left\lceil \frac{U_0 + 2jb}{b-X} \right\rceil, \quad 0 \leq j \leq N-1 \tag{31}$$

where $\lceil a \rceil$ is the smallest integer greater than or equal to a. To illustrate the implications of (31), let us assume that the initial state is $U_0 = 0$; the first bit of any codeword is then $Q(U_0) = -b$ regardless of the input. We can make the following detailed observations. For $b - \frac{2b}{N-1} < X < b$, the corresponding N-bit codeword consists of all positive bits except for the first one. For $X = b - \frac{2b}{N-1}$, a second negative bit appears at item step $N-1$, and as X decreases, this bit advances to earlier time steps $N-2, N-3, \cdots$. At $X = b - \frac{4b}{N-1}$, the second negative bit reaches time step

[3]The dynamic range of the MASH modulator has been reduced to $(-0.9b, +0.9b)$. The corresponding reduction has not been done for our cascade structure, since the requantized residue (28) used as input to the second stage would still be in the larger range $(-b, +b)$, and only the decoder of the second stage would benefit from a reduction in dynamic range: From (27), the MSE in estimating the constant input X_1 is independent of the integer P_1 defined in (26).

$\lceil (N-1)/2 \rceil$, and a third negative bit appears at item step $N-1$. As X decreases from $b - \frac{4b}{N-1}$ to $b - \frac{6b}{N-1}$, the second bit moves from $\lceil (N-1)/2 \rceil$ to $\lceil (N-1)/3 \rceil$, and the third negative bit moves from $N-1$ to $\lceil 2(N-1)/3 \rceil$. More generally, as X crosses $b = \frac{2bq}{N-1}$, the $(j+1)$st negative bit reaches time step $\lceil j(N-1)/q \rceil$ for $0 \le j < q$, and the $(q+1)$st negative bit appears at time step $N-1$. To help visualize the process, Table II shows the codewords of the single-loop encoder for $N = 12$, as the input X is swept from $+b$ to $-b$. Note the correspondence with Table I in Section II-A1).

We can further exploit equation (31) to find an upper bound on the number of codewords of the single-loop encoder. From (31) it is clear that n_{j+1} changes from one time step to the next when X passes the constant input value given by

$$\frac{2jb + U_0}{b - X} = p \Rightarrow X = \frac{pb - (2jb + U_0)}{p}, \quad 1 \le p \le N-1. \tag{32}$$

This input value is referred to as a transition point, since it marks a change in the N-bit output sequence of the encoder. For any p in the permitted range of integers in (32), we can only allow integers j that result in values of X such that $X \in (-b, +b)$ and $U_0 \in (b - X, b + X)$. For instance, we see that for $U_0 = 0$,

$$-b < \frac{pb - 2jb}{p} < +b \Rightarrow 1 \le j < p.$$

More generally, the requirements $X \in (-b, +b)$, $U_0 \in (X - b, X + b)$ imply that

$$\begin{cases} 0 \le j < p, & 1 \le p \le N-1 \quad U_0 > 0 \\ 1 \le j < p, & 1 \le p \le N-1 \quad U_0 = 0 \\ 1 \le j \le p, & 1 \le p \le N-1 \quad U_0 < 0 \end{cases}. \tag{33}$$

The number of transition points is at most the number of permitted parameter sets (j, p) in (33); there may be fewer transition points than that, since some parameter sets may correspond to the same transition point in (32). The number of codewords is the number of transition points plus one. For $U_0 = 0$ we therefore have a maximum of

$$1 + \sum_{p=1}^{N-1} (p-1) = \frac{1}{2}(N-1)(N-2) + 1$$

codewords. More generally, the maximum number of codewords the single-loop $\Sigma\Delta$ encoder can generate is

$$C_{\max} = \begin{cases} \frac{1}{2}N(N-1) + 1 & U_0 \ne 0 \\ \frac{1}{2}(N-1)(N-2) + 1 & U_0 = 0 \end{cases}.$$

Let us further consider the "degenerate case" where two parameters sets (j_1, p_1), and (j_2, p_2) lead to the same transition point. From (32), this occurs if and only if

$$\frac{p_1 b - (2j_1 b + U_0)}{p_1} = \frac{p_2 b - (2j_2 b + U_0)}{p_2} \Rightarrow \frac{U_0}{b} \Rightarrow \frac{j_2 p_1 - j_1 p_2}{p_2 - p_1}.$$

The right-hand side of this expression is rational, so degeneration can only occur when U_0/b is rational. If that is not the case, all transition points are distinct, and the number of codewords is exactly $\frac{1}{2}N(N-1) + 1$.

