OPTIMAL DECODING FOR SIGMA DELTA MODULATORS

by

Søren Hein and Avideh Zakhor

Memorandum No. UCB/ERL M90/51

12 June 1990

COVER PAGE

OPTIMAL DECODING FOR SIGMA DELTA MODULATORS

by

Søren Hein and Avideh Zakhor

Memorandum No. UCB/ERL M90/51

12 June 1990

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

OPTIMAL DECODING FOR SIGMA DELTA MODULATORS

by

Søren Hein and Avideh Zakhor

Memorandum No. UCB/ERL M90/51

12 June 1990

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

OPTIMAL DECODING FOR SIGMA DELTA MODULATORS

Søren Hein and Avideh Zakhor *
Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720
(415) 643-6777
e-mail: shein@united.berkeley.edu and avz@united.berkeley.edu

Abstract

In [1] we subjected the single and double loop Sigma Delta ($\Sigma\Delta$) encoders to time domain analysis, and described an optimal table look-up decoding principle for constant inputs under the assumption of known initial integrator states. This paper introduces a simple implementation technique, dubbed *zooming*, for optimal decoders. The technique is applicable to all the current popular $\Sigma\Delta$ encoder structures, including single and double loop encoders, the MASH encoder [2] and the interpolative encoder proposed in [3].

1 Introduction

Sigma Delta ($\Sigma\Delta$) modulators as A/D converters have recently received considerable attention both in industry and in the communications and signal processing literature. Their theoretical attraction lies in the trade-off provided between sampling rate and resolution of the in-loop quantizer — specifically, they achieve the same resolution as a multi-bit quantizer operating at the Nyquist rate by employing a one-bit quantizer operating at many times the Nyquist rate. In practice, the nonlinearity resulting from unevenly spaced quantization levels in the multi-bit quantizer is detrimental, and a one-bit quantizer is often preferred for its extreme ease of implementation and inherent linearity of the two levels.

$\Sigma\Delta$ modulators generally require few and simple components, and are resistant towards circuit imperfections. Furthermore, they obliterate the need for stringent analog anti-aliasing filtering, and relegate the strict processing demands to the digital domain. They are thus attractive for VLSI applications in which analog and digital signals occur on the same chip and require conversion.

In a previous paper [1] we introduced a general time domain technique for analyzing $\Sigma\Delta$ modulators as A/D converters under certain assumptions; these assumptions include

- One-bit in-loop quantizer, given by

$$Q(x) = \begin{cases} -b & x \leq 0 \\ +b & x > 0 \end{cases} \tag{1}$$

where $B = (-b, +b)$ is the *full dynamic range.*

*This work was supported by NSF Grant MIP-8911017 and Joint Services Electronics Project.

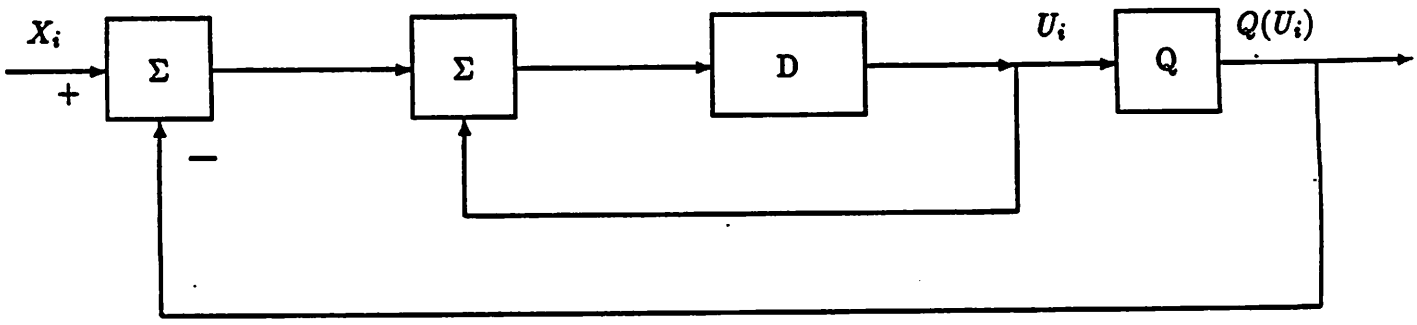

Figure 1: *Single loop ΣΔ encoder.*

- Constant input. The input X is assumed to be a constant in the *dynamic range* $D \subset B$. This reflects the fact that we are focusing on *data acquisition* applications.

  In practice, the full dynamic range is seldom used. One reason is to avoid the danger of exceeding the dynamic range; another is that the largest estimation errors are generally made when the input is close to $\pm b$ [1, 3]. Therefore we restrict the dynamic range to $D = (-kb, +kb)$, where k is chosen to be 0.9.

- Known initial integrator states. We assume for convenience that these are all initialized to zero before the encoder is started. For data acquisition it is reasonable to assume that the problem is to arrive at an estimate of the input given N quantizer outputs, where N is the *oversampling ratio*; the encoder can then be reset after each estimation cycle.

In [1] we decoupled the modulator into the *encoder* and *decoder* parts and investigated the encoder separately. The idea was to view the encoder as a *source coder* or non-uniform quantizer, dividing the dynamic range into intervals separated by *transition points*; each interval corresponds to a distinct N-bit output sequence or *codeword*. The optimal performance in terms of minimizing the MSE is achieved by a decoder which takes a codeword as input, and outputs the midpoint of the corresponding interval. Such a decoder is highly nonlinear.

In [1] we indicated that the optimal decoder could in principle be implemented using a table in the form of a PLA. In practice this is not feasible, as the table would be prohibitively large. Here we present a general technique, called *zooming*, which takes a codeword as input and outputs the corresponding upper and lower bound on the input interval producing the codeword.

2 Single loop modulator

2.1 Theory

Figure 1 shows the discrete-time model of the ideal single loop ΣΔ encoder. It consists of two summers Σ, a delay element D and a one-bit quantizer Q. The inner loop is a discrete integrator which operates on the difference between the input and the quantizer output; thus the encoder seeks to minimize the integrated difference between the input and the output. It is assumed that the state variable at time zero is $U_0 = 0$. The length in bits of the codewords,

i.e., the oversampling ratio, is denoted N. It is seen that

$$U_n = \sum_{i=0}^{n-1} [X_i - Q(U_i)] \;=\; \sum_{i=0}^{n-1} X_i - \sum_{i=0}^{n-1} Q(U_i), \;\; 1 \le n \le N-1 \tag{2}$$

An important observation which proves useful for more complicated encoders is that the input and the quantizer output are both processed by simple linear filters derivable from the open-loop encoder. In this case, both the input and the output are filtered by the in-loop integrator; for the input, this is seen by removing Q and its feedback connection, and for the output it is seen by removing Q and the input. Assuming that the input is a constant X, $X_i = X$, $0 \le i \le N-1$, the first sum in (2) equals nX. For any given codeword, we can easily find the second sum by digital integration.

The first sum in (2) is greater or less than the second sum depending on whether $Q(U_n) = +b$ or $-b$. Equation (2) then leads to the following bounds:

$$X > \bar{X}_n \;\text{ if }\; Q(U_n) = +b; \qquad X \le \bar{X}_n \;\text{ if }\; Q(U_n) = -b \tag{3}$$

where

$$\bar{X}_n = \frac{1}{n} \sum_{i=0}^{n-1} Q(U_i) \tag{4}$$

As $U_0 = 0$, we have $Q(U_0) = -b$. It is seen from Figure 1 that

$$U_1 = U_0 + X_0 - Q(U_0) = X + b > 0 \tag{5}$$

so $Q(U_1) = +b$ regardless of the input. Hence the first informative bit is $Q(U_2)$. The *zoomer* is the decoder which uses the succession of lower and upper bounds from (3) to arrive at overall lower and upper bounds on the X-interval generating the codeword. This is done using two registers L and U, initialized to $-b$ and $+b$, respectively. Sweeping n from 2 to $N-1$, the zoomer maintains the greatest lower bound and the least upper bound in the registers. This extracts all information from the codeword, and thus the resulting bounds are the tightest possible. After finishing the process, the decoder outputs $(L+U)/2$. The zoomer has a large linear component, but the conditional register updating is nonlinear.

