An Analysis of Quantization Noise in $\Delta\Sigma$ Modulation and its Application to Parallel $\Delta\Sigma$ Modulation

Thesis by
Ian Galton

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1992

(Defended April 3, 1992)

Acknowledgment

In many ways, this acknowledgment is the hardest part of the thesis to write. As time goes by my work will become dated, but what I have gained from my family, friends, and colleagues will remain. Largely by the examples and efforts of those around me, I have grown intellectually and as a person.

I have been extremely fortunate to have had Edward Posner for my advisor. In addition to having extraordinary insights into both science and people, he is one of the kindest people I know. Although one of the most productive members of the Caltech faculty, he seems constantly to be going out of his way for the benefit of those around him. He generously provided me with the resources necessary for my work yet gave me true academic freedom. He not only permitted, but encouraged me to control the direction of my work. Nevertheless, his door was always open to me and I greatly enjoyed his astute perceptions and wisdom. It is unlikely that I shall ever again be quite so free in the pursuit of my research.

Professor Joel Franklin has also profoundly enriched my experience at Caltech. At a time when I was thinking of returning to industry without pursuing my PhD, he steered me toward working for Edward Posner. Without his encouragement and guidance, it is likely that I would not have stayed at Caltech. He has been a great teacher, advisor, and friend to me.

Similarly, I have benefited greatly from my association with Professor P. P. Vaidyanathan. Throughout my years at Caltech, he has taught me both in and out of the classroom. By his actions, he has set many standards that I will try to follow in my own career. If I can one day teach and perform research at anywhere near his level of quality, then I shall be in very good shape indeed.

I am also grateful to John Miller and Bhusan Gupta. From the beginning, John

has patiently listened to every baked and half-baked research idea that I have come up with on the way to this thesis. He has often saved me from wasting time on bad ideas by pointing out flaws, yet has always been a source of encouragement. Bhusan, has similarly provided me with excellent feedback regarding my work. He has implemented a VLSI prototype based on the work in this thesis and has taught me much about the field of analog IC design.

Most of all, I am grateful to my wife, Kerry. Although she has often wondered what I find so interesting about my work, she has never questioned my desire to pursue it. She has been a constant source of encouragement and support. She has shared my frustrations as if they were her own and has been the champion of my successes. Without her, my stay at Caltech would have been considerably less enjoyable. More importantly, the happiness that she and my baby daughter, Riley, give me puts everything that I do into perspective.

Preface

This thesis is a collection of three papers that are intended for separate publication in various *IEEE* journals. Although the papers are self-contained, they are closely related in that they all deal with a class of analog-to-digital conversion systems known as $\Delta\Sigma$ modulators. The first two papers provide general analyses of existing $\Delta\Sigma$ modulator architectures and the third paper applies the results to develop a new A/D converter architecture consisting of $\Delta\Sigma$ modulators that operate in parallel.

The papers assume some knowledge of $\Delta\Sigma$ modulators. All of the necessary information can be found in *Oversampling Delta-Sigma Data Converters Theory, Design and Simulation*, Edited by J.C. Candy, G.C. Temes, New York, IEEE Press, 1992. In particular, the tutorial introduction paper makes for easy reading and presents most of the necessary background.

Abstract

The trend toward digital signal processing in communication systems has resulted in a large demand for fast, accurate analog-to-digital (A/D) converters, and advances in VLSI technology have made $\Delta\Sigma$ modulator based A/D converters attractive solutions. However, because they are non-linear systems, they have proven difficult to analyze. Rigorous analyses have been previously performed only for a small number of artificial input sequences and then only for the simplest of $\Delta\Sigma$ modulator architectures. This thesis consists of three self-contained papers addressing these and related problems. The first two papers extend the repertoire of tractable input sequences for most of the known $\Delta\Sigma$ modulator architectures. The third paper applies the results from the first two papers to develop a scalable architecture for parallel $\Delta\Sigma$ Modulation.

The first paper concentrates on the first-order $\Delta\Sigma$ modulator and develops rigorous results for a large class of input sequences. Under the assumptions that some circuit noise is present and that the input sequence does not cause overload, a simple autocorrelation expression is developed that is only locally dependent upon the input sequence. Ergodic properties are derived and various examples are presented.

In the second paper, a rigorous analysis of the granular quantization noise in a general class of $\Delta\Sigma$ modulators is developed. Again under the assumption that some circuit noise is present, the joint statistics of the granular quantization noise sequences are determined and the sequences are shown to be correlation ergodic. The exact results developed for the granular quantization noise are shown to approximately hold for the overall quantization noise if the quantizers in the $\Delta\Sigma$ modulator overload occasionally.

The third paper develops a scalable A/D converter architecture consisting of

multiple $\Delta\Sigma$ modulators. By combining M $\Delta\Sigma$ modulator based A/D converters, each with an oversampling ratio of N, an effective oversampling ratio of approximately NM is achieved with only an M-fold increase in the quantization noise power. In particular, the special case of $N = 1$ allows for full-rate analog to digital conversion. Unlike most other approaches to trading modulator complexity for accuracy, the system retains the robustness of the individual $\Delta\Sigma$ modulators to circuit imperfections.

Contents

Granular Quantization Noise in the First-Order $\Delta\Sigma$ Modulator

Ian Galton

Abstract— $\Delta\Sigma$ modulators are attractive candidates for oversampling analog-to-digital (A/D) converters because they are amenable to VLSI implementation and have low component sensitivity. However, because they are nonlinear systems, they have proven difficult to analyze. Rigorous analyses have been performed only for a small number of artificial input sequences such as constant, sinusoidal, and Gaussian white noise input sequences [1]-[5]. By allowing for the inevitable presence of small amounts of noise in the $\Delta\Sigma$ modulator circuitry, a general framework is developed which extends the repertoire of tractable input sequences to general stochastic sequences in addition to handling many input sequences for which results have been previously presented. Under the assumptions that some circuit noise is present and that the input sequence does not cause overload, a simple autocorrelation expression is developed that is only locally dependent upon the input sequence. Ergodic properties are derived and various examples are presented.

I. Introduction

The first-order $\Delta\Sigma$ modulator [6] is the simplest of a class of systems generally referred to as $\Delta\Sigma$ modulators that employ sampled-data filters and coarse quantizers within feedback loops. They are widely used in high-precision oversampling A/D converters because they are well-suited to VLSI implementation and tend to be robust with respect to nonideal components. Accordingly they have received much attention from both academic and industrial researchers. Nevertheless, most of the previously published rigorous theoretical analyses of $\Delta\Sigma$ modulators apply only to a small set of input sequences. In the current work, we concentrate on the first-order

The author is with the Electrical Engineering Department, California Institute of Technology, 116-81, Pasadena, CA 91125; email address: galton@systems.caltech.edu

This work was supported by a grant from Pacific Bell.

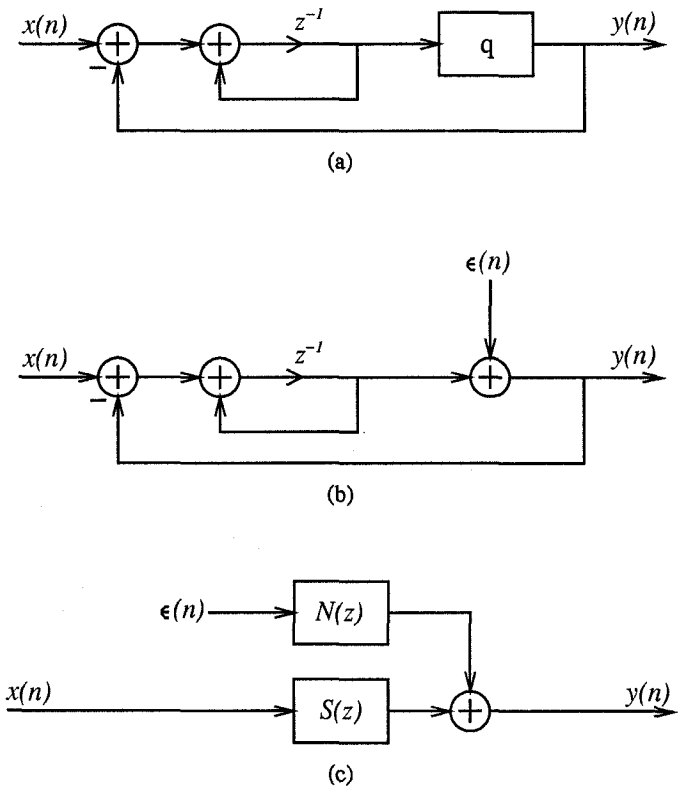


Figure 1: a) The first order $\Delta\Sigma$ modulator. b) An equivalent form of the system with the quantizer represented as an additive quantization noise source. c) An equivalent form of the system showing the different filters that act on the input and quantization noise sequences, respectively.

$\Delta\Sigma$ modulator and provide rigorous results for a large class of input sequences.

The first-order $\Delta\Sigma$ modulator consists of a sampled-data integrator, a uniform midrise quantizer [7], and a negative feedback loop surrounding the integrator and quantizer as shown in Figure 1a. The system operates on a sampled-data input, $x(n)$, and produces a quantized output, $y(n)$. The quantizer can be interpreted as an additive quantization noise source as depicted in Figure 1b. A straightforward linear systems analysis shows that the input sequence sees the one-sample delay $S(z) = z^{-1}$ while the quantization noise sequence sees the highpass filter $N(z) = 1 - z^{-1}$. Thus, as shown in Figure 1c, the output consists of two components: a component corresponding to the input sequence and a component corresponding to the quantization noise sequence.

Note that $N(z)$ is a highpass filter with a zero at zero frequency. This causes the

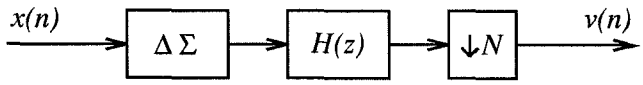


Figure 2: A $\Delta\Sigma$ modulator based oversampling A/D converter.

spectral energy of the quantization error at the output of the $\Delta\Sigma$ modulator to be weighted toward the high frequency end of the spectrum for most input sequences [8]. It is this property of the $\Delta\Sigma$ modulator that makes it useful in oversampling A/D converters.

An oversampling A/D converter consists of a $\Delta\Sigma$ modulator followed by a lowpass decimation filter as shown in Figure 2. The input to the $\Delta\Sigma$ modulator, $x(n)$, is obtained by sampling a bandlimited analog signal at a rate Nf where N is a positive integer and f is the Nyquist rate. Therefore, the spectrum of $x(n)$ is nonzero only on $(-\frac{\pi}{N}, \frac{\pi}{N})$ where 2π corresponds to the sampling rate. Provided N is sufficiently large, the spectral energy of the quantization error will fall mostly outside of $(-\frac{\pi}{N}, \frac{\pi}{N})$. The lowpass filter, removes the out-of-band quantization noise and the decimator reduces the output sequence to the Nyquist rate.

Although conceptually simple, the system has proven difficult to analyze because of the nonlinearity introduced by the quantizer. As will be shown, the quantization noise has a complicated structure that is globally dependent upon the input sequence. If two input sequences differ at just one sample time, say $n = n_0$, then the corresponding quantization noise sequences will appear very different for all $n > n_0$.

The quantizer imposes the following nonlinearity on its input:

$$\mathrm{q}(x) = \begin{cases} \Delta\left\lfloor \frac{x}{\Delta} \right\rfloor + \frac{\Delta}{2} & \text{if } -\gamma \leq x < \gamma; \\ \gamma - \frac{\Delta}{2} & \text{if } x \geq \gamma; \\ -\gamma + \frac{\Delta}{2} & \text{if } x < -\gamma; \end{cases} \tag{1}$$

where γ is usually an integer multiple of Δ. When the input to the quantizer has absolute value greater than γ the quantizer is said to *overload*. It is desirable to

avoid the overload condition because the resulting distortion tends to be severe and difficult to characterize [1]-[5], [9], [10]. Most of the existing $\Delta\Sigma$ modulator analyses including ours assume that the overload condition is avoided. Since the quantizer will not overload provided the $\Delta\Sigma$ modulator input sequence is bounded in absolute value by $\gamma - \frac{\Delta}{2}$ [5], this is not an unreasonable assumption. Furthermore, simulations show that if the overload condition occurs but does so only rarely then the performance of the $\Delta\Sigma$ modulator is not significantly degraded [11]. We can therefore expect any exact results obtained under the no-overload assumption to approximately hold if the overload condition has a low frequency of occurrence.

Even under the no-overload restriction, the system does not yield to a straightforward analysis. Most analyses rely on approximations [6], [8], [12], or apply only to specific input sequences such as constant [1],[2], sinusoidal [3], or Gaussian white noise sequences [4].

In the current work, we develop rigorous results by assuming that the input sequence contains an additive independent identically distributed (iid) random component. The assumption is not very restrictive because the random component can have an arbitrarily small variance. Moreover, since thermal noise in the analog front-end of the $\Delta\Sigma$ modulator can be modeled as an iid random sequence, the assumption is reasonable in practice. The approach has the benefit that it can be applied to a large class of input sequences. We develop a simple expression for the autocorrelation of the quantization error, $R_{ee}(n,p)$, and prove that the error is correlation ergodic. The autocorrelation expression is convenient because it is only locally dependent on the input sequence. This property makes tractable many desired input sequences that can not be handled using previously presented theory such as the class of arbitrary stochastic sequences respecting the no-overload constraint.

In Section II, we derive the theory outlined above and in Section III we apply

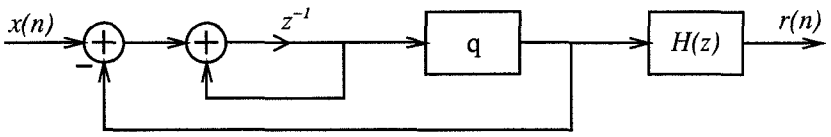


Figure 3: A first-order $\Delta\Sigma$ modulator followed by the filter $H(z)$.

it to specific input sequences. By considering constant and sinusoidal inputs, the theory is shown to contain many of the existing results concerning the first-order $\Delta\Sigma$ modulator as special cases, although new observations are also presented. In particular, for a sinusoidal input we develop a closed-form expression for the quasi-stationary autocorrelation of the quantization error. Additional classes of sequences, which heretofore have not been rigorously analyzed in conjunction with the $\Delta\Sigma$ modulator, are then considered. We also present simulation results to support our theoretical analysis.

II. Theoretical Analysis

Instead of considering the $\Delta\Sigma$ modulator in isolation, our system of study will be the $\Delta\Sigma$ modulator followed by a causal, stable, linear time-invariant digital filter with transfer function $H(z)$ and impulse response $h(n)$ as shown in Figure 3. The reason for not considering the $\Delta\Sigma$ modulator in isolation is that in practice it is almost always followed by a filter and, as we will show, the statistics of the output are dependent upon the filter. Since we could choose $H(z) = 1$, the isolated $\Delta\Sigma$ modulator is a special case of our system

We will distinguish between the *quantization noise sequence*, $\epsilon(n)$, and the *quantization error sequence*, $e(n)$. As shown in Figure 1b, the quantization noise sequence is the difference between the output and the input of the quantizer. It is the noise injected into the system by the quantizer. The quantization error sequence is the component of the output of the system in Figure 3 corresponding to the quantization noise. As mentioned above, the $\Delta\Sigma$ modulator subjects the quantization

noise sequence to the filter $N(z) = 1 - z^{-1}$. Thus the quantization error sequence is equivalent to the output of the filter $(1 - z^{-1})H(z)$ when driven by the quantization noise sequence. From the argument leading to Figure 1c, it follows that we can write the output of the system in Figure 3 as

$$r(n) = w(n) + e(n), \tag{2}$$

where $w(n)$ can be interpreted as the response of the filter $z^{-1}H(z)$ to the input sequence, $x(n)$.

As alluded to above, we will assume that the input sequence seen by the $\Delta\Sigma$ modulator consists of a *desired input sequence*, $x_d(n)$, plus an *input noise sequence*, $\{\eta_n\}$:

$$x(n) = x_d(n) + \eta_n. \tag{3}$$

We require that the η_n are independent and identically distributed with a distribution that has a density. The desired input sequence is the sampled-data signal that is to be converted into a digital sequence by the $\Delta\Sigma$ modulator (e.g., the music signal, the video signal, etc.), and the input noise sequence is an unrelated sequence that is assumed to be present in the analog front-end of the $\Delta\Sigma$ modulator. The assumption is realistic in practice because thermal noise which is ubiquitous in analog circuitry can be accurately modeled as an iid random sequence in sampled data systems.

An Expression For The Quantization Error Sequence

In the calculations to follow, we will consider the $\Delta\Sigma$ modulator to have been "turned on" at a specific time in the past. For all $n \leq a$ we will take the input sequence and all storage elements in the $\Delta\Sigma$ modulator and filter to be zero. In some cases we will consider the system in the limit as $a \to -\infty$. This corresponds to a system that has been running since the beginning of time.

Gray [5] has shown that the quantization noise sequence can be written as

$$\epsilon(n) = \frac{\Delta}{2} - \Delta\left\langle \tfrac{1}{\Delta} \sum_{i=1}^{n-a} \left[x(n-i) + \tfrac{\Delta}{2}\right] \right\rangle \tag{4}$$

provided $n > a$.† For convenience, we will take $\epsilon(n) = 0$ whenever $n \leq a$.

Since $H(z)$ is causal, its impulse response, $h(n)$, is zero for all $n < 0$. Therefore, for $n > a$ we can write the quantization error sequence as

$$e(n) = \sum_{k=0}^{\infty} \left[h(k) - h(k-1)\right] \epsilon(n-k). \tag{5}$$

Again, for convenience, we will take $e(n) = 0$ whenever $n \leq a$. Combining these two equations gives

$$e(n) = \Delta \sum_{k=0}^{n-a-1} \left[h(k) - h(k-1)\right] \left\langle \tfrac{1}{\Delta} \sum_{i=1}^{n-k-a} \left[x(n-k-i) + \tfrac{\Delta}{2}\right] \right\rangle. \tag{6}$$

Although (6) is an exact formula for the quantization error sequence, it does not give great insight into the long-term behavior of the quantization error sequence. In particular, the specific quantization error sequence obtained for a given input sequence is globally dependent upon each value of the input sequence. For example, consider two input sequences, $x_1(n)$ and $x_2(n)$, which are identical except for their first value at time $n = a$. That is, suppose

$$x_2(n) = \begin{cases} x_1(n), & \text{if } n \neq a; \\ x_1(n) + \beta, & \text{if } n = a; \end{cases}$$

for some nonzero $\beta \in \mathbf{R}$ (such that the no-overload condition is maintained). Then the quantization error sequence associated with $x_1(n)$ is

$$e_1(n) = \Delta \sum_{k=0}^{n-a-1} \left[h(k) - h(k-1)\right] \left\langle \tfrac{1}{\Delta} \sum_{i=1}^{n-k-a} \left[x_1(n-k-i) + \tfrac{\Delta}{2}\right] \right\rangle,$$

while the quantization error sequence associated with $x_2(n)$ is

$$e_2(n) = \Delta \sum_{k=0}^{n-a-1} \left[h(k) - h(k-1)\right] \left\langle \beta \tfrac{1}{\Delta} + \tfrac{1}{\Delta} \sum_{i=1}^{n-k-a} \left[x_1(n-k-i) + \tfrac{\Delta}{2}\right] \right\rangle.$$

† The angle brackets denote the fractional part operator. This operator is defined as: $\langle x \rangle = x - \lfloor x \rfloor$ for all $x \in \mathbf{R}$.

Because of the presence of β, each term of the first sum in the equation for $e_2(n)$ differs in a complicated fashion from the corresponding term in $e_1(n)$; the two quantization error sequences typically look very different.

Quantization Noise Statistics

Chou and Gray [13] have investigated the statistics of the quantization noise sequence of the first-order $\Delta\Sigma$ modulator under the assumptions that overload is avoided and that the input sequence consists of a deterministic sequence plus a a so-called *dither sequence* that is iid with a density. Mathematically the dither sequence assumption is equivalent to (3); our input noise sequence plays the role of the dither sequence. The reason that we do not refer to the input noise sequence as a dither sequence is that the term dither is usually applied to sequences that are intentionally introduced. From a practical point of view, we are making the opposite assumption. We consider the presence of the input noise sequence to be an inevitable result of the $\Delta\Sigma$ modulator having an analog front-end. The practical consequence of our distinction is that the results presented in this paper hold regardless of whether a dither sequence is intentionally added. Nevertheless, the results presented by Chou and Gray can be applied directly to our system.

In particular, they proved that the quantization noise sequence converges in distribution to a random variable that is uniformly distributed on $(-\frac{\Delta}{2}, \frac{\Delta}{2}]$ and is independent of the desired input sequence. In order to extend their work, it is convenient to begin by stating this result in a slightly different form. We do this in the following lemma.

Lemma 1: For each $r = 1, 2, \ldots$, let

$$U_r = \left\langle \mu_r + \sum_{i=1}^{r} c\eta_i \right\rangle,$$

where $\{\mu_r\}$ is any deterministic sequence, c is any real number, and $\{\eta_i\}$ is a se-

quence of independent, identically distributed random variables whose distribution has a density. Then as $r \to \infty$, U_r converges in distribution to a random variable U that is uniformly distributed on $[0,1)$.

Proof: The proof is essentially the same as that presented in [13].

■

The following lemma generalizes this result to stochastic sequences $\{\mu_r\}$.

Lemma 2: Let $\{U_r\}$, c, and $\{\eta_i\}$ be as defined in the hypothesis of Lemma 1. Let $\{\mu_r\}$ be any stochastic sequence that is independent of $\{\eta_i\}$. Then as $r \to \infty$, U_r converges in distribution to a random variable U that is uniformly distributed on $[0,1)$.

Proof: The moments of U_r are defined as $\mathrm{E}(U_r^n)$, for $n = 1, 2, \ldots$. Because the support of U_r is restricted to $[0,1)$, each moment exists and has absolute value less than or equal to one. Thus, the distribution of U_r is uniquely determined by its moments and it is sufficient to show that the moments of U_r converge to the corresponding moments of U as $r \to \infty$.†

Since the sequences $\{\eta_k\}$ and $\{\mu_k\}$ are independent, for any integer n we can write

$$\mathrm{E}(U_r^n) = \mathrm{E}\Big[\mathop{\mathrm{E}}_{\{\eta_k\}}(U_r^n)\Big].$$

By Lemma 1,

$$\mathop{\mathrm{E}}_{\{\eta_k\}}(U_r^n) \to \mathrm{E}(U^n) \tag{7}$$

as $r \to \infty$. However, in order to be sure that

$$\mathrm{E}(U_r^n) \to \mathrm{E}(U^n)$$

as $r \to \infty$ we must show that the convergence in (7) is uniform with respect to the variables $\{\mu_k\}$.

† See, for example, Theorems 30.1 and 30.2 in [14].

Note that U_r only depends on μ_r, the r^{th} element in the sequence $\{\mu_k\}$. Moreover, $\underset{\{\eta_k\}}{\mathrm{E}}(U_r^n)$ is a continuous function of μ_r. Therefore, $\sup\limits_{\{\mu_k\}} \underset{\{\eta_k\}}{\mathrm{E}}(U_r^n)$ is achieved by some value of μ_r. It follows that there exists a sequence $\{\mu_k'\}$ such that

$$\sup_{\{\mu_k\}} \left| \underset{\{\eta_k\}}{\mathrm{E}}(U_r^n) - \mathrm{E}(U^n) \right| = \left| \underset{\{\eta_k\}}{\mathrm{E}}(U_r^n) \Big|_{\mu_r=\mu_r'} - \mathrm{E}(U^n) \right|$$

Since the convergence in Lemma 1 holds for any deterministic sequence $\{\mu_k\}$, it must hold for the particular sequence $\{\mu_k'\}$. Thus

$$\lim_{r\to\infty} \sup_{\{\mu_k\}} \left| \underset{\{\eta_k\}}{\mathrm{E}}(U_r^n) - \mathrm{E}(U^n) \right| = 0$$

which implies that the convergence in (7) is uniform with respect to $\{\mu_k\}$.

■

In accordance with the usual definitions, we will take the mean and autocorrelation of the quantization noise sequence to be

$$M_\epsilon(n) = \lim_{a\to-\infty} \mathrm{E}\big[\epsilon(n)\big],$$

and

$$R_{\epsilon\epsilon}(n,p) = \lim_{a\to-\infty} \mathrm{E}\big[\epsilon(n)\epsilon(n+p)\big],$$

respectively. We will take the cross correlation of the quantization noise sequence and the desired input sequence to be

$$R_{x_d\epsilon}(n,p) = \lim_{a\to-\infty} \mathrm{E}\big[x_d(n)\epsilon(n+p)\big].$$

We will take the mean, autocorrelation, and cross correlation of the quantization error sequence, namely $M_e(n)$, $R_{ee}(n,p)$, and $R_{x_de}(n,p)$, to be analogously defined.

The following theorem is an extension of a result proven by Chou and Gray. They proved the result under the restriction that the desired input sequence is deterministic. The current result holds for deterministic and stochastic desired input sequences.

Theorem 3: For deterministic or stochastic desired input sequences, $M_\epsilon(n)$ and $R_{x_d\epsilon}(n,p)$ are zero. Consequently, $M_e(n)$ and $R_{x_de}(n,p)$ are also zero.

Proof: If $M_\epsilon(n)$ and $R_{x_d\epsilon}(n,p)$ are zero, then by the linearity of $H(z)$ it follows that $M_e(n)$ and $R_{x_de}(n,p)$ are zero. Therefore it is sufficient to show that $M_\epsilon(n)$ and $R_{x_d\epsilon}(n,p)$ are zero.

From (4) for each $n > a$ we can write $\epsilon(n) = \frac{\Delta}{2} - \Delta U_{n-a}$ where U_{n-a} corresponds to U_r in Lemma 1 with $c = \frac{1}{\Delta}$ and

$$\mu_r = \frac{1}{\Delta}\sum_{i=1}^{r}\left[x_d(n-i) + \tfrac{\Delta}{2}\right]$$

From Lemma 2,

$$\lim_{r\to\infty} \mathrm{E}(U_r) = \tfrac{1}{2}, \tag{8}$$

which implies that $M_\epsilon(n) = 0$. Moreover, (8) holds regardless of the value taken on by $x_d(n)$. Hence,

$$\lim_{r\to\infty} \mathrm{E}\left[x_d(n)U_r\right] = \tfrac{1}{2}\,\mathrm{E}\left[x_d(n)\right],$$

which implies that $R_{x_d\epsilon}(n,p) = 0$.

■

Assuming for now that autocorrelation functions for $e(n)$ and $w(n)$ exist, Theorem 3 in conjunction with (2) implies that the autocorrelation of the output of the systems of Figure 3 can be written as

$$R_{rr}(n,p) = R_{ww}(n,p) + R_{ee}(n,p).$$

Therefore, the significance of Theorem 3 is that the autocorrelation of the quantization error sequence, if it exists, characterizes the second order statistics of the quantization error.

The following theorem shows that $R_{ee}(n,p)$ indeed exists and provides a convenient expression for its evaluation.