B. Interval Widths for Single-Loop Encoder

We assume $U_0 = 0$. Setting $p = N-1$ in (5), we see that there are transition points at

$$X = b\left(1 - \frac{2j}{N-1}\right), \quad 1 \le j \le N-1.$$

Thus no quantization interval can be larger than $2b/(N-1)$. On the other hand, the largest transition point is obtained by setting $(j, p) = (1, N-1)$ in (5), since $b - X = 2jb/p$ is minimized by this choice. Therefore the interval near $X = +b$ has width $d_1 = 2b/(N-1)$, that is, it is a quantization interval of maximal width. The same holds for the quantization interval near $X = -b$, so $d_1 = d_C$.

Consider next the intervals close to $X = 0$. Equation (5) shows that $(j, p) = (1, 2)$ leads to a transition point at $X = 0$ for any oversampling ratio $N \ge 3$. The smallest positive transition point is found by maximizing the denominator and minimizing the denominator of (5) while keeping it positive. This leads to $(j, p) = ((N-2)/2, N-1)$ and $((N-3)/2, N-1)$ for N even and N odd, respectively, and the interval widths are $b/(N-1)$ and $2b/(N-1)$. By symmetry, the same holds on the other side of $X = 0$. We thus have $d_{C/2} = D_{C/2+1} \ge b/(N-1)$.

Consider now two pairs (j_1, p_1) and (j_2, p_2) satisfying (33). The distance between the corresponding transition points is

$$\frac{p_1 - 2j_1 b}{p_1} - \frac{p_2 - 2j_2 b}{p_2} = 2b\frac{p_1 j_2 - p_2 j_1}{p_1 p_2}.$$

If $p_1 = p_2$, the smallest absolute distance is $2b/(N-1)$. If $p_1 \ne p_2$, the largest product of p_1 and p_2 is $(N-1)(N-2)$, and the smallest positive values of the numerator is 1. These extremes are simultaneously achieved by choosing $(j_1, p_1) = (1, N-1)$ and $(j_2, p_2) = (1, N-2)$. Therefore, the width of the shortest quantization interval is $d_{min} = 2b/[(N-1)(N-2)]$, and one interval of this minimal width is adjacent to the interval near $X = +b$. Another interval of the same width is located near $X = -b$, so $d_{C-1} = d_2 = d_{\min}$.

Finally, let us assume that $U_0 > 0$. The largest transition point in (5) occurs at $(j, p) = (0, N-1)$. The quantization interval near $X = +b$ thus has width $d_1 = U_0 b/(N-1)$ which is of order $O(N^{-1})$.

Appendix B
Derivations for Double-Loop Encoder

Appendixes A and B contain derivations to justify the assertions of Sections III-A1) and III-B1), respectively.

A. Transition Points of Double-Loop Encoder

Equation (16) can be rewritten

$$U_n = U_0 + \sum_{i=1}^{n} (U_i - U_{i-1})$$

$$= U_0 + \sum_{i=0}^{n-1} V_i - \sum_{i=0}^{n-1} Q(U_i) \quad (34)$$

$$V_n = V_0 + \sum_{j=1}^{n} (V_j - V_{j-1})$$

$$= V_0 + nX - \sum_{j=1}^{n} Q(U_j). \quad (35)$$

Summing the left-hand side of (35) we get

$$\sum_{i=0}^{n-1} V_i = nV_0 + \frac{1}{2}n(n-1)X - \sum_{i=1}^{n-1}(n-1)Q(U_i), \quad n \geq 1. \quad (36)$$

Inserting (36) in (34) we find

$$U_n = U_0 + nV_0 + \frac{1}{2}n(n-1)X - Q(U_0) - \sum_{i=1}^{n-1}(n-i+1)Q(U_i), \quad n \geq 1. \quad (37)$$