For specificity, Figure 2 shows a flowchart for the single-loop zoomer algorithm. It embodies an initialization phase, an update of running sums, and an update of either the lower or the upper bound. The variable S is the cumulative sum of quantizer outputs, and $\bar{X}$ is the quantity calculated in (4). When $n = N-1$, the zoomer terminates and outputs its estimate.

2.2 Performance comparison

This subsection provides some quantitative measure of the performance of the zoomer compared to linear decoders. The linear decoder under consideration is the asymptotically optimal N-tap FIR filter derived by Gray [4] with unity DC gain and tap coefficients

$$h_k = 6\frac{(k+1)(N-k)}{N(N+1)(N+2)}, \;\; 0 \le k \le N-1 \tag{6}$$

Two performance measures are used, *viz.*, the mean squared error (MSE) and the worst-case estimation error, or equivalently, the signal-to-noise ratio (SNR) and the worst-case resolution

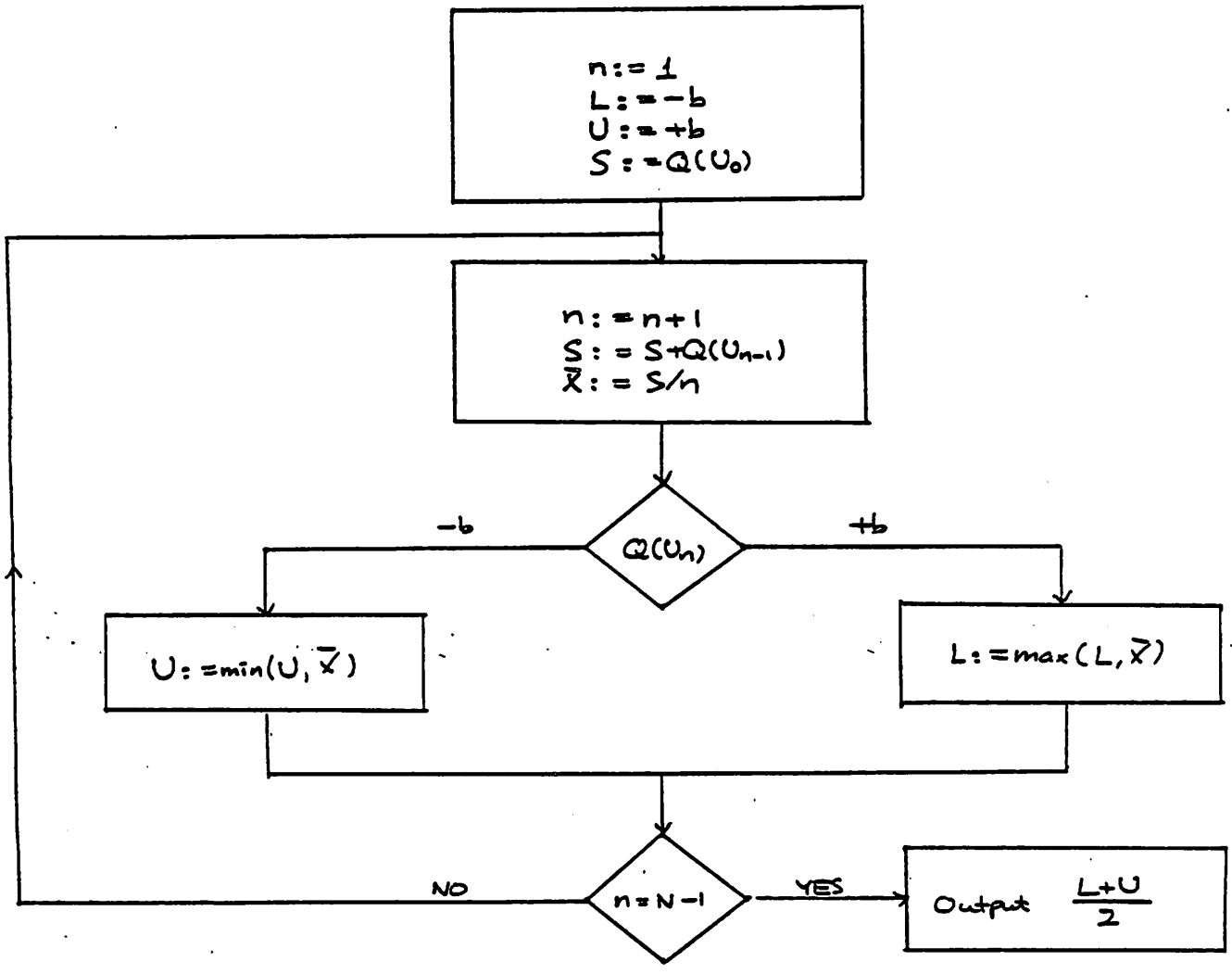

Figure 2: *Flowchart for the single-loop zoomer algorithm.*

in bits. To define these, we first introduce some notation. The number of codewords is denoted by C, and the decoder estimate of the ith codeword is denoted by $\hat{X}_i$. This estimate is found as the average of the interval bounds determined by the zoomer. We denote the length of the interval corresponding to the ith codeword by d_i. Finally, the decoder estimate as a function of the random variable X is denoted by $\hat{X}$. We assume that the constant input X is uniformly distributed on the dynamic range D. This is an analytical convenience, but it is not cardinal to our technique. Any piecewise continuous probability density function for X on D can be incorporated in the analysis below. The performance measures are defined as follows:

- The MSE is given by

$$\text{MSE} = E\left[(X - \hat{X})^2\right] \tag{7}$$

The MSE contribution from the ith interval is

$$\text{MSE}_i = E\left[(X - \hat{X})^2 \mid X \in I_i\right] = \frac{d_i^2}{12} \tag{8}$$

The total MSE is found by taking the weighted sum of these errors,

$$\text{MSE} = \sum_{i=1}^{C} \frac{d_i}{|D|} \cdot \text{MSE}_i = \sum_{i=1}^{C} \frac{d_i^3}{24kb} \tag{9}$$

where $|D|$ is the width of the dynamic range. The average input power is

$$E\left[X^2\right] = \int_{-kb}^{+kb} \frac{1}{2kb} x^2 dx = \frac{(kb)^2}{3} \tag{10}$$

Defining $\text{SNR} = 10\log_{10}\left[E(X^2)/MSE\right]$, we thus have

$$\text{SNR} = 10\log_{10} \frac{8(kb)^3}{\sum_{i=1}^{C} d_i^3} \tag{11}$$

- The worst-case error is given by

$$\varepsilon = \max_{1 \le i \le C} \left\{\frac{d_i}{2}\right\} \tag{12}$$

The worst-case resolution in bits is

$$R = \log_2 \frac{|D|}{2\varepsilon} = \log_2 \frac{kb}{\varepsilon} \tag{13}$$

Figures 3 and 4 show computer simulation results obtained in accordance with these definitions. It is seen that for a given oversampling ratio, the zoomer reduces the MSE by a factor of about 6 and the worst-case error by a factor of 2. The zoomer requires half the oversampling ratio of the FIR filter to obtain essentially the same performance. This translates into shorter data acquisition times. Alternatively, for a fixed oversampling ratio, the zoomer gives an extra bit of worst-case resolution, and it improves the SNR by 7-8 dB or slightly more than one bit.