Theorem 4: The autocorrelation of the quantization noise sequence can be written as

$$R_{\epsilon\epsilon}(n,p) = \mathrm{E}\big[\varepsilon(n, n+p)\big], \tag{9}$$

where

$$\varepsilon(n,m) = \begin{cases} \frac{\Delta}{12}^2, & \text{if } n = m; \\ \frac{1}{2}\left[\frac{\Delta}{2} - \Delta\left\langle \frac{1}{\Delta}\sum_{i=m}^{n-1}\big[x(i) + \frac{\Delta}{2}\big]\right\rangle\right]^2 - \frac{\Delta}{24}^2, & \text{if } n > m; \\ \frac{1}{2}\left[\frac{\Delta}{2} - \Delta\left\langle \frac{1}{\Delta}\sum_{i=n}^{m-1}\big[x(i) + \frac{\Delta}{2}\big]\right\rangle\right]^2 - \frac{\Delta}{24}^2, & \text{if } n < m. \end{cases} \tag{10}$$

Consequently, the autocorrelation of the quantization error sequence can be written as

$$R_{ee}(n,p) = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\big[h(j) - h(j-1)\big]\big[h(k) - h(k-1)\big]R_{\epsilon\epsilon}(n-k, n+p-j). \tag{11}$$

Proof: As in the proof of Theorem 3, write $\epsilon(n) = \frac{\Delta}{2} - \Delta U_{n-a}$. Then,

$$\mathrm{E}\big[\epsilon(n)\epsilon(n+p)\big] = \frac{\Delta^2}{4} - \frac{\Delta^2}{2}\mathrm{E}(U_{n-a}) - \frac{\Delta^2}{2}\mathrm{E}(U_{n+p-a}) + \Delta^2\,\mathrm{E}(U_{n-a}U_{n+p-a}).$$

From Lemma 2, it follows that

$$\lim_{a\to-\infty}\mathrm{E}\big[\epsilon(n)\epsilon(n+p)\big] = -\frac{\Delta^2}{4} + \Delta^2\lim_{a\to-\infty}\mathrm{E}(U_{n-a}U_{n+p-a}).$$

Therefore, to prove (9) it is sufficient to show that

$$\lim_{a\to-\infty}\mathrm{E}(U_{n-a}U_{n+p-a}) = \frac{1}{\Delta^2}\mathrm{E}\big[\varepsilon(n, n+p)\big] + \tfrac{1}{4}. \tag{12}$$

If $p = 0$, (12) holds as a direct consequence of Lemma 2. Therefore, it is sufficient to prove that (12) holds for $p \geq 1$.

The fractional part operator has the property that for any $x, y \in \mathbf{R}$, $\langle x + y\rangle = \big\langle\langle x\rangle + y\big\rangle$. It follows that for $p \geq 1$

$$U_{n+p-a} = \left\langle U_{n-a} + \frac{1}{\Delta}\sum_{i=0}^{p-1}\big[x(n+i) + \tfrac{\Delta}{2}\big]\right\rangle.$$

Therefore, by Lemma 2

$$\lim_{a\to-\infty} \mathrm{E}(U_{n-a}U_{n+p-a}) = \mathrm{E}\left[U\left\langle U + \tfrac{1}{\Delta}\sum_{i=0}^{p-1}\left[x(n+i)+\tfrac{\Delta}{2}\right]\right\rangle\right], \qquad (13)$$

where U is uniformly distributed on $[0,1)$. The expectation on the right side of (13) can be evaluated in closed form. However, the algebra is messy so it has been relegated to the appendix as Lemma A1. Applying Lemma A1 to (13) for $p \geq 1$ gives

$$\lim_{a\to-\infty} \mathrm{E}(U_{n-a}U_{n+p-a}) = \tfrac{1}{3} + \tfrac{1}{2}\,\mathrm{E}\left\langle \tfrac{1}{\Delta}\sum_{i=0}^{p-1}\left[x(n+i)+\tfrac{\Delta}{2}\right]\right\rangle^2 - \tfrac{1}{2}\,\mathrm{E}\left\langle \tfrac{1}{\Delta}\sum_{i=0}^{p-1}\left[x(n+i)+\tfrac{\Delta}{2}\right]\right\rangle.$$

This can be rearranged as (12) so the proof of (9) is complete.

Combining (5) and the autocorrelation definition gives

$$R_{ee}(n,n+p) = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\left[h(j)-h(j-1)\right]\left[h(k)-h(k-1)\right]\lim_{a\to-\infty}\mathrm{E}_{\eta}\left[\epsilon(n-j)\epsilon(n+p-k)\right],$$

which is equivalent to (11). The term by term multiplication of the series for $e(n)$ and $e(n+p)$ and the interchange of the limit and the sums are justified because the impulse response, $h(n)$, is absolutely summable (because $H(z)$ is stable). Hence, (11) is a direct consequence of (9) and the stability of the filter.

■

Several observations can be made regarding (9). Note that $\varepsilon(n,m)$ is formally a constant offset plus the squared quantization error of a uniform midrise quantizer operating upon a finite partial sum of the input sequence. Thus the quantization error autocorrelation is the weighted sum of the mean squared errors of multiple uniform quantizers operating on various partial sums of the input sequence.

Another observation which we anticipated in the introduction is that the quantization error autocorrelation is only locally dependent upon the input sequence. That is, for a given p, the dependence of $R_{ee}(n,p)$ upon the set of input values $\{x(k) : k < n - N\}$ can be made arbitrarily small by increasing N. This is a consequence of the impulse response of the filter, $h(n)$, being absolutely summable. If

$H(z)$ is an FIR filter, then the stronger assertion can be made that $R_{ee}(n,p)$ is only dependent upon a finite number of values from the input sequence for a given p.

Suppose we were given M systems all operating on the same desired input sequence at the same time. Each system would produce a quantization error sequence $e_i(n)$ that would differ from the other quantization error sequences because of the random variables η_n. In this case, by the law of large numbers,

$$\frac{1}{M}\sum_{i=0}^{M-1} e_i(n)e_i(n+p) \approx R_{ee}(n,n+p) \tag{14}$$

provided M is large (and $a \ll n$).

If Theorem 4 were useful only to the extent that we could predict the average behavior of many identical systems per (14), it would be of limited use. It is more often the case in practice that we are interested in the long-term time-average behavior of $e(n)e(n+p)$. The question therefore arises as to what bearing the conditional autocorrelation has on the time-average behavior of $e(n)e(n+p)$.

Such relationships between time and ensemble averages are usually referred to as ergodic properties [15], [16]. The following theorem and corollary present ergodic results which greatly extend the utility of the Theorem 4.

Theorem 5: The following equations

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\epsilon(n) = 0 \tag{15}$$

and

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} x_d(n)\epsilon(n+p) = 0 \tag{16}$$

hold in probability. Moreover, whenever one of the limits exist,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\epsilon(n)\epsilon(n+p) = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} R_{\epsilon\epsilon}(n,p) \tag{17}$$

holds in probability. In particular the limits exist if the desired input sequence is quasi-stationary.

Proof: As in Theorem 3, we begin by writing $\epsilon(n) = \frac{\Delta}{2} - \Delta U_{n-a}$. The proof of Lemma 2 indicates that

$$\underset{\{\eta_k\}}{\mathrm{E}}(U_r) \rightarrow \mathrm{E}(U)$$

uniformly with respect to the variables $\{\mu_k\}$ as $r \rightarrow \infty$. Applying Lemma A2 (presented in the appendix) with $U_n - \frac{1}{2}$ playing the role of X_n indicates that as $N \rightarrow \infty$

$$\frac{1}{N}\sum_{n=0}^{N-1} U_n \rightarrow \tfrac{1}{2}$$

in probability. Therefore, (15) holds in probability. The argument that (16) holds in probability is almost identical.

To show that (17) holds in probability provided either limit exists, it is sufficient to show that

$$\lim_{N\rightarrow\infty} \frac{1}{N}\sum_{n=0}^{N-1}\left[\epsilon(n)\epsilon(n+p) - R_{\epsilon\epsilon}(n,p)\right] = 0 \tag{18}$$

holds in probability. Without loss of generality, we will assume that $a = 0$ and that $p \geq 0$. Then, $\epsilon(n) = \frac{\Delta}{2} - \Delta U_n$ and a sufficient condition for (18) to hold in probability is that both

$$\frac{1}{N}\sum_{n=0}^{N-1}\left[U_n - \tfrac{1}{2}\right] = 0 \tag{19}$$

and

$$\frac{1}{N}\sum_{n=0}^{N-1}\left[U_nU_{n+p} - \frac{1}{\Delta^2}R_{\epsilon\epsilon}(n,p) - \tfrac{1}{4}\right] = 0 \tag{20}$$

hold in probability. But we have already shown that (19) holds in probability so we can conclude that (18) holds in probability if (20) holds in probability.

Define

$$X_n = U_nU_{n+p} - \frac{1}{\Delta^2}R_{\epsilon\epsilon}(n,p) - \tfrac{1}{4}.$$

The proof of Theorem 4 can be slightly modified to show that for any positive integer j

$$\underset{\{\eta_n : n>j\}}{\mathrm{E}}(X_k) \rightarrow 0 \tag{21}$$

as $k \to \infty$ regardless of the desired input sequence or the values taken on by the variables $\{\eta_0, \ldots \eta_j\}$. As above, we would like to apply Lemma A2 and conclude that (20) holds in probability. To do this, we must show that the convergence in (21) is uniform with respect to the variables $\{\eta_0, \ldots \eta_j\}$ and the desired input sequence, $x_d(n)$. This is equivalent to showing that the convergence is uniform with respect to the variables $\{\mu_n\}$ as defined in Theorem 1.

By the definition of uniform convergence, it is sufficient to show that the sequence

$$\{r_k\} = \Big\{ \sup_{\{\mu_n\}} \big[\mathop{\mathrm{E}}_{\{\eta_n : n>j\}} (X_k) \big] : k = 0, 1, \ldots \Big\}, \tag{22}$$

converges to zero as $k \to \infty$. We will do this by showing that the $p+1$ subsequences of $\{r_k\}$ defined as $\{r_k^{(m)}\} = \{r_{k(p+1)+m} : k = 0, 1, \ldots\}$, $0 \le m \le p$, each converge to zero as $k \to \infty$. Since these sequences together contain the elements of $\{r_k\}$, this is equivalent to showing that $\{r_k\}$ converges.

By the definition of U_k, it follows that, for a given k and p, $\mathop{\mathrm{E}}_{\{\eta_n : n>j\}} (X_k)$ only depends on the μ_k and μ_{k+p}. Moreover, it is a continuous function of these variables. Hence,

$$\mathop{\mathrm{E}}_{\{\eta_n : n>j\}} (X_k) \Big|_{\mu_k = \alpha, \mu_{k+p} = \beta} = \sup_{\{\mu_n\}} \big[\mathop{\mathrm{E}}_{\{\eta_n : n>j\}} (X_k) \big]$$

for some constants α and β. Consequently, for each m there exists a particular sequence $\{\mu_n^{(m)}\}$ such that

$$r_k^{(m)} = \mathop{\mathrm{E}}_{\{\eta_n : n>j\}} (X_{k(p+1)+m}) \Big|_{\{\mu_n\} = \{\mu_n^{(m)}\}}.$$

Since $\mathop{\mathrm{E}}_{\{\eta_n : n>j\}} (X_{k(p+1)+m}) \to 0$ for every sequence $\{\mu_n\}$ it must converge to zero for the particular sequence $\{\mu_n^{(m)}\}$. Hence, for each $0 \le m \le p$, $r_k^{(m)} \to 0$ as $k \to \infty$. This completes the proof that (17) holds in probability provided either limit exists.

If the desired input sequence is quasi-stationary then the quantization noise sequence is also quasi-stationary [13] and so the limit on the right side of (17) exists. In this case, since (18) holds in probability, (17) must hold in probability.

■

Corollary 6: The following equations

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} e(n) = 0 \tag{23}$$

and

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} x_d(n)e(n+p) = 0 \tag{24}$$

hold in probability. Moreover, whenever one of the limits exist,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} e(n)e(n+p) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} R_{ee}(n,p) \tag{25}$$

holds in probability. In particular the limits exist if the desired input sequence is quasi-stationary.

Proof: Each equation follows formally by expanding $e(n)$ with (5) and applying Theorem 5. In each case, the various limits and sums can be interchanged because the impulse response of $H(z)$ is absolutely summable.

■

There are various ergodic theorems that have been or can be applied to the first-order $\Delta\Sigma$ modulator for specific classes of input sequences [16]-[19]. However, the published ergodic theorems do not apply to the class of input sequences considered in the current work. As an example of why it is necessary to consider whether the quantization noise has ergodic properties, suppose that the input sequence $x(n)$ is a random variable uniformly distributed on $(-\frac{1}{2^n}, \frac{1}{2^n}]$ for $n = 1, 2, \ldots$. In this case, it is easy to verify that ergodic results such as those presented in the Theorem 5 do not hold.

We now develop procedures for applying the theory presented thus far to arbitrary input sequences. Recall that our theory requires the input sequence to contain

the random variables, η_n. Therefore, even if the desired input sequence is deterministic, the actual input sequence is stochastic. As argued in the introduction, this assumption is realistic in practice. However, many of the existing results for the $\Delta\Sigma$ modulator are in terms of purely deterministic signals. So that we can later compare our theory to existing work, we first develop a systematic approach to annexing the deterministic case into our theory. We then consider the more important class of arbitrary stochastic input sequences with known statistics. Finally, we present a procedure for obtaining approximate results when the statistics of the desired input sequence are not fully known. To avoid cluttering the development, specific examples are deferred to the next section.

Deterministic Input Sequences

Most of the treatments concerning deterministic input sequences involve the evaluation of the quasi-stationary autocorrelation $R_\epsilon(p)$. In this case,

$$R_\epsilon(p) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \big[\epsilon(n)\epsilon(n+p)\big],$$

since the input sequence does not contain a random component (see [5] and [20] for a discussion of quasi-stationary processes).

To circumvent the restriction that the input sequence contain the random variables, η_n, we take the limiting case as the distribution function of the η_n approaches a unit step function at the origin (i.e., as the η_n converge in distribution to a random variable that is zero with probability one). From Theorem 5 and the definition of $R_\epsilon(p)$,

$$R_\epsilon(p) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} R_{\epsilon\epsilon}(n,p) \tag{26}$$

in probability. Consider a sequence of probability distribution functions $\{P_k(x)\}$ such that

$$\lim_{k\to\infty} P_k(x) = \begin{cases} 1 & \text{if } x \geq 0; \\ 0 & \text{otherwise.} \end{cases}$$

Let $R_{\epsilon\epsilon}(n,p)\,\big|_{P_k(x)}$ be the value of $R_{\epsilon\epsilon}(n,p)$ corresponding to η_n with distribution function $P_k(x)$. From (9) and (10), we have

$$\lim_{k\to\infty} R_{\epsilon\epsilon}(n,p)\,\Big|_{P_k(x)} = \begin{cases} \frac{\Delta}{12}^2, & \text{if } p=0; \\ \frac{1}{2}\left[\frac{\Delta}{2} - \Delta\left\langle \frac{p}{2} + \frac{1}{\Delta}\sum_{i=n}^{n+p-1} x_d(n)\right\rangle\right]^2 - \frac{\Delta}{24}^2, & \text{if } p>0; \\ \frac{1}{2}\left[\frac{\Delta}{2} - \Delta\left\langle \frac{p}{2} + \frac{1}{\Delta}\sum_{i=n+p}^{n-1} x_d(n)\right\rangle\right]^2 - \frac{\Delta}{24}^2, & \text{if } p<0. \end{cases} \tag{27}$$

Define

$$\widehat{R}_\epsilon(p) = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\lim_{k\to\infty} R_{\epsilon\epsilon}(n,p)\,\big|_{P_k(x)}.$$

Applying (27) gives

$$\widehat{R}_\epsilon(p) = \lim_{N\to\infty}\frac{\Delta^2}{2N}\sum_{n=0}^{N-1}\left[\frac{1}{2} - \left\langle\frac{p}{2} + \frac{1}{\Delta}\sum_{i=n}^{n+|p|-1} x_d(i)\right\rangle\right]^2 - \frac{\Delta}{24}^2. \tag{28}$$

Care must be taken to properly interpret $\widehat{R}_\epsilon(p)$. It is tempting to consider it to *be* the autocorrelation of the deterministic input signal, $x_d(n)$, without any contribution from the η_n. As we shall see, in some cases this interpretation is valid and in other cases it is not. In general, what can be said is that given a deterministic input signal, for any $\epsilon > 0$ there is an uncountable infinity of deterministic signals each of which has a mean squared difference from the original signal of less than ϵ and a quasi-stationary autocorrelation equal to $\widehat{R}_\epsilon(p)$. In cases where $R_\epsilon(p) = \widehat{R}_\epsilon(p)$, the existence of the limit in (27) implies that the ideal result for the purely deterministic input sequence is approximately valid if some noise is present. It becomes increasingly accurate as the noise level is reduced. Of course, this can be determined from simulations and observations of actual systems, but the argument above provides a theoretical basis for the behavior. In cases where $R_\epsilon(p) \neq \widehat{R}_\epsilon(p)$, we should be wary of applying the deterministic analysis to a physical $\Delta\Sigma$ modulator implementation. By adding the slightest amount of noise per (3), the resulting autocorrelation will equal $\widehat{R}_\epsilon(p)$ in probability. In this sense, the purely deterministic result is not a physically stable solution.

Stochastic Input Sequences

In conjunction with existing results concerning uniform quantizers, our theory can be used to handle stochastic desired input sequence respecting the no-overload constraint. The first task is to evaluate the autocorrelation function, $R_{ee}(n,p)$, of the quantization error sequence.

As is evident from Theorem 4 and the observations following it, to evaluate this expression we must evaluate the mean squared quantization error corresponding to various quantized partial sums of the input sequence. It is in solving this part of the problem that we benefit from existing results concerning uniform quantizers. If the resulting expression for $R_{ee}(n,p)$ is not dependent upon n, then the quantization error sequence is wide-sense stationary and we are done. Otherwise, we may perform a time-average of $R_{ee}(n,p)$ to obtain the quasi-stationary autocorrelation function. In either case, Corollary 6 ensures that the resulting function, if it exists, converges to the time-average of $e(n)e(n+p)$ in probability.

For a given input sequence, the success of our approach depends upon evaluating the mean squared quantization error of partial sums of the input sequence. Fortunately, considerable attention has been devoted to analyzing the effect of uniform quantization upon stochastic sequences [5], [21]-[24]. In particular, Sripad and Snyder [24] have derived an exact expression for the probability density function of a quantized sequence. If the statistics of $x(n)$ are known, then, using Sripad and Snyder's expression, $R_{ee}(n,p)$ can be evaluated easily. In particular, if the filter has length M and we know all $2M$ and lower joint probability distribution functions of the input sequence, we can calculate the mean squared quantization error, $\sigma_e^2 = R_{ee}(n,0)$. If we know the $2M+N$ and lower joint probability distribution functions of the input sequence, we can calculate $R_{ee}(n,p)$ for all $|p| \leq N$. As will be demonstrated in the next section, the first step is to apply Sripad and Snyder's expression to calculate the probability distributions of the quantized partial sums.

It is then straightforward to evaluate $R_{\epsilon\epsilon}(n,p)$.

Approximate Analysis

Sometimes the statistics of the input sequence are not known or are only partially known. For many such input sequences, our theory gives rise to approximate analyses. As with deterministic and stochastic input sequences, we benefit from having reduced the problem to one of evaluating the mean squared error of a uniform quantizer operating on partial sums of the input sequence.

If the input to a uniform midrise quantizer is sufficiently "busy" or "active" on a scale that is larger than the quantization step size, Δ, it is common to approximate the quantization noise as uniformly distributed on $[-\frac{\Delta}{2}, \frac{\Delta}{2})$ [5], [21]-[24]. In many cases, the partial sums of such a sequence also satisfy this property. Indeed, even if the individual members of the sequence do not satisfy the property, it is possible partial sums of several members do satisfy the property.

Note that (10) is an offset plus the squared quantization error of a partial sum of the desired input sequence. If all the partial sums are busy in the sense described above, it follows that

$$R_{\epsilon\epsilon}(n,p) \approx \begin{cases} \frac{\Delta}{12}^2, & \text{if } p = 0; \\ 0, & \text{otherwise.} \end{cases}$$

In this case, it follows from (11) that

$$R_{ee}(n,p) \approx \frac{\Delta}{6}^2 \Big[\big(h(n) * h(-n)\big)(p) - \big(h(n) * h(-n)\big)(p+1) \Big].$$

In a $\Delta\Sigma$ modulator based oversampling A/D converter, it is not likely that the individual members of the desired input sequence are busy on a scale that is larger than Δ. However, the desired input sequence might be busy on a smaller scale. In this case, sequences of partial sums containing many terms might be busy on a scale that is larger than Δ. Thus, if we know only enough of the low order statistics of the desired input sequence to evaluate (10) for all the cases where the uniform

quantization noise approximation is not valid, we can obtain an good approximation to $R_{\epsilon\epsilon}(n,p)$.

Of course, the accuracy of the uniform quantization noise approximation is highly dependent upon the nature of the input sequence and must be assessed on an individual basis. Fortunately, there exists a large body of work addressing issues of applicability and accuracy associated with the approximation.

III. Application to Specific Input Sequences

Constant-Amplitude Input Sequences

Although constant-amplitude input sequences have been considered by several people, Gray [2] was the first person to perform an exact analysis. With $\Delta = 1$ and for an input sequence $x(n) = x$, where x is an irrational number bounded in absolute value by $\frac{1}{2}$, he showed that the quantization error sequence has a quasi-stationary autocorrelation given by

$$R_\epsilon(p) = \tfrac{1}{12} - \tfrac{1}{2}\langle p(\tfrac{1}{2}+x)\rangle\big(1 - \langle p(\tfrac{1}{2}+x)\rangle\big). \tag{29}$$

An equivalent result can be obtained from our theory. From (28), for any $x \in (-\frac{1}{2}, \frac{1}{2})$

$$\begin{aligned}\widehat{R}_\epsilon(p) &= \lim_{N\to\infty} \frac{1}{2N} \sum_{n=0}^{N-1} \left[\tfrac{1}{2} - \left\langle p\big(\tfrac{1}{2}+x\big)\right\rangle\right]^2 - \tfrac{1}{24} \\ &= \tfrac{1}{12} - \tfrac{1}{2}\langle p(\tfrac{1}{2}+x)\rangle\Big(1 - \langle p(\tfrac{1}{2}+x)\rangle\Big)\end{aligned}$$

which agrees with (29). Therefore, provided x is irrational, $R_\epsilon(p) = \widehat{R}_\epsilon(p)$.

The form of $\widehat{R}_\epsilon(p)$ is not dependent on whether x is rational or irrational. However, (29) does not hold for rational x [2]. It follows that $R_\epsilon(p) \neq \widehat{R}_\epsilon(p)$ if x is rational. Since it is not possible to generate a perfectly constant rational voltage, there is little practical significance to this discrepancy. Adding the slightest amount of noise to a rational constant input in accordance with (3) causes the quasi-stationary autocorrelation to equal $\widehat{R}_\epsilon(p)$ with probability one. Hence, the purely

deterministic result is not a physically stable solution in the case of a rational constant input.

Sinusoidal Input Sequences

Because the $\Delta\Sigma$ modulator is not a linear system, its overall performance can not be characterized solely with respect to sinusoidal inputs. Nevertheless, sinusoidal inputs are often used to test and partially characterize A/D converters. Therefore, sinusoidal input sequences have received considerable attention in the $\Delta\Sigma$ modulator literature. Gray, Chou, and Wong [3] were the first people to perform an exact analysis. As in the constant-input case, they showed that the quantization noise sequence is quasi-stationary and derived an exact expression for the quasi-stationary autocorrelation function. Unlike the constant input case their expression is not in closed-form. In contrast, our theory does yield a closed-form result.

Suppose $x_d(n) = A\cos n\omega_0$ where $|A| < \gamma - \frac{\Delta}{2}$. From (28), we have

$$\widehat{R}_\epsilon(p) = \lim_{N\to\infty} \frac{1}{2N} \sum_{n=0}^{N-1} \left[\tfrac{\Delta}{2} - \Delta \left\langle \frac{p}{2} + \tfrac{1}{\Delta} \sum_{i=n}^{n+|p|-1} A\cos i\omega_0 \right\rangle \right]^2 - \tfrac{\Delta}{24}^2.$$

After some trigonometric manipulation, this becomes

$$\widehat{R}_\epsilon(p) = \lim_{N\to\infty} \frac{1}{2N} \sum_{n=0}^{N-1} \left[\tfrac{\Delta}{2} - \Delta \left\langle \tfrac{1}{\Delta} \left\{ B(\omega_0, p) \sin[\omega_0 n + \theta(\omega_0, p)] + \tfrac{p\Delta}{2} \right\} \right\rangle \right]^2 - \tfrac{\Delta}{24}^2,$$

where

$$B(\omega_0, p) = \frac{\sin\big(\omega_0 |p|/2\big)}{\sin(\omega_0/2)}, \tag{30}$$

and

$$\theta(\omega_0, p) = \tfrac{1}{2}\big(|p| - 1\big)\omega_0 - \frac{\pi}{2}.$$

Note that for each even p, $\widehat{R}_\epsilon(p)$ is equal to a constant plus the average squared error of a quantizer operating on a sinusoid. Similarly, for each odd p it is equal to a constant plus the average squared error of a quantizer operating on a sinusoid offset by $\frac{\Delta}{2}$. Closed-form expressions for these quantities have been derived by Clavier,

Panter, and Grieg [25] and reformulated by Gray [5]. Our theory has thus reduced the problem to one that can be solved by applying results from another problem that has a known solution.

When $\omega_0/2\pi$ is an irrational number, we can apply the results to obtain

$$\widehat{R}_\epsilon(p) = \begin{cases} \frac{\Delta^2}{12} & p = 0; \\ \Delta^2\left\{(R-\frac{1}{2})^2 + \frac{\zeta^2}{2} - \frac{2\zeta}{\pi} - \frac{2}{\pi}\sum_{k=1}^{R-1}\left[2k\sin^{-1}\left(\frac{k}{\zeta}\right)+2\sqrt{\zeta^2-k^2}\right]\right\} & p \text{ even}; \\ \Delta^2\left\{S^2 + \frac{\zeta^2}{2} - \frac{4\zeta}{\pi} - \frac{2}{\pi}\sum_{k=1}^{R-1}\left[(2k+1)\sin^{-1}\left(\frac{k+\frac{1}{2}}{\zeta}\right)+2\sqrt{\zeta^2-(k+\frac{1}{2})^2}\right]\right\} & p \text{ odd}; \end{cases}$$

where $\zeta = \frac{1}{\Delta}B(\omega_0, p)$, $R = \lceil\zeta\rceil$, and $S = \lfloor\zeta + \frac{1}{2}\rfloor$ (when comparing these results to those in [5], note that various algebraic errors have been corrected). Similar results apply to the less important case in which $\omega_0/2\pi$ is a rational number.

Although not an intuitive result, the expression for $\widehat{R}_\epsilon(p)$ is certainly a closed-form expression and is simple to evaluate. From (30) it follows that $B(\omega_0, p) \leq |p|$ for all ω_0. Thus, each sum has at most $\frac{p}{\Delta} + 1$ terms. Since the autocorrelation is most interesting for values of p near the origin and since Δ is most often unity, the sums rarely involve many terms.

Once again, it is interesting to answer the question of whether $\widehat{R}_\epsilon(p)$ and $R_\epsilon(p)$ are equal. Because the expression for $R_\epsilon(p)$ is a double infinite summation of Bessel functions, a quantitative comparison of the two functions for all values of p is difficult. A simpler approach is to compare the functions when $p = 0$. In this case,

$$R_\epsilon(0) = \frac{\Delta^2}{12} - \Delta^2\sum_{l=0}^{\infty}\frac{1}{(\pi 2l)^2}(-1)^l J_0\big(2\pi l\zeta/\sin(\omega_0/2)\big),$$

whereas $\widehat{R}_\epsilon(0) = \frac{\Delta^2}{12}$. Clearly, the two expressions are not equal. A similar but more involved analysis shows that they are not equal for any finite values of p. As in the case of rational constant inputs, that $R_\epsilon(p)$ and $\widehat{R}_\epsilon(p)$ differ indicates that the purely deterministic result is not a physically stable solution. For example,

the slightest amount of noise added according to (3) causes the second term in the expression for $R_\epsilon(0)$ to vanish in probability.

A Simple Class of Stochastic Input Sequences

In the following, we will assume that the variance of the input noise sequence is so small that we can ignore its effect when evaluating (9). This is not a necessary assumption, but it makes the calculations simpler. In an actual $\Delta\Sigma$ modulator, the assumption is equivalent to assuming that the circuit noise floor is significantly below the quantization noise floor.

As a first test of our theory for stochastic input sequences, suppose that $x_d(n)$ is a sequence of independent random variables with characteristic functions satisfying:

$$\Phi_{x_d(n)}(2\pi n/\Delta) = 0$$

for all $n \neq 0$. For example, a sequence satisfies this property if each member has a mean of β_n and is uniformly distributed on $[\beta_n - \frac{\Delta}{2}, \beta_n + \frac{\Delta}{2})$ (respecting the no-overload constraint). Such a sequence might be created by adding a stochastic dither sequence, d_n, satisfying the characteristic function equation above, to a deterministic sequence, β_n.