For the single-loop encoder, we now used the fact (3). Unfortunately, a similar fact does not hold for the double-loop encoder, necessitating a less detailed manipulation of (37). Its last two terms involving quantizer outputs are more conveniently written

$$Q(U_0) + \sum_{i=1}^{n-1}(n-i+1)Q(U_i) = \frac{1}{2}n(n+1)b - 2jb \quad (38)$$

where j is an integer in the range $0 \leq j \leq \frac{1}{2}n(n+1)$. The extreme sequences consisting of all $+b$'s or all $-b$'s correspond to $j = 0$ and $j = N-1$, respectively. There are values of j between 0 and $N-1$ for which no N-bit codeword exists, but for any codeword there exists a j between 0 and $N-1$. Inserting (38) in (37), we find

$$U_n = U_0 + nV_0 + \frac{1}{2}n(n-1)X - \frac{1}{2}n(n+1)b + 2jb. \quad (39)$$

The quantizer transition points are the values of $X \in (-b, +b)$ which result in $U_n = 0$ at some time step n between 1 and $N-1$. Unfortunately, the variable j in (39) has no simple interpretation as in the single-loop case. Hence, we can only solve (39) for values of X that *might be* transition points. Counting the number of parameter sets (j, n) for which $U_n = 0$ has a solution $X \in (-b, +b)$ will result in an upper bound on the number of transition points. Setting U_n in (39) to zero, we obtain

$$X = \frac{\frac{1}{2}n(n+1)b - (2jb + U_0 + nV_0)}{\frac{1}{2}n(n-1)}, \quad 2 \leq n \leq N-1. \quad (40)$$

For $n = 0$ and $n = 1$ there are no transition points, since U_0 and U_1 are independent of X. Equation (40) is analogous to (5) for the single-loop encoder.

We now determine an upper bound on the number of codewords. Considering that X must be in the range $(-b, +b)$, it follows from (40) that j must satisfy

$$\frac{n - (U_0 + nV_0)/b}{2} < j < \frac{n^2 - (U_0 + nV_0)/b}{2}. \quad (41)$$

The number of integer values of j satisfying (41) is $\leq \frac{1}{2}(n^2 - n)$. Therefore, the number of codewords is upper bounded by

$$C_{\max} \leq 1 + \sum_{n=2}^{N-1} \frac{1}{2}(n^2 - n) = \frac{1}{6}N(N-1)(N-2) + 1, \quad N \geq 3.$$

B. Interval Widths for Double-Loop Encoder

We assume for convenience that the initial states are $U_0 = V_0 = 0$, and show that the largest positive transition point occurs at $X = (N-4)b/(N-2)$. Equations (16) together with the initial states imply that

$$Q(U_0) = -b \quad U_1 = +b \quad V_1 = X - b$$
$$Q(U_1) = +b \quad U_2 = X - b < 0 \quad Q(U_2) = -b.$$

This states that the first three output bits are $(-b, +b, +b)$ regardless of the input. At times $n \geq 3$, the quantizer produces only positive bits until U_n becomes negative—the closer X is to $+b$, the longer it takes to generate the first negative bit. We will find the value of X which results in the first negative bit appearing at time $N-1$; this input corresponds to the most positive transition point. Assuming $Q(U_3) = \cdots = Q(U_{N-2}) = +b$, we find from (37) that

$$U_{N-1} = \frac{1}{2}(N-1) \cdot [(N-2)X - (N-4)b].$$

Setting U_{N-1} to zero, we conclude that the most positive transition point is at $X = (N-4)b/(N-2)$. The corresponding quantization interval has width $2b/(N-2)$. By symmetry, a similar interval exists from $-b$ to $-(N-4)b/(N-2)$.

VIII. APPENDIX C
PROOFS OF THEOREMS 1, 2 AND 3

Proof of Theorem 1: From 25, the number of negative bits A on an N-bit codeword is given by

$$A = N - P + \Theta.$$

The averaging decoder estimate is therefore

$$\hat{X} = \frac{(N-A) - A}{N}b = -b + (P - \Theta)\Delta$$

so

$$X - \hat{X} = \Theta\Delta - R. \quad (42)$$

On the other hand, we can use (21) in (22) to show that

$$X = -b + (P - \Theta)\Delta + \frac{U_N - U_0}{N}$$

so

$$X - \hat{X} = \frac{U_N - U_0}{N}.$$

TABLE II
CODEWORDS AND TRANSITION POINTS FOR THE SINGLE LOOP ENCODER AT AN OVERSAMPLING RATIO OF $N = 12$; IT IS ASSUMED THAT $U_0 = 0$. AS X CROSSES A TRANSITION POINT FROM ABOVE, THE OUTPUT SEQUENCE CHANGES FROM THE CODEWORD ON THE LINE ABOVE TO THE ONE ON THE LINE CONTAINING THE TRANSITION POINT. TRANSITION POINTS ARE NORMALIZED BY b, AND THE BITS ARE SHOWN AS + AND −