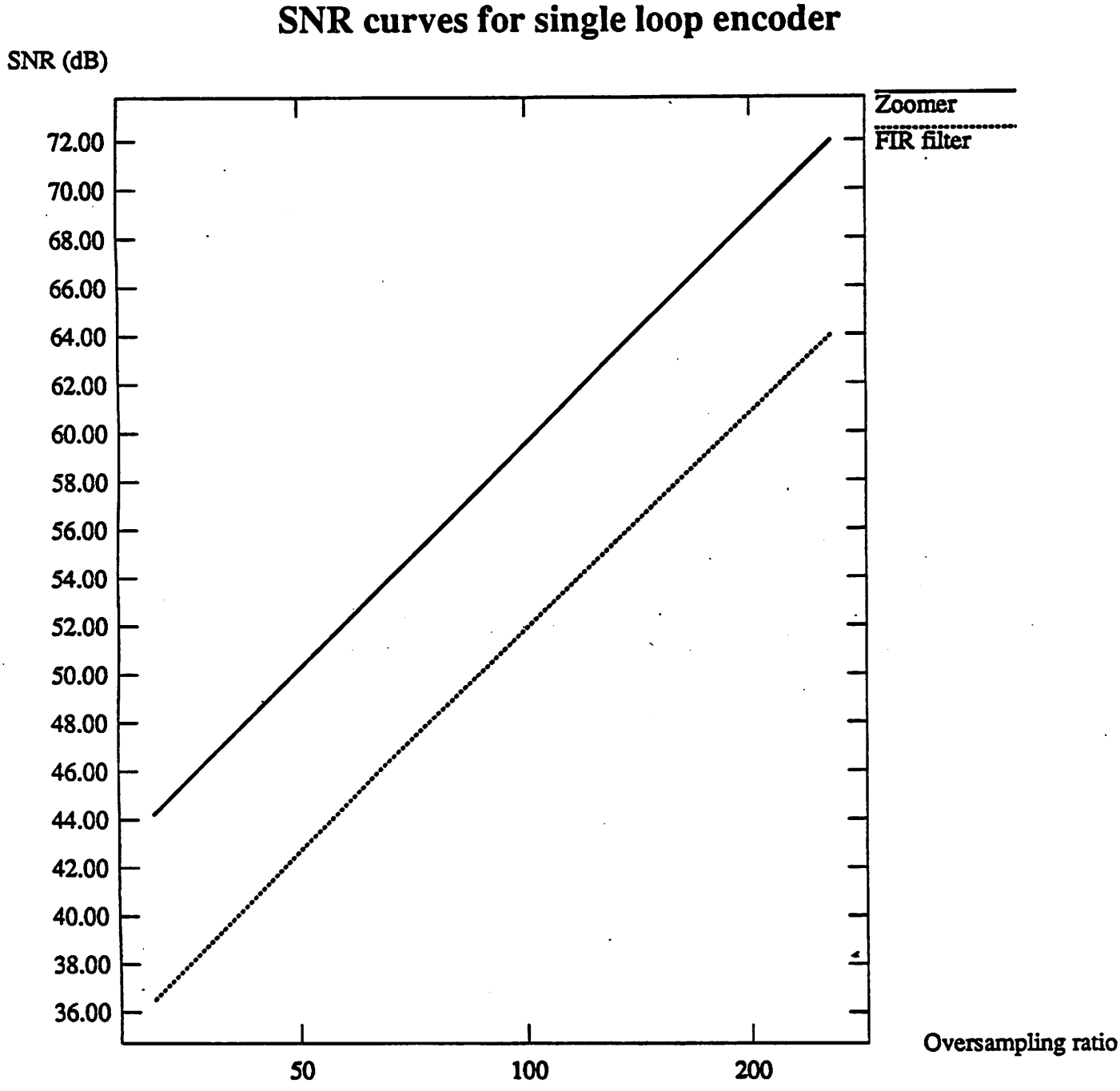

Figure 3: *SNR as a function of oversampling ratio for the zoomer and the asymptotically optimal FIR filter for single loop decoding.*

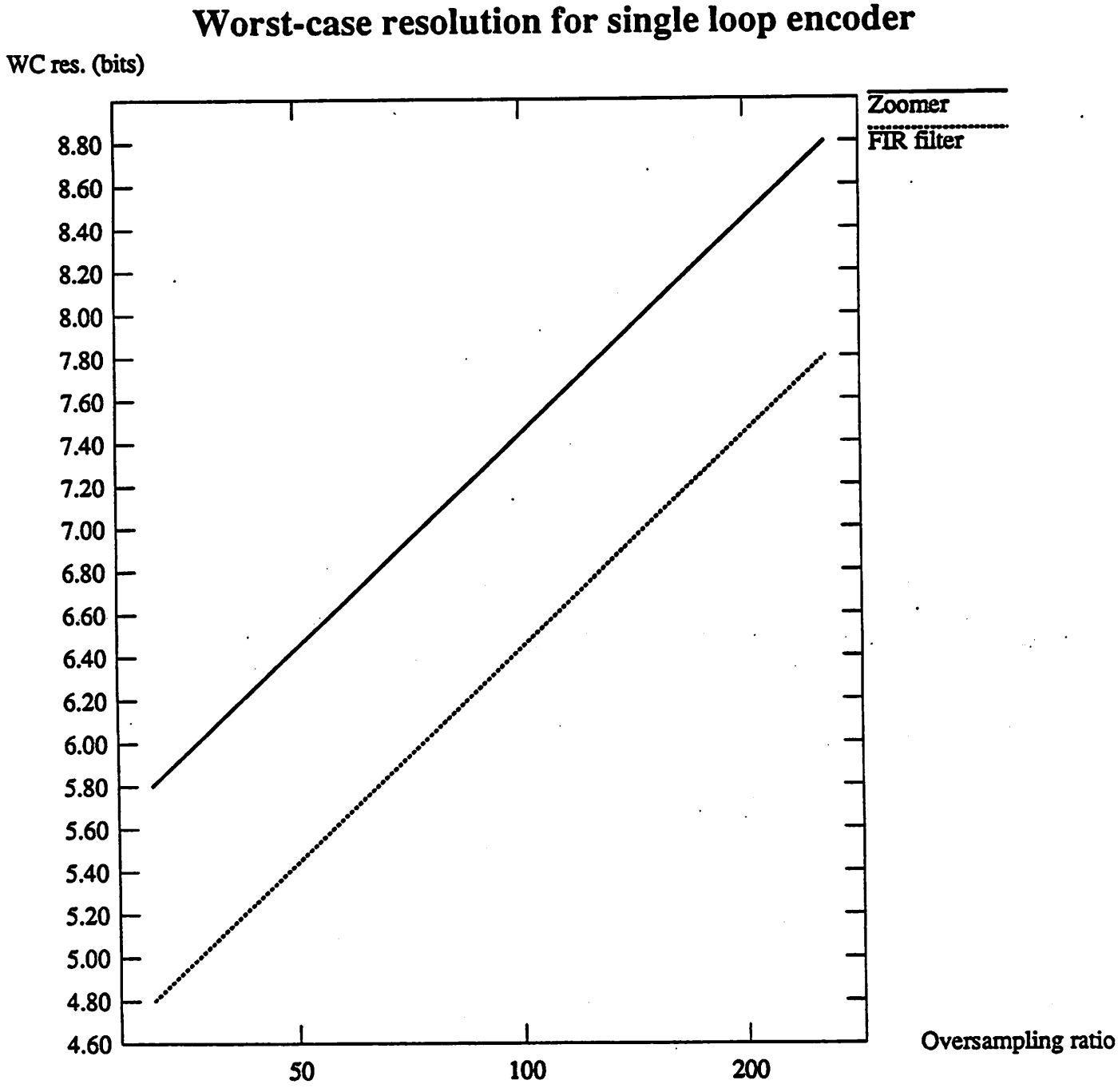

Figure 4: *Worst case resolution in bits as a function of oversampling ratio for the zoomer and the asymptotically optimal FIR filter for single loop decoding.*