Since the members of the sequence are independent, adding them together corresponds to multiplying their characteristic functions. Hence, any partial sum of the form

$$S_{n,m} = \sum_{i=m}^{n-1} \left[x(i) + \tfrac{\Delta}{2}\right]$$

has a characteristic function satisfying $\Phi_{S_{n,m}}(2\pi n/\Delta) = 0$ for all $n \neq 0$. This is a necessary and sufficient condition for the error produced by quantizing $S_{n,m}$ to be uniformly distributed [24],[26]. Applying this result to evaluate (10) gives

$$\mathop{\mathrm{E}}_{\{x_d(k)\}}\left[\varepsilon(n,m)\right] = \begin{cases} \frac{\Delta}{12}^2 & \text{if } n = m; \\ 0 & \text{if } n \neq m. \end{cases}$$

For the case of a $\Delta\Sigma$ modulator with a one-bit quantizer (i.e., the case of $\gamma = 1$ and $\Delta = 1$), Chou and Gray [13] have presented an equivalent result. They pointed out that the result is of limited use because in order to satisfy the no-overload constraint, the input sequence must have zero mean (i.e., $\beta_n = 0$ must hold for all n). However, if a multibit quantizer is used, the restriction (when each member of the input sequence is uniformly distributed on $[\beta_n - \frac{\Delta}{2}, \beta_n + \frac{\Delta}{2})$) is that they have means satisfying $-\gamma + \Delta < \beta_n < \gamma - \Delta$ for all n. Since $\Delta = 2\gamma/(2^R - 1)$, where R is the number of quantizer bits, this is frequently not a severe restriction. For example, if an 8 bit quantizer with $\gamma = \frac{1}{2}$ is used in the $\Delta\Sigma$ modulator then the allowed range of the input sequence means is $[-\frac{\Delta}{2} + \frac{1}{255}, \frac{\Delta}{2} - \frac{1}{255}]$ which is 99.6% of the full dynamic range of the input.

Gaussian Input Sequences

As outlined in the previous section, the general procedure for determining $R_{ee}(n,p)$ involves calculating the mean-squared quantization error of various partial sums of the input sequence. In the example just considered, this was particularly simple because all the partial sums had uniformly distributed quantization error. For arbitrary non-stationary stochastic input sequences, the method is still straightforward, but the calculations can be tedious because each partial sum may have a different distribution. In such cases, the calculations are most easily performed using a computer.

To illustrate the general method but nevertheless obtain results that can be verified by hand, we consider the case of a stationary Gaussian desired input sequence. Because all partial sums of such sequences have Gaussian distributions, we need not explicitly determine the distribution of each partial sum so the tedium mentioned above is avoided.

The specifics of the example are as follows. Let $x_d(n)$ be a stationary Gaussian

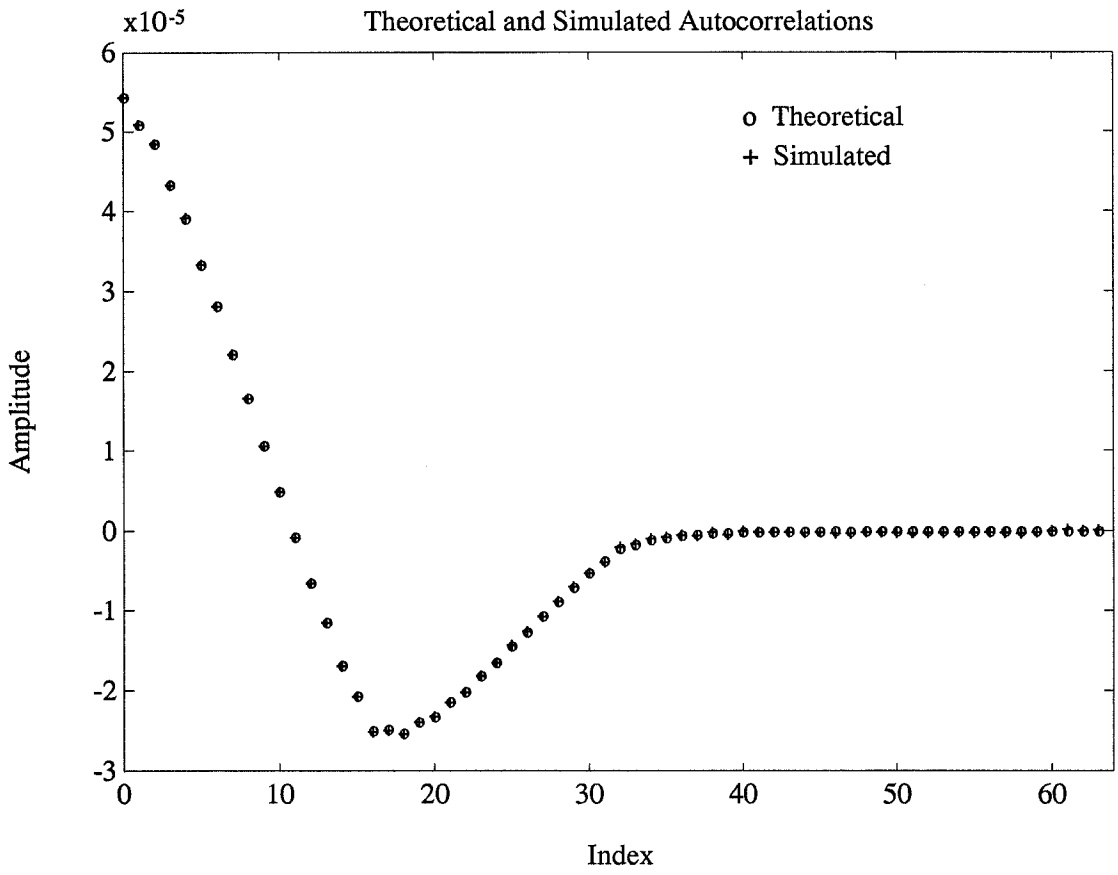


Figure 3: The autocorrelation function as predicted by theory and as obtained by simulation. In this example, $M = 16$, $\sigma^2 = 0.05$, and $\alpha = -0.8$. Each point of the simulation was generated by averaging two-million consecutive points of the form $e(n)e(n+p)$.

sequence with autocorrelation $R_{x_d x_d}(p) = \sigma^2 \alpha^{|p|}$ where σ^2 is the variance of the sequence and $|\alpha| < 1$. Let the $\Delta\Sigma$ modulator have $\gamma = \infty$ and $\Delta = 1$, and let $H(z)$ have the form: $H(z) = F^2(z)$, where

$$F(z) = \frac{1}{M} \sum_{n=0}^{M-1} z^{-1}.$$

It is necessary to have $\gamma = \infty$ so that the overload condition is avoided. If γ were finite, the $\Delta\Sigma$ modulator would be sure to overload sooner or later because Gaussian distributions do not have finite support. However, the example can be applied as a good approximation when γ is finite provided $\gamma \gg \sigma^2$ because in such cases the overload condition is rare.

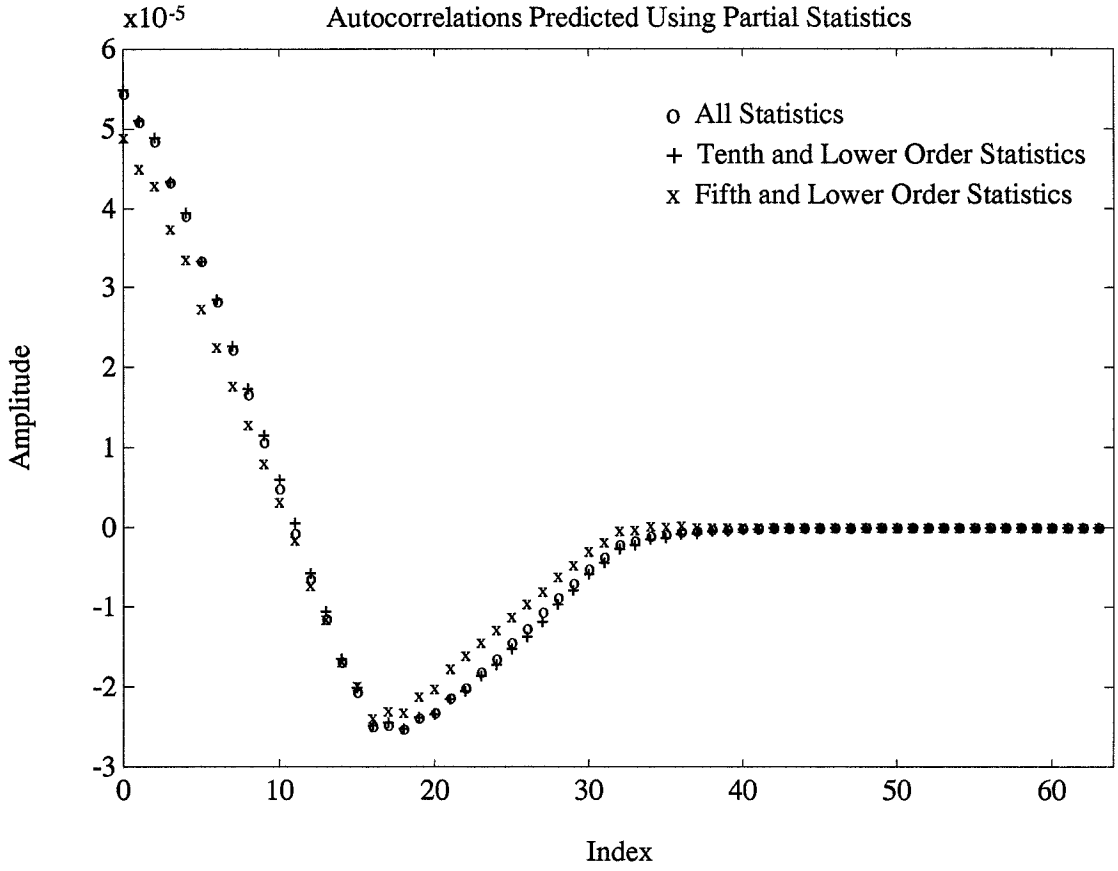


Figure 4: The theoretically predicted autocorrelation function of Figure 3 and approximations to it as predicted by the theory when only tenth and lower order and fifth and lower order statistics are known.

Using the easily derived fact that each partial sum, S_N, of N consecutive samples of the desired input sequence has a Gaussian distribution with variance

$$\sigma_N^2 = \sigma^2 \frac{N(\alpha - \alpha^{-1}) - 2\alpha^N + 2}{2 - \alpha - \alpha^{-1}},$$

it is straightforward to evaluate (10) using standard techniques (see, e.g., [24]). Figure 3 shows sixty-four points of the autocorrelation so calculated for the case of $M = 16$, $\sigma^2 = 0.05$, and $\alpha = -0.8$. The figure also shows the autocorrelation as found by computer simulation.

As shown in the previous section, if the statistics of the desired input sequence are only partially known, it is often possible to obtain approximate results. We

illustrate this by applying the approximation to the previous example.

For arbitrary stochastic desired input sequences, in order to calculate $R_{\epsilon\epsilon}(n,p)$ using Theorem 4, it is necessary to know all $|p|$ and lower order statistics of the desired input sequence. For example, if we only know the fifth and lower order statistics, we can only calculate $R_{\epsilon\epsilon}(n,p)$ when $|p| \leq 5$. In this case, to apply the approximation of the previous section, we would set $R_{\epsilon\epsilon}(n,p) = 0$ whenever $|p| > 5$. Although we know all the statistics of the Gaussian desired input sequence considered in the previous example, we can nevertheless apply the approximation and compare the approximate results to the exact result.

Figure 4 shows the results of the approximation for the cases where $R_{\epsilon\epsilon}(n,p)$ is only calculated for $|p| \leq 5$ and for $|p| \leq 10$. The curves are labeled in terms of the order of statistics that would generally be required to obtain the corresponding approximations if the input sequence were not Gaussian. The approximation is quite good in both cases.

IV. Conclusion

We have presented a unified approach to analyzing the quantization error of the first-order $\Delta\Sigma$ modulator. The approach handles many of the previously analyzed input sequences in addition to a large class of new input sequences. By averaging over the arbitrarily small amount of circuit noise assumed to be present at the analog input to the $\Delta\Sigma$ modulator we have derived a simple expression for the autocorrelation of the quantization error. Each term in the expression is formally equal to the quantization error of a non-overloaded uniform quantizer operating upon a finite partial sum of consecutive input sequence samples. Hence, existing results concerning uniform quantizers are directly applicable in evaluating the autocorrelation expression for specific input sequences. In particular, if the statistics of the desired input sequence are known, then the autocorrelation can be calculated using

standard techniques. If only partial statistics are known, an approximate result can be obtained. The theory is also applicable to deterministic input sequences and has been applied to obtain a new closed-form result for sinusoidal input sequences. We have presented ergodic results which assert that under mild conditions the autocorrelation equals the time-average autocorrelation in probability. We have applied the theory to various input sequences, some of which have been previously considered and some of which are new. Simulation results have been presented that closely support the theory.

Appendix: Supporting Lemmas

Lemma A1: Let α be a random variable that is uniformly distributed on $[0,1)$. Then for any $x, y \in \mathbf{R}$,

$$\mathop{\mathrm{E}}_{\alpha}\big(\langle\alpha+x\rangle\langle\alpha+y\rangle\big) = \tfrac{1}{3} + \tfrac{1}{2}\big(\langle x-y\rangle^2 - \langle x-y\rangle\big).$$

Proof: We will prove the lemma in two steps. In the first step we derive the relation

$$\mathop{\mathrm{E}}_{\alpha}\big(\langle\alpha+x\rangle\langle\alpha+y\rangle\big) = \tfrac{1}{3} + \tfrac{1}{2}\big(\langle x\rangle - \langle y\rangle\big)^2 - \tfrac{1}{2}\big|\langle x\rangle - \langle y\rangle\big|. \tag{31}$$

In the second step we show the surprising result that given $u, v \in \mathbf{R}$,

$$\big(\langle u+v\rangle - \langle u\rangle\big)^2 - \big|\langle u+v\rangle - \langle u\rangle\big| = \langle v\rangle^2 - \langle v\rangle. \tag{32}$$

The lemma follows by combining (31) and (32) with $u = y$ and $v = x - y$.

To prove (31) we proceed as follows. By the properties of the fractional part

operator,

$$\begin{aligned}\mathop{\mathrm{E}}_{\alpha}\big(\langle\alpha+x\rangle\langle\alpha+y\rangle\big) &= \int_0^1 \langle\alpha+x\rangle\langle\alpha+y\rangle d\alpha \\ &= \int_0^1 \big\langle\alpha+\langle x\rangle\big\rangle\big\langle\alpha+\langle y\rangle\big\rangle d\alpha \\ &= \int_0^{u_1} \big(\alpha+\langle x\rangle\big)\big(\alpha+\langle y\rangle\big) d\alpha \\ &\quad + \int_{u_1}^{u_2} \big(\alpha+\max\{\langle x\rangle,\langle y\rangle\}-1\big)\big(\alpha+\min\{\langle x\rangle,\langle y\rangle\}\big) d\alpha \\ &\quad + \int_{u_2}^{1} \big(\alpha+\langle x\rangle-1\big)\big(\alpha+\langle y\rangle-1\big) d\alpha,\end{aligned}$$

where $u_1 = \min\{1-\langle x\rangle, 1-\langle y\rangle\}$ and $u_2 = \max\{1-\langle x\rangle, 1-\langle y\rangle\}$. Expressing the integrands in terms of u_1 and u_2 gives

$$\begin{aligned}\mathop{\mathrm{E}}_{\alpha}\big(\langle\alpha+x\rangle\langle\alpha+y\rangle\big) &= \int_0^{u_1} \big(\alpha-u_1+1\big)\big(\alpha-u_2+1\big) d\alpha \\ &\quad + \int_{u_1}^{u_2} \big(\alpha-u_1\big)\big(\alpha-u_2+1\big) d\alpha \\ &\quad + \int_{u_2}^{1} \big(\alpha-u_1\big)\big(\alpha-u_2\big) d\alpha.\end{aligned}$$

Expanding the integrands and collecting terms gives

$$\begin{aligned}\mathop{\mathrm{E}}_{\alpha}\big(\langle\alpha+x\rangle\langle\alpha+y\rangle\big) &= \int_0^1 \big(\alpha^2-\alpha u_1-\alpha u_2+u_1u_2\big) d\alpha \\ &\quad + \int_0^{u_2} \big(\alpha-u_1\big) d\alpha \\ &\quad + \int_0^{u_1} \big(\alpha+1-u_2\big) d\alpha.\end{aligned}$$

Evaluating the integrals, collecting terms, and expanding u_1 and u_2 gives

$$\begin{aligned}\mathop{\mathrm{E}}_{\alpha}\big(\langle\alpha+x\rangle\langle\alpha+y\rangle\big) &= \tfrac{1}{3} + \tfrac{1}{2}u_1 - \tfrac{1}{2}u_2 + \tfrac{1}{2}(u_1-u_2)^2 \\ &= \tfrac{1}{3} + \tfrac{1}{2}\big(\langle x\rangle-\langle y\rangle\big)^2 - \tfrac{1}{2}\max\{\langle x\rangle,\langle y\rangle\} + \tfrac{1}{2}\min\{\langle x\rangle,\langle y\rangle\}\end{aligned}$$

from which (31) follows.

It remains to prove (32). Because $\langle u+v\rangle = \big\langle\langle u\rangle+v\big\rangle$, without loss of generality we can assume $u \in [0,1)$. For convenience, define,

$$f(v,u) = \big(\langle u+v\rangle - \langle u\rangle\big)^2 - \big|\langle u+v\rangle - \langle u\rangle\big|.$$

Choose any $v \in \mathbf{R}$. Then there exists some integer P such that $v \in [P, P+1)$. Hence we must have either $v+u \in [P, P+1)$ or $v+u \in [P+1, P+2)$. In the first case,

$$\begin{aligned} f(v,u) &= (u+v-P-u)^2 - |u+v-P-u| \\ &= (v-P)^2 - (v-P) \\ &= \langle v \rangle^2 - \langle v \rangle. \end{aligned}$$

In the second case,

$$\begin{aligned} f(v,u) &= \big(u+v-(P+1)-u\big)^2 - \big|u+v-(P+1)-u\big| \\ &= v^2 - 2v(P+1) + (P+1)^2 - (P+1) + v \\ &= (v-P)^2 - (v-P) \\ &= \langle v \rangle^2 - \langle v \rangle. \end{aligned}$$

Hence, $f(v,u) = \langle v \rangle^2 - \langle v \rangle$ for all $u, v \in \mathbf{R}$.

■

Lemma A2: For each $k = 0, 1, \ldots$, let X_k be a deterministic function of the two random sequences $\{\chi_0, \ldots, \chi_k\}$ and $\{\eta_0, \ldots, \eta_k\}$, where the η_n are independent random variables that are independent of the χ_n. Suppose that the distribution of each X_k has its support restricted to $[-\beta, \beta]$ where $\beta \in \mathbf{R}$, and that for each non-negative integer j, as $k \to \infty$

$$\underset{\{\eta_n : n > j\}}{\mathrm{E}}(X_k) \to 0$$

uniformly with respect to the variables $\{\eta_0, \ldots, \eta_j\}$ and $\{\chi_0, \chi_1, \ldots\}$. Then

$$\frac{1}{N} \sum_{n=0}^{N-1} X_n \to 0$$

in probability as $N \to \infty$.

Proof: Define the random variable, S_N, as

$$S_N = \frac{1}{N} \sum_{n=0}^{N-1} X_n.$$

For each $\epsilon > 0$, Chebyshev's inequality [14] gives

$$\text{Prob}\{|S_N| \geq \epsilon\} \leq \frac{\text{E}(S_N^2)}{\epsilon^2}. \tag{33}$$

By the linearity of the expectation operator,

$$\text{E}(S_N^2) = \frac{1}{N^2} \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} \text{E}(X_j X_k). \tag{34}$$

Since the η_n are independent and X_j is independent of the variables $\{\eta_n : n > j\}$, for each $k > j$ we can write

$$\text{E}(X_j X_k) = \text{E}\big[X_j \underset{\{\eta_n : n > j\}}{\text{E}}(X_k)\big].$$

Since

$$\underset{\{\eta_n : n > j\}}{\text{E}}(X_k) \to 0$$

uniformly as $k \to \infty$, it follows that $\text{E}(X_j X_k) \to 0$ as $|k - j| \to \infty$. In particular, this means that there exists a positive integer M such that

$$\big|\text{E}(X_j X_k)\big| < \frac{\epsilon^3}{2} \qquad \text{whenever} \qquad |j - k| \geq M. \tag{35}$$

Since the support of each X_k is restricted to $[-\beta, \beta]$, even when $|j - k| < M$ it follows that

$$\big|\text{E}(X_j X_k)\big| \leq \beta^2. \tag{36}$$

Choose

$$N' = \max\left\{\left\lceil \frac{(2M-1)\beta^2}{\epsilon^3/2} \right\rceil, M\right\}.$$

Dividing the terms on the right side of (34) into two groups corresponding to $|j-k| < M$ and $|j-k| \geq M$ and applying the upper bounds (36) and (35), respectively with $N \geq N'$ gives $\text{E}(S_N^2) < \epsilon^3$. From (33), it follows that for each $N \geq N'$

$$\text{Prob}\{|S_N| \geq \epsilon\} \leq \epsilon$$

This implies that $S_N \to 0$ in probability as $N \to \infty$.

■

References

1. R. M. Gray, "Oversampled sigma-delta modulation," *IEEE Trans. Commun.*, vol. COM-35, pp. 481-489, May 1987.
2. R. M. Gray, "Spectral analysis of quantization noise in a single-loop sigma-delta modulator with dc input," *IEEE Trans. Commun.*, vol. COM-37, pp. 588-599, June 1989.
3. R. M. Gray, W. Chou, and P. W. Wong, "Quantization noise in single-loop sigma-delta modulation with sinusoidal inputs," *IEEE Trans. Inform. Theory*, vol. 35, no. 9, pp. 956-968, Sept. 1989.
4. P. W. Wong and R. M. Gray, "Sigma-delta modulation with I.I.D. Gaussian inputs," *IEEE Trans. Inform. Theory*, vol. 36, no. 4, pp. 784-798, July 1990.
5. R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inform. Theory*, vol. 36, no. 6, pp. 1220-1244, Nov. 1990.
6. H. Inose, Y. Yasuda, J. Murakami, " A telemetering system code modulation—Δ–Σ modulation," *IRE Trans. Space Elect. Telemetry*, col. SET-8, pp. 204-209, Sept. 1962.
7. N. S. Jayant, P. Noll, *Digital Coding of Waveforms Principles and Applications to Speech and Video.* Englewood Cliffs, NJ: Prentice-Hall, 1984.
8. 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.
9. E. N. Protonotarios, "Slope overload noise in differential pulse code modulation systems," *Bell Syst. Tech. J.*, vol. 46, no. 9, pp. 2119-2177, Nov. 1967.
10. D. J. Goodman and L. J. Greenstein, "Quantizing noise of ΔM/PCM encoders," *Bell Syst. Tech. J.*, vol. 52, pp. 183-204, Feb. 1973.
11. J. C. Candy, "A use of double integration in sigma-delta modulation," *IEEE*

Trans. Commun., vol. COM-33, pp. 249-258, Mar. 1985.

12. S. H. Ardalan and J. J. Paulos, "An analysis of nonlinear behavior in delta-sigma modulators," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 593-603, June 1987.
13. W. Chou, and R. M. Gray, "Dithering and its effects on sigma-delta and multistage sigma-delta modulation," *IEEE Trans. Inform. Theory*, vol. 37, no. 3, pp. 500-513, May 1991.
14. P. Billingsley, *Probability and Measure.* New York: John Wiley and Sons, 1986.
15. A. M. IAglom, *An Introduction to The Theory of Stationary Random Functions.* Englewood Cliffs, NJ: Prentice Hall, 1962.
16. K. Petersen, *Ergodic Theory.* Cambridge: Cambridge Univ. Press, 1983.
17. J. C. Kieffer, "Analysis of dc input response for a class of one-bit feedback encoders," *IEEE Trans. Commun.*, vol. COM-38, pp. 337-340, Mar. 1990.
18. L. Kuipers, H. Niederreiter, *Uniform Distribution of Sequences*, New York: John Wiley & Sons, 1974.
19. D. F. Delchamps, "Exact asymptotic statistics for sigma-delta quantization noise," Proceedings *Twenty-Eighth Annual Allerton Conference on Communication, Control, and Computing*, October, 1990
20. L. Ljung, *System Identification: Theory for the User.* Englewood Cliffs, NJ: Prentice-Hall, 1987.
21. W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, pp. 446-472, Jul. 1948.
22. B. Widrow, "A study of rough amplitude quantization by means of nyquist sampling theory," *IRE Trans. Circuit Theory*, vol. CT-3, pp. 266-276, 1956.
23. B. Widrow, "Statistical analysis of amplitude quantized sampled data systems," *Trans. Amer. Inst. Elect. Eng., Pt II: Applications and Industry*, vol. 79, pp. 555-568, 1960.

24. A. B. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-25, pp. 442-448, Oct. 1977.

25. A. G. Clavier, P. F. Panter, and D. D. Grieg, "Distortion in a pulse count modulation system," *AIEE Trans.*, vol. 66, pp. 989-1005, 1947.

26. L. Schuchman, "Dither signals and their effects on quantization noise," *IEEE Trans. Commun. Technol.*, vol. COM-12, pp. 162-165, Dec. 1964.

Granular Quantization Noise in a Class of $\Delta\Sigma$ Modulators

Ian Galton

Abstract— The trend toward digital signal processing in communication systems has resulted in a large demand for fast accurate analog-to-digital (A/D) converters and advances in VLSI technology have made $\Delta\Sigma$ modulator based A/D converters attractive solutions. However, rigorous theoretical analyses have only been performed for the simplest $\Delta\Sigma$ modulator architectures. Existing analyses of more complicated $\Delta\Sigma$ modulators usually rely on approximations and computer simulations. In this paper, a rigorous analysis of the granular quantization noise in a general class of $\Delta\Sigma$ modulators is developed. Under the assumption that some circuit noise is present, the joint statistics of the granular quantization noise sequences are determined and the sequences are shown to be correlation ergodic. The exact results developed for the granular quantization noise are shown to approximately hold for the overall quantization noise if the quantizers in the $\Delta\Sigma$ modulator overload occasionally.

I. Introduction

Although $\Delta\Sigma$ modulator based A/D converters employ complicated digital circuitry, they require minimal analog circuitry and can generally be implemented without the trimmed components and precise reference voltages required in other types of A/D converters. Accordingly, they are well suited to implementation using fine-line VLSI processes optimized for high-speed digital applications. With the growing demand for highly accurate A/D converters and recent advances in VLSI technology, $\Delta\Sigma$ modulators have received considerable attention from both industrial and academic researchers [1] [2].

Many $\Delta\Sigma$ modulator variations have been developed [3]. Most operate on a

The author is with the Electrical Engineering Department, California Institute of Technology, 116-81, Pasadena, CA 91125; email address: galton@systems.caltech.edu

This work was supported by a grant from Pacific Bell.

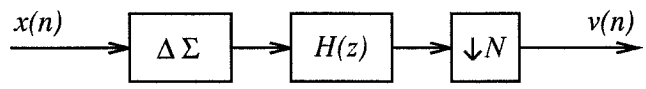


Figure 1: A $\Delta\Sigma$ modulator based oversampling A/D converter.

sampled-data input signal $x(n)$ and produce a quantized sampled-data output signal $y(n)$. A typical $\Delta\Sigma$ modulator architecture consists of linear combinations of sampled-data filters and coarse quantizers surrounded by feedback loops. Usually, the idea is to highpass filter the additive noise sequences introduced by the quantizers while simply delaying or lowpass filtering the input sequence.