Transition point	0	1	2	3	4	5	6	7	8	9	10	11
1	−	+	+	+	+	+	+	+	+	+	+	+
9/11	−	+	+	+	+	+	+	+	+	+	+	−
4/5	−	+	+	+	+	+	+	+	+	+	−	+
7/9	−	+	+	+	+	+	+	+	+	−	+	+
3/4	−	+	+	+	+	+	+	+	−	+	+	+
5/7	−	+	+	+	+	+	+	−	+	+	+	+
2/3	−	+	+	+	+	+	−	+	+	+	+	+
7/11	−	+	+	+	+	+	−	+	+	+	+	−
3/5	−	+	+	+	+	−	+	+	+	+	−	+
5/9	−	+	+	+	+	−	+	+	+	−	+	+
1/2	−	+	+	+	−	+	+	+	−	+	+	+
5/11	−	+	+	+	−	+	+	+	−	+	+	−
3/7	−	+	+	+	−	+	+	−	+	+	+	−
2/5	−	+	+	+	−	+	+	−	+	+	−	+
1/3	−	+	+	−	+	+	−	+	+	−	+	+
3/11	−	+	+	−	+	+	−	+	+	−	+	−
1/4	−	+	+	−	+	+	−	+	−	+	+	−
1/5	−	+	+	−	+	−	+	+	−	+	−	+
1/7	−	+	+	−	+	−	+	−	+	+	−	+
1/9	−	+	+	−	+	−	+	−	+	−	+	+
1/11	−	+	+	−	+	−	+	−	+	−	+	−
0	−	+	−	+	−	+	−	+	−	+	−	+
−1/11	−	+	−	+	−	+	−	+	−	+	−	−
−1/9	−	+	−	+	−	+	−	+	−	−	+	−
−1/7	−	+	−	+	−	+	−	−	+	−	+	−
−1/5	−	+	−	+	−	−	+	−	+	−	−	+
−1/4	−	+	−	+	−	−	+	−	−	+	−	+
−3/11	−	+	−	+	−	−	+	−	−	+	−	−
−1/3	−	+	−	−	+	−	−	+	−	−	+	−
−2/5	−	+	−	−	+	−	−	+	−	−	−	+
−3/7	−	+	−	−	+	−	−	−	+	−	−	+
−5/11	−	+	−	−	+	−	−	−	+	−	−	−
−1/2	−	+	−	−	−	+	−	−	−	+	−	−
−5/9	−	+	−	−	−	+	−	−	−	−	+	−
−3/5	−	+	−	−	−	−	+	−	−	−	−	+
−7/11	−	+	−	−	−	−	+	−	−	−	−	−
−2/3	−	+	−	−	−	−	−	+	−	−	−	−
−5/7	−	+	−	−	−	−	−	−	+	−	−	−
−3/4	−	+	−	−	−	−	−	−	−	+	−	−
−7/9	−	+	−	−	−	−	−	−	−	−	+	−
−4/5	−	+	−	−	−	−	−	−	−	−	−	+
−9/11	−	+	−	−	−	−	−	−	−	−	−	−

Since U_0 is uniformly distributed by assumption, and (22) can be shown to imply that U_N is also uniformly distributed, we have $E[(X - \hat{X})|X] = 0$. By (42), this implies

$$0 = E[(\Theta\Delta - R)|X] = \Delta \cdot P[\Theta = 1|X] - R$$

which shows the theorem.

Proof of Theorem 2: Using (42),

$$\begin{aligned} E\left[\left(X - \hat{X}\right)^2 \,\middle|\, X\right] &= \sum_{\theta=0}^{1} P[\Theta = \theta] \cdot (\theta\Delta - R)^2 \\ &= \left(1 - \frac{R}{\Delta}\right)(-R)^2 + \frac{R}{\Delta}(\Delta - R)^2 \\ &= R(\Delta - R). \end{aligned}$$

Proof of Theorem 3: The random variable R is uniformly distributed on $[0, \Delta)$, so $ER = \frac{\Delta}{2}$ and $ER^2 = \frac{\Delta^2}{3}$. Now

$$\begin{aligned} E\left[\left(X - \hat{X}\right)^2\right] &= E\left[E\left[\left(X - \hat{X}\right)^2 \,\middle|\, X\right]\right] \\ &= E[R(\Delta - R)] \\ &= \frac{\Delta}{2} \cdot \Delta - \frac{\Delta^2}{3} \\ &= \frac{\Delta^2}{6}. \end{aligned}$$

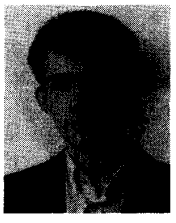
REFERENCES

[1] R. M. Gray, "Oversampled sigma–delta modulation," *IEEE Trans. Commun.*, vol. COM-35, pp. 481–489, May 1987.