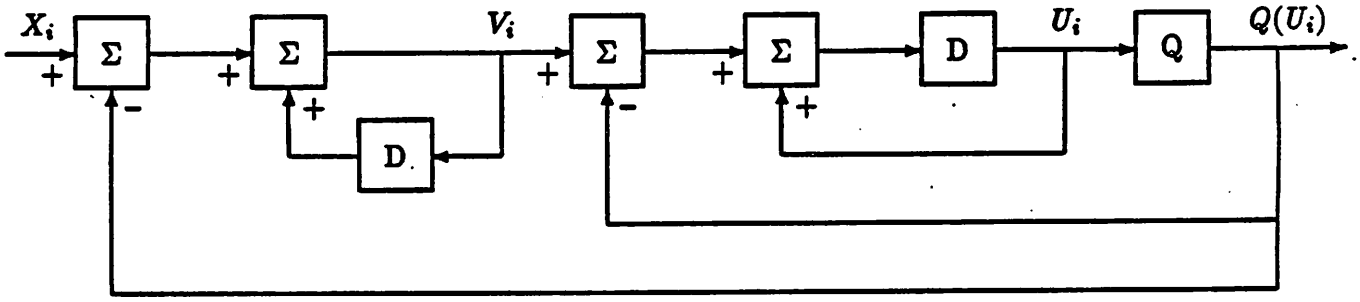

Figure 5: *Double loop ΣΔ encoder.*

A direct and fair comparison of these results to those obtainable with other types of A/D converters is difficult, but for purposes of illustration we briefly consider the popular Dual Slope converter which is cheap and robust. We assume that the required number of clock cycles can reasonably be compared with the oversampling ratio of a ΣΔ modulator.

A Dual Slope converter using M clock cycles has the effect of dividing the dynamic range into $M/2$ intervals of length $4b/M$ each. This leads to $\log_2 \frac{M}{2}$ bits of worst-case resolution, and the MSE equals $4b^2/3M^2$. To match the worst-case error of the single-loop zoomer at $N = 128$, the Dual Slope converter thus requires $M \approx 450$, and to match the MSE, it requires $M \approx 3000$.

3 Double loop modulator

3.1 Theory

The analysis of the double loop encoder proceeds in a fashion similar to the one in subsection 2.1. Figure 5 shows the discrete-time model of the ideal double loop ΣΔ encoder. The inner loop consists of two cascaded discrete integrators, and the quantizer output is fed back two places. The governing difference equations are

$$\begin{aligned} U_n &= U_{n-1} + V_{n-1} - Q(U_{n-1}); \quad U_0 = 0 \\ V_n &= V_{n-1} + X_n - Q(U_n); \quad V_0 = 0 \end{aligned} \tag{14}$$

It can be shown [1] that

$$U_n = b + \sum_{i=1}^{n-1}(n-i)X_i - \sum_{i=1}^{n-1}(n-i+1)Q(U_i), \quad 2 \le n \le N-1 \tag{15}$$

As in subsection 2.1 the two summation terms in (15) can be accounted for as results of open-loop filtering of previous samples. The first summation can be obtained by applying X to the open-loop digital filter resulting when the output feedback paths and the quantizer are deleted, leaving the cascaded integrators. Assuming that the input is constant, $X_i = X$, $1 \le i \le N-1$, this sum equals $n(n-1)X/2$. For the second summation in (15), the pertinent open-loop

filter can be obtained by removing the input and the quantizer. More explicitly, if the second summation is denoted by W_n,

$$W_n = \sum_{i=1}^{n-1}(n-i+1)Q(U_i), \quad n \geq 2; \quad W_1 = 0 \tag{16}$$

we have

$$W_n = W_{n-1} + S_n + Q(U_{n-1}), \quad n \geq 2 \tag{17}$$

where

$$S_n = \sum_{i=1}^{n-1} Q(U_i) \tag{18}$$

The first sum in (15) is greater or less than the second sum, depending on whether $Q(U_n) = +b$ or $-b$. Equation (15) thus leads to the following bounds:

$$X > \bar{X}_n \text{ if } Q(U_n) = +b; \qquad X \leq \bar{X}_n \text{ if } Q(U_n) = -b \tag{19}$$

where

$$\bar{X}_n = \frac{-b + \sum_{i=1}^{n-1}(n-i+1)Q(U_i)}{\frac{1}{2}n(n-1)}, \quad n \geq 2 \tag{20}$$

As $Q(U_0) = -b$, $Q(U_1) = +b$ regardless of the input, the first informative bit is $Q(U_2)$. The zoomer for the double loop encoder is the obvious generalization of the single loop zoomer which updates the bounds L and U according to (19,20) for $2 \leq n \leq N-1$.

To be specific, Figure 6 shows a flowchart for the double-loop zoomer algorithm. The variable S is the cumulative sum of the quantizer outputs given by (18), and W holds the result of the summation (16). P is the denominator in the bound fraction (20).

3.2 Performance comparison

This subsection compares the performance of the double loop zoomer to that of linear decoders. There is no parallel in the literature to the asymptotically optimal FIR filter (6) for single loop modulation. The linear decoder under consideration here is chosen to be the N-tap sinc^3 filter which is believed to be close to optimal. To heuristically support this we mention that for M-stage MASH encoders, it is shown in [6] that sinc^{M+1} decoders perform very well. Further, the optimal FIR filter for the single loop encoder is quite close in shape to a sinc^2 filter.

The performance measures considered are again signal-to-noise ratio and worst-case resolution.

Figures 7 and 8 show computer simulation results for the signal-to-noise ratio and the worst-case resolution. It is seen that the zoomer is far superior to the sinc^3 filter. The SNR and worst-case resolution achieved by the sinc^3 filter at at oversampling ratio of 256 are reached by the zoomer at $N \approx 100$ and 64 respectively. This translates into shorter data acquisition times. At $N = 256$, the zoomer ideally achieves 5 bit better worst-case resolution than the FIR filter. At the same oversampling ratio, the SNR is ideally improved by 38 dB or more than 6 bits.

A tentative comparison with a Dual Slope converter, similar to the one in subsection 2.2, can be made. To match the worst-case resolution of the double-loop zoomer at an oversampling

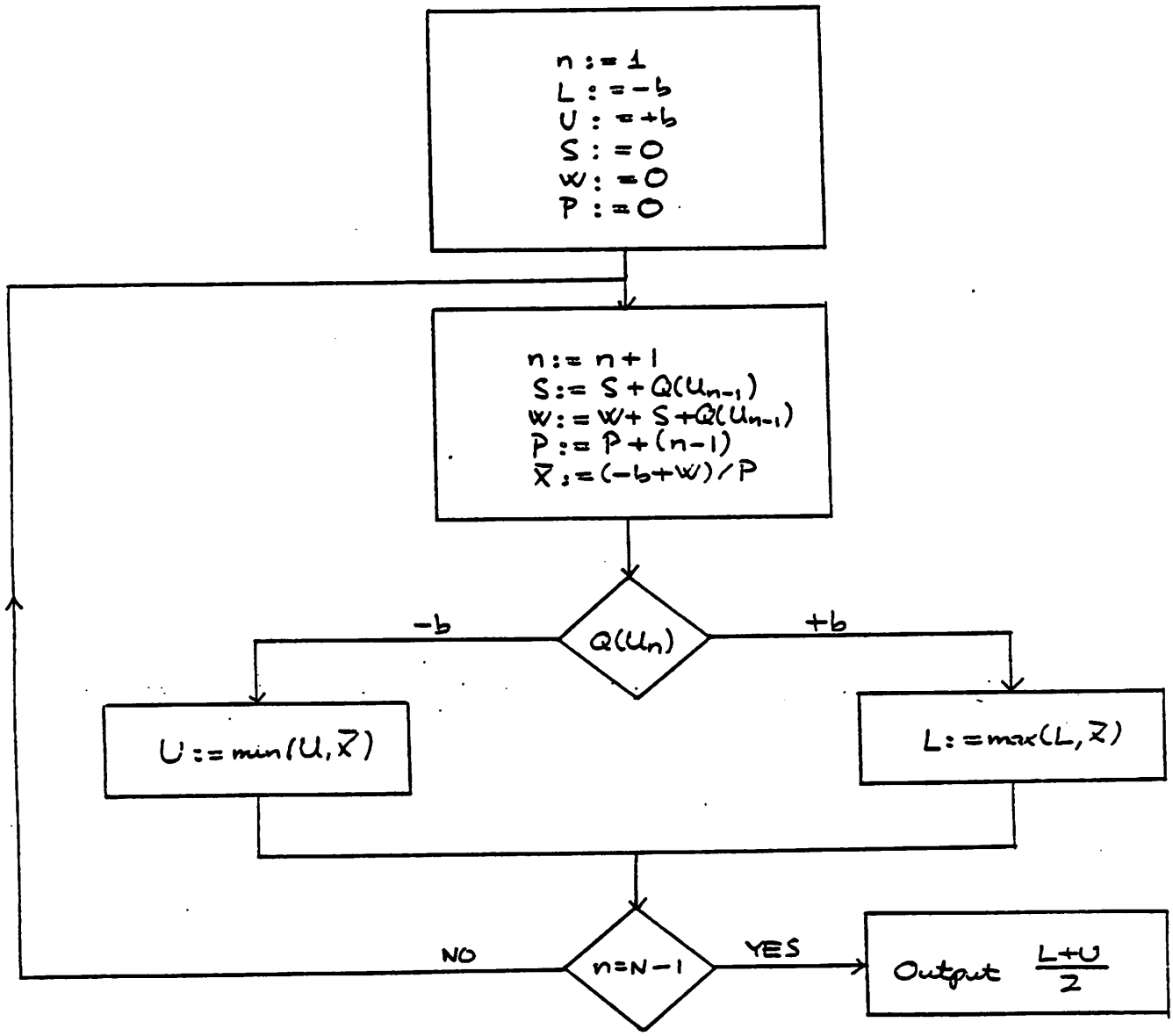

Figure 6: *Flowchart for the double-loop zoomer algorithm.*

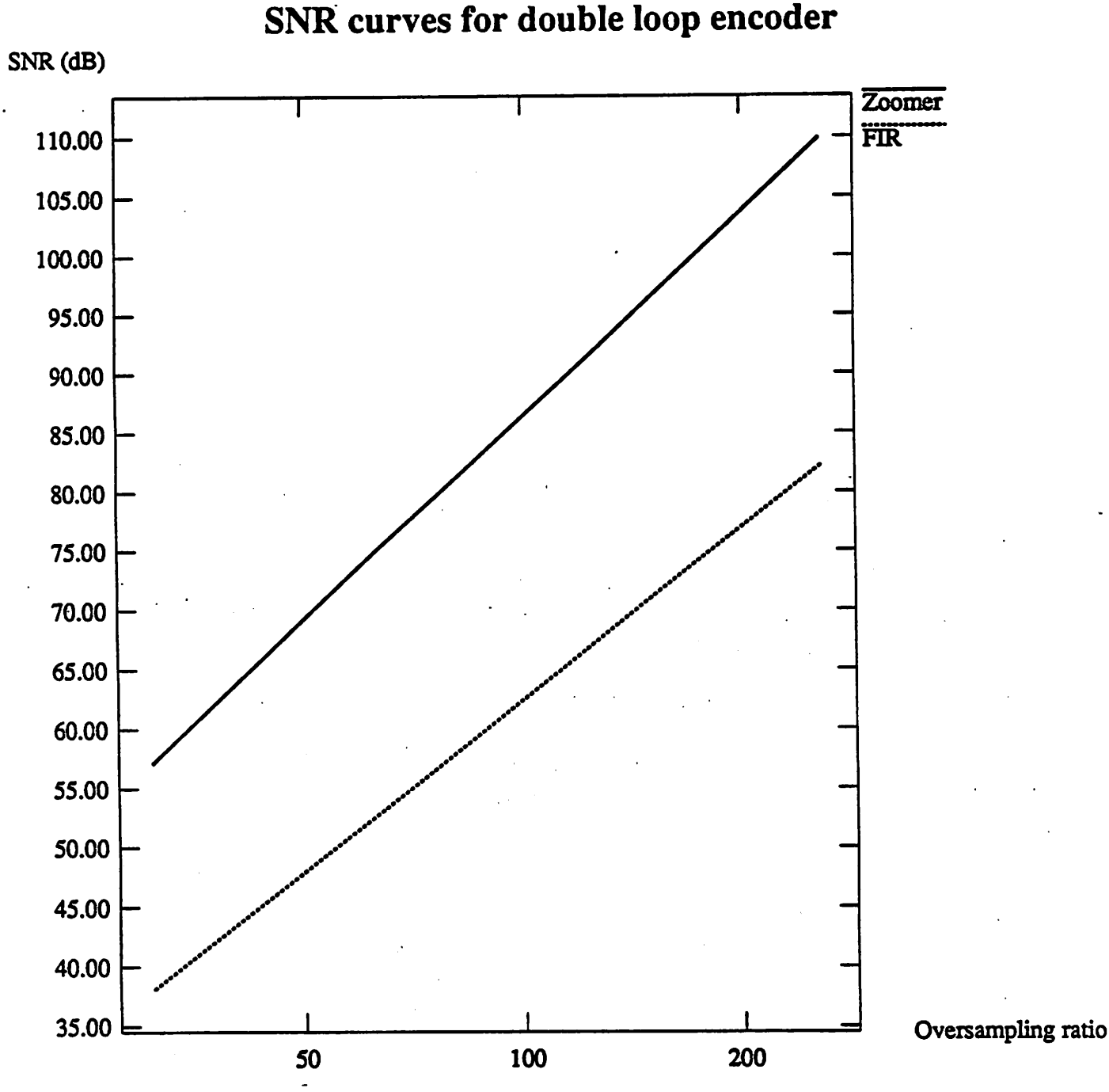

Figure 7: *SNR as a function of oversampling ratio for the zoomer and the asymptotically optimal FIR filter for double loop decoding.*

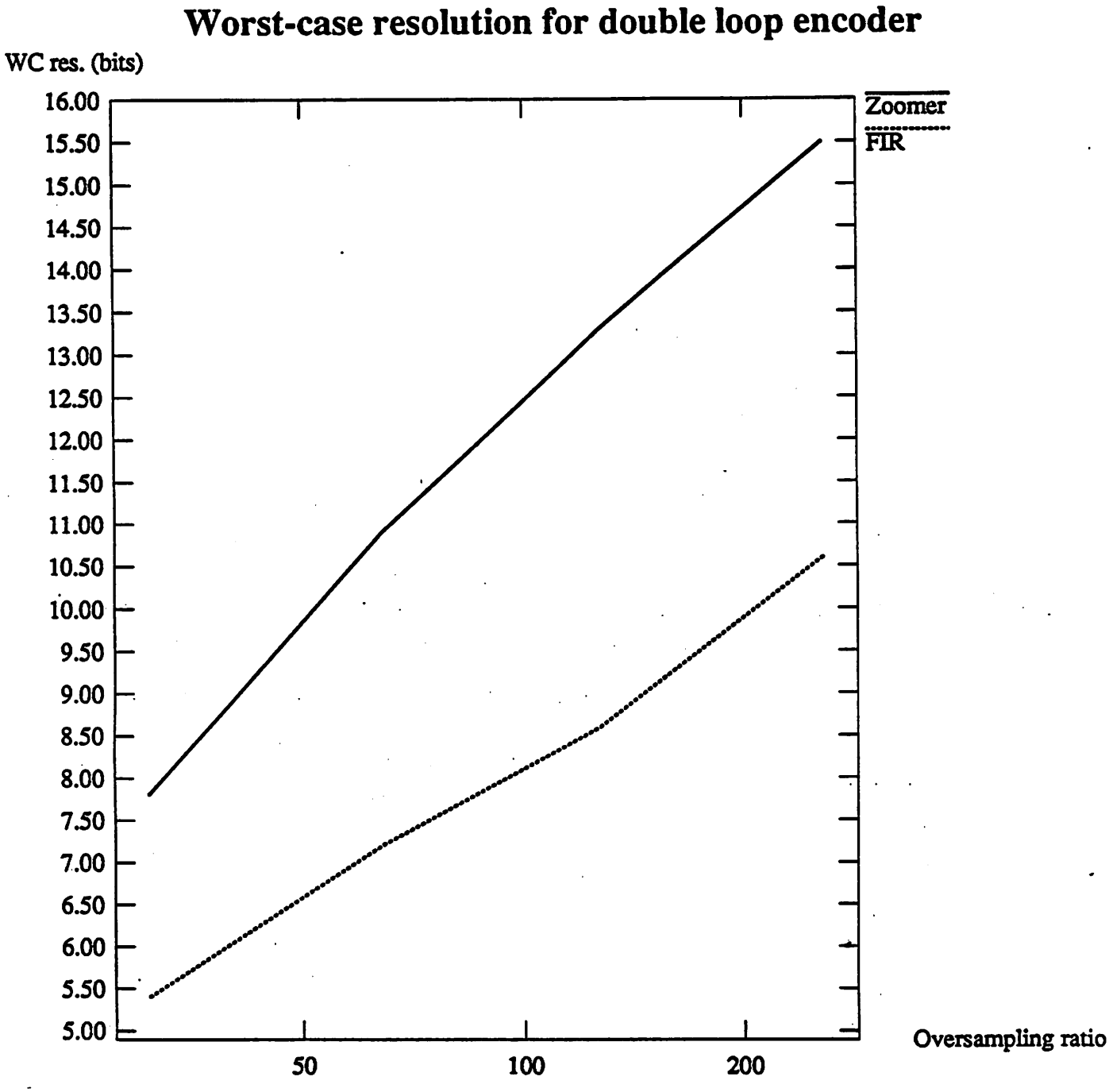

Figure 8: *Worst-case resolution in bits as a function of oversampling ratio for the zoomer and the asymptotically optimal FIR filter for double loop decoding.*

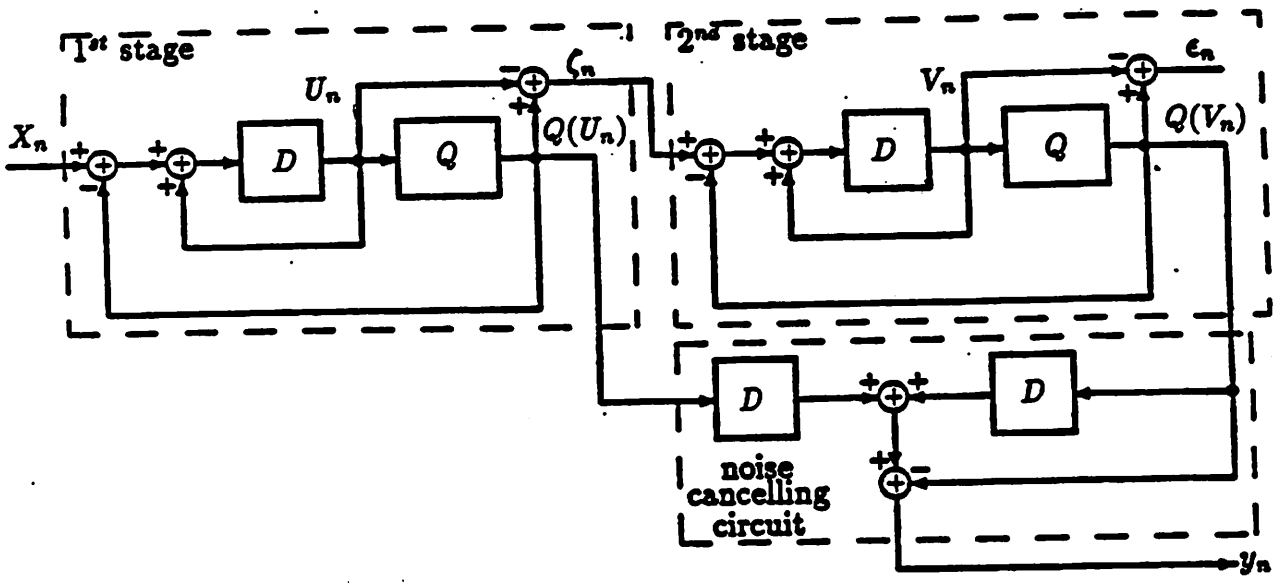

Figure 9: *Two-stage ΣΔ encoder; from [5].*

ratio of $N = 128$, the Dual Slope converter requires $M \approx 20,000$ clock cycles, and to match the MSE performance, about 90,000 clock cycles.

4 Two-stage MASH modulator

4.1 Theory

Figure 9 shows the discrete-time model of the two-stage MASH encoder. The MASH architecture was originated by Uchimura *et al.* [2] and has been extensively analyzed by Wong, Chou and Gray in several papers, including [5] and [6]. The figure is copied from [5] with slightly modified notation. Specifically, we have translated (u, v, q, x, y) in [5] into (U, V, Q, X, Y) for consistency.

The encoder consists of two single loop stages, of which the first is fed with the input, and the second is fed with the quantization error sequence of the first stage. In addition, the figure shows a simple noise cancelling circuit. This has the effect of eliminating the direct appearance of the first stage quantization error in the output sequence $\{Y_n\}$. It should be noted that although this is a desirable characteristic, the circuit might in general be throwing away information present in the separate stage outputs. Here we will adopt the viewpoint that the noise cancelling circuit is really part of a decoder, and the decoder should not be limited to operate on the Y_n sequence obtained by irreversibly collapsing the two output sequences into one. We will therefore work directly with $\{Q(U_i)\}$ and $\{Q(V_i)\}$.

The difference equations governing the two-stage encoder are

$$U_n = U_{n-1} + X_{n-1} - Q(U_{n-1}); \quad U_0 = 0$$

$$V_n = V_{n-1} - U_{n-1} + Q(U_{n-1}) - Q(V_{n-1}); \quad V_0 = 0 \tag{21}$$

These can be solved to yield

$$U_n = \sum_{i=0}^{n-1} X_i - \sum_{i=0}^{n-1} Q(U_i), \quad n \geq 1 \tag{22}$$

$$V_n = -\sum_{i=0}^{n-2}(n-1-i)X_i + \sum_{i=0}^{n-1}(n-i)Q(U_i) - \sum_{i=0}^{n-1} Q(V_i), \quad n \geq 2 \tag{23}$$

In (23), the first summation can be interpreted as the result of filtering the X sequence with a cascade of two integrators. This is also seen directly from Figure 9 by removing the quantizers and their feedback. The third summation in (23) can be explained by removing the first stage and deleting the feedback from the second quantizer. Finally, the second summation in (23) is explained as follows: the first stage integrates the outputs of the first quantizer, and this sequence is fed to a second integrator.

As before we assume that the input is the constant X. At time n, (22) and (23) each provide potential new bounds suitable for a zoomer. Specifically, (22) gives

$$X > \bar{X}_n^{(1)} \text{ if } Q(U_n) = +b; \qquad X \leq \bar{X}_n^{(1)} \text{ if } Q(U_n) = -b; \quad n \geq 1 \tag{24}$$

where

$$\bar{X}_n^{(1)} = \frac{1}{n}\sum_{i=0}^{n-1} Q(U_i) \tag{25}$$

Equation (23) gives

$$X < \bar{X}_n^{(2)} \text{ if } Q(V_n) = +b; \qquad X \geq \bar{X}_n^{(2)} \text{ if } Q(V_n) = -b; \quad n \geq 2 \tag{26}$$

where

$$\bar{X}_n^{(2)} = \frac{\sum_{i=0}^{n-1}(n-i)Q(U_i) - \sum_{i=0}^{n-1} Q(V_i)}{\frac{1}{2}n(n-1)} \tag{27}$$

To be specific, Figure 10 shows a flowchart for the two-stage zoomer algorithm. Variables S and T hold the cumulative sums of the two quantizer outputs $Q(U_n)$ and $Q(V_n)$. W contains the first summation of (27). P is the denominator of (27). $\bar{X}^{(1)}$ and $\bar{X}^{(2)}$ correspond to the quantities (25) and (27).

4.2 Performance comparison

We will compare the MASH zoomer to the N-tap sinc^3 filter. It is shown in [5] that this filter has the same performance dependence on oversampling ratio as the ideal lowpass filter; furthermore, it is stated in [5] that no sinc^M filter, $M > 3$, will achieve a better trade-off with oversampling ratio N.

Figures 11 and 12 show SNR and worst-case resolution as functions of N. It is seen that the zoomer outperforms the sinc^3 filter by 20-30 dB of SNR and 2-3 bits of worst-case resolution. For the depicted range of oversampling ratios, this translates into a reduction by a factor of 2-3 in data acquisition times to achieve a given performance.

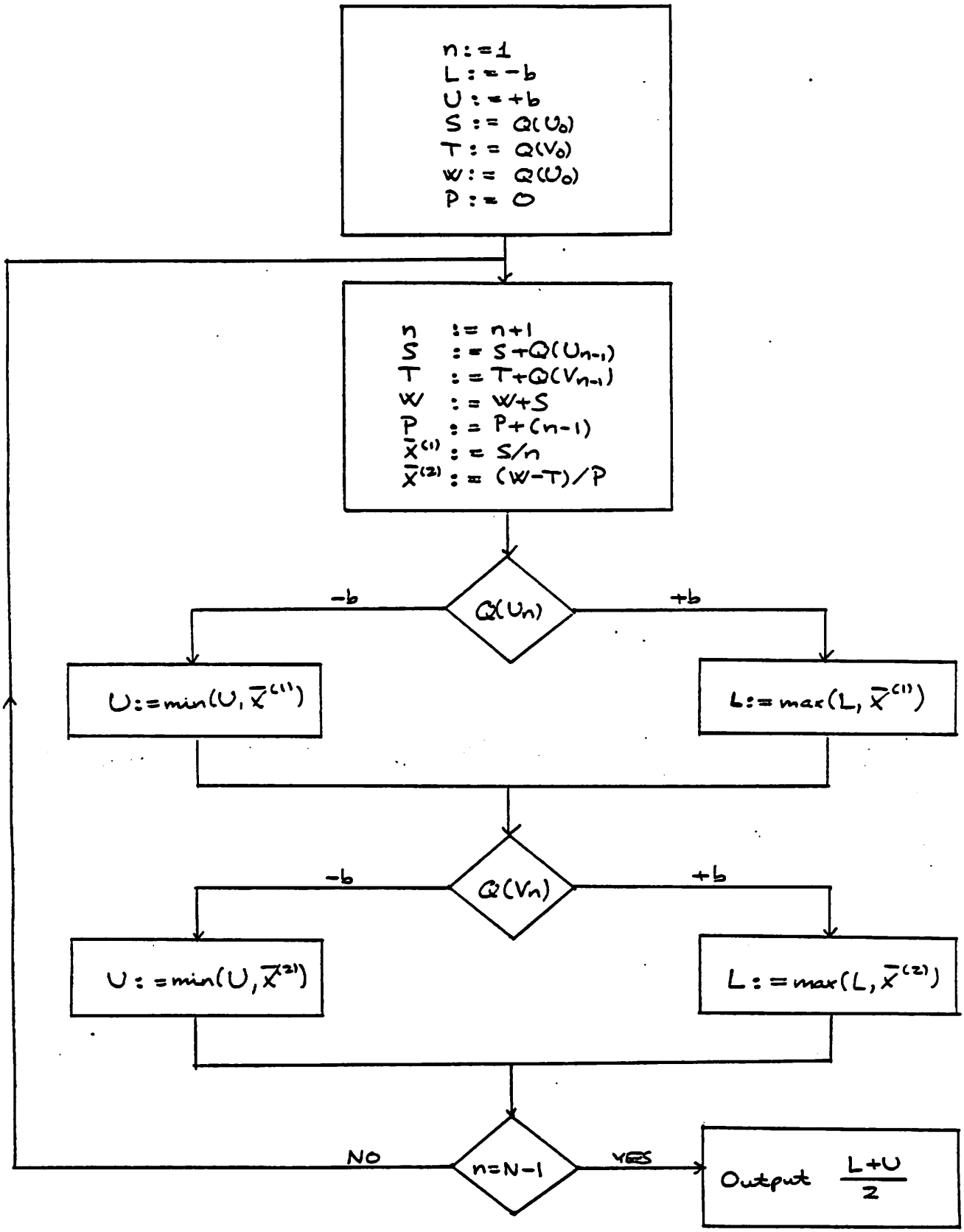

Figure 10: *Flowchart for the two-stage zoomer algorithm.*

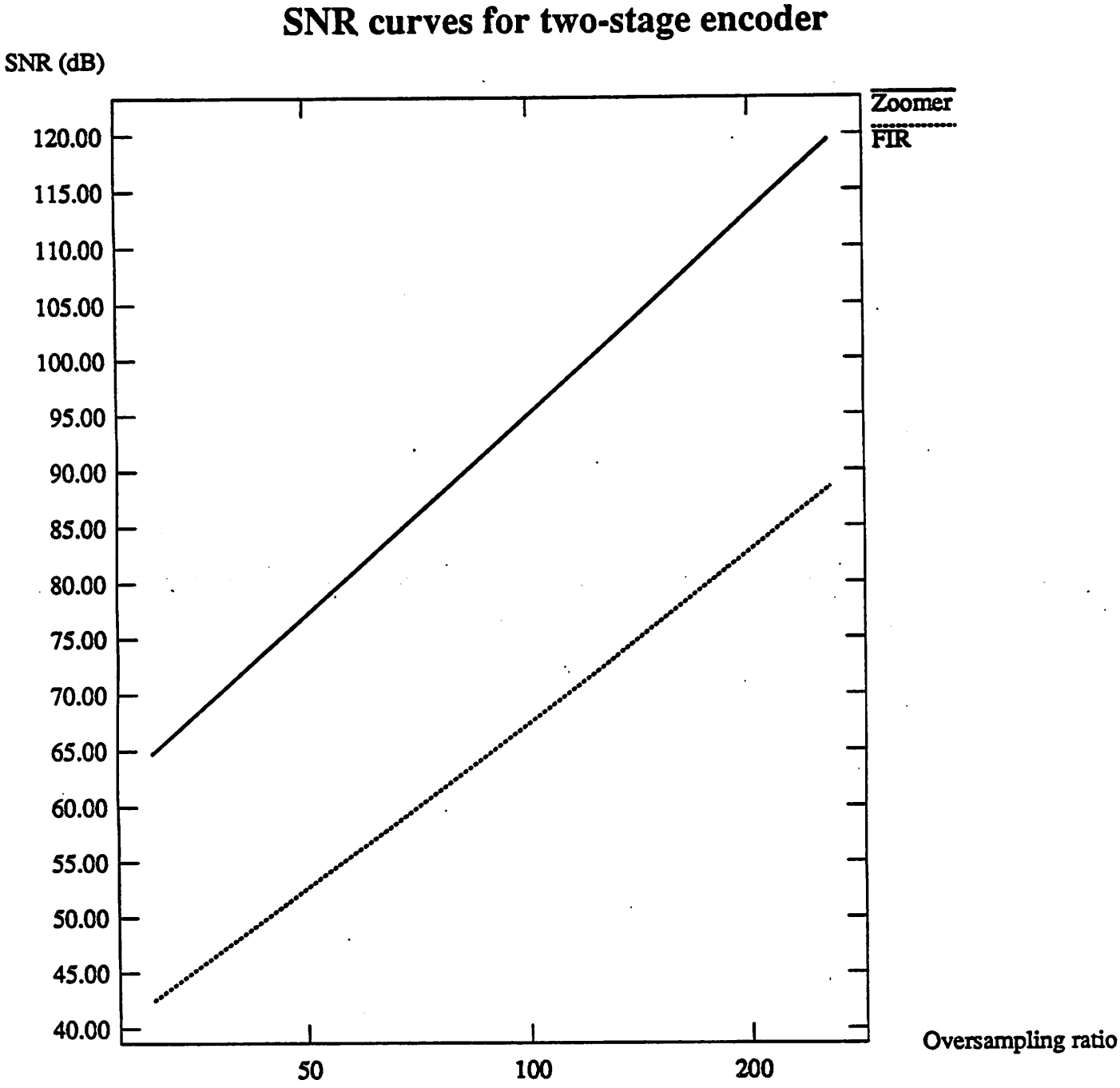

Figure 11: *SNR as a function of oversampling ratio for the zoomer and the sinc^3 filter for two-stage decoding.*

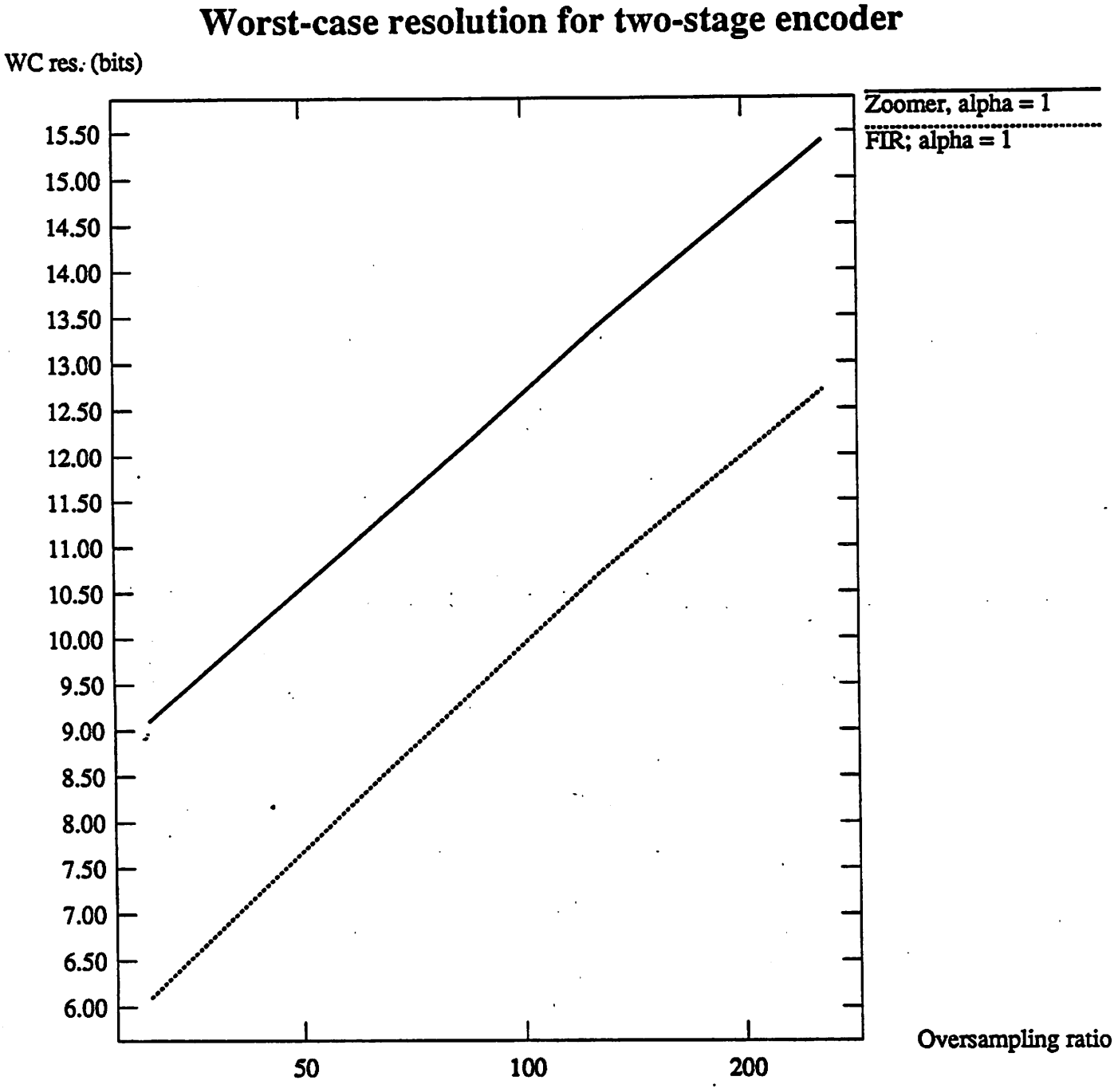

Figure 12: *Worst-case resolution in bits as a function of oversampling ratio for the zoomer and the sinc^3 filter for two-stage decoding.*

5 Conclusions

We have introduced a general technique for optimal decoding of the output of ideal $\Sigma\Delta$ modulators, under the assumptions of constant input and known initial integrator states. The technique is based on deriving a succession of upper and lower bounds on the input interval. The bounds are given as fractions. The numerator of a bound fraction is the result of filtering a quantizer output sequence up to some time n with a filter closely related to the open-loop linear part of the encoder. The denominator is the result of passing an all-1 sequence through another filter easily derivable from the encoder.

The optimal decoder is highly nonlinear, as might be expected from the nonlinear nature of the encoder. Our results indicate that under ideal circumstances, substantial reductions in MSE and worst-case error can be achieved. This translates into substantial reductions in data acquisition times.

References

[1] S. Hein and A. Zakhor, "Lower Bounds on the MSE of the Single and Double Loop Sigma Delta Modulators", *Proc. Int. Conf. Circuits and Systems*, May 1990.

[2] 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. Acoustics, Speech and Signal Proc.*, vol. 36 no. 12, pp. 1899-1905, Dec. 1988.

[3] 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*, Submitted for publication, Massachusetts Institute of Technology, 21 July 1989.

[4] R. M. Gray, "Spectral Analysis of Quantization Noise in a Single-Loop Sigma-Delta Modulator with dc Input", *IEEE Trans. Comm.*, vol. 37 no. 6, pp. 588-599, June 1989.

[5] P. W. Wong and R. M. Gray, "Two-stage Sigma-Delta Modulation", Submitted for publication, Stanford University, 17 May 1989.

[6] W. Chou, P. W. Wong and R. M. Gray, "Multistage Sigma-Delta Modulation", *IEEE Trans. Info. Theory*, vol. 35 no. 4, pp. 784-796, July 1989.

PERFORMANCE-CONSTRAINED PHYSICAL DESIGN OF ANALOG AND MIXED ANALOG/DIGITAL CIRCUITS

by

Umakanta Choudhury

Memorandum No. UCB/ERL M92/38

16 April 1992