A $\Delta\Sigma$ modulator based oversampling A/D converter consists of a $\Delta\Sigma$ modulator, a lowpass filter, and an N-sample decimator as shown in Figure 1. Typically, $x(n)$ corresponds to a continuous-time bandlimited signal that has been sampled at N times the Nyquist rate causing the spectrum of $x(n)$ to be nonzero only on $[-\frac{\pi}{N}, \frac{\pi}{N}]$. Since the $\Delta\Sigma$ modulator highpass filters the quantization noise sequences, for large N most of the spectral energy of the quantization error at the $\Delta\Sigma$ modulator output falls outside of $[-\frac{\pi}{N}, \frac{\pi}{N}]$ and can in principle be removed without distorting the input sequence by subsequent lowpass filtering. The decimator reduces the filtered output sequence to the Nyquist rate.

The filters applied to input sequence and to the quantization noise sequences by a given $\Delta\Sigma$ modulator can generally be determined using linear systems theory. However, quantization is a nonlinear process that destroys information so characterizing the quantization noise sequences has proven difficult. Most of the existing theoretical results assume that the quantizers in the $\Delta\Sigma$ modulator do not overload. The results generally fall into three categories: 1) approximate characterizations, 2) rigorous characterizations for specific deterministic input sequences, and 3) rigorous characterizations for input sequences that contain an independent identically dis-

tributed (iid) random component but are otherwise arbitrary. In the first category, the most common approach is to assume that the quantization noise is white. While the approach can be applied to any input sequence and any $\Delta\Sigma$ modulator, it does not always provide accurate results [4]. Results in the second category are interesting but are limited to first-order $\Delta\Sigma$ modulators and cascades of first-order $\Delta\Sigma$ modulators operating on simple input sequences such as constants and sinusoids [4]. In the third category, various results have been developed for the first-order $\Delta\Sigma$ modulator [5] [6] and cascades of first-order $\Delta\Sigma$ modulators [5]. The approach has the benefit that it can be used to obtain rigorous results for a large class of input sequences [6]. The amplitude of the iid random component can be arbitrarily small, so the assumption tends not to be restrictive in practice [6]. For example, thermal noise at the analog input of a $\Delta\Sigma$ modulator can be modeled as an iid random sequence.

The current work extends the earlier results that characterize quantization noise for input sequences containing an iid random component. In particular, results are developed for a generic $\Delta\Sigma$ modulator of which most of the previously published $\Delta\Sigma$ modulators are special cases. The statistics of the quantization noise sequences are evaluated and the various second-order correlations are shown to be ergodic. In addition to generalizing the earlier work to a larger class of $\Delta\Sigma$ modulators, the current work weakens the previous assumptions made about the iid random component of the input sequence, and does not require that the quantizers never overload.

The remainder of the paper is divided into four main sections. The generic $\Delta\Sigma$ modulator is developed in Section II. In Section III, a granular quantization noise sequence expression is derived. In Section IV the statistics of the granular quantization noise sequences are determined and the second-order correlations are deduced and shown to be ergodic. In Section V it is shown that if the quantizers overload

relatively infrequently then the results of the previous section are approximately valid if applied to the overall quantization noise sequences.

II. A Generic $\Delta\Sigma$ modulator

Three common $\Delta\Sigma$ modulators are shown in Figure 2. The *first-order* $\Delta\Sigma$ modulator shown in Figure 2a consists of a sampled-data integrator, a uniform midrise quantizer, and a negative feedback path. Figure 2b shows a *second-order double-loop* $\Delta\Sigma$ modulator that differs from the first-order $\Delta\Sigma$ modulator in that it contains a second integrator and feedback loop. A *third-order cascaded* $\Delta\Sigma$ modulator is shown in Figure 2c. It is a cascade of the other two $\Delta\Sigma$ modulators followed by a digital filter $\mathbf{U}(z)$.

Most of the published $\Delta\Sigma$ modulator variations, including those shown in Figure 2, can be represented, from a signal processing point of view, as special cases of the generic $\Delta\Sigma$ modulator shown in Figure 3. The system consists of a linear time invariant (LTI) digital system, $\mathbf{T}(z)$, followed by a bank of quantizers followed by another LTI digital system, $\mathbf{U}(z)$. A feedback path joins the output of the quantizer bank to the input of $\mathbf{T}(z)$.

In analyzing the generic $\Delta\Sigma$ modulator, we will make use of the *matrix transfer functions* of $\mathbf{T}(z)$ and $\mathbf{U}(z)$. The behavior of any multi-input multi-output LTI system can be represented mathematically by a matrix transfer function [7]. For example, $\mathbf{T}(z)$ can be represented as a $K \times (K+1)$ matrix of z-transforms $T_{j,k}(z)$, $1 \le j \le K$, $1 \le k \le K+1$. For a given j and k, $T_{j,k}(z)$ is the transfer function of the system joining the k^{th} input node to the j^{th} output node. If we define

$$\mathbf{r}(n) = \begin{bmatrix} r_1(n) \\ \vdots \\ r_k(n) \end{bmatrix} \qquad \text{and} \qquad \mathbf{s}(n) = \begin{bmatrix} s_1(n) \\ \vdots \\ s_k(n) \end{bmatrix},$$

where $r_k(n)$ and $s_k(n)$ are the sequences denoted in Figure 3, and define $\mathbf{R}(z)$ and $\mathbf{S}(z)$ to be the vectors obtained by z-transforming the elements of $\mathbf{r}(n)$ and $\mathbf{s}(n)$,

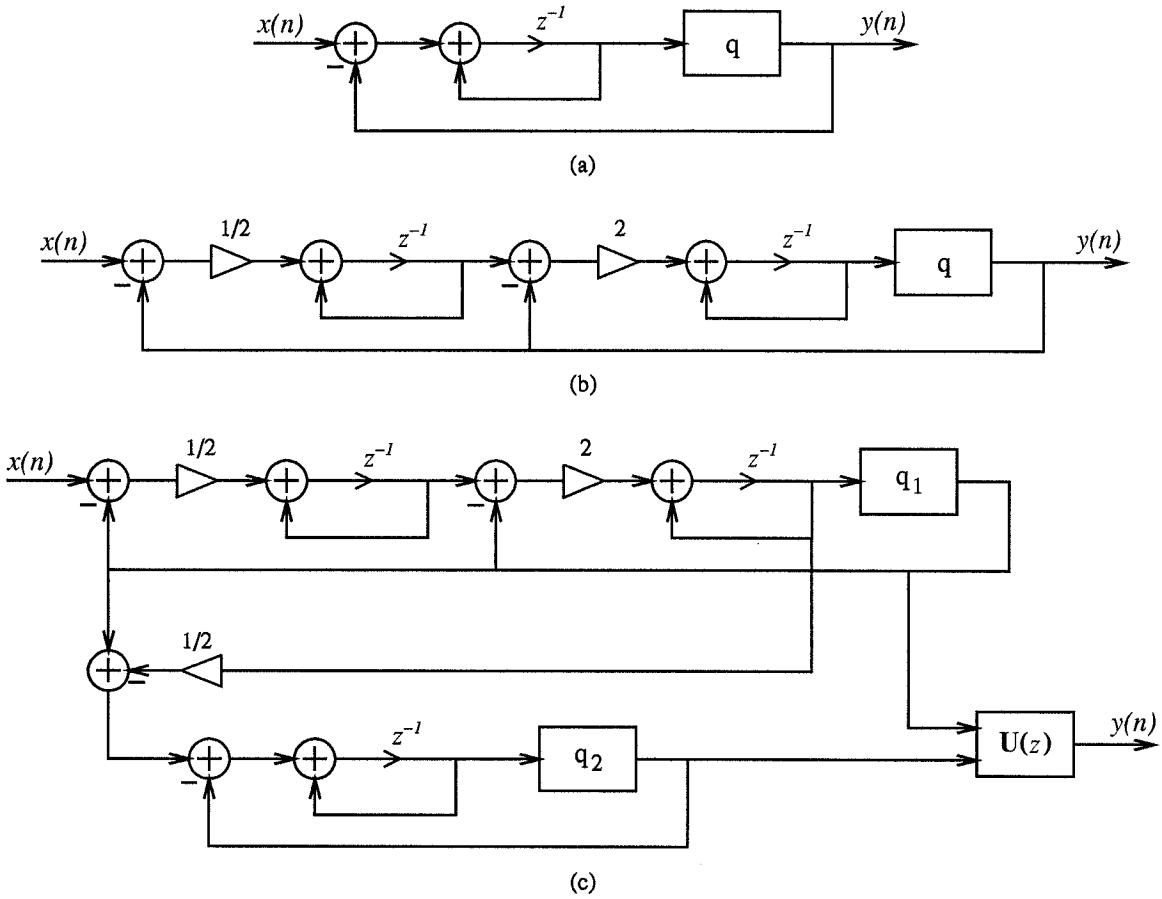


Figure 2: a) The first-order $\Delta\Sigma$ modulator. b) The second-order double-loop $\Delta\Sigma$ modulator. c) A cascade $\Delta\Sigma$ modulator that consists of a second-order double-loop $\Delta\Sigma$ modulator and a first-order $\Delta\Sigma$ modulator.

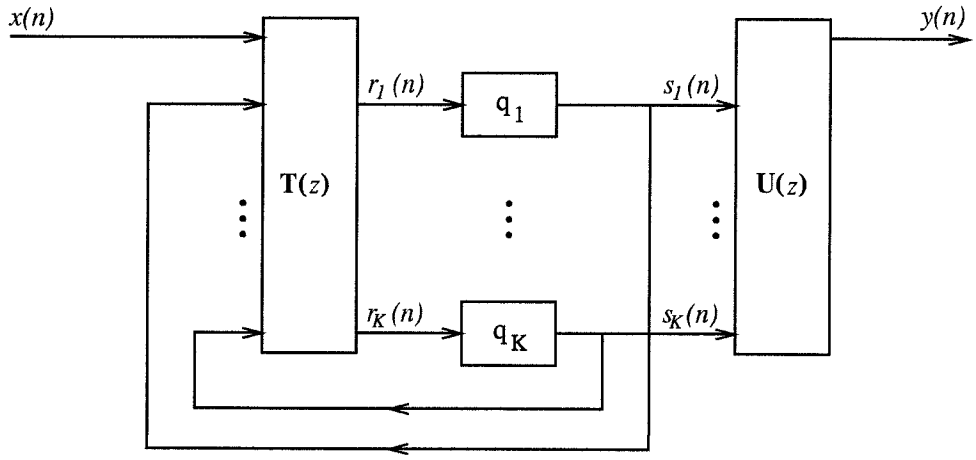


Figure 3: A generic $\Delta\Sigma$ modulator architecture.

respectively, then $\mathbf{R}(z)$ is related to $\mathbf{S}(z)$ as $\mathbf{R}(z) = \mathbf{T}(z)\mathbf{S}(z)$. Similarly, $\mathbf{U}(z)$ can be represented by a $1 \times K$ vector of z-transforms $U_k(z)$, $1 \leq k \leq K$. Thus, $Y(z) = \mathbf{U}(z)\mathbf{S}(z)$ where $Y(z)$ is the z-transform of $y(n)$ in Figure 3.

It will be useful to partition $\mathbf{T}(z)$ into a $K \times 1$ vector, $\mathbf{F}(z)$, and a $K \times K$ matrix, $\mathbf{G}(z)$:

$$\mathbf{F}(z) = \begin{bmatrix} F_1(z) \\ \vdots \\ F_K(z) \end{bmatrix} \qquad \text{and} \qquad \mathbf{G}(z) = \begin{bmatrix} G_{1,1}(z) & \dots & G_{1,K}(z) \\ \vdots & \ddots & \vdots \\ G_{K,1}(z) & \dots & G_{K,K}(z) \end{bmatrix}$$

where $F_k(z) = T_{k,1}(z)$ and $G_{j,k}(z) = T_{j,k+1}(z)$. Therefore,

$$\mathbf{T}(z) = \Big[\, \mathbf{F}(z) \,\Big|\, \mathbf{G}(z) \,\Big].$$

We will denote the impulse responses (i.e., the inverse z-transforms) of $F_k(z)$ and $G_{j,k}(z)$ as $f_k(n)$ and $g_{j,k}(n)$, respectively.

We can interpret the bank of quantizers as a single vector quantization device operating on the vector $\mathbf{r}(n)$ and producing the vector $\mathbf{s}(n)$. With $q_k(\cdot)$ denoting the functional operation of the k^{th} quantizer in Figure 3, define the vector-valued vector function $\mathbf{q}(\cdot)$ as

$$\mathbf{q}(\cdot) = \begin{bmatrix} q_1(\cdot) \\ \vdots \\ q_K(\cdot) \end{bmatrix}.$$

With these definitions, we can rearrange the generic $\Delta\Sigma$ modulator as shown in Figure 4. The scalar input sequence is converted into a vector by $\mathbf{F}(z)$. The vector is added to the vector output of $\mathbf{G}(z)$ and then quantized. The quantized vector is converted into the scalar output by $\mathbf{U}(z)$ and also applied to the input of $\mathbf{G}(z)$. The system is equivalent to that of Figure 3 in the sense that $x(n)$, $y(n)$, $\mathbf{r}(n)$ and $\mathbf{s}(n)$ are the same in both systems.

The results developed in this paper are applicable to any $\Delta\Sigma$ modulator that can be written in the form of Figure 3 or, equivalently, Figure 4, and that satisfies the following conditions.

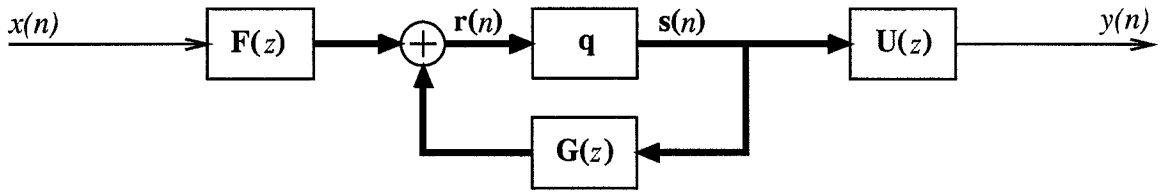


Figure 4: A rearrangement of the generic $\Delta\Sigma$ modulator that is equivalent to the generic $\Delta\Sigma$ modulator in terms of its input-output relationship and the signals seen by its quantizers.

Condition 1: The quantizers $q_1, \ldots, q_k$ are uniform midrize quantizers with step sizes $\Delta_1, \ldots, \Delta_K$, and no-overload ranges $(-\gamma_1, \gamma_1], \ldots, (-\gamma_K, \gamma_K]$, where for each k, γ_k is an integer multiple of Δ_k.

Condition 2: For each k, the impulse response $f_k(n)$ does not converge to zero as $n \to \infty$ and for each $j \neq k$, the difference $\frac{1}{\Delta_k} f_k(n) - \frac{1}{\Delta_j} f_j(n)$ does not converge to zero as $n \to \infty$.

Condition 3: For each j, k, the impulse response $g_{j,k}(n)$ only takes on values that are integer multiples of Δ_j/Δ_k.

For some of the results we will also assume that the $\Delta\Sigma$ modulator satisfies the following condition:

Condition 4: For each k and each $p \neq 0$, the difference $f_k(n) - f_k(n+p)$ does not converge to zero as $n \to \infty$.

Most of the common $\Delta\Sigma$ modulator variations satisfy Conditions 1–3. With the notable exception of $\Delta\Sigma$ modulators that contain a first-order $\Delta\Sigma$ modulator operating directly on the input sequence (e.g., as shown in Figure 2a), most also satisfy Condition 4. Throughout the paper, we will tacitly assume that the $\Delta\Sigma$ modulator satisfies Conditions 1–3. However, all results that specifically assume Condition 4 will be so noted. In cases where Condition 4 is not satisfied because the $\Delta\Sigma$ modulator contains a first-order $\Delta\Sigma$ modulator, applicable results that are analogous to those that assume Condition 4 in this paper have been developed in

[6].

Each of the $\Delta\Sigma$ modulators shown in Figure 2 are special cases of the generic $\Delta\Sigma$ modulator. Each satisfies Conditions 1–3 and those in Figures 2b and 2c also satisfy Condition 4. The transfer functions $\mathbf{T}(z)$, $\mathbf{F}(z)$, and $\mathbf{G}(z)$ can be written for each of the three $\Delta\Sigma$ modulators by inspection. In the case of the first-order $\Delta\Sigma$ modulator of Figure 2a, we have

$$\mathbf{T}(z) = \begin{bmatrix} \frac{z^{-1}}{1-z^{-1}} & -\frac{z^{-1}}{1-z^{-1}} \end{bmatrix},$$

$$F(z) = \frac{z^{-1}}{1-z^{-1}} \qquad \text{and} \qquad G(z) = -\frac{z^{-1}}{1-z^{-1}}. \tag{1}$$

In the case of the second-order $\Delta\Sigma$ modulator shown in Figure 1b, we have

$$\mathbf{T}(z) = \begin{bmatrix} \frac{z^{-1}}{1-z^{-1}} & -\frac{z^{-2}}{(1-z^{-1})^2} - \frac{2z^{-1}}{1-z^{-1}} \end{bmatrix},$$

$$F(z) = \frac{z^{-1}}{1-z^{-1}} \qquad \text{and} \qquad G(z) = -\frac{z^{-2}}{(1-z^{-1})^2} - \frac{2z^{-1}}{1-z^{-1}}. \tag{2}$$

Finally, in the case of the cascaded $\Delta\Sigma$ modulator shown in Figure 2c, we have

$$\mathbf{T}(z) = \begin{bmatrix} \frac{z^{-2}}{(1-z^{-1})^2} & -\frac{z^{-2}}{(1-z^{-1})^2} - \frac{2z^{-1}}{1-z^{-1}} & 0 \\ \frac{1}{2}\frac{z^{-3}}{(1-z^{-1})^3} & \frac{1}{2}\frac{z^{-2}}{(1-z^{-1})^2} + \frac{2z^{-1}}{1-z^{-1}} & -\frac{z^{-1}}{1-z^{-1}} \end{bmatrix},$$

$$\mathbf{F}(z) = \begin{bmatrix} \frac{z^{-2}}{(1-z^{-1})^2} \\ \frac{1}{2}\frac{z^{-3}}{(1-z^{-1})^3} \end{bmatrix} \quad \text{and} \quad \mathbf{G}(z) = \begin{bmatrix} -\frac{z^{-2}}{(1-z^{-1})^2} - \frac{2z^{-1}}{1-z^{-1}} & 0 \\ \frac{1}{2}\frac{z^{-2}}{(1-z^{-1})^2} + \frac{2z^{-1}}{1-z^{-1}} & -\frac{z^{-1}}{1-z^{-1}} \end{bmatrix}. \tag{3}$$

For the first two of these examples, it is easy to verify that $U(z) = 1$. In the third example, $U(z)$ is the, as yet unspecified, system shown in Figure 2c.

We can interpret the vector quantizer in Figure 4 as a device that adds the vector $\epsilon(n) = \mathbf{s}(n) - \mathbf{r}(n)$ to its input as shown in Figure 5a. We will refer to $\epsilon(n)$ as the *quantization noise vector*. For each k, $1 \leq k \leq K$, $\epsilon_k(n)$, the k^{th} element of $\epsilon(n)$, is the quantization noise introduced by the k^{th} quantizer.

Through straightforward matrix manipulations, the system can be rearranged as shown in Figure 5b where

$$S(z) = \mathbf{U}(z)\big(\mathbf{I} - \mathbf{G}(z)\big)^{-1}\mathbf{F}(z), \tag{4}$$

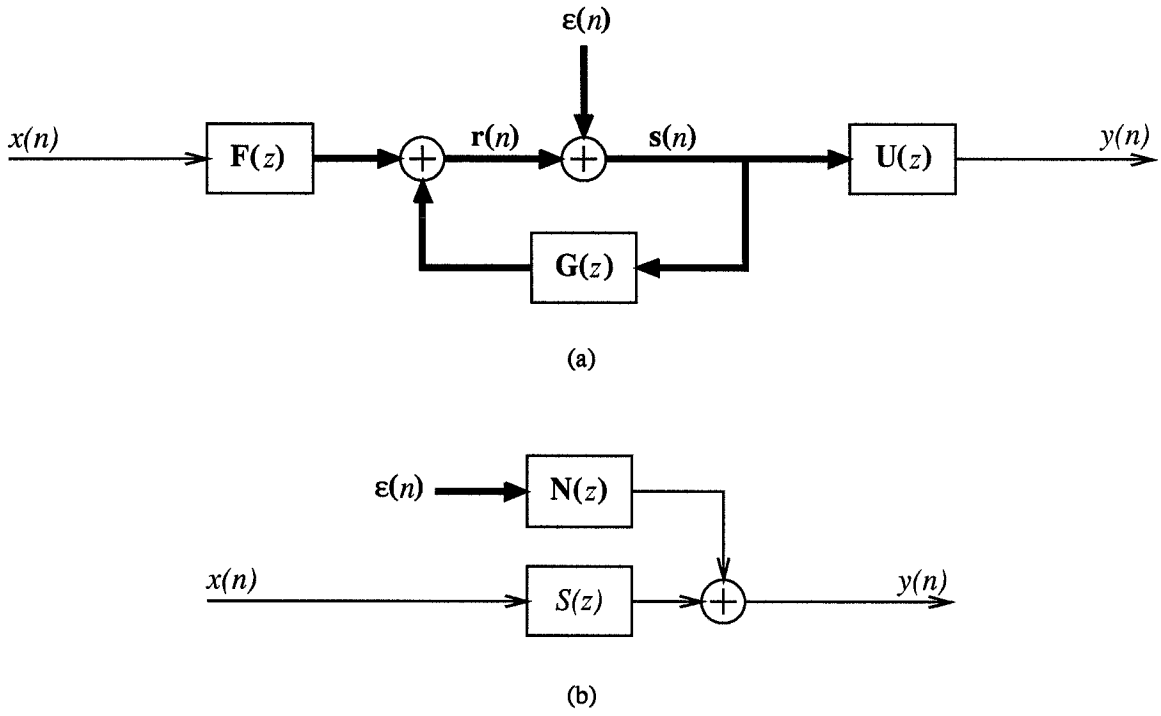


Figure 5: a) The system of Figure 4 with the quantizer bank replaced with the equivalent additive noise source. b) An equivalent form of the system showing the different filters that act on the input and quantization noise, respectively.

and

$$\mathbf{N}(z) = \mathbf{U}(z)\big(\mathbf{I} - \mathbf{G}(z)\big)^{-1}. \tag{5}$$

Note that $S(z)$ is a scalar transfer function, whereas $\mathbf{N}(z)$ is a $1 \times K$ vector of transfer functions, $N_k(z)$.

The input sequence to the $\Delta\Sigma$ modulator sees the filter $S(z)$ while the quantization noise sequences see the filter $\mathbf{N}(z)$. Therefore, we will refer to $S(z)$ as the *signal filter* and to $\mathbf{N}(z)$ as the *noise filter*. In most $\Delta\Sigma$ modulators, $\mathbf{T}(z)$ and $\mathbf{U}(z)$ are chosen such that $S(z)$ is a pure delay while $\mathbf{N}(z)$ is a highpass filter.

For example, from (1) it is easy to verify that the signal and noise filters for the first-order $\Delta\Sigma$ modulator of Figure 2a are $S(z) = z^{-1}$ and $N(z) = 1 - z^{-1}$, respectively. Similarly, from (2) the signal and noise filters for the second-order $\Delta\Sigma$ modulator of Figure 2b are $S(z) = z^{-2}$ and $N(z) = (1 - z^{-1})^2$, respectively. Finally,

if we choose

$$\mathbf{U}(z) = \left[z^{-1}\left(1 + (1 - z^{-1})^2\right) \quad 2(1 - z^{-1})^2 \right]$$

then from (3) it can be verified that the signal and noise filters for the cascaded $\Delta\Sigma$ modulator of Figure 2c are $S(z) = z^{-3}$ and

$$\mathbf{N}(z) = \begin{bmatrix} 0 & (1 - z^{-1})^3 \end{bmatrix},$$

respectively.

Up to this point, we have simply defined the generic $\Delta\Sigma$ modulator and have applied some basic linear systems theory. Therefore, while the material presented thus far has not previously appeared in the literature in a unified form, neither is it fundamentally new. For example, the signal and noise filters associated with the $\Delta\Sigma$ modulators of Figure 2 are well known [3]. However, in the remainder of the paper, we will use the generic $\Delta\Sigma$ modulator framework as a starting point to generalize previous results and develop new results regarding the statistics and ergodicity of the quantization noise introduced by the bank of quantizers.

III. An Expression For The Granular Quantization Noise

Throughout the remainder of the paper, we will consider the $\Delta\Sigma$ modulator to have been "turned on" at a specific time in the past which we will denote as a. For all $n \leq a$ we will take the input sequence and all storage elements in the $\Delta\Sigma$ modulator to be zero. Whenever we consider quantization error sequences, we will tacitly take them to be zero for all $n < a$. In some cases we will consider the system in the limit as $a \to -\infty$. This corresponds to a system that has been running since the beginning of time.

At this point it is convenient to differentiate between *granular* quantization noise, and *overload* quantization noise. If the input to a quantizer at time n is within the no-overload range of the quantizer, the quantization noise at time n is defined

to be granular quantization noise. If the input exceeds the limits of the no-overload range, the quantization noise at time n is said to contain overload quantization noise.

Any uniform quantizer with a finite no-overload range is functionally equivalent to the cascade of a non-overloadable quantizer (i.e., a quantizer with an infinite no-overload range) followed by an amplitude limiter. For example, in the generic $\Delta\Sigma$ modulator, we can think of q_k as the cascade of a non-overloadable uniform midrise quantizer Q_k, with step size Δ_k, followed by an amplitude limiter L_k. For each input x, the output of the amplitude limiter would be

$$\mathrm{L}_k(x) = \begin{cases} x & \text{if } x \in (-\gamma_k, \gamma_k]; \\ -\gamma_k + \frac{\Delta_k}{2} & \text{if } x \leq -\gamma_k; \\ \gamma_k - \frac{\Delta_k}{2} & \text{if } x > \gamma_k. \end{cases}$$

With this view-point, the granular quantization noise is introduced by the non-overloadable quantizer and the overload quantization noise is introduced by the amplitude limiter. Figure 6a shows a version of the generic $\Delta\Sigma$ modulator where the bank of quantizers, $\mathbf{q}$, has been replaced by the equivalent cascade of non-overloadable quantizers $\mathbf{Q}$, and amplitude limiters $\mathbf{L}$.

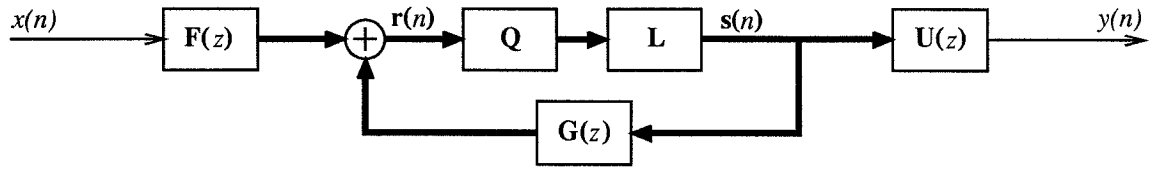
Let us define the granular and overload quantization noise vectors as

$$\epsilon_g(n) = \begin{bmatrix} \epsilon_{g_1}(n) \\ \vdots \\ \epsilon_{g_K}(n) \end{bmatrix} \qquad \text{and} \qquad \epsilon_o(n) = \begin{bmatrix} \epsilon_{o_1}(n) \\ \vdots \\ \epsilon_{o_K}(n) \end{bmatrix},$$

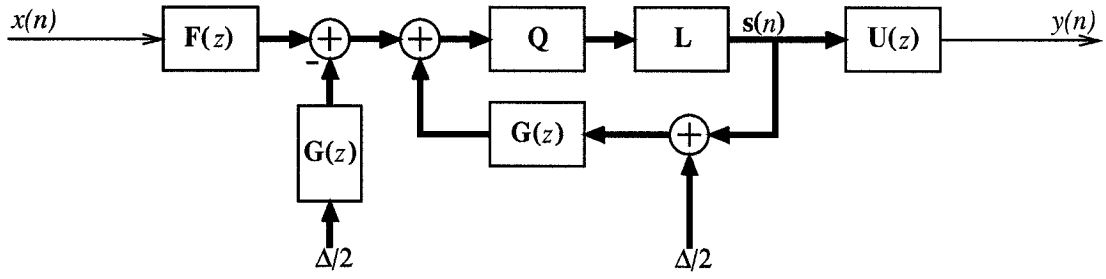
respectively, where for each k, $\epsilon_{g_k}(n)$ is the difference between the output and the input of the non-overloadable quantizer Q_k, and $\epsilon_{o_k}(n)$ is the difference between the output and the input of the amplitude limiter L_k. Therefore, the overall quantization noise vector is $\epsilon(n) = \epsilon_g(n) + \epsilon_o(n)$.

Since Q_k is a non-overloadable uniform midrise quantizer with step size Δ_k, the granular quantization noise introduced by q_k can be written as

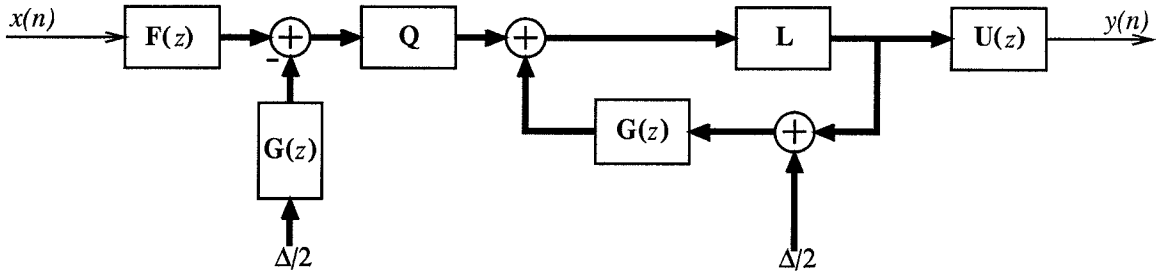
$$\epsilon_{g_k}(n) = \frac{\Delta_k}{2} - \Delta_k \left\langle \frac{r_k(n)}{\Delta_k} \right\rangle. \tag{6}$$



(a)



(b)



(c)

Figure 6: a) The system of Figure 4 with the quantizer bank replaced with the equivalent cascade of non-overloadable quantizers and amplitude limiters. b) An equivalent form of the system in which $\Delta/2$ has been effectively added and subtracted from the system. c) An equivalent form of the system in which the non-overloadable quantizers have been moved to the left of the feedback loop.

The problem with (6) is that the relationship between $r_k(n)$ and $x(n)$ is not yet clear. In order to make use of (6) we must rewrite it in terms of the input sequence $x(n)$.

It is convenient to redraw Figure 6a such that the bank of non-overloadable quantizers is not contained within a feedback loop. To do this, we proceed as follows. Define the vector Δ as

$$\Delta = \begin{bmatrix} \Delta_1 \\ \vdots \\ \Delta_K \end{bmatrix}.$$

The system shown in Figure 6b is equivalent to that of Figure 6a because $\frac{\Delta}{2}$ has

been effectively added and subtracted from the system. Because the quantizers are midrise quantizers, $\mathbf{s}(n) + \frac{\Delta}{2}$ is a vector whose k^{th} element is an integer multiple of Δ_k for each index k and each time n. By Condition 3, the output of $\mathbf{G}(z)$ also has this property. Since the quantization noise introduced by a non-overloadable uniform quantizer is unaffected if its input changes by integer multiples of its step size, the quantization noise produced by $\mathbf{Q}$ is independent of the feedback loop in Figure 6b. For this reason, we can move $\mathbf{Q}$ to the left of the feedback loop to obtain an equivalent system as shown in Figure 6c.

From Figure 6c it is straightforward to rewrite (6) as

$$\epsilon_{g_k}(n) = \frac{\Delta_k}{2} - \Delta_k \left\langle \frac{1}{\Delta_k} \sum_{m=0}^{n-a} \left[f_k(m)x(n-m) + \sum_{l=1}^{K} \Delta_l g_{k,l}(m) \right] \right\rangle. \tag{7}$$

As a consequence of Condition 3, the second sum in (7) only takes on values that are integer multiples of $\Delta_k/2$.

Equation (7) is a generalization of the quantization noise expression for the first-order $\Delta\Sigma$ modulator found by Gray [4]. It is the starting point for the statistical analysis of the quantization noise performed in the remainder of the paper.

IV. Granular Quantization Noise Statistics

In this section we consider the statistics of the granular quantization noise introduced by the quantizers in the generic $\Delta\Sigma$ modulator. Of course, the notion of statistical behavior requires that underlying events be random and we have not as yet placed any such requirements upon the input sequence. We therefore assume the input sequence to be of the following form:

$$x(n) = x_d(n) + \eta_n, \tag{8}$$

where $x_d(n)$ is a bounded stochastic or deterministic sequence and $\{\eta_n\}$ is a sequence of independent identically distributed (iid) random variables that are independent of $x_d(n)$ and whose distribution is not a lattice distribution [8]. We will refer to $x_d(n)$

as the *desired input sequence* and to $\{\eta_n\}$ as the *input noise sequence.* The desired input sequence is the sampled data signal that is to be converted into a digital sequence by the $\Delta\Sigma$ modulator (e.g., the music signal, the video signal, etc.), and the input noise sequence is an unrelated sequence that is assumed to be present in the analog input circuitry. For example, the input noise sequence might correspond to thermal noise which is ubiquitous in analog circuitry and in sampled-data systems can be modeled accurately as an iid random sequence.

Throughout this section we will consider only granular quantization noise $\epsilon_g(n)$. If the quantizers never overload, then $\epsilon_o(n) = \mathbf{0}$, and the results of this section directly apply to the overall quantization noise $\epsilon(n)$. However, we do not require that the quantizers never overload. Indeed, in the next section we suppose the quantizers do overload and show that if the overload times are relatively infrequent, then the results presented in this section for granular quantization noise are approximations when applied to the overall quantization noise.

Theorem 1: For each k, $1 \leq k \leq K$,

(i) The quantization error of the k^{th} quantizer, $\epsilon_{g_k}(n)$, converges in distribution as $a \rightarrow -\infty$ to a random variable $\epsilon'_{g_k}(n)$ that is uniformly distributed on $\left(-\frac{\Delta_k}{2}, \frac{\Delta_k}{2}\right]$.

(ii) For each $j \neq k$, $\epsilon'_{g_j}(n)$ and $\epsilon'_{g_k}(n)$ are independent.

(iii) Provided Condition 4 is satisfied, for each $p \neq 0$, $\epsilon'_{g_k}(n)$ and $\epsilon'_{g_k}(n+p)$ are independent.

(iv) Each member of the desired input sequence, $x_d(n)$, is independent of $\epsilon'_{g_k}(n)$.

Proof: For each $j \neq k$, from (7) we can write $\epsilon_{g_j}(n) = \frac{\Delta_j}{2} - \Delta_j U_{n-a}$ and $\epsilon_{g_k}(n) = \frac{\Delta_k}{2} - \Delta_k V_{n-a}$ where U_{n-a} and V_{n-a} are random variables satisfying the hypothesis

of Lemma A1.† From Lemma A1, as $a \to -\infty$ $\epsilon_{g_j}(n)$ and $\epsilon_{g_k}(n)$ converge in distribution to $\epsilon'_{g_j}(n) = \frac{\Delta_j}{2} - \Delta_j U$ and $\epsilon'_{g_k}(n) = \frac{\Delta_k}{2} - \Delta_k V$ where U and V are independent random variables that are uniformly distributed on $[0, 1)$. Hence, $\epsilon'_{g_j}(n)$ and $\epsilon'_{g_k}(n)$ are uniformly distributed on $\left(-\frac{\Delta_j}{2}, \frac{\Delta_j}{2}\right]$ and $\left(-\frac{\Delta_k}{2}, \frac{\Delta_k}{2}\right]$, respectively, and are independent. This proves parts (i) and (ii).

The proof of part (iii) is similar to that of part (ii). The proof of part (iv) is a direct consequence of Lemma A2.

■

In accordance with the usual definitions, we will take the mean of the quantization noise from the k^{th} quantizer to be

$$\lim_{a \to -\infty} \mathrm{E}\big[\epsilon_{g_k}(n)\big],$$

the autocorrelation of the quantization noise from the k^{th} quantizer to be

$$R_{\epsilon_{g_k}\epsilon_{g_k}}(n, p) = \lim_{a \to -\infty} \mathrm{E}\big[\epsilon_{g_k}(n)\epsilon_{g_k}(n + p)\big],$$

the cross correlation of the quantization noise from the j^{th} and k^{th} quantizers to be

$$R_{\epsilon_{g_j}\epsilon_{g_k}}(n, p) = \lim_{a \to -\infty} \mathrm{E}\big[\epsilon_{g_j}(n)\epsilon_{g_k}(n + p)\big],$$

and the cross correlation of the quantization noise from the k^{th} quantizer and the desired input sequence to be

$$R_{x_d\epsilon_{g_k}}(n, p) = \lim_{a \to -\infty} \mathrm{E}\big[x_d(n)\epsilon_{g_k}(n + p)\big].$$

The following corollary is an immediate consequence of Theorem 1 and the fact that for each k and n, $\epsilon_{g_k}(n)$ is bounded.

Corollary 2: The mean of each granular quantization noise sequence is zero for all integers n. The cross-correlations, $R_{\epsilon_{g_j}\epsilon_{g_k}}(n, p)$ with $j \neq k$, and $R_{x_d\epsilon_{g_k}}(n, p)$, are

† Lemmas A1, A2, A3, and A4 are presented in the appendix.

both zero for all integers n and p. Moreover, provided Condition 4 is satisfied, the autocorrelation $R_{\epsilon_{g_k}\epsilon_{g_k}}(n,p)$ is not a function of n and can be written as†

$$R_{\epsilon_{g_k}\epsilon_{g_k}}(p) = \delta_p \frac{\Delta_k^2}{12}.$$

By virtue of (4) and (5) we can consider the output of the $\Delta\Sigma$ modulator to be $y(n) = w(n) + e(n)$ where $w(n)$ is a filtered version of $x(n)$ and $e(n)$ is a filtered version of the quantization noise. Suppose the quantizers never overload and that the $\Delta\Sigma$ modulator satisfies Condition 4. Then, since the granular quantization noise sequences are white and independent of each other we can write the power spectral density of $e(n)$ as

$$S_{ee}(e^{j\omega}) = \sum_{k=0}^{K} \frac{\Delta_k^2}{12} \left|N_k(e^{j\omega})\right|^2,$$

where $N_k(z)$ is the k^{th} element of $\mathbf{N}(z)$ in (5). Suppose further that the desired input sequence is wide-sense stationary or, more generally, quasi-stationary [9]. Then, because the desired input sequence and the granular quantization noise sequences are independent, we have

$$S_{yy}(e^{j\omega}) = S_{x_d x_d}(e^{j\omega})\left|S(e^{j\omega})\right|^2 + S_{ee}(e^{j\omega}).$$

Thus, Theorem 1 and Corollary 2 provide a simple means of evaluating the statistical performance of the $\Delta\Sigma$ modulator with respect to granular quantization noise. The question arises as to whether analogous assertions can be made regarding the distribution of values taken on by a single instance of the granular quantization noise vector. For example, in a $\Delta\Sigma$ modulator satisfying Condition 4 do $\epsilon_{g_k}(n)$ and

† The function δ_p is the *Kronecker Delta* defined as

$$\delta_p = \begin{cases} 1 & \text{if } p = 0; \\ 0 & \text{otherwise.} \end{cases}$$

$\epsilon_{g_k}(n+1)$ for $n = a, a+1, \ldots$ take on values that are independent and uniformly distributed? Simulations such as that shown in Figure 7, and theoretical results [10] [11] [12] indicate that for specific cases the question may be answered in the affirmative. Nevertheless, general results analogous to Theorem 1 that relate to the time distributions of the granular quantization noise vector are not known to the author. We can, however, prove that the statistical averages in Corollary 2 converge in probability to the corresponding time averages. In particular, the following theorem establishes that for each j, k, and p the time averages of $\epsilon_{g_k}(n)$, $x_d(n)\epsilon_{g_k}(n+p)$ and $\epsilon_{g_j}(n)\epsilon_{g_k}(n+p)$ converge in probability to the corresponding statistical averages. Since each of the terms are bounded, convergence in probability implies convergence in the mean [8]. Sequences whose time-average correlations converge in some sense to the corresponding statistical correlations are usually referred to as *correlation ergodic*. Therefore, the theorem asserts that the granular quantization noise vector is correlation ergodic.

Theorem 3: As $N \to \infty$,

$$\frac{1}{N}\sum_{n=0}^{N-1} \epsilon_{g_k}(n) \to 0 \tag{9}$$

and

$$\frac{1}{N}\sum_{n=0}^{N-1} x_d(n)\epsilon_{g_k}(n+p) \to 0 \tag{10}$$

in probability. Moreover, provided Condition 4 is satisfied,

$$\frac{1}{N}\sum_{n=0}^{N-1} \epsilon_{g_j}(n)\epsilon_{g_k}(n+p) \to R_{\epsilon_{g_j}\epsilon_{g_k}}(p) \tag{11}$$

in probability as $N \to \infty$.

Proof: Because the proofs of (9), (10), and (11) are similar, we will only state the proof of (10).

For each $n = 0, 1, \ldots$, let $X_n = x_d(n-p)\epsilon_{g_k}(n)$. Then, it is sufficient to show

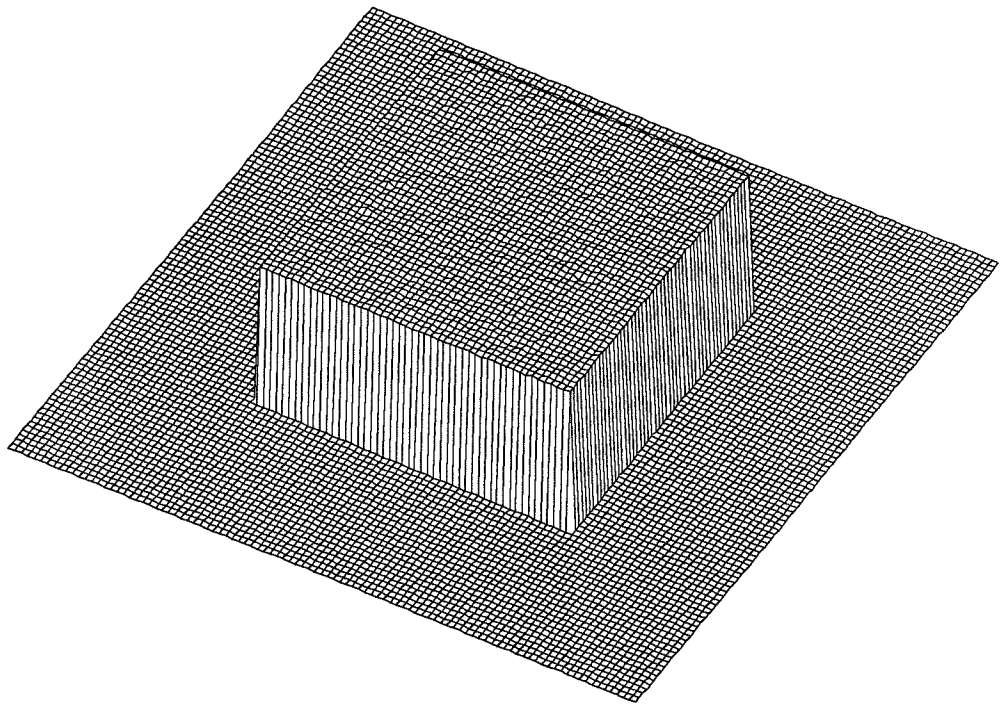


Figure 7: The density of values taken on by $\epsilon(n)$ and $\epsilon(n+1)$, $n = 0, 1, \ldots$, in a non-overloading second-order double-loop $\Delta\Sigma$ modulator. The desired input sequence was a unit-amplitude sinusoid, the input noise sequence had a variance of 10^{-6}, and the quantizer had unity step-size. The density is plotted on the interval $[-2, 2]^2$.

that

$$\frac{1}{N}\sum_{n=0}^{N-1} X_n \to 0 \tag{12}$$

holds in probability. As in the proof of Theorem 1, we can write $\epsilon_{g_k}(n) = \frac{\Delta_k}{2} - \Delta_k U_{n-a}$. Lemma A3 implies that for each $j \geq a$

$$\mathop{\mathrm{E}}_{\{\eta_n : n \geq j\}}\left[U_{n-a}\right] \to \tfrac{1}{2}$$

uniformly with respect to the variables $\eta_a, \eta_{a+1}, \ldots, \eta_{j-1}$ and with respect to the desired input sequence $x_d(n)$ as $n \to \infty$. Equivalently, since the desired input sequence is independent of the input noise sequence,

$$\mathop{\mathrm{E}}_{\{\eta_n : n \geq j\}}\left[X_n\right] \to 0$$

uniformly with respect to these variables as $n \to \infty$. Hence, Lemma A4 implies that (12) holds in probability.

■

V. The Effect of Quantizer Overload

As discussed in Section III, the quantization noise vector is made up of granular and overload quantization noise vectors:

$$\epsilon(n) = \epsilon_g(n) + \epsilon_o(n).$$

Generally, overload quantization noise is highly correlated to the input sequence, is difficult to characterize mathematically, and tends to spoil the performance of $\Delta\Sigma$ modulators [13]. In practical $\Delta\Sigma$ modulators, it is often best to choose the no-overload ranges of the quantizers so as to avoid overload altogether. However, in some practical systems, it is convenient to use coarse quantizers that occasionally overload. For example, the second-order $\Delta\Sigma$ modulator of Figure 2b is often used with a one-bit quantizer. In this configuration, input sequences that overload the $\Delta\Sigma$ modulator can be found with arbitrarily small maximum amplitude. Even so, for many applications the system achieves acceptable performance because the overload condition is rare for small amplitude input sequences.

In cases where the quantizers overload, the results of the previous section still apply to the granular quantization noise, but do not provide insight into the behavior of overload quantization noise. Figure 8 shows the simulated density of values taken on by the quantization noise terms $\epsilon(n)$ and $\epsilon(n+1)$, $n = 0, 1, \ldots$, for a second-order double-loop $\Delta\Sigma$ modulator that occasionally overloads. The simulation parameters were the same as those of Figure 7 except that an overloadable quantizer was used. In comparing the two figures, we see that overload complicates the structure of the

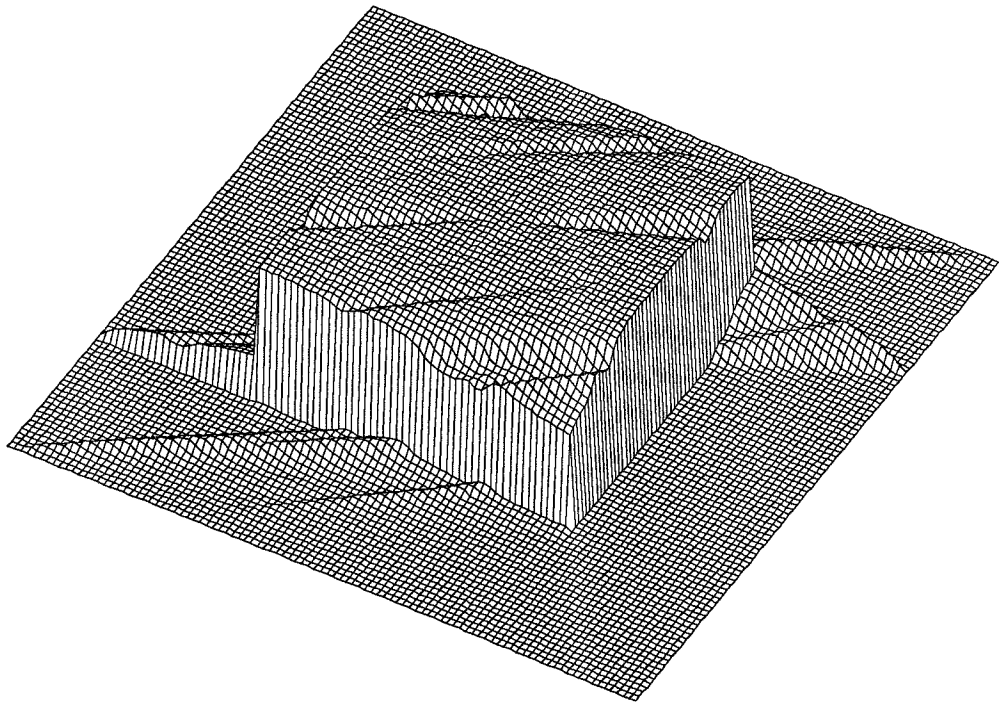


Figure 8: The density of values taken on by $\epsilon(n)$ and $\epsilon(n+1)$, $n = 0, 1, \ldots$, in an overloading second-order double-loop $\Delta\Sigma$ modulator. The desired input sequence was a unit-amplitude sinusoid, the input noise sequence had a variance of 10^{-6}, and the quantizer had unity step-size. The density is plotted on the interval $[-2, 2]^2$.

joint distribution. Nevertheless, it is evident that in this case,

$$\frac{1}{N}\sum_{n=0}^{N-1}\epsilon(n)\epsilon(n+1) \approx 0$$

for large N.

This approximate behavior is characteristic of $\Delta\Sigma$ modulators in which overload has a low frequency of occurrence. As can be deduced from Figure 6c, the overload noise $\epsilon_o(n)$ originates in the bank of limiters, **L**, and is subjected to the noise filter $\mathbf{N}(z)$. In many systems, including all those in Figure 2, the noise filter has a finite impulse response. In such cases, if the quantizers overload infrequently the overall quantization error, $e(n)$, would only infrequently be a function of over-

load quantization noise. In this sense, the results of the previous section would be approximately valid as written. In other words, the effects of quantizer overload are not catastrophic provided the times at which the quantizers overload are infrequent.

VI. Conclusion

We have defined and analyzed a generic $\Delta\Sigma$ modulator architecture. The results provide a unified framework for analyzing a large class of $\Delta\Sigma$ modulators because most of the known $\Delta\Sigma$ modulators are special cases of the generic system. Assuming that a small amount of circuit noise is present in the analog front end of the $\Delta\Sigma$ modulator, we have performed a statistical analysis of the granular quantization noise for arbitrary input sequences. In particular, we have shown that in most $\Delta\Sigma$ modulators with orders greater than one, each quantization noise sequence converges in distribution to a sequence of random variables that are uniformly distributed and independent of each other and the input sequence as the run-time increases to infinity. This behavior is markedly different from that of the first-order $\Delta\Sigma$ modulator as developed in [6]. We have also shown that the granular quantization noise is correlation ergodic. Unlike most other theoretical treatments, we do not require that the quantizers never overload.

Appendix: Supporting Lemmas

Lemma A1: For each $p = 0, 1, \ldots$, let

$$U_p = \left\langle \mu_p + \sum_{k=0}^{p} c_k \eta_k \right\rangle \qquad \text{and} \qquad V_p = \left\langle \nu_p + \sum_{k=0}^{p} d_k \eta_k \right\rangle,$$

where $\{\eta_k\}$ is a sequence of independent, identically distributed random variables whose distribution is not a lattice distribution, $\{c_k\}$ and $\{d_k\}$ are any real sequences that do not converge to zero and whose difference does not converge to zero, and $\{\mu_p\}$ and $\{\nu_p\}$ are any two sequences of random variables each of which is indepen-

dent of every η_k. Then as $p \to \infty$, U_p and V_p converge in distribution to random variables U and V that are each uniformly distributed on $[0,1)$ and are independent.

Proof: For each $p = 0, 1, \ldots$, let

$$X_p = \mu_p + \sum_{k=0}^{p} c_k \eta_k.$$

Then $U_p = \langle X_p \rangle$. Let $\Phi_\eta(t)$ be the characteristic function of each η_k, and let $\Phi_{\mu_p}(t)$ be the characteristic function of μ_p. Since the η_k are independent of each other and of each μ_p, we can write

$$\Phi_{X_p}(t) = \Phi_{\mu_p}(t) \prod_{k=0}^{p} \Phi_\eta(c_k t).$$

All characteristic functions are one at $t = 0$, and less than or equal to one for $t \neq 0$. However, since the distribution of the η_k is not a lattice distribution, we are assured that $\Phi_\eta(t)$ is strictly less than one for all $t \neq 0$.†. Therefore,

$$\lim_{p \to \infty} \Phi_{X_p}(t) = \begin{cases} 1, & \text{if } t = 0; \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Let P_{X_p} be the probability measure of X_p. Given $A \subset \mathbf{R}$, let

$$P_{\langle X_p \rangle}(A) = \sum_{k=-\infty}^{\infty} P_{X_p}\Big(\big(A \cap [0,1)\big) + k\Big).$$

By the definition of the fractional part operator, it follows that $P_{\langle X_p \rangle}$ must equal the probability measure of $\langle X_p \rangle$ for every set on which the sum converges. But by the countable additivity of P_{X_p} and the fact that if A is measurable so is $\cup_{k=-\infty}^{\infty}\big(\big(A \cap [0,1)\big) + k\big)$, the sum converges for all sets on which P_{X_p} is defined. Therefore $P_{\langle X_p \rangle}$ is the probability measure of $\langle X_p \rangle$.

The characteristic function of $\langle X_p \rangle$ is

$$\begin{aligned} \Phi_{\langle X_p \rangle}(t) &= \int_{-\infty}^{\infty} e^{jtx} P_{\langle X_p \rangle}(dx) \\ &= \int_0^1 e^{jtx} \sum_{k=-\infty}^{\infty} P_{X_p}(dx + k). \end{aligned}$$

† See, for example, Problem 26.1 in [8].

Hence, for all integers m,

$$\begin{aligned}\Phi_{\langle X_p \rangle}(2\pi m) &= \int_0^1 \sum_{k=-\infty}^{\infty} e^{j2\pi m(x+k)} P_{X_p}(dx+k) \\ &= \int_{-\infty}^{\infty} e^{j2\pi mx} P_{X_p}(dx) \\ &= \Phi_{X_p}(2\pi m).\end{aligned}$$

From (13) this implies

$$\lim_{p\to\infty} \Phi_{U_p}(2\pi m) = \begin{cases} 1, & \text{if } m = 0; \\ 0, & \text{if } m \text{ is a nonzero integer.} \end{cases} \tag{14}$$

Define $\Phi_U(t)$ as

$$\Phi_U(t) = e^{-jt/2} \frac{\sin(t/2)}{t/2}.$$

Then

$$\lim_{p\to\infty} \Phi_{U_p}(2\pi m) = \Phi_U(2\pi m).$$

Taking the inverse Fourier transform of $\Phi_U(t)$ shows it to be the characteristic function of a random variable U that is uniformly distributed on $[0,1)$. Moreover, by definition the support of U_p is restricted to $[0,1)$. Therefore $\Phi_U(t)$ and $\Phi_{U_p}(t)$ are uniquely determined by their samples at $t = 2\pi m$, $m = 0, \pm 1, \ldots$.†

A necessary and sufficient condition for U_p to converge in distribution to U is that $\Phi_{U_p}(t)$ converge to $\Phi_U(t)$ for each $t \in \mathbf{R}$.‡ Therefore, by the argument above, any subsequence of $\{U_p\}$ that converges in distribution at all must converge in distribution to U. Since the sequence of probability measures associated with $\{U_p\}$ is tight (a consequence of the sequence having bounded support) it follows that there exists a subsequence of $\{U_p\}$ that converges in distribution to U which further implies that U_p converges in distribution to U.††

It remains to show that U and V are independent. By the definition of the fractional part operator, for each $x, y \in \mathbf{R}$, we have $\langle x+y \rangle = \langle \langle x \rangle + y \rangle$. Therefore,

† See, for example, Theorem XIX.6.1 in [14].

‡ See, for example, Theorem 26.3 in [8].

†† See, for example, Theorem 25.10 and its corollary in [8].

we can rewrite U_p as

$$U_p = \left\langle V_p + \mu_p - \nu_p + \sum_{k=0}^{p}(c_k - d_k)\eta_k \right\rangle.$$

For each p, define the random variable W_p to be U_p under the constraint that $V_p = \alpha$ where α is any constant on $[0,1)$. That is, let $W_p = U_p \mid_{V_p=\alpha}$

By the same argument used to show that U_p converges in distribution to U, it follows that W_p converges in distribution to a random variable W that is uniformly distributed on $[0,1)$. But the distribution of each W_p is, by definition, the conditional distribution of U_p given that $V_p = \alpha$. Therefore, the distribution of W is the conditional distribution of U given that $V = \alpha$. Since U and W have the same distribution, it follows that U and V are independent.

■

The first part of the previous lemma is an extension of a result proven by Chou and Gray [5]. They proved that U_p converges in distribution to U under the restrictions that $c_k = 1$ for all k, that the η_n are iid with distributions having densities, and that the μ_k are deterministic.

Lemma A2: Let U be as in Lemma A1, and let X be any random variable that is independent of each η_k. Then X and U are independent.

Proof: The proof is almost identical to the portion of the proof of Lemma A1 that shows U and V are independent.

■

Lemma A3: Let U_p and V_p be as defined in Lemma A1. Then, as $p \to \infty$

$$\underset{\{\eta_n : n \geq 0\}}{\mathrm{E}}\left[U_p\right] \to \tfrac{1}{2}, \tag{15}$$

and

$$\underset{\{\eta_n : n \geq 0\}}{\mathrm{E}}\left[U_p^2\right] \to \tfrac{1}{3}, \tag{16}$$

where the convergence is uniform with respect to $\mu_0, \mu_1, \ldots$. Moreover, as $p \to \infty$

$$\underset{\{\eta_n : n \geq 0\}}{\mathrm{E}}\big[U_p V_p\big] \to \tfrac{1}{4}, \tag{17}$$

where the convergence is uniform with respect to $\mu_0, \mu_1, \ldots$ and $\nu_0, \nu_1, \ldots$.

Proof: By definition, U_p and V_p have bounded support. Hence U_p^2 can be considered a bounded continuous function of U_p and $U_p V_p$ can be considered a bounded continuous function of U_p and V_p. It follows from Lemma A1 that as $p \to \infty$[†]

$$\underset{\{\eta_n : n \geq 0\}}{\mathrm{E}}\big[U_p\big] \to \mathrm{E}\big[U\big],$$

$$\underset{\{\eta_n : n \geq 0\}}{\mathrm{E}}\big[U_p^2\big] \to \mathrm{E}\big[U^2\big],$$

and

$$\underset{\{\eta_n : n \geq 0\}}{\mathrm{E}}\big[U_p V_p\big] \to \mathrm{E}\big[UV\big].$$

Since U is uniformly distributed on $[0,1)$, it is easy to verify that $\mathrm{E}\big[U\big] = \frac{1}{2}$ and $\mathrm{E}\big[U^2\big] = \frac{1}{3}$. Since U and V are independent, $\mathrm{E}\big[UV\big] = \frac{1}{4}$.

It remains to show that the convergence in (15), (16), and (17) is uniform. Since the proofs of the three cases are similar, we will only prove that the convergence in (17) is uniform.

Note that for each integer $p \geq 0$, $\underset{\{\eta_n : n \geq 0\}}{\mathrm{E}}\big[U_p V_p\big]$ is a function of only the p^{th} members of the sequences $\{\mu_k\}$ and $\{\nu_k\}$. Since it is a bounded piecewise continuous function of μ_p and ν_p, there exist sequences $\{\mu'_k\}$ and $\{\nu'_k\}$ such that for every $p \geq 0$

$$\sup_{\{\mu_k\},\{\nu_k\}} \left| \underset{\{\eta_n : n \geq 0\}}{\mathrm{E}}\big[U_p V_p\big] - \tfrac{1}{4} \right| = \left| \underset{\{\eta_n : n \geq 0\}}{\mathrm{E}}\big[U_p V_p\big] \Big|_{\mu_p = \mu'_p, \nu_p = \nu'_p} - \tfrac{1}{4} \right|.$$

Since (17) holds for all sequences $\{\mu_k\}$ and $\{\nu_k\}$, it must hold for the specific sequences $\{\mu'_k\}$ and $\{\nu'_k\}$. Therefore,

$$\lim_{p \to \infty} \sup_{\{\mu_k\},\{\nu_k\}} \left| \underset{\{\eta_n : n \geq 0\}}{\mathrm{E}}\big[U_p V_p\big] - \tfrac{1}{4} \right| = 0$$

---

† See, for example, Theorem 29.1 in [8].

which implies that the convergence in (17) is uniform with respect to $\{\mu_k\}$ and $\{\nu_k\}$.

■

Lemma A4: For each $k = 0, 1, \ldots$, let X_k be a deterministic function of the two random sequences $\{\chi_0, \ldots, \chi_k\}$ and $\{\eta_0, \ldots, \eta_k\}$, where the η_n are independent random variables that are independent of the χ_n. Suppose that the distribution of each X_k has its support restricted to $[-\beta, \beta]$ where $\beta \in \mathbf{R}$, and that for each non-negative integer j, as $k \to \infty$

$$\mathop{\mathrm{E}}_{\{\eta_n : n > j\}}(X_k) \to 0$$

uniformly with respect to the variables $\{\eta_0, \ldots, \eta_j\}$ and $\{\chi_0, \chi_1, \ldots\}$. Then

$$\frac{1}{N} \sum_{n=0}^{N-1} X_n \to 0$$

in probability as $N \to \infty$.

Proof: This lemma corresponds to Lemma A2 in [6]. See [6] for the proof.

■

References

1. F. Goodenough, "High-resolution ADCs up dynamic range in more applications," *Electronic Design*, pp. 65-79, April 11, 1991.
2. *Oversampling Delta-Sigma Data Converters Theory, Design and Simulation*, Edited by J. C. Candy, G. C. Temes, New York, IEEE Press, 1992.
3. J. C. Candy, G. C. Temes, "Oversampling Methods for A/D and D/A Conversion," *Oversampling Delta-Sigma Data Converters Theory, Design and Simulation*, Edited by J. C. Candy, G. C. Temes, New York: IEEE Press, pp. 1-25, 1992.
4. R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inform. Theory*, vol. 36, no. 6, pp. 1220-1244, Nov. 1990.

5. W. Chou, and R. M. Gray, "Dithering and its effects on sigma-delta and multistage sigma-delta modulation," *IEEE Trans. Inform. Theory*, vol. 37, no. 3, pp. 500-513, May 1991.

6. I. Galton, "Granular quantization noise in the first-order $\Delta\Sigma$ modulator," Submitted to *IEEE Trans. Inform. Theory*, Nov. 1991.

7. P. P. Vaidyanathan, *Multirate Systems and Filter Banks.* Englewood Cliffs, NJ: Prentice-Hall, 1992.

8. P. Billingsley, *Probability and Measure.* New York: John Wiley and Sons, 1986.

9. L. Ljung, *System Identification: Theory for the User.* Englewood Cliffs, NJ: Prentice-Hall, 1987.

10. J. C. Kieffer, "Analysis of DC input response for a class of one-bit feedback encoders," *IEEE Trans. Commun.*, vol. COM-38, pp. 337-340, Mar. 1990.

11. D. F. Delchamps, "Exact asymptotic statistics for sigma-delta quantization noise," Proceedings *Twenty-Eighth Annual Allerton Conference on Communication, Control, and Computing*, October, 1990

12. L. Kuipers, H. Niederreiter, *Uniform Distribution of Sequences.* New York: John Wiley & Sons, 1974.

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

14. W. Feller, *An Introduction to Probability Theory and Its Applications.* Volume II, New York: John Wiley & Sons, 1966.

Parallel $\Delta\Sigma$ Modulation

Ian Galton

Abstract— Although $\Delta\Sigma$ modulators are popular candidates for low to moderate rate analog to digital converters, the requirement that their inputs be oversampled has discouraged their application to higher rate converters. Even in current applications, their accuracy is limited by the maximum available oversampling ratio. An approach to relaxing the oversampling requirement through parallel $\Delta\Sigma$ modulation is presented. By combining M $\Delta\Sigma$ modulators, each with an oversampling ratio of N, an effective oversampling ratio of approximately NM is achieved with only an M-fold increase in the quantization noise power. In particular, the special case of $N = 1$ allows for full-rate analog to digital conversion. The individual $\Delta\Sigma$ modulators can be any from a large class of popular $\Delta\Sigma$ modulators. Unlike most other approaches to trading modulator complexity for accuracy, the system retains the robustness of the individual $\Delta\Sigma$ modulators to circuit imperfections.

I. Introduction

Primarily because of advances in VLSI technology, $\Delta\Sigma$ modulator based A/D converters have become popular in applications requiring high precision. Although they employ complicated digital circuitry, their relatively simple analog circuitry tends to be robust with respect to component inaccuracies and noise [1]. They generally do not require the trimmed components or precise reference voltages necessary in conventional A/D converters. Since fine-line VLSI technology is more amenable to high density, high-speed digital circuitry than to accurate analog circuitry, $\Delta\Sigma$ modulator based converters are attractive candidates for VLSI implementation.

There are many types of $\Delta\Sigma$ modulators. From a signal processing point of

The author is with the Electrical Engineering Department, California Institute of Technology, 116-81, Pasadena, CA 91125; email address: galton@systems.caltech.edu

This work was supported by a grant from Pacific Bell.

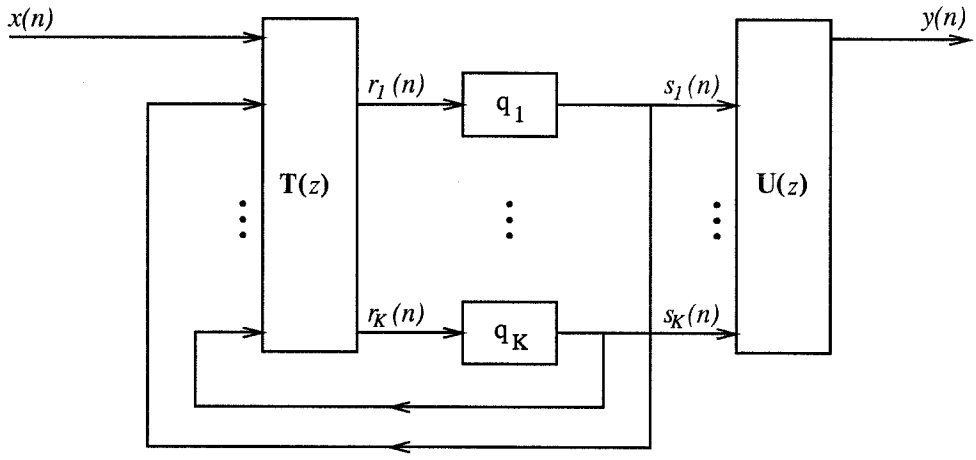


Figure 1: A generic $\Delta\Sigma$ modulator architecture.

view, most can be represented as special cases of the generic $\Delta\Sigma$ modulator shown in Figure 1 [2]. The system operates on a sampled-data input sequence, $x(n)$, and produces a quantized sampled-data output sequence, $y(n)$. It consists of a linear time-invariant (LTI) digital system, $\mathbf{T}(z)$, followed by a bank of quantizers followed by another LTI digital system, $\mathbf{U}(z)$. A feedback path joins the output of the quantizer bank to the input of $\mathbf{T}(z)$. In an actual implementation, $\mathbf{T}(z)$ would be a sampled-data analog circuit, the bank of quantizers would be a bank of low resolution A/D converters, $\mathbf{U}(z)$ would be a digital circuit, and the feedback path would contain a bank of digital-to-analog (D/A) converters†.

A $\Delta\Sigma$ modulator based oversampling A/D converter consists of a $\Delta\Sigma$ modulator, a lowpass filter, and an N-sample decimator as shown in Figure 2. The filter and decimator are together referred to as a *decimation filter.* Typically, the sampled-data input sequence corresponds to a continuous-time signal sampled at a rate Nf, where N is a positive integer referred to as the *oversampling ratio* and f is the Nyquist rate. This ensures that the spectrum of the input sequence is restricted

† Unfortunately, the term *digital* is not applied consistently throughout the literature. In signal processing literature a *digital filter* is a sampled data system whereas in the VLSI literature a *digital circuit* is a circuit in which voltage levels are assumed to have discrete values.

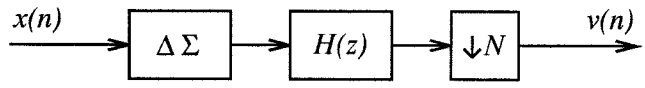


Figure 2: A conventional $\Delta\Sigma$ modulator based A/D converter with an oversampling ratio of N.

to $(-\frac{\pi}{N}, \frac{\pi}{N})$. The decimation filter reduces the rate of the output sequence to f.

The main idea behind all the $\Delta\Sigma$ modulator variations is simple. As shown in [2], the output of any $\Delta\Sigma$ modulator that fits the paradigm of Figure 1 is the sum of two components. The first component is a filtered version of the input sequence and the second component is a filtered version of the quantization noise sequences introduced by the quantizers. In most cases, the filter applied to the input sequence is just an L-sample delay while those applied to the quantization noise sequences are highpass filters. If the input sequence occupies the portion of the spectrum below the passbands of the filters applied to the quantization noise, the output of the $\Delta\Sigma$ modulator can be lowpass filtered to remove much of the remaining quantization noise without greatly distorting the input sequence component. It is for this reason that $\Delta\Sigma$ modulators find application in oversampled A/D converters. Oversampling ensures that the input sequence occupies only a low frequency portion of the spectrum, and decimation filtering removes the out-of-band quantization noise and reduces the output sample rate to the Nyquist rate.

The oversampling requirement is the essential drawback of $\Delta\Sigma$ modulator based converters; the circuitry must be designed to operate at a significantly higher rate than the system produces output samples. The greater the required accuracy of the A/D converted sequence, the larger the necessary oversampling ratio. Hence, accuracy is limited by circuit speed.

The proliferation of $\Delta\Sigma$ modulator architectures is a result of the continuing search for systems that require smaller oversampling ratios for a given level of

accuracy. Most of the research has emphasized designing the filters and topology of the $\Delta\Sigma$ modulator to increase the frequency band over which the quantization noise is attenuated. Because of the non-linearity introduced by the quantizers and the requirement that the topology of the system be amenable to VLSI implementation, this has proven to be a difficult problem. In particular, it is difficult to choose the architecture so as to minimize the required oversampling ratio while maintaining stability and a high tolerance to circuit imperfections.

We propose an alternative approach in which multiple $\Delta\Sigma$ modulator based converters are operated in parallel in such a way that an effective oversampling ratio is achieved that is significantly higher than the actual oversampling ratio. We call the architecture the $\Pi\Delta\Sigma$ modulator (the Π is a mnemonic for "parallel"). The primary advantage of the $\Pi\Delta\Sigma$ modulator is that it combines M $\Delta\Sigma$ modulator based converters with an oversampling ratio of N and achieves an accuracy commensurate with an oversampling ratio of approximately NM aside from an M-fold increase in the quantization noise power. For example, second-order $\Delta\Sigma$ modulators provide approximately 2.5 bits of accuracy for every doubling of the oversampling ratio, N [3]. Hence, for every doubling of M, the $\Pi\Delta\Sigma$ modulator would provide an additional 2 bits of accuracy (the M-fold increase in quantization noise power is responsible for the .5 bit difference between the two values). Another advantage of the $\Pi\Delta\Sigma$ modulator is that it retains the robustness properties of the individual $\Delta\Sigma$ modulators.

In the special case of $N = 1$, the $\Pi\Delta\Sigma$ modulator operates as a full-rate A/D converter; the input sample rate equals the output sample rate. The only other known practical A/D converter architecture with this property is the flash converter [4]. As will be demonstrated with simulations, the full-rate $\Pi\Delta\Sigma$ modulator compares favorably with the flash converter.

The chief drawback of the $\Pi\Delta\Sigma$ modulator is that it requires a large amount

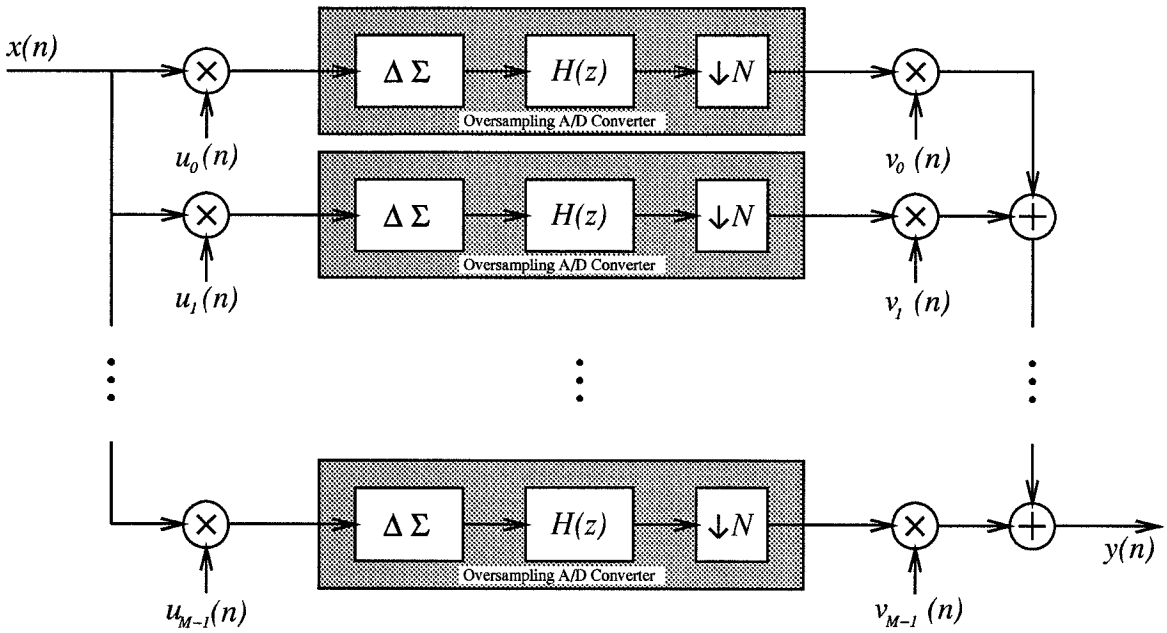


Figure 3: The $\Pi\Delta\Sigma$ modulator architecture.

of digital processing. The oversampling converter on each channel typically requires a filter of length $N(2M-1)$. Although the filters are simple integer FIR filters in which explicit multiplications can be avoided, they are likely to occupy most of the circuit space required by the $\Pi\Delta\Sigma$ modulator.

II. Architecture

The $\Pi\Delta\Sigma$ modulator architecture is shown in Figure 3. It consists of M *channels* that operate on the sampled-data input sequence in parallel. Each contains two binary multipliers capable of multiplying their inputs by plus or minus one and a $\Delta\Sigma$ modulator based oversampling A/D converter. The r^{th} channel multiplies the sampled-data input sequence by the internally generated sequence $u_r(n)$, performs an oversampled A/D conversion of the product, and multiplies the resulting digital sequence by the internally generated sequence $v_r(n)$. The output sequence is the digital sum of the output sequences from all the channels.

The sequences $u_r(n)$ and $v_r(n)$, $0 \leq r \leq M-1$, are referred to as *Hadamard*

modulation sequences and the process of multiplying by a Hadamard modulation sequence is referred to as *Hadamard modulation.* Each sequence is derived from an $M \times M$ Hadamard matrix, $\mathbf{H}$.† If $m(j,k)$, $0 \leq j,k \leq M-1$, is the element on the j^{th} row and k^{th} column of $\mathbf{H}$, then $u_r(n)$ and $v_r(n)$ are defined as follows:

$$u_r(n) = m\Big(r, \Big\lfloor \frac{n+L-1}{N} \Big\rfloor \bmod M\Big), \tag{1}$$

and

$$v_r(n) = m\big(r, n \bmod M\big), \tag{2}$$

where L is the signal delay of the $\Delta\Sigma$ modulators.‡ Figure 4 shows a legal set of Hadamard modulation sequences for the case of $M = 4$, $N = 3$, and $L = 1$. Since Hadamard matrices of a given size are not unique, other legal Hadamard modulation sequences exist.

Since Hadamard matrices consist solely of plus and minus ones, the Hadamard modulation sequences also consist solely of plus and minus ones. Hence, the multipliers need only pass or invert the sign of their input depending upon whether the current value of the modulation sequence is one or minus one, respectively. For the first multiplier on each channel, this requires the capability of analog sign inversion, and for the second multiplier it requires the capability of digital sign inversion. Indeed, the reason for using Hadamard modulation is that it simplifies the design of the multipliers. Although modulation sequences generated from any unitary matrix will work in the $\Pi\Delta\Sigma$ modulator framework, Hadamard sequences are the only such sequences consisting exclusively of plus and minus ones.

The use of Hadamard modulation, however, imposes a restriction on the number of channels, M. Specifically, M must be chosen such that there exists an $M \times M$

† A Hadamard matrix, $\mathbf{H}$, consists exclusively of plus and minus ones and has the property that $\mathbf{H}^T\mathbf{H} = M\mathbf{I}$ where $\mathbf{I}$ is the identity matrix [5]

‡ The brackets: $\lfloor\ \rfloor$, denote the floor function. For each $x \in \mathbf{R}$, $\lfloor x \rfloor$ equals the greatest integer less than or equal to x. For example, $\lfloor 3.2 \rfloor = 3$, and $\lfloor -3.2 \rfloor = -4$.

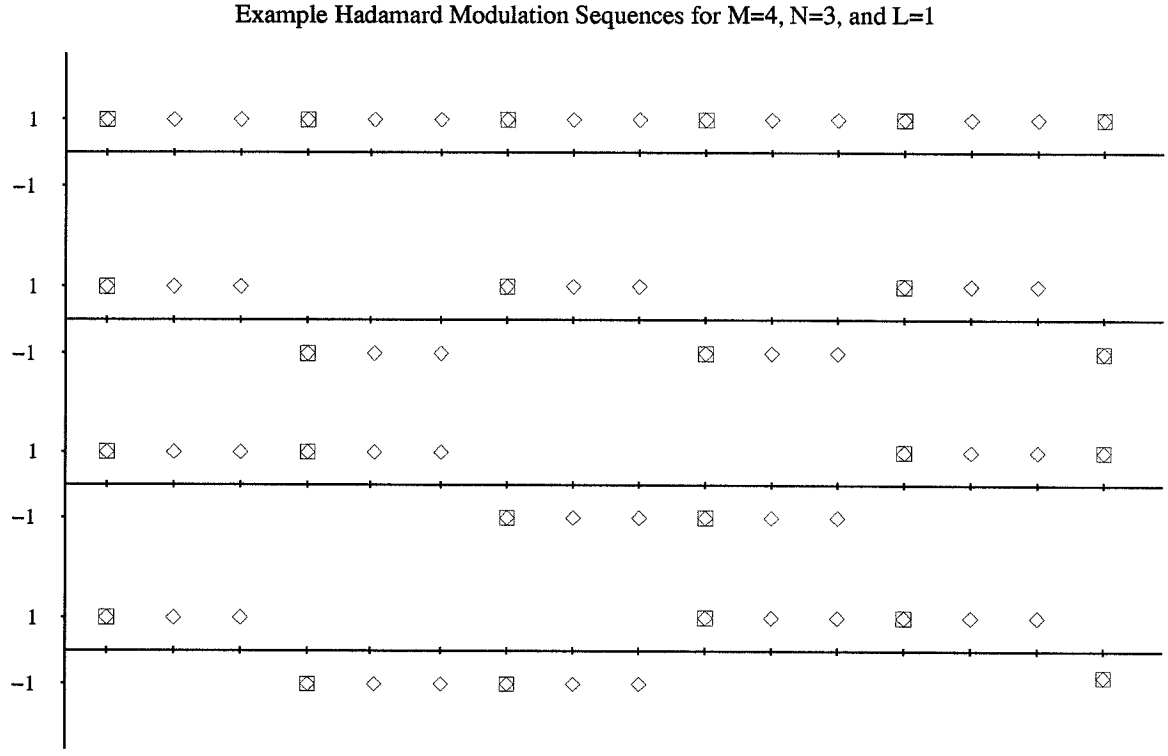


Figure 4: Example Hadamard modulation sequences for the $\Pi\Delta\Sigma$ modulator with $M = 4$, $N = 3$, and $L = 1$. The diamonds correspond to $u_r(n)$ and the squares correspond to $v_r(n)$. The graphs are plotted against time; the tick marks represent input sequence sample times. Each graph corresponds to one channel.

Hadamard matrix. A sufficient condition for this to occur is that M be a non-negative power of two. Several simple circuits for generating Hadamard modulation sequences when M is any non-negative power of two have been presented [6]-[9]. Hadamard matrices also exist for which M is not a power of two. A necessary condition is that M be a multiple of four [5]. Indeed, Hadamard matrices for every multiple of four less than 428 are known and mathematicians have conjectured (but not proven) that such matrices exist for all multiples of four.

The A/D converter on each channel consists of a $\Delta\Sigma$ modulator, a lowpass digital filter, and an N-sample decimator. The implementation of the A/D converters is the central design problem associated with the $\Pi\Delta\Sigma$ modulator. One question that must be answered is what type of $\Delta\Sigma$ modulator should be used. The answer depends in large measure on the application and is not greatly influenced by the use

of the $\Pi\Delta\Sigma$ modulator framework. However, for the sake of analysis we will place some restrictions on the class of $\Delta\Sigma$ modulators to be considered in the remainder of the paper. Specifically, we will assume that for each j, k, $1 \leq j, k \leq K$: 1) the k^{th} quantizer is a uniform midrise quantizer with step-size Δ_k and no-overload range $[-\gamma_k, \gamma_k)$, 2) the impulse response joining $x(n)$ to $r_k(n)$ does not converge to zero, 3) the impulse response joining $x_j(n)$ to $r_k(n)$ is a sequence of integer multiples of Δ_k/Δ_j, and 4) the equivalent filter applied to the input sequence is just an L-sample delay. Most of the commonly used $\Delta\Sigma$ modulators satisfy these requirements in principle although, in practice, leaky summing nodes and filter gain errors can give rise to systems that slightly violate the assumptions. Nevertheless, we will show that the theoretical results are approximately valid when the $\Delta\Sigma$ modulators suffer from such imperfections. We also impose the restriction that if different types of $\Delta\Sigma$ modulators are used in the same $\Pi\Delta\Sigma$ modulator, they must have the same signal delay, L.

Another question associated with the design of the oversampling A/D converters is what frequency response should the decimation filter, $H(z)$, have. Again, the answer is largely dependent upon the application, although the $\Pi\Delta\Sigma$ modulator framework does impose a restriction upon the length of the filter. Specifically, we require that the filter have a length no greater than $N(2M-1)$. In general, $H(z)$ should be designed as if the $\Delta\Sigma$ modulator and the filter were to be used in isolation with an oversampling ratio of NM. The reason for this filter choice and the length restriction will become clear from the analysis performed in the next section.

III. Analysis

The idea behind the $\Pi\Delta\Sigma$ modulator is simple. As in conventional $\Delta\Sigma$ modulator based converters the goal is to filter out as much of the quantization noise as possible without significantly distorting the input sequence. As described above, the

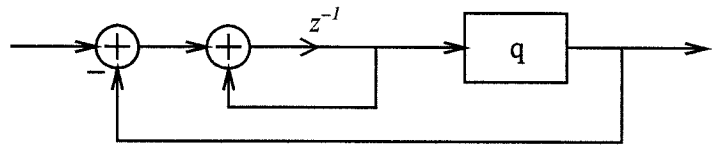


Figure 5: The first-order $\Delta\Sigma$ modulator.

$\Delta\Sigma$ modulator applies a highpass filter to the quantization noise without changing the spectral content of the input sequence. In a conventional $\Delta\Sigma$ modulator based converter, the signal and the highpass filtered noise are lowpass decimation filtered together. In contrast, the $\Pi\Delta\Sigma$ modulator effectively applies a different lowpass filter to the input sequence than it does to the highpass filtered noise. The input sequence sees a filter with a wide passband while the highpass filtered noise sees a filter with a narrow passband. Hence, more of the quantization noise is removed than in the conventional system.

For example, consider a sixteen channel $\Pi\Delta\Sigma$ modulator employing the single-loop $\Delta\Sigma$ modulator shown in Figure 5 and an oversampling ratio of fifteen. Take the filters, $H(z)$, to have the triangular impulse response of length $N(2M-1) = 465$ given by

$$h(n) = \begin{cases} \frac{1}{233^2}(n+1), & \text{if } 0 \leq n \leq 232; \\ \frac{1}{233^2}(464-n), & \text{if } 233 \leq n \leq 464; \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

From Figure 5 it is easy to verify that each $\Delta\Sigma$ modulator filters the quantization noise by $1-z^{-1}$, while subjecting the $\Delta\Sigma$ modulator input sequence to only a delay, z^{-1}. The frequency response of this noise filter is shown as the dashed-dotted curve in Figure 6. The frequency response of $H(z)$ is shown as the solid curve in Figure 6. Together, these two filters attenuate the quantization noise on each channel. As will be shown, aside from adding quantization noise, the overall effect of the $\Pi\Delta\Sigma$ modulator on the input sequence is to apply a filter, $H'(z)$, with the response shown as the dashed curve in Figure 6. Because the input sequence is oversampled by a

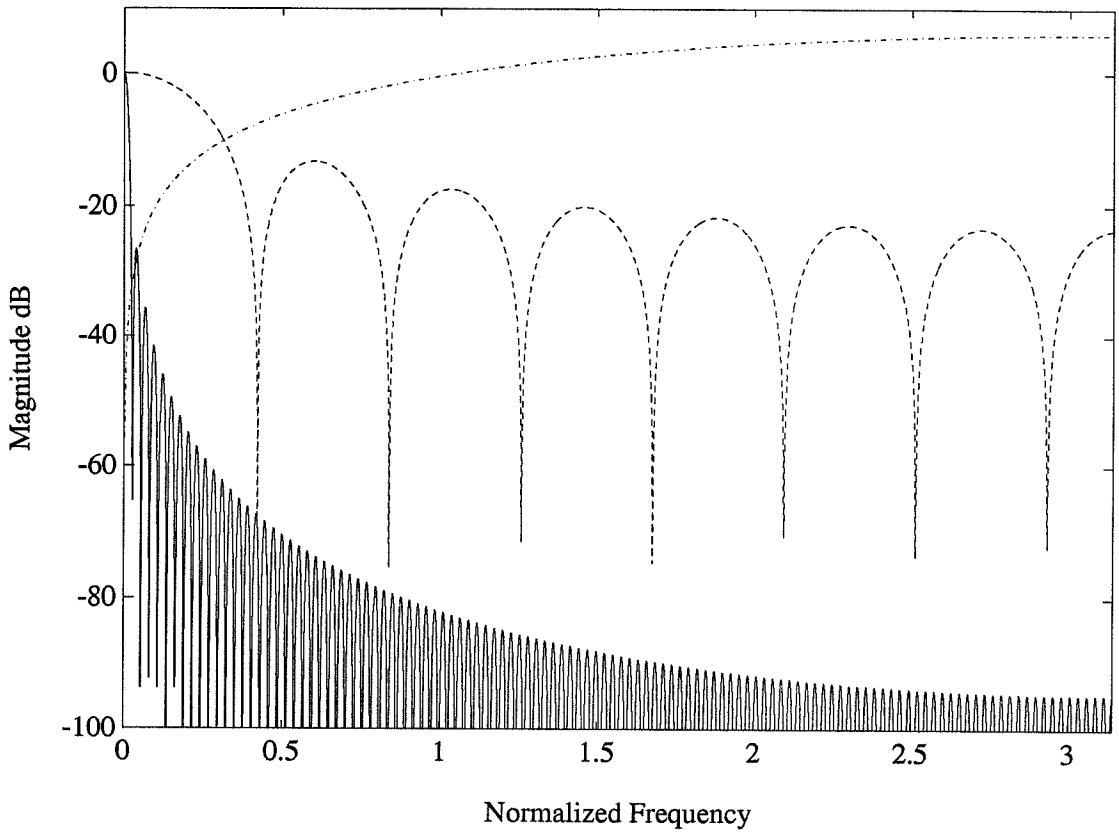


Figure 6: Filtering performed by the $\Pi\Delta\Sigma$ modulator. The dashed-dotted curve is filter response applied to the quantization noise on each channel by the $\Delta\Sigma$ modulator. The solid curve is the filter response applied to the quantization noise on each channel by the decimation filter, $H(z)$. The dashed curve is the overall filter response applied to the input sequence by the $\Pi\Delta\Sigma$ modulator.

factor of fifteen, its energy is restricted to the frequency band $(-\frac{\pi}{15}, \frac{\pi}{15})$. This is sufficiently narrow that $H'(z)$ does not distort the input sequence beyond repair. Hence, the quantization noise on each channel is highly attenuated, while the input sequence is preserved.

As we will show, although the outputs of the channels are summed, the quantization noise does not add coherently. In the example above, summing the channels increases the noise power by a factor of sixteen. This raises the noise floor by about 12dB or, equivalently, two bits of precision are lost. The increase in noise power is more than made up for by the reduction in quantization noise achieved by each channel.

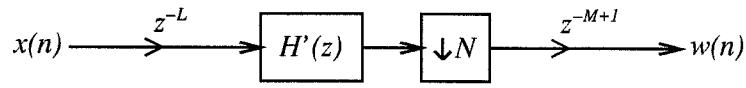


Figure 7: A system equivalent to the ΠΔΣ modulator when the quantizers are replaced by identity operators.

In the example, two results are alluded to: 1) the component of the ΠΔΣ modulator output not corresponding to quantization noise is just the input sequence passed through a filter, $H'(z)$, with a passband wider than $(-\frac{\pi}{N}, \frac{\pi}{N})$, and 2) the quantization noise from the channels does not add coherently. Our analysis consists of presenting and proving mathematically precise formulations of these results.

The Filter Applied to The Input Sequence

We first define $H'(z)$ and prove that the overall effect of the ΠΔΣ modulator aside from generating quantization noise is to filter the input sequence by $H'(z)$. The simplest approach is to analyze a system that is equivalent to the ΠΔΣ modulator except that the quantization noise sequences are all zero. Such a system can be obtained by simply replacing the quantizers in the ΠΔΣ modulator with identity operators. The following theorem formally states the result. It also indicates that the impulse response of $H'(z)$ is equal to the center N samples of the impulse response of $H(z)$.

Theorem 1: The system obtained by replacing the quantizers in the ΠΔΣ modulator with identity operators is equivalent to the system shown in Figure 7. The impulse response of $H'(z)$ is

$$h'(n) = \begin{cases} Mh(n + NM - N), & \text{if } 0 \leq n \leq N-1; \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

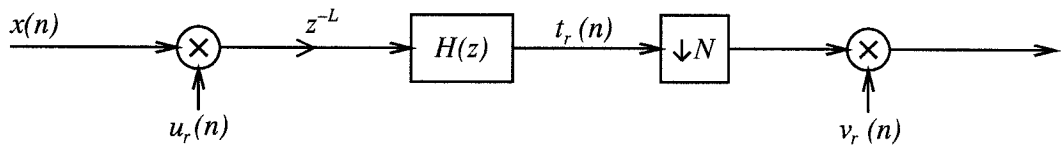


Figure 8: The r^{th} channel of the $\Pi\Delta\Sigma$ modulator with the quantizers replaced by identity operators.

Proof: By definition, each $\Delta\Sigma$ modulator in the $\Pi\Delta\Sigma$ modulator reduces to an L-sample delay when its quantizers are replaced by identity operators. Figure 8 shows the r^{th} channel of the $\Pi\Delta\Sigma$ modulator with the $\Delta\Sigma$ modulator replaced by an L-sample delay. From the figure and the definition of $u_r(n)$, it follows that

$$t_r(n) = \sum_{k=0}^{N(2M-1)-1} h(k)x(n-L-k)v_r\left(\left\lfloor \frac{n-1-k}{N} \right\rfloor\right).$$

Let $w'(n)$ be the output of the system obtained by replacing the quantizers in the $\Pi\Delta\Sigma$ modulator with identity operators. Applying the N-sample decimator, Hadamard modulating, and summing the $t_r(n)$ for $0 \le r \le M-1$ gives

$$\begin{aligned} w'(n) &= \sum_{r=0}^{M-1} t_r(nN)v_r(n) \\ &= \sum_{k=0}^{N(2M-1)-1} h(k)x(Nn-L-k) \sum_{r=0}^{M-1} v_r(n)v_r\left(\left\lfloor \frac{nN-1-k}{N} \right\rfloor\right). \end{aligned}$$

Let

$$A(k) = \sum_{r=0}^{M-1} v_r(n)v_r\left(\left\lfloor \frac{nN-1-k}{N} \right\rfloor\right)$$

when $0 \le k \le N(2M-1)-1$ and let $A(k) = 0$ for all other values of k. Since $v_r(n)$ is defined as the r^{th} row of an $M \times M$ Hadamard matrix repeated periodically, it follows that

$$\sum_{r=0}^{M-1} v_r(n)v_r(n+m) = \begin{cases} M, & \text{if } m \text{ is a multiple of } M; \\ 0, & \text{otherwise.} \end{cases}$$

Hence, A(k) must either equal zero or M depending upon the value of k. In the range $0 \le k \le N(2M-1)-1$, we have

$$n-(2M-1) \le \left\lfloor \frac{nN-1-k}{N} \right\rfloor \le n-1.$$

Therefore, $A(k) = 0$ unless

$$\left\lfloor \frac{nN - 1 - k}{N} \right\rfloor = n - M,$$

in which case it equals M. This occurs when $NM - N \leq k \leq NM - 1$. It follows that

$$w'(n) = M \sum_{k=NM-N}^{NM-1} h(k)x(Nn - L - k).$$

Applying (4) gives

$$w'(n) = \sum_{k=0}^{N-1} h'(k)x\big(N(n - M + 1) - L - k\big).$$

By inspection, this is equal to $w(n)$, the output of the system shown in Figure 7.

■

Independence of the Channels

We now turn to the problem of formulating and proving mathematically precise versions of the assertion that the quantization noise from the M channels of the $\Pi\Delta\Sigma$ modulator does not add coherently. In particular, making two assumptions about the input sequence, we will show that the statistical and time average cross correlations of the quantization error sequences produced by different channels of the $\Pi\Delta\Sigma$ modulator are zero. Hence, the mean squared quantization error of the $\Pi\Delta\Sigma$ modulator is the sum of the mean squared quantization errors of the individual channels.

The first of the two assumptions is that analog circuit noise gives rise to an *actual input sequence*, $x(n)$, consisting of a *desired input sequence*, $x_d(n)$, plus an *input noise sequence*, $\{\eta_n\}$, where the η_n are independent identically distributed (iid) random variables that are independent of $x_d(n)$ and whose distribution is not a so-called *lattice distribution* [10]:

$$x(n) = x_d(n) + \eta_n. \tag{5}$$

The desired input sequence is the sampled-data sequence that is to be converted into a quantized sequence by the $\Pi\Delta\Sigma$ modulator. The η_n sequence arises in the circuitry of the $\Pi\Delta\Sigma$ modulator and is not related to the desired input sequence. Although it is assumed to be unavoidable, it can have an arbitrarily small variance. The assumption is essentially always valid in physical implementations [11], [2]. For example, thermal noise is ubiquitous in analog circuitry and can be modeled accurately as an iid random sequence in sampled-data systems.

The second assumption is that the input sequence does not overload the $\Delta\Sigma$ modulators. That is, we assume the input sequence is such that the quantizers in the $\Delta\Sigma$ modulators never overload. Although many useful types of $\Delta\Sigma$ modulators do overload during normal operation, it is usually the case that they overload at only a small percentage of the sample times [3]. In such cases, it has been shown that the effects of overload on $\Delta\Sigma$ modulator performance are not severe [2]. It follows that in such cases the degradation in performance experienced by the $\Pi\Delta\Sigma$ modulator should also not be severe.

In the calculations to follow, we consider the $\Pi\Delta\Sigma$ modulator to have been "turned on" at a specific time in the past. That is, we assume there exists some integer, a, such that for each $n < a$, the states of the filters within the $\Pi\Delta\Sigma$ modulator are all zero. Whenever we consider quantization error sequences, we will take them to be zero for all $n \leq a$. In some cases, we will consider the system in the limit as $a \to -\infty$. This corresponds to a system that has been running since the beginning of time.

We will write the component of the output of the $\Pi\Delta\Sigma$ modulator corresponding to quantization noise as $e(n)$ and refer to it as the *overall quantization error sequence*. Hence, $y(n) = w(n) + e(n)$. The overall quantization error sequence is the sum of the quantization error sequences arising on the individual channels. Let $e_r(n)$, $0 \leq r \leq M-1$, be the quantization error at the output of $H(z)$ on the

r^{th} channel of the $\Pi\Delta\Sigma$ modulator. From Figure 3, the overall quantization error sequence of the $\Pi\Delta\Sigma$ modulator is

$$e(n) = \sum_{r=0}^{M-1} v_r(n) e_r(Nn). \tag{6}$$

In accordance with the usual definitions, we will take the mean and autocorrelation of the overall quantization error sequence to be

$$M_e(n) = \lim_{a \to -\infty} \mathrm{E}\big[e(n)\big],$$

and

$$R_{ee}(n,p) = \lim_{a \to -\infty} \mathrm{E}\big[e(n)e(n+p)\big],$$

respectively. We will take the cross correlation of the overall quantization error sequence and the desired input sequence to be

$$R_{x_d e}(n,p) = \lim_{a \to -\infty} \mathrm{E}\big[x_d(n)e(n+p)\big].$$

The following theorem asserts that the overall quantization error is not statistically biased or correlated to the input sequence.

Theorem 2: $M_e(n) = 0$ and $R_{x_d e}(n,p) = 0$

Proof: Because of the linearity of the limit and expectation operators,

$$\lim_{a \to -\infty} \mathrm{E}\big[e(n)\big] = \sum_{r=0}^{M-1} v_r(n) \lim_{a \to \infty} \mathrm{E}\big[e_r(Nn)\big].$$

The result follows from application of the corresponding result for $\Delta\Sigma$ modulators proved in [2]. The proof that $R_{x_d e}(n,p) = 0$ similarly follows from the corresponding result for $\Delta\Sigma$ modulators proved in [2].

■

Theorem 2 implies that the second order statistics of the overall quantization error sequence are completely characterized by $R_{ee}(n,p)$. Provided the autocorrelations exist,

$$R_{yy}(n,p) = R_{ww}(n,p) + R_{ee}(n,p).$$

The following theorem and corollary assert that $R_{ee}(n,p)$ is the sum of the autocorrelations of the individual channel quantization error sequences. As shown in [11] and [2], these autocorrelations exist so it follows that $R_{ee}(n,p)$ also exists. The theorem states that the expectation over the random variables, η_n, of the product of the quantization error from different channels converges to zero uniformly with respect to the desired input sequence as the run-time of the $\Pi\Delta\Sigma$ modulator increases to infinity. The corollary deduces the form of the overall quantization error autocorrelation.

Theorem 3: Whenever $r \neq q$, for each integer $b \geq a$,

$$\mathop{\mathrm{E}}_{\{\eta_k : k>b\}} \left[e_r(n) e_q(n+p) \right] \rightarrow 0$$

uniformly with respect to the desired input sequence, $x_d(n)$, and the variables $\{\eta_a, \ldots, \eta_b\}$ as either $n \rightarrow \infty$ or as $b \rightarrow -\infty$.

Proof: We will prove the result for $b \rightarrow -\infty$. The proof for $n \rightarrow \infty$ is similar.

Applying the results presented in [2], for each $n \geq a$ we can write

$$e_r(n) = \sum_{k=0}^{K} \sum_{j=a}^{n} g_{r,k}(n-j) \epsilon_{r,k}(a,j), \tag{7}$$

where[†]

$$\epsilon_{r,k}(a,j) = \frac{1}{2} - \left\langle \sum_{i=a}^{j-L} f_{r,k}(j-i) u_r(i) x(i) + \delta(j) \right] \Big\rangle. \tag{8}$$

Here $g_{r,k}(n)$ is an absolutely summable sequence, $f_{r,k}(n)$ is a sequence that does not converge to zero, $\delta(j)$ is a sequence that takes on values of $\frac{1}{2}$ and zero, and L is the signal delay of the $\Delta\Sigma$ modulators in the $\Pi\Delta\Sigma$ modulator.

Because $g_{r,k}(n)$ is absolutely summable and the $\epsilon_{r,k}(a,j)$ terms form a bounded set, we need only show that the expectation with respect to η of the product of any

[†] The angle brackets denote the fractional part operator. This operator is defined as: $\langle x \rangle = x - \lfloor x \rfloor$ for all $x \in \mathbf{R}$.

term of $e_r(n)$ with any term of $e_q(m)$ converges to zero uniformly with respect to the desired input sequence as $a \to -\infty$.

Choose any non-negative integers n, m, k_r, and k_q. For each $p = 0, 1, \ldots$, define

$$X_p = \frac{1}{2} - \left\langle a_p + \sum_{i=n-p}^{n-L} f_{r,k_r}(n-i)u_r(i)\eta_i \right\rangle,$$

and

$$Y_p = \frac{1}{2} - \left\langle b_p + \sum_{i=m-p}^{m-L} f_{q,k_q}(m-i)u_q(i)\eta_i \right\rangle,$$

where $\{a_k\}$ and $\{b_k\}$ are any real sequences. In particular, from (8) we could choose the $\{a_k\}$ and $\{b_k\}$ such that

$$X_p = \epsilon_{r,k_r}(n-p, n),$$

and

$$Y_p = \epsilon_{q,k_q}(m-p, m).$$

Hence, it is sufficient to show that $\mathrm{E}[X_p, Y_p] \to 0$ uniformly with respect to $\{a_k\}$ and $\{b_k\}$ as $p \to \infty$.

Recall that the actual input sequence, $x(n)$, is the sum of the desired input sequence, $x_d(n)$, and the input noise sequence, $\{\eta_n\}$. By definition, X_p and Y_p are piecewise continuous functions of each η_n with range $(-\frac{1}{2}, \frac{1}{2}]$. Therefore, they are random variables with distributions supported on $(-\frac{1}{2}, \frac{1}{2}]$.

Applying Lemma A1 (presented in the appendix), it follows that X_p and Y_p converge in distribution to random variables X and Y respectively, where X and Y are statistically independent and uniformly distributed on $(-\frac{1}{2}, \frac{1}{2}]$. Since X_p and Y_p are piecewise continuous and bounded, this implies

$$\begin{aligned} \lim_{p\to\infty} \mathrm{E}[X_p Y_p] &= \mathrm{E}[XY] \\ &= 0 \end{aligned} \tag{9}$$

for every choice of the desired input sequence.†

It remains to show that the convergence is uniform with respect to $\{a_k\}$ and $\{b_k\}$. By definition, $\mathrm{E}[X_pY_p]$ is a function of only the p^{th} members of the sequences $\{a_k\}$ and $\{b_k\}$. Since X_pY_p is a piecewise continuous function of a_p and b_p, there exist sequences $\{a'_k\}$ and $\{b'_k\}$ such that for every p

$$\mathrm{E}[X_pY_p]\,\big|_{a_p=a'_p,b_p=b'_p} = \sup_{\{a_k\},\{b_k\}} \mathrm{E}[X_pY_p].$$

Since (9) holds for all sequences $\{a_k\}$ and $\{b_k\}$ it must hold for the specific sequences $\{a'_k\}$ and $\{b'_k\}$. It follows that

$$\lim_{p\to\infty} \sup_{\{a_k\},\{b_k\}} \mathrm{E}[X_pY_p] = 0,$$

which implies that the convergence in (9) is uniform with respect to $\{a_k\}$ and $\{b_k\}$.

■

Corollary 4: The autocorrelation of the overall error sequence, can be written as

$$R_{ee}(n,p) = \sum_{r=0}^{M-1} v_r(n)v_r(m)R_{e_re_r}(nN,pN),$$

where $R_{e_re_r}(n,p)$ is the autocorrelation of $e_r(n)$.

Proof: From (6) and the definition of the autocorrelation,

$$\begin{aligned} R_{ee}(n,p) = &\sum_{r=0}^{M-1} v_r(n)v_r(n+p)R_{e_re_r}(nN,pN) \\ &+ \sum_{\substack{r=0 \\ r\neq q}}^{M-1}\sum_{q=0}^{M-1} v_r(n)v_q(n+p) \lim_{a\to-\infty} \mathrm{E}\big[e_r(nN)e_q((n+p)N)\big]. \end{aligned} \tag{10}$$

Therefore, we need only show that the second sum in (10) is zero. Since the input noise sequence is independent of the desired input sequence, we can write

$$\mathrm{E}[e_r(n)e_r(m)] = \mathop{\mathrm{E}}_{\{x_d(k)\}}\Big\{\mathop{\mathrm{E}}_{\eta}[e_r(n)e_r(m)]\Big\}.$$

† See, for example, Theorem 29.1 of [10].

By Theorem 3 the inner expectation converges to zero uniformly as $a \to -\infty$. It follows that $\lim_{a\to-\infty} \mathrm{E}[e_r(n)e_q(m)] = 0$ whenever $r \neq q$. Therefore, the second sum in (10) is zero.

■

Of course, the statistical averages are of little practical use if they do not relate to the corresponding time averages. Therefore, the next problem is to prove that the overall quantization error sequence is *mean* and *correlation ergodic.* That is, we wish to establish relationships between time averages of $e(n)$, $x_d(n)e(n+p)$, and $e(n)e(n+p)$ and the functions $M_e(n)$, $R_{x_de}(n)$, and $R_{ee}(n,p)$, respectively. Note that we are forced to average $R_{ee}(n,p)$ over n because, in general, $e(n)$ is not wide-sense stationary [11].

Theorem 5: The following equations

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} e(n) = 0 \tag{11}$$

and

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} x_d(n)e(n+p) = 0 \tag{12}$$

hold in probability. Moreover, whenever one of the limits exist,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} e(n)e(n+p) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} R_{ee}(n,p) \tag{13}$$

holds in probability.

Proof: The proofs that (11) and (12) hold in probability follow directly from (6) and the corresponding results in [11] and [2] for $\Delta\Sigma$ modulators.

From (6) we can write

$$\frac{1}{N}\sum_{n=0}^{N-1} e(n)e(n+p) = \frac{1}{N}\sum_{n=0}^{N-1}\sum_{r=0}^{M-1} v_r(n)v_r(n+p)e_r\big(nN\big)e_r\big((n+p)N\big) + \frac{1}{N}\sum_{n=0}^{N-1}\sum_{\substack{r=0 \\ r\neq q}}^{M-1}\sum_{q=0}^{M-1} v_r(n)v_r(n+p)e_r\big(nN\big)e_q\big((n+p)N\big). \tag{14}$$

As shown in [11] and [2], whenever either limit exists the following equation holds in probability:

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} e_r(n)e_r(n+p) = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} R_{e_re_r}(n,p).$$

This with Corollary 4 implies that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\sum_{r=0}^{M-1} v_r(n)v_r(n+p)e_r\big(nN\big)e_r\big((n+p)N\big) = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} R_{ee}(n,p)$$

holds in probability whenever either limit exists. Therefore, we need only show that the second term in (14) is zero in probability, or equivalently that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} e_r(n)e_q(n+p) = 0 \tag{15}$$

holds in probability when $r \neq q$. But this follows directly from Theorem 3 and Lemma A2.

■

The Effective Oversampling Ratio

We have shown: 1) that the output of the ΠΔΣ modulator consists of an overall quantization error sequence plus a filtered version of the input sequence, 2) that the autocorrelation of the overall quantization error sequence is the sum of the autocorrelations of the quantization error sequences from each channel, and 3) that the overall quantization error sequence is mean and correlation ergodic. It remains to explain the claim that the ΠΔΣ modulator achieves an effective oversampling

ratio approximately M times greater than that of its oversampling A/D converters aside from the M-fold increase in quantization noise power.

Recall that in a conventional $\Delta\Sigma$ modulator based converter, for an oversampling ratio of say N, the lowpass decimation filter must be chosen to remove as much of the out of band quantization noise as possible under the constraint that it have a passband of at least $(-\frac{\pi}{N}, \frac{\pi}{N})$. If its passband is too narrow, the desired component of the output, namely the delayed version of the input sequence, is distorted beyond repair. However, the filters $H(z)$ in the oversampling A/D converters of the $\Pi\Delta\Sigma$ modulator are not subject to this constraint. Instead, they are subject to two milder constraints. The first is that they must have a length no greater than $N(2M-1)$. The second is that the filter $H'(z)$, whose impulse response is the center N samples of the impulse response of $H(z)$, must have a passband of at least $(-\frac{\pi}{N}, \frac{\pi}{N})$. The length constraint is not unreasonable for decimation filters in conventional $\Delta\Sigma$ modulator based converters with oversampling ratios of NM [3]. Moreover, it is easy to verify that most such filters satisfy the passband constraint on $H'(z)$. The filter with the impulse response (3) is one such example.

IV. Sensitivity to Nonidealities

In the analysis of the preceding section, we took the $\Pi\Delta\Sigma$ modulator to consist of ideal components and assumed that the $\Delta\Sigma$ modulators never overload. In practice, however, components are never ideal and the $\Delta\Sigma$ modulators sometimes do overload. It is important to have an idea how the performance of the $\Pi\Delta\Sigma$ modulator deteriorates in the face of these nonidealities.

Generally, there are three classes of nonideal component behavior that degrade the accuracy of the $\Pi\Delta\Sigma$ modulator. One is analog circuit noise. As shown in the preceding section, some analog circuit noise is actually beneficial because it gives rise to the channel independence property of Theorem 3. However, excessive cir-

cuit noise degrades the accuracy of the $\Delta\Sigma$ modulators. Another class of nonideal component behavior increases the quantization error. For example, quantizers with non-uniform step sizes can increase the overall quantization error. Finally, nonideal analog filters can cause the $\Delta\Sigma$ modulators to distort the component of the output not corresponding to quantization error and circuit noise. Although the distortion can be nonlinear, it can often be accurately modeled as linear distortion [3], [12]. Therefore, aside from adding quantization and circuit noise, a nonideal $\Delta\Sigma$ modulator applies the filter $z^{-L} + D(z)$ to the input sequence where $D(z)$ is an unintentionally added term that we refer to as the *distortion filter.*

The $\Pi\Delta\Sigma$ modulator tends to be robust with respect to circuit noise originating in the analog multipliers and $\Delta\Sigma$ modulators and to nonideal components that increase quantization error. The reason is that the additional error contributed by each channel tends to either be independent from that of the other channels (as in the case of thermal noise), or becomes uncorrelated because of the Hadamard modulation performed prior to summing the channels. Hence, the performance degradation on each channel adds in power to produce the overall performance degradation. That is, if P_r is the power of the error resulting from the nonidealities on the r^{th} channel, the overall power of the additional error is just $\sum_{r=0}^{M-1} P_r$.

The behavior of the $\Pi\Delta\Sigma$ modulator with respect to nonideal analog filters that distort the component of the output not corresponding to circuit noise and quantization error is more difficult to characterize. For each r, let $D_r(z)$ be the distortion filter applied by the $\Delta\Sigma$ modulator on the r^{th} channel. Define $e_d(n)$ to be the component of the $\Pi\Delta\Sigma$ modulator output corresponding to these distortion filters. Proceeding as in the proof of Theorem 1, we can write

$$e_d(n) = \sum_{k=0}^{\infty} x(Nn - L - k)\widehat{A}(n, k), \tag{16}$$

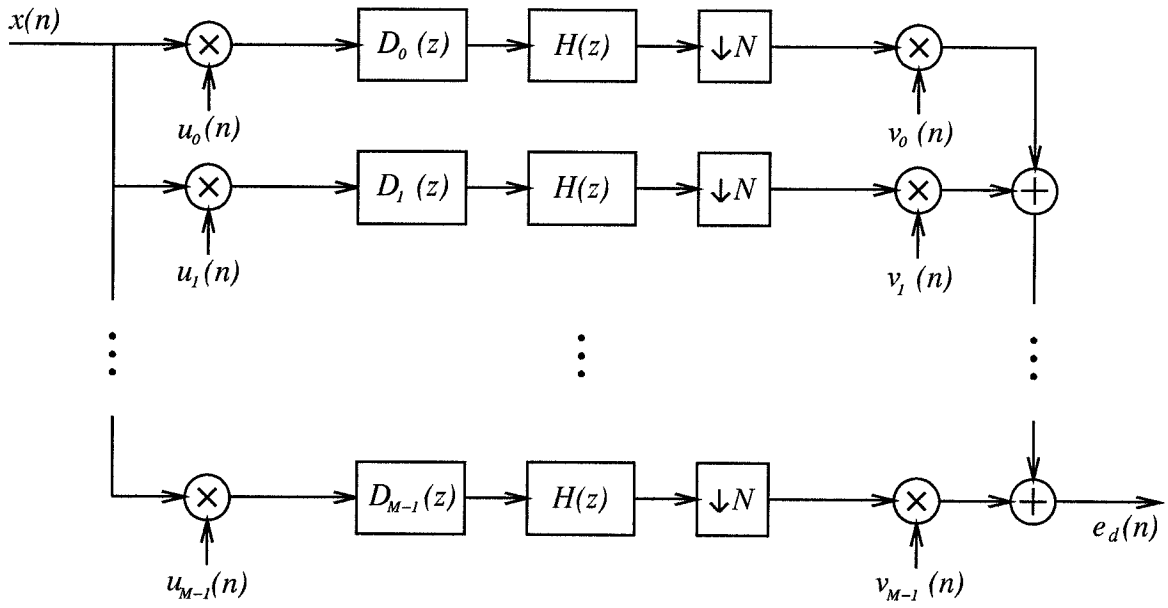


Figure 9: The effect of the distortion filters, $D_r(z)$, in the $\Delta\Sigma$ modulators.

where

$$\widehat{A}(n,k) = \sum_{r=0}^{M-1} (d_r * h)(k) v_r(n) v_r\Big(n - \Big\lceil \frac{k+1}{N} \Big\rceil\Big), \tag{17}$$

and $(d_r * h)(k)$ is the convolution of the impulse responses of $H(z)$ and $D_r(z)$. Since $v_r(n)$ is periodic with period M, it follows that $\widehat{A}(n,k)$ is periodic in n with period M. Hence, $\widehat{A}(n,k)$ can be thought of as the impulse response of a linear periodically time varying filter as shown in Figure 9. In general, the power of the distortion error depends upon the nature of the distortion filters. For example, from both Figure 9 and (17), it is evident that if $D_r(z)$ does not pass significant energy in the passband of $H(z)$, namely $(-\frac{\pi}{NM}, \frac{\pi}{NM})$, then the power of $e_d(n)$ will tend to be low.

It is often the case that the $\Delta\Sigma$ modulators will have similar distortion filters. For example, if the $\Delta\Sigma$ modulators contain leaky integrators of the same form and are otherwise essentially ideal, each will have a distortion filter of the same form. In the extreme, $D_r(z) = D(z)$ for each $0 \le r \le M-1$, in which case $\widehat{A}(n,k)$ reduces to

$$\widehat{A}(n,k) = \begin{cases} M(d*h)(k), & \text{if } NM - N \le k \bmod NM \le NM - 1; \\ 0, & \text{otherwise.} \end{cases}$$

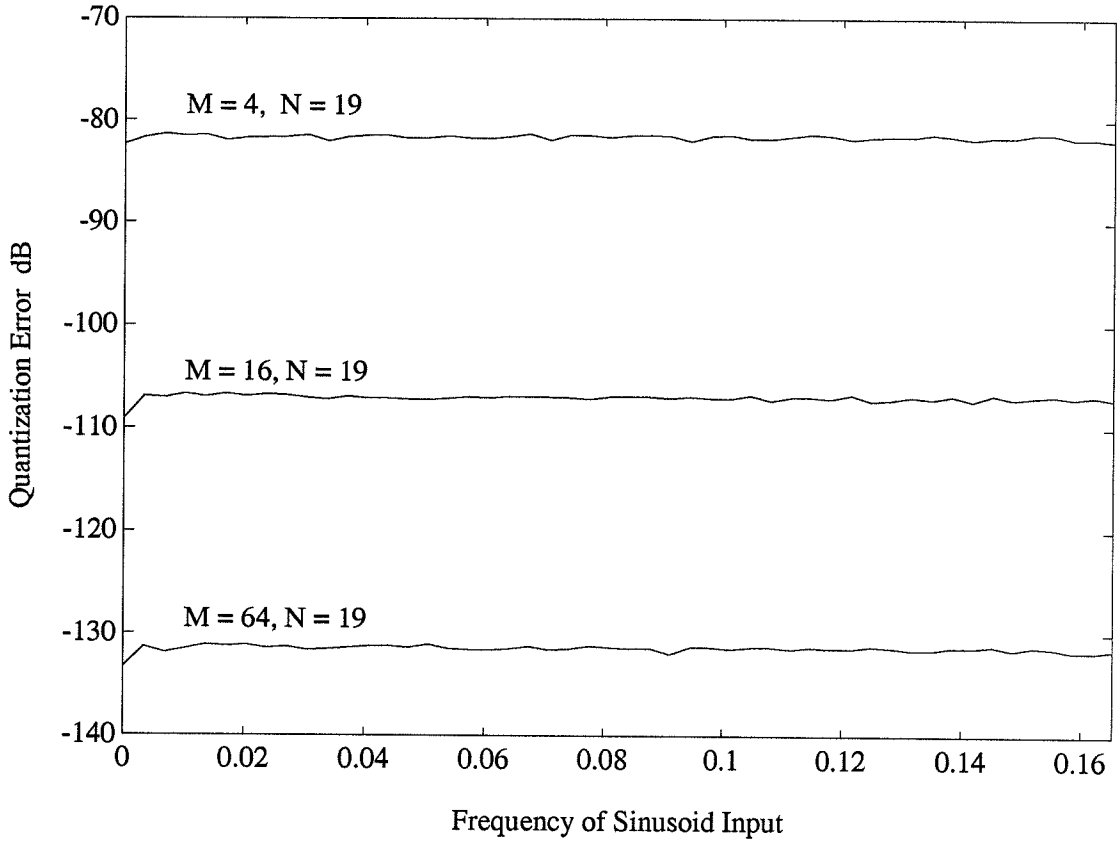


Figure 10: The quantization error power of the $\Pi\Delta\Sigma$ modulator with a sinusoidal input versus the frequency of the sine wave. Data for four, sixteen, and sixty-four channel $\Pi\Delta\Sigma$ modulators each with an oversampling ratio of nineteen are shown. Second-order double-loop $\Delta\Sigma$ modulators with four level quantizers and $\Delta = 1$ are used. The input has an amplitude of 0.5.

Hence, if all the distortion filters are equal, $\widehat{A}(n,k)$, loses its dependence on n and becomes a linear time invariant filter. In this case, the distortion could be compensated by digital filtering following the $\Pi\Delta\Sigma$ modulator.

V. Simulations

In this section we present $\Pi\Delta\Sigma$ modulator simulation results. In each simulation, second-order double-loop $\Delta\Sigma$ modulators [3] with four level quantizers for which $\Delta = 1$ were used. Random numbers of variance $8.3 \cdot 10^{-6}$ were added to the inputs to simulate the effect of the η_n in (5).

Figure 10 shows simulations of four, sixteen, and sixty-four channel $\Pi\Delta\Sigma$ mod-

ulators with oversampling ratios of nineteen and sine wave inputs of amplitude 0.5. The quantization error power of each $\Pi\Delta\Sigma$ modulator is plotted against the frequency of the sine wave. The frequency range shown is $[0, \frac{\pi}{19})$ which corresponds to full bandwidth after decimation by nineteen.

We see that the quantization error power is not dependent upon the frequency of the input. If the $\Pi\Delta\Sigma$ modulator were a linear system, this would fully characterize the mean-squared quantization error performance. However, since the $\Pi\Delta\Sigma$ modulator is not a linear system, it is possible that the quantization error power might be different for other types of input sequences. Nevertheless, simulations with other non-overloading input sequences including finite sums of sinusoids and various colored random sequences do not indicate such a dependence. The same results were obtained for other oversampling ratios including the special case of $N = 1$.

It is customary to refer to the accuracy of a $\Delta\Sigma$ modulator based converter in terms of the number of bits that a uniform quantizer would require to generate the same quantization error power [1]. A frequently used formula relating bits of accuracy to the quantization error power of a non-overloaded uniform quantizer is

$$R = \tfrac{1}{2}\log_2 \frac{\gamma^2}{3\sigma^2},$$

where R is the number of bits, $(-\gamma, \gamma]$ is the no-overload range of the quantizer, and σ^2 is the quantization error power [13]. Taking $\gamma = 1$ and applying this formula to the simulation results shown in Figure 10 indicates that for an oversampling ratio of nineteen, accuracies of approximately 13, 17, and 21 bits are achieved by $\Pi\Delta\Sigma$ modulators with 4, 16, and 64 channels, respectively. Hence, for each doubling of M, the accuracy of the $\Pi\Delta\Sigma$ modulator is increased by approximately two bits.

Figure 11 shows simulations of four, sixteen, and sixty-four channel full-rate $\Pi\Delta\Sigma$ modulators operating on a sine wave of fixed frequency, arbitrarily chosen as $\omega = 3.71$. The quantization error power of each $\Pi\Delta\Sigma$ modulator is plotted against the amplitude of the sine wave. We see that when the amplitude of the input is

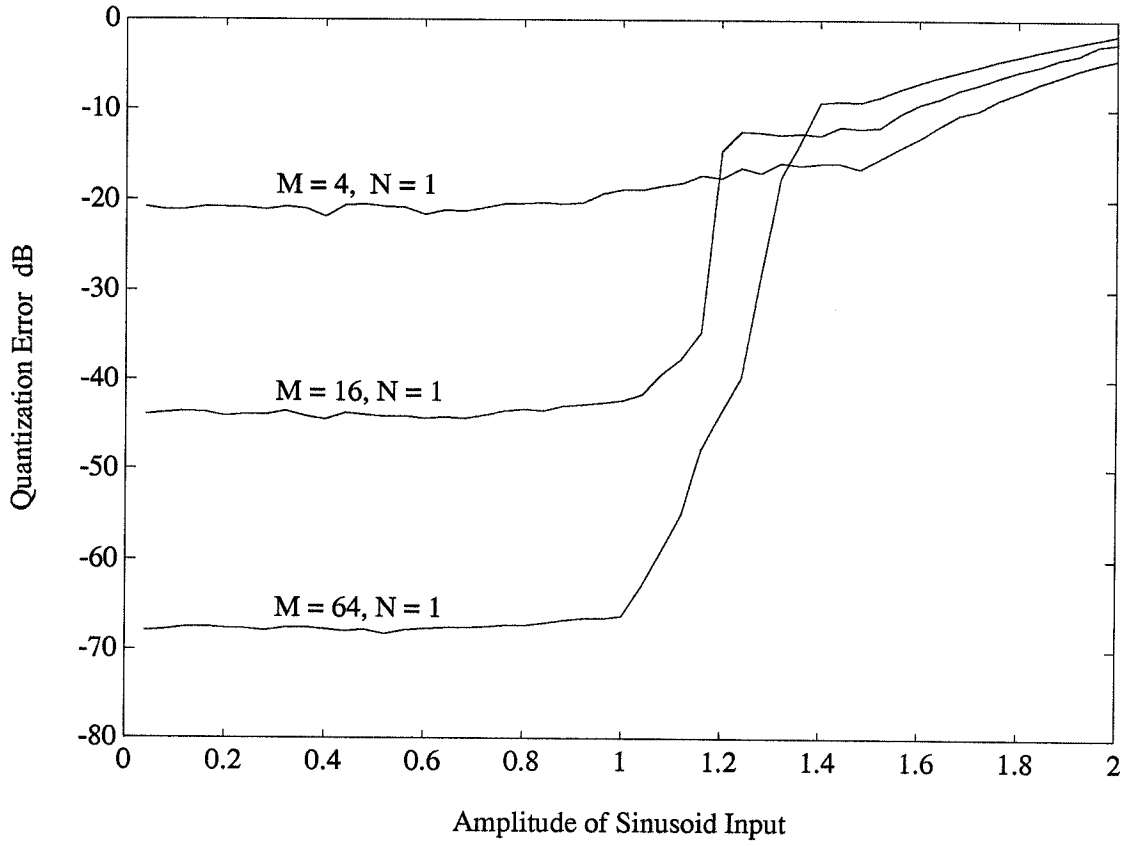


Figure 11: The quantization error power of the full-rate ($N = 1$) $\Pi\Delta\Sigma$ modulator with a sinusoidal input versus amplitude of the sine wave. Data for four, sixteen, and sixty-four channel $\Pi\Delta\Sigma$ modulators are shown. Second-order double-loop $\Delta\Sigma$ modulators with four level quantizers and $\Delta = 1$ are used. The input has a frequency of 3.71.

bounded in absolute value by 1, the quantization error power is not a function of amplitude. However, the quantization error power increases strongly with input amplitudes greater than 1 as a result of quantizer overload. For $\Delta\Sigma$ modulators of the type simulated, overload occurs at input amplitudes of about 1 [3], so it is not surprising that this is also the case with the $\Pi\Delta\Sigma$ modulator. Simulations of $\Pi\Delta\Sigma$ modulators with non-unity oversampling ratios show similar behavior.

Proceeding as above, for non-overloading inputs, accuracies of approximately 2.5, 6.5, and 10.5 bits are achieved by full-rate $\Pi\Delta\Sigma$ modulators with 4, 16, and 64 channels, respectively. Again, we see that for each doubling of M, the accuracy of the $\Pi\Delta\Sigma$ modulator is increased by approximately two bits as predicted by theory.

VI. Conclusion

A new A/D converter architecture called the $\Pi\Delta\Sigma$ modulator has been presented. The architecture represents a generalization of conventional $\Delta\Sigma$ modulator based A/D converter architectures. It combines M $\Delta\Sigma$ modulator based converters operating in parallel and achieves an accuracy equivalent to that which one of the converters would achieve if operated alone with an oversampling ratio of NM, minus $\frac{1}{2}\log_2 M$ bits. For example, if its component $\Delta\Sigma$ modulators provide K additional bits of accuracy for every doubling of the oversampling ratio, the corresponding $\Pi\Delta\Sigma$ modulator provides $K - \frac{1}{2}$ additional bits of accuracy for every doubling of M. Therefore, the $\Pi\Delta\Sigma$ modulator can be thought of as a device that performs oversampling in parallel. For a given type of $\Delta\Sigma$ modulator, the $\Pi\Delta\Sigma$ modulator approach allows designers to trade circuit speed against area to achieve the desired level of A/D conversion accuracy. In particular, for the special case of $N = 1$ the $\Pi\Delta\Sigma$ modulator operates as a full-rate A/D converter.

A theoretical analysis has been presented that relates the autocorrelation of the overall quantization error of the $\Pi\Delta\Sigma$ modulator to the autocorrelations of the quantization errors of the component $\Delta\Sigma$ modulators. Therefore, existing $\Delta\Sigma$ modulator theory and experience can be brought to bear on the $\Pi\Delta\Sigma$ modulator. Although the theoretical results assume ideal operating conditions and components, a qualitative analysis asserting that the $\Pi\Delta\Sigma$ modulator is robust with respect to nonideal operating conditions and components has been presented.

Simulations supporting the theoretical analysis have been presented and discussed. In particular, they show that $\Pi\Delta\Sigma$ modulators using common second-order double-loop $\Delta\Sigma$ modulators can achieve excellent performance with relatively few channels. The simulations also indicate that for non-overloading inputs, the accuracy of the $\Pi\Delta\Sigma$ modulator does not depend upon the nature of the input sequence.

While the implementation of robust practical $\Delta\Sigma$ modulator based A/D con-

verters is non-trivial, good solutions exist and the field is rapidly maturing. A plethora of academic and commercial implementations have been reported [14], [15]. This progress bodes well for the $\Pi\Delta\Sigma$ modulator because its implementation requirements are similar. Aside from its $\Delta\Sigma$ modulator based converters, the $\Pi\Delta\Sigma$ modulator consists of components that are well understood and amenable to VLSI implementation. Moreover, the $\Pi\Delta\Sigma$ modulator is as robust with respect to circuit nonidealities as are its component $\Delta\Sigma$ modulator based converters. The main impediment to the implementation of the $\Pi\Delta\Sigma$ modulator is that it requires a large amount of digital circuitry.

Appendix: Supporting Lemmas

The following Lemma presents the underlying reason why the quantization error from the different channels of the $\Pi\Delta\Sigma$ modulator does not add coherently. The random variables η_n cause the quantization error from the channels to be random processes that are asymptotically independent. That is, in a $\Pi\Delta\Sigma$ modulator that has been running since the beginning of time, the random processes corresponding to the channel quantization error are independent.

Lemma A1: Let $\mathbf{U}_p = \left(U_p^{(0)}, U_p^{(1)}, \ldots, U_p^{(M-1)}\right)$, $p = 0, 1, \ldots$, be a sequence of random variables defined on $\mathbf{R}^k$ where

$$U_p^{(r)} = \left\langle a_r(p) + \sum_{k=0}^{p} h_r(k) u_r(k) \eta_k \right\rangle,$$

$a_r(n)$ is any real sequence, $h_r(n)$ is any sequence that does not converge to zero, $u_r(n)$ is as defined in (1), and $\{\eta_n\}$ is the sequence of iid random variables introduced in (5). Then $\mathbf{U}_p$ converges in distribution to a random variable $\mathbf{U}$ that is uniformly distributed on $[0,1)^M$ as $p \to \infty$.

Proof: We will prove the result for the case where $u_r(n)$ corresponds to $N = 1$

and $L = 1$ in (1). No additional theory is required to prove the general case, but the notation is more confusing.

Define $\mathbf{s}_p = (s_p^{(0)}, \ldots, s_p^{(M-1)})^T$ where

$$s_p^{(r)} = a_r(p) + \sum_{k=0}^{p} h_r(k) u_r(k) \eta_k.$$

Let $\mathbf{y}_p = (y_p^{(0)}, \ldots, y_p^{(M-1)})^T$ be the vector defined by $\mathbf{y}_p = \frac{1}{M}\mathbf{H}\mathbf{s}_p$, where $\mathbf{H} = \{m_{j,k}\}$ is the $M \times M$ Hadamard matrix used to generate $u_r(n)$.

Since $u_r(n) = m_{r,(n \bmod M)}$, it follows that

$$y_p^{(r)} = a'_r(p) + \sum_{k=0}^{\lfloor (p+r)/M \rfloor} h_r(Mk + r) \eta_{Mk+r}, \tag{18}$$

where $a'_r(p)$ is a sequence of real numbers. Note that the elements of $\mathbf{y}_p$ each contain a mutually exclusive subset of the random variables η_n, so they must be statistically independent. Thus, the M-dimensional probability measure of $\mathbf{y}_p$ is

$$P_{\mathbf{y}_p} = P_{y_p^{(0)}} \times P_{y_p^{(1)}} \times \cdots \times P_{y_p^{(M-1)}},$$

where the $P_{y_p^{(r)}}$ are the one-dimensional probability measures associated with the elements of $\mathbf{y}_p$.

We wish to determine the M-dimensional characteristic function of $\mathbf{s}_p$:

$$\Phi_{\mathbf{s}_p}(\mathbf{t}) = \int_{\mathbf{R}^M} e^{j\mathbf{t}\cdot\mathbf{s}_p} P_{\mathbf{s}_p}(d\mathbf{s}_p).$$

This can be done by considering $P_{\mathbf{y}_p}$ and performing a change of variables. Let $T : \mathbf{R}^M \to \mathbf{R}^M$ be the mapping associated with the matrix $\mathbf{H}$. Then T^{-1} is the mapping associated with the matrix $\frac{1}{M}\mathbf{H}$. It follows that $P_{\mathbf{s}_p} = P_{\mathbf{y}_p} T^{-1}$. Applying the change of variables formula from analysis gives†

$$\int_{\mathbf{R}^M} e^{j\mathbf{t}\cdot\mathbf{s}_p} P_{\mathbf{s}_p}(d\mathbf{s}_p) = \int_{\mathbf{R}^M} e^{j\mathbf{t}\cdot T\mathbf{y}_p} P_{\mathbf{y}_p}(d\mathbf{y}_p).$$

† See for example, Theorem 16.12 of [10].

Hence,

$$\begin{aligned}\Phi_{\mathbf{s}_p}(\mathbf{t}) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{j\mathbf{t}\cdot\mathbf{H}\mathbf{y}_p} P_{y_p^{(0)}}(dy_p^{(0)}) \cdots P_{y_p^{(M-1)}}(dy_p^{(M-1)}) \\ &= \Phi_{y_p^{(0)}}(\mathbf{m_0}\cdot\mathbf{t}) \cdots \Phi_{y_p^{(M-1)}}(\mathbf{m_{M-1}}\cdot\mathbf{t}),\end{aligned}$$

where $\mathbf{m_r}$ is the r^{th} column of the Hadamard matrix.

Let $\Phi_\eta(t)$ be the characteristic function of the random variables η_n. Because the η_n are independent, from (18), the magnitude of the characteristic function of $y_p^{(p)}$ is

$$\left|\Phi_{y_p^{(r)}}(t)\right| = \prod_{k=0}^{\lfloor (p+r)/M \rfloor} \left|\Phi_\eta\big(th_r(Mk+r)\big)\right|.$$

By hypothesis, none of the η_n have lattice distributions. This implies that $|\Phi_\eta(t)| < 1$ for all $t \neq 0$.† Since $h_r(n)$ does not converge to zero, and all characteristic functions equal 1 at the origin,

$$\lim_{p\to\infty} \Phi_{y_p^{(r)}}(t) = \begin{cases} 1, & \text{if } t = 0; \\ 0, & \text{otherwise.} \end{cases}$$

Since $\mathbf{m_r}\cdot\mathbf{t} = 0$ for all r, $0 \le r \le M-1$, iff $\mathbf{t} = \mathbf{0}$ (because $\mathbf{H}$ is invertible), it follows that

$$\lim_{p\to\infty} \Phi_{\mathbf{s}_p}(\mathbf{t}) = \begin{cases} 1, & \text{if } \mathbf{t} = \mathbf{0}; \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

As shown in [2], for any random variable X the characteristic functions of X and $\langle X \rangle$ are equal on the set of points $\{2\pi m : m = 0, \pm 1, \ldots\}$. That is,

$$\Phi_{\langle X\rangle}(2\pi m) = \Phi_X(2\pi m)$$

for every integer m. The result can easily be extended to multiple dimensions by simply modifying the notation in the proof to accommodate the extra dimensions. The extension implies that

$$\Phi_{\mathbf{U}_p}(2\pi m_0, \ldots, 2\pi m_{M-1}) = \Phi_{\mathbf{s}_p}(2\pi m_0, \ldots, 2\pi m_{M-1})$$

† See, for example, Problem 26.1 of [10].

for every set of integers $\{m_0, \ldots, m_{M-1}\}$. Consequently, (19) implies that

$$\lim_{p\to\infty} \Phi_{\mathbf{U}_p}(2\pi m_0, \ldots, 2\pi m_{M-1}) = \begin{cases} 1, & \text{if } (m_0, \ldots, m_{M-1}) = \mathbf{0}; \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Define

$$\Phi_{\mathbf{U}}(\mathbf{t}) = \prod_{k=0}^{M-1} e^{-jt_k/2} \frac{\sin(t_k/2)}{t_k/2}.$$

Then for every set of integers $\{m_0, \ldots, m_{M-1}\}$

$$\lim_{p\to\infty} \Phi_{\mathbf{U}_p}(2\pi m_0, \ldots, 2\pi m_{M-1}) = \Phi_{\mathbf{U}}(2\pi m_0, \ldots, 2\pi m_{M-1}).$$

Taking the inverse Fourier transform of $\Phi_{\mathbf{U}}(\mathbf{t})$ shows it to be the characteristic function of $\mathbf{U}$.

By definition, $\mathbf{U}$ and $\mathbf{U}_p$ each have their support restricted to $[0,1)^M$. Therefore, $\Phi_{\mathbf{U}}(\mathbf{t})$ and $\Phi_{\mathbf{U}_p}(\mathbf{t})$ are each uniquely determined by their samples at all M-tuples of integers $(m_0, \ldots, m_{M-1})$.

A necessary and sufficient condition for $\mathbf{U}_p$ to converge in distribution to $\mathbf{U}$ is that $\Phi_{\mathbf{U}_p}(\mathbf{t})$ converge to $\Phi_{\mathbf{U}}(\mathbf{t})$ for each $\mathbf{t} \in \mathbf{R}^M$. Therefore, by the argument above, any subsequence of $\mathbf{U}_p$ that converges in distribution at all must converge in distribution to $\mathbf{U}$. Since the sequence of probability measures associated with $\{\mathbf{U}_p\}$ is tight (a consequence of the sequence having bounded support), it follows that there exists a subsequence of $\{\mathbf{U}_p\}$ that converges in distribution to $\mathbf{U}$, which further implies that $\mathbf{U}_p$ converges in distribution to $\mathbf{U}$.†

■

The next Lemma is proven in [11] but is restated here for convenience. It is the mechanism behind each of the ergodic results in Theorem 5.

Lemma A2: For each $k = 0, 1, \ldots$, let X_k be a deterministic function of the two random sequences $\{\chi_0, \ldots, \chi_k\}$ and $\{\eta_0, \ldots, \eta_k\}$, where the η_n are independent

† See, for example, Theorem 25.10 and its corollary in [10].

random variables that are independent of the χ_n. Suppose that the distribution of each X_k has its support restricted to $[-\beta, \beta]$ where $\beta \in \mathbf{R}$, and that for each non-negative integer j, as $k \to \infty$

$$\mathop{\mathrm{E}}_{\{\eta_n : n > j\}}(X_k) \to 0$$

uniformly with respect to the variables $\{\eta_0, \ldots, \eta_j\}$ and $\{\chi_0, \chi_1, \ldots\}$. Then

$$\frac{1}{N} \sum_{n=0}^{N-1} X_n \to 0$$

in probability as $N \to \infty$.

References

1. J. C. Candy, G. C. Temes, "Oversampling methods for A/D and D/A conversion," *Oversampling Delta-Sigma Data Converters Theory, Design and Simulation*, New York: IEEE Press, pp. 1-25, 1992.
2. I. Galton, "Granular quantization noise in a class of $\Delta\Sigma$ modulators," Submitted to *IEEE Trans. Inform. Theory*, Mar. 1992
3. J. C. Candy, "A use of double integration in sigma-delta modulation," *IEEE Trans. Commun.*, vol. COM-33, pp. 249-258, Mar. 1985.
4. D. Seitzer, G. Pretzl, N. A. Hamdy, *Electronic Analog-to-Digital Converters: Principles, Circuits, Devices, Testing.* New York: J. Wiley, 1983.
5. J. H. van Lint, R. M. Wilson, *A Course in Combinatorics.* Cambridge: Cambridge University Press, To appear in 1992.
6. F. H. Harmuth, *Sequency Theory.* New York: Academic Press, 1977.
7. L. C. Fernandez, K.R. Rao, "Design of a synchronous walsh-function generator," *IEEE Trans. Electromagnetic Compatibility*, vol. EMC-19, no. 4, pp. 407-409, Nov. 1977.
8. F. Kitai, C. K. Yuen, "Walsh function generators," *Applications of Walsh functions and sequency theory*, New York: IEEE Press, pp. 297-315, 1974.

9. P. W. Besslich, "Walsh function generators for minimum orthogonality error," *Trans. Electromagnetic Compatibility*, vol. EMC-15, no. 4, pp. 177-179, Nov. 1973.

10. P. Billingsley, *Probability and Measure.* New York: John Wiley and Sons, 1986.

11. I. Galton, "Granular quantization noise in the first-order $\Delta\Sigma$ modulator," Submitted to *IEEE Trans. Inform. Theory*, Nov. 1991.

12. B. E. Boser, B.A. Wooley, "The design of sigma-delta modulation analog-to-digital converters," *IEEE J. Solid-State Circuits*, vol. SC-23, pp. 1298-1308, Dec. 1988.

13. N. S. Jayant, P. Noll, *Digital Coding of Waveforms Principles and Applications to Speech and Video.* Englewood Cliffs, NJ: Prentice-Hall, 1984.

14. *Oversampling Delta-Sigma Data Converters Theory, Design and Simulation*, Edited by J. C. Candy, G. C. Temes, New York, IEEE Press, 1992.

15. F. Goodenough, "High-resolution ADCs up dynamic range in more applications," *Electronic Design*, pp. 65-79, April 11, 1991.