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

[3] ——, "Decimation for sigma–delta modulation," *IEEE Trans. Commun.*, vol. COM-34, pp. 72–76, Jan. 1986.

[4] A. Zakhor and A. V. Oppenheim, "Sampling and reconstruction schemes for multidimensional signals," *Proc. IEEE,* vol. 78, pp. 31–55, Jan. 1990.
[5] R. M. Gray, "Spectral analysis of quantization noise in a single loop sigma delta modulator with dc input," *IEEE Trans. Commun.,* vol. 37, pp. 588–599, June 1989.
[6] K. C.-H. Chao, S. Nadeem, W. L. Lee, and C. G. Sodini, "A higher order topology for interpolative modulators for oversampling A/D converters," *IEEE Trans. Circ. Syst.,* vol. 37, pp. 309–318, Mar. 1990.
[7] A. N. Netravali, "Optimum digital filters for interpolative A/D converters," *Bell Syst. Tech. J.,* vol. 56, no. 9, pp. 1629–1641, Nov. 1977.
[8] K. Uchimura, T. Hayashi, T. Kimura, and A. Iwata, "Oversampling A-to-D and D-to-A converters with multistage noise shaping modulators," *IEEE Trans. Acoust., Speech Signal Processing,* vol. 36, pp. 1899–1905, Dec. 1988.
[9] G. W. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers.* New York: Oxford Univ. Press, 1979.
[10] R. M. Gray and A. H. Gray, Jr., "Asymptotically optimal quantizers," *IEEE Trans. Inform. Theory,* vol. IT-23, pp. 143–144, Feb. 1977.
[11] J. C. Candy and O. J. Benjamin, "The structure of quantization noise from sigma delta modulation," *IEEE Trans. Commun.,* vol. COM-29, pp. 1316–1323, Sept. 1981.
[12] P. Wong and R. M. Gray, "Two-stage sigma delta modulation," submitted for publication.



Søren Hein was born in May 1968 in Copenhagen, Denmark. He received the M.Sc. degree in electrical engineering from the Technical University of Denmark in January 1989. He is currently pursuing the Ph.D. degree in electrical engineering at the University of California at Berkeley.

His research interests include algorithmic aspects of oversampled A/D conversion and signal processing. he has also worked on error-correction coding for satellite communications.

Khalid M. Ibrahim was born in Baghdad, Iraq, in 1959. He received the B.S. and M.S. degrees in electrical engineering from the University of Baghdad in 1980 and 1983, respectively.

From 1983 to 1985, he worked with the University of Baghdad in the field of instrumentation, spread-spectrum communication and coding theory. From 1985 to 1988, he worked for the Scientific Research Council, Space Research Center in the field of instrumentation, image processing and parallel processing. In 1989, he worked with the University of California at Berkeley in the field of signal reconstruction from partial information. Since 1990, he has been with Aeromatrics Inc., CA, where he is presently a scientist working in the field of signal and image processing and parallel processing for laser Doppler applications.

Avideh Zakhor (M'87) received the B.S. degree from California Institute of Technology, Pasadena, and the S.M. and Ph.D. degrees from Massachusetts Institute of Technology, Cambridge, all in electrical engineering, in 1983, 1985, and 1987 respectively. Her doctoral thesis was on sampling and reconstruction schemes for multidimensional signals.

In 1988, she joined the Faculty at U.C. Berkeley where she is currently Assistant Professor in the Department of Electrical Engineering and Computer Sciences. Her research interests are in the general area of signal processing and its applications to images and video, and biomedical data. She has been a consultant to a number of industrial organizations in the areas of signal processing, communications and medical imaging, and has two pending patents on MRI signal processing, one on sigma delta modulators, and one on phase shifting mask design for optical lithography.

Dr. Zakhor was a General Motors scholar from 1982 to 1983, received the Henry Ford Engineering Award and Caltech Prize in 1983, was a Hertz fellow from 1984 to 1988, received the Presidential Young Investigators (PYI) award, IBM junior faculty development award, and Analog Devices junior faculty development award in 1990, and ONR young investigator award in 1991. She is a member of the image and multidimensional digital signal processing committee and Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING.