Chapter 2 in *Delta-Sigma Data Converters*
Edited by S. Norsworthy, R. Schreier, and G. Temes

Quantization Noise in $\Delta\Sigma$ A/D Converters

Robert M. Gray
Information Systems Laboratory
Department of Electrical Engineering
Stanford University
Stanford, CA 94305-4055.

November 16, 1995

1 Introduction

The heart of a $\Delta\Sigma$ modulator and any other analog-to-digital converter (ADC) is a quantizer, a device which maps real numbers into a finite set of possible representative values, often as few as two. Any analysis of the behavior of a $\Delta\Sigma$ modulator must include consideration of the behavior of the quantizer. The quantization operation is inherently nonlinear and hence rigorous analysis is complicated even in the simplest of systems. When quantizers are incorporated into linear systems with feedback such as $\Delta\Sigma$ modulators and bang-bang control systems, the analysis becomes even more difficult. Simulations cannot capture all aspects of possible system behavior and are not always reproducible as different random number generators are used and care is not always taken to ensure that sample functions are long enough for sample averages to be close to expectations with high probability. As a result, various methods based on approximations have been widely used, even in some applications where they were known to give misleading or outright incorrect results. Often, however, approximate methods have quite successfully predicted some aspects of system behavior, as many of the other chapters of this book will attest. These approximations are usually implicitly or explicitly based on either the asymptotic results of Bennett [1] or on the exact results of Widrow [2] as extended by Sripad and Snyder [3]. We shall see, however, that the underlying conditions assumed by these results can be and usually are invalid in typical $\Delta\Sigma$ modulators and there are not any good guidelines for determining when the approximations might nonetheless yield good results. A common reason for using the approximation methods in spite of their shortcomings is that simulations are inadequate and that exact analysis is thought to be either impossible or prohibitively difficult. Successful applications of the approximations do not lessen the puzzlement of engineers who use such methods to quantify the behavior of a system, and then find that the system exhibits bizarre artifacts not suggested by the theory. They often suspect the system rather than the theory of being flawed. A classical example is the ability of approximate methods to correctly predict the signal-to-noise ratio of a single-loop $\Delta\Sigma$ modulator while producing a completely incorrect prediction of the quantization error spectrum that fails to include audible objectionable tones in the final output.

The goal of this chapter is to discuss in some detail the most common approximations, their underlying justification in quantization systems, and the common errors made in their application. The hope is to give practicing engineers a degree of skepticism in their use of these methods and to prepare them for unpleasant surprises. We also consider a variety of examples where the approximations are not needed and exact descriptions of system behavior can be found by combining linear systems methods with a few nonlinear system techniques. The mathematical methods are not particularly deep; most come from Fourier analysis and probability theory, but much of the algebra and calculus is not particularly pretty and is left to the references. By "exact" it is meant that exact solutions are found to the nonlinear difference equations modeling $\Delta\Sigma$ modulators. Certainly real system will have many variations that are not yet included in the basic equations, but the given nonlinear equations will be solved without recourse to linearizing approximations.

Unfortunately the exact solutions do not extend to all $\Delta\Sigma$ architectures of practical interest, but they do provide a rich collection of types of behavior and of important attributes of systems which determine that behavior. These examples can provide useful insight to more complex systems. The theory also provides some surprising results, including the fact that some of the common approximations hold exactly in some systems even though the underlying conditions usually assumed for those approximations are violated.

Because the quantizer plays the key role in a $\Delta\Sigma$ modulation, we begin by looking at quantization error in a simple quantizer that is used without feedback or linear filtering. This leads in several steps to the more complicated example of $\Delta\Sigma$ modulation.

2 Uniform Quantization

The basic common component to most analog-to-digital converters is a uniform quantizer. It is assumed that the quantizer has an even number, say M, of levels and that the distance between the output levels (the *bin width*) is Δ. The special case of $M = 2$ is common in $\Delta\Sigma$ modulators, but the theory here and later holds for any even M.

The M quantization levels are equally interspersed in the interval $B = [a, a + M\Delta]$, where often the region is chosen to be symmetric around the origin, i.e., $a = -M\Delta/2$, and each level y_k is the center of its quantization cell $R_k = [a+k\Delta, a+(k+1)\Delta)$; $k = 0, 1, \ldots, M$. Any input u in this range will map into a quantized value $q(u) = y_k$ if $u \in R_k$, i.e., the quantizer mapping is a minimum distance (nearest neighbor) mapping. For a given number of levels M, there are only two free parameters in a uniform quantizer: the offset a and the bin width Δ.

The quantizer error is defined as $\epsilon = q(u) - u$. If u is in the region B, then the maximum error resulting is $\Delta/2$. Outside this region the input is mapped into the nearest quantization level, but the error is greater than $\Delta/2$ and the quantizer is said to *overload* or *saturate*. We will refer to the interval B as the *no-overload region* of the quantizer. Some of the methods described hold only when when u is in the no-overload region with probability 1.

If the input is described by a probability distribution, then the the performance of a quantizer is often measured by the mean squared error, the expected error energy $E(\epsilon^2)$. A uniform quantizer is said to be *optimal* for an input probability distribution if for the given M the parameters a and Δ are chosen to minimize the mean squared error.

If the input is a sequence of samples u_n, then we will be interested in the error sequence $\epsilon_n = q(u_n) - u_n$, and we will wish to see if it resembles random noise in some way. A trivial rewriting of this formula yields the so-called "additive noise model" of quantization

$$q_n = q(u_n) = u_n + \epsilon_n$$

expressing the output of the quantizer as its input plus a "noise" term. There is no genuine modeling here, this is simply a convenient definition of the quantization error such that the output can be written as the sum of the input and a "noise" term. The modeling enters when assumptions are made about the statistical behavior of ϵ and its dependence on the input signal, as will be seen.

3 The Additive White Noise Approximation

Many of the original results and insights into the behavior of quantization error are due to Bennett [1], and much of the work since then has its origins in that classic paper. Bennett first developed conditions under which quantization noise could be reasonably modeled as additive white noise. Unfortunately, much of the literature assumes more than was proved by Bennett and often uses Bennett's approximations when his results do not apply. Subsequently Sripad and Snyder [3] extended Widrow's approach [2] and found necessary and sufficient conditions for certain aspects of the approximation to hold exactly. We here explore the approximations and these underlying conditions in order to consider their suitability for use in analyzing $\Delta\Sigma$ modulators.

A common statement of the approximation is that the quantization error ϵ_n has the following properties, which we refer to collectively as the "input-independent additive white noise approximation":

Property 1: ϵ_n is statistically independent of the input signal u_k for all n, k (strong version) or ϵ_n is uncorrelated with the input signal u_n (weak version),

Property 2: ϵ_n is uniformly distributed in $[-\Delta/2, \Delta/2]$,

Property 3: ϵ_n is an independent identically distributed (i.i.d.) sequence (strong version) or ϵ_n has a flat power spectral density (it is "white") (weak version).

These approximations enormously simplify system analysis because they replace a deterministic nonlinearity by a stochastic linear system, thereby permitting the use of linear systems methods to analyze an otherwise linear system containing a quantizer. These properties have been used in the wide majority of published analyses of systems containing quantizers in the communications, control, and signal processing literature. Most often the approximation is made without reference and with only small (if any) mention of its possible limitations. The natural questions that arise are

Question 1: Are the approximations good under ordinary conditions; that is, do they accurately model the true behavior of quantization error?

Question 2: If the approximations are not good, is it still possible for them to yield good predictions of actual system behavior?

It is easy to demonstrate a negative answer to Question 1 if the strong form of the approximation is considered: the quantizer error is a deterministic function of the input and hence cannot be statistically independent of the input. There is hope, however, that a weaker form of independence of uncorrelation (or linear independence) holds in the sense that $E[u_n \epsilon_k] = E[u_n]E[\epsilon_k]$ for all n, k, where E denotes expectation or probabilistic average. This property is sufficient to ensure that second order analysis involving output correlations and spectra can be carried out without the complexity of cross-terms. This allows the common "noise shaping" interpretation of $\Delta\Sigma$ modulators because it implies that the quantization noise can be filtered without thereby also changing the input signal. This approximation is almost universally made in the analysis of oversampled ADC's, yet, as we shall see, it can be incorrect.

Question 2 is not so easily answered. Engineering mathematics often uses ideas such as impulses and flicker $(1/f)$ noise that are physically impossible, yet yield perfectly good predictions of real system behavior when carefully used and suitably interpreted. The answer to this question will vary depending on the system and a goal of this chapter is to provide a feel for examples where answers based on the white noise approximation can be trusted and where they cannot be.

If we back off on the complete input-independent additive noise model by eliminating the first property, Bennett's theory provides a motivation for approximating quantization noise by Properties 2 and 3, which will be seen to hold under specific conditions (first proved by Bennett). When these properties hold approximately, we shall refer to the approximation as the "additive white noise approximation," dropping the "input-independent" modifier. We shall also discuss the non-asymptotic results of Sripad and Snyder which provide conditions under which several of the common approximations hold in an exact sense. Unfortunately, we shall see that the conditions of both Bennett and of Sripad and Snyder do not hold in typical $\Delta\Sigma$ modulators and hence the additive white noise approximation is not justified mathematically. Perhaps surprisingly, we shall later find that in spite of this fact, the additive white noise approximation in fact holds exactly for ideal multistage and higher order $\Delta\Sigma$ architecures provided the quantizers do not overload.

Property 2 might be true if the quantizer cannot overload, but it is clearly false if overload occurs with nonzero probability. Property 3 is also plausible, but must be demonstrated for a particular system. A simple but important example where both Properties 2 and 3 hold exactly is in the case where the input signal is itself i.i.d. and is uniformly distributed over the no-overload range. In this example it is easy to see that the quantization error is uniformly distributed over $(-\Delta/2, \Delta/2)$ and has zero mean and variance $\Delta^2/12$, the ubiquitous result for the mean squared error of a uniform quantizer with bin width Δ. It should be pointed out that even in the case of a uniformly distributed input signal, Property 1 is not true even in the weak sense, i.e., the input signal and quantization error are not uncorrelated. It is straightforward to show that $E(u_n \epsilon_k) = -\sigma_{\epsilon_n}^2 \delta_{n-k}$, where δ_l is the Kronecker delta function, 1 if $l = 0$ and 0 otherwise. In words, the correlation between input and error at a common sample time is as large as the error variance, it is not 0! In contrast, if the input probability density is a triangle on the no-overload range (the convolution of two uniform densities covering half the no-overload range), then again the error is uniformly distributed and white, but now $E(u_n \epsilon_k) = 0$ for all n and k, i.e., the input and the quantizer are indeed uncorrelated. The point is that the weak version of Property 1 might or might not hold, depending on the input density and the particular quantizer.

Bennett argued more generally that Properties 2 and 3 are approximately true (along with some other properties) if certain underlying conditions hold. The uniform white quantization noise assumption subsequently gained a wide popularity, largely due to the work of Widrow [2] who provided a sufficient condition for the quantizer error to be uniformly distributed. For completeness we quote the basic results of Bennett and sketch their proof.

Bennett's Theorem

Suppose that the following conditions (Bennett's conditions) hold:

1. The input is in the no-overload region,
2. M asymptotically large,
3. Δ asymptotically small, and
4. the joint probability density function (pdf) of the input signal at different sample times is smooth.

Then the error sequence has the following properties:

1. The sequence $\{\epsilon_n\}$ is approximately uniformly distributed, that is, has marginal pdf

$$f_\epsilon(\alpha) \approx \begin{cases} \frac{1}{\Delta} & \text{if } \alpha \in (-\frac{\Delta}{2}, \frac{\Delta}{2}) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

This in turn implies that $E[\epsilon_n] \approx 0$,

$$\sigma^2_{\epsilon_n} \approx \frac{\Delta^2}{12}, \tag{2}$$

and

$$E[q_n] \approx E[u_n], \tag{3}$$

i.e., the expectation of the quantized output is approximately the same as that of the input. In statistical terms, the quantized value is an unbiased estimator of the input.

2. The sequence $\{\epsilon_n\}$ is approximately an independent identically distributed (i.i.d.) random process.

Sketch of proof of Bennett's theorem: First consider marginal distribution of the error. Define as usual the cumulative distribution function (cdf) $F_{\epsilon_n}(\alpha) = \Pr(\epsilon_n \leq \alpha)$; $\alpha \in (-\Delta/2, \Delta/2)$ and the pdf $f_{\epsilon_n}(\alpha) = dF_{\epsilon_n}(\alpha)/d\alpha$. Referring to the definitions of q and ϵ we can write

$$F_{\epsilon_n}(\alpha) = \sum_{k=0}^{M-1} \Pr(\epsilon_n \leq \alpha \text{ and } u_n \in R_k).$$

Since the pdf is assumed to be smooth, the mean value theorem of calculus implies that

$$\Pr(\epsilon_n \leq \alpha \text{ and } u_n \in R_k) = \int_{-(M/2+k)\Delta}^{-(M/2+k)\Delta+\alpha} f_{u_n}(\beta)\, d\beta \approx f_{u_n}(y_k)\alpha.$$

Using the Riemann sum approximation to an integral yields

$$F_{\epsilon_n}(\alpha) \approx \frac{\alpha}{\Delta} \sum_{k=0}^{M-1} f_{u_n}(y_k)\Delta \approx \frac{\alpha}{\Delta} \int_{-M/2}^{M/2} f_{u_n}(u) du \approx \frac{\alpha}{\Delta},$$

and hence

$$f_{\epsilon_n}(\alpha) \approx \frac{1}{\Delta} \text{ for } \alpha \in (-\frac{\Delta}{2}, \frac{\Delta}{2}).$$

To prove that the error sequence is approximately memoryless, a similar idea is applied to vectors of error samples:

$$\Pr(\epsilon_l \leq \alpha_l;\ l = n, \ldots, n+k-1) = \sum_{i_1,\ldots,i_k} \Pr(\epsilon_l \leq \alpha_l \text{ and } u_l \in R_{i_l};\ l = n, \ldots, n+k-1),$$

For $\alpha_l \in (-\Delta/2, \Delta/2)$, $l = 1, \ldots, k$

$$\Pr(\epsilon_l \leq \alpha_l \text{ and } u_l \in R_{i_l};\ l = n, \ldots, n+k-1) =$$

$$\int_{-(M/2+k)\Delta}^{-(M/2+k)\Delta+\alpha_1} \cdots \int_{-(M/2+k)\Delta}^{-(M/2+k)\Delta+\alpha_k} f_{u_n,\ldots,u_{n+k-1}}(\beta_1,\ldots,\beta_k)\,d\beta_1\cdots d\beta_k$$
$$\approx f_{u_n,\ldots,u_{n+k-1}}(y_{i_1},\ldots,y_{i_k})\alpha_1\cdots\alpha_k,$$

whence

$$f_{\epsilon_n,\ldots,\epsilon_{n+k-1}}(\alpha_1,\ldots,\alpha_k) \approx \frac{1}{\Delta^k}.$$

This immediately implies that

$$R_\epsilon(n,k) = E[\epsilon_n\epsilon_k] \approx \sigma_\epsilon^2\delta_{n-k}.$$

Bennett did not explicitly treat the issue of the correlation of quantizer error and input, but his basic method of calculus approximations can be applied to the task. The analysis is somewhat more delicate, however, since higher order terms can make a difference as pointed out by Kollár [4]. It is easy to argue that $R_{q,\epsilon}(n,k) \approx 0$ when $n \neq k$, so we focus on the case of $n = k$. Following [4], we can approximate the input density in the quantization bin R_l for output y_l by a Taylor series expansion as $f_u(y_l+\delta) \approx f_u(y_l) + f_u'(y_l)\delta$, where the higher order terms can be shown to be negligible in comparison. This leads to the approximation that

$$\begin{aligned}
R_{q,\epsilon}(n,k) &= E[q(U_n)\epsilon_k] = \sum_{l=0}^{M-1} y_l E[\epsilon_k | q(u_n) = y_k]\Pr(q(u_n) = y_k) \\
&= \sum_{l=0}^{M-1} y_l \int_{-\Delta/2}^{\Delta/2} (-\delta)[f_u(y_l) + f_u'(y_l)\delta]\,d\delta \\
&= -\sum_{l=0}^{M-1} \frac{\Delta^2}{12} y_l f_u'(y_l) \\
&\approx \frac{\Delta^2}{12}\int_a^{a+M\Delta} y f_u'(y)\,dy. \qquad (4)
\end{aligned}$$

Integrating by parts then yields

$$R_{q,\epsilon}(n,n) \approx \frac{\Delta^2}{12}(1 - (a+M\Delta)f_u(a+M\Delta) - af_u(a)). \qquad (5)$$

The behavior thus depends on the behavior of the input density near the borders of the no-overload range $[a, a+M\Delta]$. If the density is zero at the edge of the no-overload range, then $R_{q,\epsilon}(n,n) \approx \Delta^2/12 = R_\epsilon(n,n)$, which implies in turn that

$$\begin{aligned}
R_{u,\epsilon}(n,k) &= E[u_n\epsilon_k] = E[(q(u_n) - \epsilon_n)\epsilon_k] \qquad (6) \\
&= E[q(U_n)\epsilon_k] - R_\epsilon(n,k) = 0,
\end{aligned}$$

i.e., $R_{u,\epsilon}(n,k) \approx 0$ and the input and quantization error are uncorrelated. If, however, the density is not 0 at the borders of the no-overload zone (as is the case with a uniform density on the no-overload region), then the signal and quantizer error are not uncorrelated. Bucklew and Gallagher [5] have shown that if the uniform quantizer is optimal, that is, if a and Δ are chosen to minimize mean squared quantization error, then one will have exactly $R_{q,\epsilon}(n,k) = 0$ and hence $R_{u,\epsilon}(n,k) = -\sigma_\epsilon^2 \delta_{n-k}$ and the input and quantizer error have correlation equal to minus the quantizer error energy. Optimal choice of Δ will involve a shrinking of the binwidth and a resulting overload of the quantizer, but the Bennett approximations still hold [5]. We note that this same result holds if the Lloyd-Max optimal nonuniform quantizer is used (see, e.g., [6] and pp. 180–181 of [8]).

An alternative to the asymptotic (large number of quantizer levels) analysis of Bennett is the exact approach of Sripad and Snyder [3], which is a variation of the characteristic function method that will be used here. Sripad and Snyder demonstrated necessary and sufficient condition for the various properties to hold. The conditions are stated in terms of the characteristic function

$$\Phi_u(\nu) = E[e^{j\nu u_n}] \tag{7}$$

of the input random variable and the joint characteristic function $\Phi_{u_n,u_k}(\nu,\mu) = E[e^{j(\nu u_n + \mu u_k}]$. The input process $\{u_n\}$ is assumed to be stationary so that the characteristic function does not depend on n.

• A necessary and sufficient condition for the quantizer error to be uniformly distributed on $[-\Delta/2, \Delta/2]$ is that

$$\Phi_u(\frac{2\pi k}{\Delta}) = 0,\ k = \pm 1, \pm 2, \ldots. \tag{8}$$

• A necessary and sufficient condition for ϵ_n and ϵ_k to be independent and uniformly distributed on $[-\Delta/2, \Delta/2]$ is that $\Phi_{u_n,u_k}(\frac{2\pi l}{\Delta}, \frac{2\pi m}{\Delta}) = 0$, for all l, m for which $(l,m) \neq (0,0)$.

• A sufficient condition for ϵ_n and u_n to be uncorrelated is that

$$\Phi_u(\frac{2\pi k}{\Delta}) = \dot{\Phi}_u(\frac{2\pi k}{\Delta});\ k = \pm 1, \pm 2, \ldots, \tag{9}$$

where $\dot{\Phi}$ is the derivative of Φ_u with respect to its argument.

The condition that Φ_u have zero value for all integral multiples of $2\pi/\Delta$ is satisfied by a random variable u having a uniform density over $[-\Delta/2, \Delta/2]$ as well as for any density formed by adding an independent random variable x to u to form $x+u$, since the resulting product of characteristic functions will inherit the zeros of Φ_u. It is also satisfied for a uniform density on $[-A, A]$ if $2A/\Delta$ is an integer M. The sufficient condition for uncorrelated signal and quantization error is satisfied, for example, by a triangular density on $[-A, A]$ if $2A/\Delta$ is an integer M. It should be pointed out that in this case the uniform quantizer is not optimal so that this result is consistent with Kollár's development, but differs from Bucklew and Gallagher.

It should be noted that the Sripad and Snyder conditions yield exact results rather than approximations, but the binwidth Δ is essentially assumed to be fixed. Furthermore, because the derivation involves a Fourier series expansion of the quantizer error probability density function on $[-\Delta/2, \Delta/2]$, the derivation implicitly assumes that the quantizer does not overload, i.e., that all of the nonzero probability density resides in the no-overload region.

We have no seen a variety of conditions under which various aspects of the white noise approximation are true approximately or exactly. The question no is whether these conditions are relevant to $\Delta\Sigma$. First consider the Bennett conditions.

1) Is the input is in the no-overload region?

Often this is not known, a problem especially true if the quantizer is inside a feedback loop. This must be verified for a particular $\Delta\Sigma$ modulator architecture and is in fact known only for a few.

2) Is M asymptotically large?

This is almost never true in $\Delta\Sigma$, where typically $M = 2$, which is not large by any stretch of the imagination.

3) Is Δ asymptotically small?

This is almost never true in $\Delta\Sigma$, where typically Δ is as large as the allowed input signal range.

4) Is the joint density of the input signal at different sample times smooth?

This is never true in $\Delta\Sigma$ where the input to the quantizer includes a discrete component due to the feedback from the quantizer. Thus the pdf of the quantizer input has a continuous component and an impulsive component, violating the smoothness condition used to prove the Bennett theorem.

If the Bennett theorem cannot be made to apply, the next alternative is to test the Sripad and Snyder conditions. This is not easily done, however, because in a $\Delta\Sigma$ modulator the input signal to the quantizer includes both an original signal and a fed back output of the quantizer, as well as linear filtering. Hence the Sripad and Snyder conditions cannot be tested without solving for the quantizer input density. This is in fact close to the method that will be used to approach the problem.

Given the above observations, the white noise approximation is at best suspicious and at worst simply wrong in $\Delta\Sigma$ analysis. Not surprisingly, it was found early in some oversampled ADCs (such as the simple single loop $\Delta\Sigma$ modulator) that the quantizer noise was not at all white and that the noise contained discrete spikes whose amplitude and frequency depended on the input [9, 10]. Perhaps surprisingly, however, simulations and actual circuits of higher order $\Delta\Sigma$ modulators and of interpolative coders often exhibited quantizer noise that appeared to be approximately white. Unfortunately, these systems also often exhibited unstable behavior not predicted by the white noise analysis.

The approximation can be modified to attempt to better approximate quantization error. Common approaches involve replacing the quantizer by a linear gain using describing function analysis. (See, for example, [11, 12, 13, 14, 15].) The remaining error can then be approximated by input independent white noise, an approach introduced by Booton [16, 17] and applied to $\Delta\Sigma$ modulators (with additional linear filtering permitted in the loops) as in Ardalan and Paulos [18].

One can improve the approximations by using higher order terms in various expansions of the quantizer nonlinearity, but this approach has not been noticeably successful in the ADC application, primarily because of its difficulty. In addition, traditional power series expansions are not well suited to the discontinuous nonlinearities of quantizers. (See Arnstein [19] and Slepian [20] for series expansion solutions for quantization noise in delta modulators and DPCM.)

Another approach to the analysis of quantization error is to modify the system by adding a small random signal or dither at the input of the quantizer. By suitably choosing the dither signal, one can in some cases force the quantization error to satisfy all aspects of the white noise approximation, but at the possible cost of corrupting the signal and reducing the allowed no-overload range [21, 22, 23, 24].

The approach taken here is a variation of the classical characteristic function method of Rice [25] and the transform method of Davenport and Root [26], who represented memoryless nonlinearities using Fourier or Laplace transforms. A similar application of Fourier analysis was made to quantization noise analysis by Clavier, Panter, and Grieg [27, 28], whose work was contemporary to Bennett's, but did not have the impact of the latter. They provided an exact analysis of the quantizer noise resulting when a uniform quantizer is driven by one or two sinusoids and thereby demonstrated both that quantization noise could behave quite unlike the predictions of the Bennett theory and that in some cases the behavior of such noise could be exactly quantified.

Subsequently the characteristic function method was applied to the study of quantization noise by Widrow [29], and his formulation has been used in the subsequent development of conditions under which the quantization noise is white, in particular by by Sripad and Snyder [3] and the classic work of Iwersen on delta modulation [31].

Combining the characteristic function method with solutions to nonlinear difference equations and some basic results from nonlinear dynamical systems theory, we can obtain an exact analysis of several interesting $\Delta\Sigma$ modulators.

4 The Characteristic Function Method

Consider now a quantizer input u, quantizer output $q(u)$, and quantizer error $\epsilon = q(u) - u$. It is convenient to normalize the quantizer output and input by the bin width Δ (or equivalently to assume $\Delta = 1$) and write after some algebra

$$e = \frac{\epsilon}{\Delta} = \frac{1}{2} - < \frac{u}{\Delta} > . \tag{10}$$

where $< r >$ denotes the fractional part of r (or $r \bmod 1$). The function $e(u)$ can be expanded as a Fourier series for u in the no-overload region as

$$e = e(u) = \sum_{l \neq 0} \frac{1}{2\pi j l} e^{2\pi j l \frac{u}{\Delta}} = \sum_{l=1}^{\infty} \frac{1}{\pi l} \sin(2\pi l \frac{u}{\Delta}). \tag{11}$$

This series will hold for almost all values of u in the no-overload region and it is a key formula for all of the subsequent analysis. One can similarly write a Fourier series for e^2 as

$$e(u)^2 = \frac{1}{12} + \sum_{l \neq 0} \frac{1}{2(\pi l)^2} e^{2\pi j l \frac{u}{\Delta}} = \frac{1}{12} + \sum_{l=1}^{\infty} \frac{1}{2(\pi l)^2} \cos(2\pi l \frac{u}{\Delta}) \tag{12}$$

Suppose now that a sequence u_n is put into the quantizer, where we require that $|u_n| \leq M\Delta/2$. We wish to study the behavior of the normalized error sequence $e_n = \epsilon_n/\Delta$. From (10)–(11) it is immediate that

$$e_n = \frac{1}{2} - < u_n > = \sum_{l \neq 0} \frac{1}{2\pi j l} e^{2\pi j l \frac{u_n}{\Delta}} = \sum_{l=1}^{\infty} \frac{1}{\pi l} \sin(2\pi l \frac{u_n}{\Delta}). \tag{13}$$

For some specific examples of sequences u_n, (13) can be used to obtain a form of Fourier series representation directly for the sequence e_n. We take a more direct route and focus on second order properties (mean, correlation, spectra) rather than a complete characterization of the sequence. Here the primary interest is the long term average behavior of the error sequence e_n. In particular we look at the average mean, second moment, and autocorrelation function. As we also wish to consider probablistic expectations when dealing with random inputs such as dithered inputs, it is useful to consider averages that include both time and probabilistic averages. A useful formalism for simultaneously considering such averages is the class of *quasi-stationary processes* considered by Ljung [32, 62]. A discrete time process e_n is said to be *quasi-stationary* if there is a finite constant C such that $E(e_n) \leq C$ for all n; and $|R_e(n,k)| \leq C$ for all n, k, where $R_e(n,k) = E(e_n e_k)$; and if for each k the limit

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} R_e(n, n+k) \tag{14}$$

exists, in which case the limit is defined as $R_e(k)$. Following Ljung we introduce some notation: Given a process x_n, define

$$\bar{E}\{x_n\} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} E(x_n), \tag{15}$$

if the limit exists. Thus for a quasi-stationary process $\{e_n\}$ the autocorrelation is given by

$$R_e(k) = \bar{E}\{e_n e_{n+k}\}, \tag{16}$$

the mean is defined by

$$m_e = \bar{E}\{e_n\} \tag{17}$$

and the average power is given by

$$R_e(0) = \bar{E}\{e_n^2\}. \tag{18}$$

Other moments are similarly defined. These moments reduce to the corresponding time averages or probabilistic averages in the special cases of deterministic or random processes, respectively.

The *power spectrum* of the process is defined in the general case as the discrete time Fourier transform of the autocorrelation:

$$S_e(f) = \sum_{n=-\infty}^{\infty} R_e(n) e^{-2\pi j f n}, \tag{19}$$

where the frequency f is normalized to lie in $[0,1]$. The usual linear system input/output relations hold for this general definition of spectrum (see Chapter 2 of Ljung [32]). In fact, the class of quasi-stationary processes can be viewed as the most general class for which the familiar formulas of ordinary linear system second-order correlation and spectral analysis remain valid.

We now proceed to apply the basic formulas (13) and (12) to find an expression for the moments. Plugging (13) into (17) and (12) into (18) and assuming that the limits can be interchanged results in

$$\bar{E}\{e_n\} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{l\neq 0} \frac{1}{2\pi j l} e^{2\pi j l \frac{u_n}{\Delta}} = \sum_{l\neq 0} \frac{1}{2\pi j l} \bar{E}\{e^{2\pi j l \frac{u_n}{\Delta}}\},$$

$$\bar{E}\{e_n^2\} = \frac{1}{12} + \sum_{l\neq 0} \frac{1}{2(\pi l)^2} \bar{E}\{e^{2\pi j l \frac{u_n}{\Delta}}\},$$

and for $k \neq 0$

$$R_e(k) = \sum_{i\neq 0} \sum_{l\neq 0} \frac{j}{2\pi i} \frac{j}{2\pi l} \bar{E}\{e^{2\pi j (i \frac{u_n}{\Delta} + l \frac{u_{n+k}}{\Delta})}\}$$

These expressions can be most easily given in terms of the one-dimensional characteristic function

$$\bar{\Phi}_u(l) = \bar{E}\{e^{2\pi j l \frac{u_n}{\Delta}}\} \tag{20}$$

and a two-dimensional characteristic function

$$\bar{\Phi}_u^{(k)}(i,l) = \bar{E}\{e^{2\pi j (i \frac{u_n}{\Delta} + l \frac{u_{n+k}}{\Delta})}\};\ k \neq 0, \tag{21}$$

as

$$\bar{E}\{e_n\} = \sum_{l\neq 0} \frac{1}{2\pi j l} \bar{\Phi}_u(l), \tag{22}$$

$$\bar{E}\{e_n^2\} = \frac{1}{12} + \sum_{l\neq 0} \frac{1}{2(\pi l)^2} \bar{\Phi}_u(l), \tag{23}$$

and for $k \neq 0$

$$R_e(k) = -\sum_{i\neq 0} \sum_{l\neq 0} \frac{1}{2\pi i} \frac{1}{2\pi l} \bar{\Phi}_u^{(k)}(i,l). \tag{24}$$

The interchange of the limits is an important technical point that must be justified in any particular application.

If the characteristic functions of (20)–(21) can be evaluated, then the moments and spectrum of the process can be computed from (22)–(24).

5 PCM Quantization Noise

First consider a purely deterministic input to a simple quantizer with M levels, a simple pulse coded modulation (PCM) system with no feedback. We will not consider in detail the example of a dc input to an ordinary uniform quantizer in any detail because the results are trivial. We consider a more interesting (and active) input, a sinusoid $u_n = A\sin(n\omega_0 + \theta)$ with a fixed initial phase θ. We assume that $A \leq M/2$ so that the quantizer is not overloaded. Define also $f_0 = \omega_0/2\pi$.

For the given purely deterministic example, the one-dimensional characteristic function can be expressed as

$$\bar{\Phi}_u(l) = \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} e^{j2\pi l\gamma\sin(n\omega_0+\theta)} = \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} e^{j2\pi l\gamma \sin(2\pi < nf_0 + \frac{\theta}{2\pi} >)}, \tag{25}$$

where the fractional part can be inserted since $\sin(2\pi u)$ is a periodic function in u with period 1.

This limit can be evaluated, but the result depends strongly on what is assumed about the input signal. As this is a delicate and often misunderstood issue that will arise often in the analysis of quantization systems, it merits some discussion. If we choose f_0 to be a rational number, say K/N in lowest terms, then clearly $\sin(n2\pi f_0 + \theta)$ is periodic with period N since $\sin((n + N)2\pi f_0 + \theta) = \sin(n2\pi f_0 + K2\pi + \theta) = \sin(n2\pi f_0 + \theta)$. In this case it is also true that $e^{j2\pi l\gamma\sin(n\omega_0+\theta)}$ is periodic and hence the limiting sum becomes a finite sum over a single period. It is also immediately true in this case that the quantizer output and the quantizer error signal are also periodic. If, however, the frequency is chosen at random according to a probability density function, then the probability of its being a rational number will be 0 and the probability of its being an irrational number will be 1. In this case the discrete time input signal is not periodic (it is an example of what is called an "almost periodic function" [33]). We henceforth consider only this case in detail because it is typical when the frequency is selected at random from a continuous distribution.

When the frequency is irrational, the above limit can be evaluated using a classical result in ergodic theory of Hermann Weyl (see, e.g., Petersen [34]): If g is an integrable function, a is an irrational number, and b is any real number, then

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} g(< an + b >) = \int_0^1 g(u)du. \tag{26}$$

This remarkable result follows since the sequence of numbers $< an + b >$ uniformly fills the unit interval and hence the sums approach an integral in the limit. Applying (26) to (25) yields

$$\bar{\Phi}_u(l) = \int_0^1 du e^{j2\pi l\gamma \sin(2\pi u)} = J_0(2\pi l\gamma), \tag{27}$$

where $\gamma = A/\Delta$ and J_m is the ordinary Bessel function of order m.
The mean and second moment of the quantizer noise can then be found using the fact $J_0(r) = J_0(-r)$:

$$\bar{E}\{e_n\} = \sum_{l\neq 0} \frac{1}{2\pi j l} J_0(2\pi l\gamma) = 0, \tag{28}$$

$$\bar{E}\{e_n^2\} = \frac{1}{12} + \frac{1}{\pi^2} \sum_{l=1}^{\infty} \frac{1}{l^2} J_0(2\pi l\gamma). \tag{29}$$

Note that the result does not depend on the frequency of the input sinusoid (provided the frequency is an irrational number) and that the time average mean is 0, which agrees with that predicted by the assumption that ϵ_n is uniformly distributed on $[-\Delta/2, \Delta/2]$. The second moment, however, differs from the value of 1/12 predicted by the uniform assumption by the right hand sum of weighted Bessel functions. Note that if $\gamma = A/\Delta$ becomes large (which with A held fixed and the no-overload assumption means that the number of quantization levels is becoming large), then $J_0(2\pi l\gamma) \to 0$ and hence the second moment converges to 1/12 in the limit.

To compute the autocorrelation of the quantization noise, we use similar steps to find the joint characteristic function $\bar{\Phi}_u^{(k)}(i, l)$ as

$$R_e(k) = \sum_{n=-\infty}^{\infty} S_n e^{2\pi jk\lambda_n}, \tag{30}$$

where $\lambda_n =< (2n-1)\omega_0/2\pi >$ are normalized frequencies in $[0, 1)$ and

$$S_n = (\frac{1}{\pi} \sum_{l=1}^{\infty} \frac{J_{2n-1}(2\pi\gamma l)}{l})^2 \tag{31}$$

are the spectral components at the frequency λ_n. Thus

$$S_e(f) = \sum_n S_n \delta(f - \lambda_n). \tag{32}$$

where $\delta(f)$ denotes a Dirac delta function.

The spectrum of the quantizer error therefore is purely discrete and consists of all odd harmonics of the fundamental frequency of the input sinusoid. The energy at each harmonic depends in a very complicated way on the amplitude of the input sinusoid. In particular, the quantizer noise is decidedly not white since it has a discrete spectrum and since the spectral energies are not flat. Thus here the white noise approximation of Bennett and of the describing function approach is invalid, even if M is large. Claasen and Jongepier [30] argued that if the spectrum analyzer has limited resolution, then one can make assumptions about the statistical behavior of the coefficients in the spectrum which leads to an approximately white spectrum. Indeed a short term FFT will look somewhat white, but higher resolution will clearly show the discrete nature of the error. In the $\Delta\Sigma$ case to be considered, even short term FFTs clearly show the spikes and the corresponding tones can be heard in audio reconstructions.

The crosscorrelation is handled in a similar fashion to obtain

$$R_{ue}(k) = \sum_{l\neq 0} \bar{E}\{u_n e^{j2\pi l \frac{u_{n+k}}{\Delta}}\} = A\cos(k\omega_0)\sum_{l=1}^{\infty}\frac{J_1(2\pi l\gamma)}{l}, \tag{33}$$

which is not equal to the product of the means since the error mean is 0. Thus the error and the input are not asymptotically uncorrelated.

The basic procedure used above of computing characteristic functions which in turn yield the quantization error moments and spectra can be used with more complicated input signals to obtain exact formulas which can be evaluated numerically.

6 Dithered PCM

We next consider a quantizer input process of the form the form $u_n = x_n + w_n$, where x_n is the possibly nonstationary original system input (such as the deterministic sinusoid previously considered) and w_n is an i.i.d. random process which is called a *dither* process. A key attribute of the dither process is that it is independent of the x_n process, that is, x_n is independent of w_k for all times n and k. We still require that the quantizer input u_n be in the no-overload region. This has the effect of reducing the allowed input dynamic range and hence limiting the overall SQNR. Dithering has long been used as a means of improving the subjective quality of quantized speech and images (see Jayant and Noll, Section 4.8, and the references therein [35]). The principal theoretical property of dithering was developed by Schuchman [22], who proved that if the quantizer does not overload and the characteristic function of the marginal probability density function of the dither signal is 0 at integral multiples of $2\pi/\Delta$, then the quantizer error $\epsilon_n = q(x_n + w_n) - (x_n + w_n)$ is independent of the original input signal x_n. It follows from Sripad and Snyder [3] that under these conditions the quantization error is also white. See, for example, [35, 24]. It is not true, however, that the quantization *noise* $q(x_n + w_n) - x_n$ is independent of signal or white (a common misconception that is still found in some texts, see [36, 24] for a discussion).

Given a stationary random process w_n, recall the definition of (7) of $\Phi_w(\alpha) = E(e^{j\alpha w_n})$, the ordinary characteristic function of w_n. Because of the independence of the processes, the one-dimensional characteristic function of (20) becomes

$$\bar{\Phi}_u(l) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} E(e^{2\pi j l \frac{1}{\Delta}(x_n+w_n)}) = \Phi_w(2\pi\frac{l}{\Delta})\bar{\Phi}_x(l). \tag{34}$$

The two-dimensional characteristic function of (21) is

$$\bar{\Phi}_u^{(k)}(i,l) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} E(e^{2\pi j \frac{1}{\Delta}[i(x_n+w_n)+l(x_{n+k}+w_{n+k})]})$$

$$= \Phi_w(2\pi\frac{1}{\Delta}i)\Phi_w(2\pi\frac{1}{\Delta}l)\bar{\Phi}_x^{(k)}(i,l); k \neq 0 \tag{35}$$

Now suppose that the marginal distribution of w_n is such that Schuchman's conditions are satisfied, that is, the quantizer is not overloaded and

$$\Phi_w(2\pi\frac{1}{\Delta}l) = 0;\ l = \pm 1, \pm 2, \pm 3, \cdots. \tag{36}$$

(Recall that $\Phi_w(0) = 1$ for any distribution.) This is the condition shown by Schuchman to be necessary and sufficient for the quantization error to be independent of the original input x_n. The principal example is a dither signal with a uniform marginal on $[-\Delta/2, \Delta/2]$ (and an input amplitude constrained to avoid overload when added to the dither) or sums of independent uniform variates. For this case we have

$$\bar{\Phi}_u(l) = \begin{cases} 1; & l = 0 \\ 0; & \pm 1, \pm 2, \ldots. \end{cases} \tag{37}$$

and for $k \neq 0$

$$\bar{\Phi}_u^{(k)}(i,l) = \begin{cases} \bar{\Phi}_x^{(k)}(0,0) = 1; & \text{i=l=0} \\ 0; & \text{otherwise.} \end{cases} \tag{38}$$

Thus in this example we have from (22)–(24) that e_n has zero mean, a second moment of 1/12, and an autocorrelation function $R_e(k) = 0$ when $k \neq 0$, that is, the quantization error is indeed white when Schuchman's condition is satisfied. This is true for a general quasi-stationary input, including the sinusoid previously considered. Observe that (38) is in fact a Sripad and Snyder condition for the generalized characteristic function $\bar{\Phi}$ for the quantizer input process. Since it is obtained by multiplying the characteristic function of the input by those for i.i.d. uniform random variables, the resulting dithered signal results in uniform white quantization error. If the input sequence is $A\sin(n\omega_0)$ as before and the dither sequence is an i.i.d. sequence of uniform random variables on $[-\Delta/2, \Delta/2]$, then the overload condition becomes $A/\Delta \leq M - 1/2$, which has effectively reduced the allowable γ.

A similar exercise shows that the error and input are uncorrelated. Additional effort is needed to prove independence (see, e.g., [24]).

Although dithering yields a quantizer error with nice statistical properties, it corrupts the signal (unless subtractive dither is used) and reduces the SQNR achievable with a given quantizer since the input amplitude must be reduced enough so that the original signal plus the dither stays within the no-overload region. This loss may be acceptable (and small) when the number of quantization levels is large. It is significant if there are only a few quantization levels. For example, if $M = 2$, then a uniform dither on $[-\Delta/2, \Delta/2]$ can only avoid overload if the signal is confined to have magnitude less than $\Delta/2$.

7 Single Loop $\Delta\Sigma$

The basic $\Delta\Sigma$ modulator can be motivated by an intuitive argument based on the dithering idea. Suppose that instead of adding an i.i.d. random process to the signal before quantization, the quantization noise itself is used as a dither signal, that is, i.i.d. signal-independent noise is replaced by deterministic signal-dependent noise which (hopefully) approximates a white signal-independent process. Reversing the noise sign for convenience and inserting a delay in the forward path of the feedback loop (to reflect the physical delay inherent in a quantizer) yields the system described by the nonlinear difference equation

$$u_n = x_{n-1} - \epsilon_{n-1} = u_{n-1} + x_{n-1} - q(u_{n-1});\ n = 1, 2, \cdots \tag{39}$$

This difference equation can be depicted is equivalent to the traditional discrete time form of a single-loop $\Delta\Sigma$ modulator, which can therefore be thought of as a deterministically dithered PCM system, an idea introduced in 1960 by C. C. Cutler in Figure 2 of [37], where he referred to the system as a quantizer "with a single step of error compensation." The name "Delta-Sigma" modulator was introduced by Inose and Yasuda in 1963 [38], who provided the first published description of its basic properties. The name was intended to reflect the fact that the system first took a difference (Delta) and then integrated (Sigma). The modern popularity of these systems, much of the original analysis, and the alternative name "Sigma-Delta" modulator is due to Candy and his colleagues [39, 40, 41, 42, 10]. The name $\Sigma\Delta$ reflects the fact that the system can also be represented as the cascade of an integrator (Sigma) and a Delta-modulator. In the author's opinion, this is a better name because the system does not really form a difference of successive samples of the input signal as suggested by the Delta Sigma name, it forms the difference between the input and a digital approximation of the previous input that is fed back. The name $\Delta\Sigma$ does not incorporate the key attribute of quantization in the system, the reverse order does. The author bows to the majority of coauthors, however, and adopts the older name.

Given the interpretation of the system as a deterministically dithered quantizer, one might hope that the deterministic dither might indeed yield a white quantization noise process, but unfortunately this circular argument does not hold for the simple single-loop system, as will be seen.

Since $u_n = q(u_n) - \epsilon_n$, (39) yields the difference equation

$$q(u_n) = x_{n-1} + \epsilon_n - \epsilon_{n-1}, \tag{40}$$

which has the intuitive interpretation that the quantizer output can be written as the input signal (delayed) plus a difference (or discrete time derivative) of an error signal. The hope is that this difference will be a high frequency term which can be removed by low pass filtering to obtain the original signal. For convenience we assume that $u_0 = 0$ and we normalize the above terms by Δ and use the definition of ϵ_n to write

$$e_n = \frac{\epsilon_n}{\Delta} = \frac{q(u_n) - u_n}{\Delta} = \frac{q(x_{n-1} - \epsilon_{n-1})}{\Delta} - \frac{x_{n-1}}{\Delta} - e_{n-1};\ n = 1, 2, \cdots. \tag{41}$$

Since $u_0 = 0$, $\epsilon_0 = \Delta/2$.

We shall assume that the input range is $[-b, b)$, that is, $-b \leq x_n < b$ for all n. Intuitively, we would like to make Δ small in order to keep the quantizer error small; but we dare not make it too small or the quantizer may overload (M is considered fixed). An easy induction argument in [43, 44] shows that the smallest value of Δ as a function of b for which overload never occurs is $\Delta = 2b/(M-1)$. In the most common case of a binary quantizer, $\Delta = 2b$. This implies that for Δ chosen in this manner, the Bennett condition of not overloading the quantizer is met for the simple single loop $\Delta\Sigma$ modulator.

To find an explicit expression for e_n in terms of the x_n, sum equation (40) from $k = 1$ to n:

$$\sum_{k=1}^{n} \frac{q(u_k)}{\Delta} = \sum_{k=1}^{n} \frac{x_{k-1}}{\Delta} + \sum_{k=1}^{n} e_k - \sum_{k=1}^{n} e_{k-1} = \sum_{k=1}^{n} \frac{x_{k-1}}{\Delta} + e_n - \frac{1}{2}.$$

Define $1(u) = 1$ if $u \geq 0$ and 0 otherwise so that $q(u)/\Delta = 1(u) - 1/2$ and we have for $n = 1, 2, \cdots$

$$y_n = \frac{1}{2} - e_n = \sum_{k=1}^{n} (\frac{x_{k-1}}{\Delta} + \frac{1}{2}) - \sum_{k=1}^{n} 1(u_{k-1}).$$

(y_n is more convenient to deal with than e_n.)

Taking the fractional part of both sides yields

$$< y_n >=< \sum_{k=1}^{n} (\frac{x_{k-1}}{\Delta} + \frac{1}{2}) > .$$

If $-b \leq x_n < b$ for all n, then $y_n \in [0, 1)$ and hence $< y_n >= y_n$. Thus $y_0 = 0$ and

$$y_n =< \sum_{k=0}^{n-1} (\frac{1}{2} + \frac{x_k}{\Delta}) >=< \frac{n}{2} + \sum_{k=0}^{n-1} \frac{x_k}{\Delta} >; \; n = 1, 2, \cdots$$

and hence

$$e_n = \frac{1}{2} - < \frac{n}{2} + \sum_{k=0}^{n-1} \frac{x_k}{\Delta} > . \tag{42}$$

Compare this with PCM case $e_n = 1/2 - < u_n/\Delta >$.

When the quantizer is put into a feedback loop with an integrator, the overall effect is to integrate the input plus a constant bias before taking the fractional part. The overall nonlinear feedback loop therefore appears as an affine operation (linear plus a bias) on the input followed by a memoryless nonlinearity.

The techniques used to find the time average moments for e_n in the memoryless quantizer case can now be used by replacing u_n by the sum

$$s_n = \sum_{k=0}^{n-1} (\frac{1}{2} + \frac{x_k}{\Delta}), \tag{43}$$

evaluating the characteristic functions of (20)–(21) and applying (22)–(24) for $\bar{\Phi}_s$ instead of $\bar{\Phi}_u$. Thus

$$\bar{\Phi}_s(l) = \bar{E}\{e^{\pi jln} e^{2\pi jl \sum_{i=0}^{n-1} x_i}\} \tag{44}$$

$$\bar{\Phi}_s^{(k)}(i,l) = e^{\pi jlk} \bar{E}\{e^{\pi j(i+l)n} e^{2\pi j(i+l) \sum_{m=0}^{n-1} x_m} e^{2\pi jl \sum_{m=n}^{n-1+k} x_m}\}. \tag{45}$$

To evaluate these limits it is necessary as in the PCM case to assume a particular form for the input signal. A simple but important signal is a dc value: $x_k = x$ for all k, where $-b \leq x < b$ is fixed. Although clearly of limited practical application, it can be considered as an approximation to a very slowly varying input, that is, to the case where $\Delta\Sigma$ modulator has a large oversampling ratio as it might for sensor measurements.

Analogous to the analysis for PCM with a sinusoidal input, there are two possible assumptions on the dc value which lead to nice solutions, but which have fundamentally different behavior and interpretation. The assumption that we shall make is that $x/2b$ is an irrational number. This assumption is physically and intuitively correct for an ADC because any truly analog random signal will be describable by a probability density function and hence with probability 1 will produce an irrational number. As in the PCM case, choosing $x/2b$ irrational will permit the evaluation of the above limits using asymptotic results from ergodic theory such as Weyl's theorem (26). If on the other hand it is assumed that $x/2b$ is a rational number, than the limits become finite sums and the output and error signals become periodic, that is, that "tones" or "limit cycles" are produced. Much of the analysis described here can be modified for rational inputs. For example, the analysis for single loop $\Delta\Sigma$ for rational dc inputs may be found in [45, 47]. We do not pursue the analysis for rational inputs here because in the author's view it is of little interest in describing the behavior of ADC systems to analyze carefully behavior resulting from zero probability inputs. For further discussion on the issue of irrational vs. rational dc values see Iwersen [46]. This presents a potential cause for confusion, however, because simulations of ADC behavior on a digital computer will necessarily produce rational input signals and hence the resulting periodic behavior will appear to disagree with the theory. The reconciliation of this apparent paradox is to make sure that the simulations well approximate the assumptions required by the theory. If a rational dc is selected with a modest denominator (e.g., a few hundred), then the assumption of an irrational input is clearly violated and the resulting signals will indeed be periodic and the various statistics poor approximations to the theory. If one instead generates a random number using a uniform random number generator on a digital computer, the resulting number will still be rational, but it will be "approximately" irrational in that with high probability the fraction in lowest terms will have a denominator that is extremely large (hundreds of thousands or millions). This means that the resulting signals will be periodic, but with extremely long periods so that spectral analyzers will not see the periodicities. All statistics computed will well match the theory in this case (provided of course that the theory is correct).

One system where the assumption of a rational dc input signal is valid is a DAC since by definition a digital input signal can take on only rational values. Hence the results developed here for irrational inputs do not apply to the analysis of $\Delta\Sigma$ modulators for DAC. The basic methods can still be used, but the asymptotic results must be replaced by appropriate finite sums and the answers will be different from those of the irrational case.

Assuming an irrational dc input x, we can replace u_n by $s_n = n\beta$, where $\beta = (1/2 + x/\Delta)$, in (20)–(21) and evaluate the characteristic functions using (26):

$$\bar{\Phi}_s(l) = \int_0^1 du e^{j2\pi u} = \begin{cases} 0; & \pm 1, \pm 2, \ldots. \\ 1; & l = 0 \end{cases} \tag{46}$$

$$\bar{\Phi}_s^{(k)}(i,l) = \begin{cases} e^{2\pi lk\beta}; & i = -l \\ e^{2\pi lk\beta} \int_0^1 du e^{j2\pi u} = 0; & \text{otherwise}, \end{cases} \tag{47}$$

Thus we have from (22)–(23) that $\bar{E}\{e_n\} = 0$ and $\bar{E}\{e_n^2\} = \frac{1}{12}$, which agrees with the uniform noise approximation, that is, these are exactly the time average moments one would expect with a sequence of uniform random variables. The second order properties, however, are quite different. From (24) and 1.443.3 of [48]

$$R_e(k) = \sum_{l \neq 0} (\frac{1}{2\pi l})^2 e^{j2\pi lk\beta} = \frac{1}{2}\frac{1}{\pi^2}\sum_{l=1}^{\infty} \frac{\cos(2\pi lk\beta)}{l^2} = \frac{1}{12} - \frac{< k\beta >}{2}(1 - < k\beta >). \tag{48}$$

This does not correspond to a white process. The exponential expansion implies that the spectrum is purely discrete having amplitude

$$S_n = \begin{cases} 0; & \text{if } n = 0 \\ \frac{1}{(2\pi n)^2}; & \text{if } n \neq 0. \end{cases} \tag{49}$$

at frequencies $< n\beta > = < n(1/2 + x/\Delta) >$. Thus the locations and hence the amplitude of spikes of the the quantizer error spectrum depend strongly on the value of the input signal.

Thus as in the simple PCM case with a sinusoidal input, the Bennett and describing function white noise approximations inaccurately predict the spectral nature of the quantizer noise process, which is neither continuous nor white.

Next consider a more "active" input $x_n = A\cos n\omega_0$, where $\omega_0/2\pi$ is assumed to be irrational and where we consider a full scale sinusoid with $|A| = b$ as an example. Define $\alpha = \gamma/2 \sin\frac{\omega_0}{2}$. The same general procedure with a lot more algebra [49] now results in

$$m_e = 0, \tag{50}$$

$$R_e(0) = \frac{1}{12} - \sum_{l=1}^{\infty} \frac{1}{(\pi 2l)^2}(-1)^l J_0(4\pi l\alpha), \tag{51}$$

$$R_y(k) = \sum_{m=-\infty}^{\infty} S_m e^{j2\pi k\lambda_m}, \tag{52}$$

where

$$S_m = \begin{cases} \frac{1}{2}; & m = 0 \\ (\frac{1}{\pi}\sum_{l=1}^{\infty} \frac{J_m(2\pi\alpha(2l-1))}{2l-1}(-1)^l)^2; & m \text{ even} \\ (\frac{1}{\pi}\sum_{l=1}^{\infty} \frac{J_m(4\pi\alpha l)}{2l}(-1)^l)^2; & m \text{ odd.} \end{cases} \tag{53}$$

and

$$\lambda_m = \begin{cases} < m\frac{\omega_0}{2\pi} - \frac{1}{2} >; & m \text{ even} \\ < m\frac{\omega_0}{2\pi} >; & m \text{ odd.} \end{cases} \tag{54}$$

With a sinusoidal input, the input and quantizer error are *not* uncorrelated.

As in the PCM case, the spectrum of y_n is purely discrete and has amplitude s_l at the frequency λ_l. This spectrum is extremely non-white since it is not continuous and not flat. The output frequencies depend on the input frequency ω_0 and comprise all harmonics of the input frequency ω_0. It is interesting to observe that not only are all harmonics of the input frequency contained in the output signal, but also all shifts of these harmonics by π (when computed in radians). These shifted harmonics are not present in the PCM case.

8 Two-Stage (Cascade or MASH) $\Delta\Sigma$ Modulation

We now turn to the two-stage or cascaded MASH $\Delta\Sigma$ modulator. Here two $\Delta\Sigma$ modulators are cascaded and the final output formed by a linear combination of the individual outputs.

In particular, suppose that the first $\Delta\Sigma$ has input x_n, integrator state v_n, output $q(v_n)$, and quantizer error signal $\zeta_n = q(v_n) - v_n$. This first $\Delta\Sigma$ is defined by the difference equation

$$v_n = v_{n-1} + x_{n-1} - q(v_{n-1});\ n = 1, 2, \cdots \tag{55}$$

The first stage error sequence ζ_n is the input to the second loop, where the integrator state is denoted u_n and the quantizer error is denoted ϵ_n. The difference equation for the second stage is then

$$u_n = u_{n-1} + \zeta_{n-1} - q(u_{n-1});\ n = 1, 2, \cdots \tag{56}$$

The quantizers are exactly as in Section 3, that is, uniform quantizers with M levels and bin width $\Delta = 2b/(M-1)$, where the input x_n is between $-b$ and b. Note that this means that the first quantizer does not overload and hence ζ_n is in the range $[-b, b)$ and hence also the second quantizer does not overload. The output of the two stage $\Delta\Sigma$ modulator is defined by

$$\psi_n = q(v_{n-1}) - q(u_n) + q(u_{n-1});\ n = 1, 2, \cdots, \tag{57}$$

a linear combination of the two quantizer outputs. As in (40) we have that $q(v_n) = x_{n-1} + \zeta_n - \zeta_{n-1}$ and $q(u_n) = \zeta_{n-1} + \epsilon_n - \epsilon_{n-1}$ and hence (57) becomes

$$\psi_n = x_{n-2} + \epsilon_n - 2\epsilon_{n-1} + \epsilon_{n-2}. \tag{58}$$

In contrast to (40), this has the interpretation of being the original signal plus a second order difference instead of the first order difference of the single loop system. Note that although this signal depends on the outputs of both stages, the first stage quantization noise cancels out and only the second stage noise remains. This fact is basic to the operation of the system. In particular, we need only know the behavior of the second stage quantization noise in order to find the behavior of the output. We shall see that the second stage noise is better behaved than the first stage noise.

Eq. (58) suggests that a natural means of producing the reproduction signal $\hat{x}$ is to pass ψ_n through a low pass filter and downsample.

Assuming that initially the integrator states (quantizer inputs) are $u_0 = v_0 = 0$, then applying (7) to both stages gives for $n = 1, 2, \cdots$

$$p_n = \frac{\psi_n}{\Delta} =< \frac{1}{2} - \frac{n}{2} + \sum_{k=0}^{n-1} \frac{x_k}{\Delta} >,$$

$$e_n = \frac{\epsilon_n}{\Delta} =< \frac{1}{2} - \frac{n}{2} + \sum_{k=0}^{n-1} p_k >= -\frac{1}{2} + < \sum_{l=0}^{n-1} l(\frac{1}{2} + \frac{x_{n-l}}{\Delta}) > . \qquad (59)$$

As in the ordinary $\Delta\Sigma$ modulator, we can modify (20)–(24) by replacing the quantizer input u_n by a sum term

$$s_n = \sum_{i=0}^{n-1} \sum_{l=0}^{i-1} (\frac{1}{2} + \frac{x_l}{\Delta}) = \sum_{l=0}^{n-1} l(\frac{1}{2} + \frac{x_{n-l}}{\Delta}). \qquad (60)$$

and then proceed exactly as before. We here illustrate the results only for the simple case of an irrational dc input $x_n = x$. Define $\beta = 1/2 + x/\Delta$ as before and we have that $s_n = \beta/2n^2 - \beta/2n$. The limits in the characteristic functions are evaluated using a form of Weyl's theorem to obtain

$$\bar{\Phi}_s(l) = \begin{cases} 0; & l \neq 0 \\ 1; & l = 0. \end{cases} \qquad (61)$$

$$\bar{\Phi}_s^{(k)}(i,l) = \begin{cases} 1; & i = l = 0 \\ 0; & \text{otherwise.} \end{cases} \qquad (62)$$

These characteristic functions are identical to those of the dithered PCM case in (37) and (38) and hence the conclusions are the same: the generalized Sripad and Snyder conditions are met and the quantization error in the second stage is indeed white and its marginal first and second moments agree with those of a uniform distribution! Since only the second stage error appears in the final reconstruction in an ideal system, the white noise approximation can safely be used for SQNR analysis. It is perhaps surprising that such a purely deterministic system with a fixed dc input can produce a sequence that appears to be uniformly distributed white noise when its first and second order moments are measured. A slight variation on the foregoing analysis can be used to prove that the second stage quantizer noise and the original input are asymptotically uncorrelated.

The production of a deterministic signal that masquerades as white noise is reminiscent of the theory of "chaos," the branch of nonlinear dynamical systems theory that focuses on transformations on points that produce sequences that appear to be random. There is, however, nothing chaotic about the quantization noise sequence. Technically, its Lyapunov exponent is zero and hence it is not chaotic. Chaos can be made to occur in variations on the basic $\Delta\Sigma$ architecture, e.g., by using an integrator in the feedback loop with gain greater than 1. See, e.g., Feely and Chua [50, 51] and Schreier[52].

It should be reemphasized that the above analysis depended critically on the underlying assumption of an irrational dc input. If the input were rational, then the asymptotic limits would be replaced by finite sums and the behavior would be different. In particular, the error and output sequences would be periodic and the system would exhibit "limit cycle" behavior. As previously discussed, the periodic behavior would be evident if a rational input signal with a modest denominator is chosen in a simulation. Choosing the input using a good random number generator, yields results that well approximate the asymptotic theory. It is also important to note that the solutions hold for the idealized system represented by the nonlinear difference equations. Real circuits would not have perfectly matched nonleaky integrators which would result in behavior departing from the theory. In particular, one would expect to see discrete frequency components in such systems even for irrational dc inputs (as one does for a single stage idealized system). The accuracy of the theory at predicting actual behavior for simulated or physical circuits depends strongly on the degree to which the simulations reflect the assumptions of the theory and the physical circuits implement the commonly used nonlinear difference equations used to describe the systems.

The analysis can be extended to the case of a sinusoidal input, but the analysis is much more complicated and the noise is not white [53] and it is not asymptotically uncorrelated with the input.

9 Second Order $\Delta\Sigma$ Modulation

Another $\Delta\Sigma$ system is the second order multiloop $\Delta\Sigma$ introduced by Candy [40] and first rigorously analyzed by He, Buzo, and Kuhlmann [54, 55, 56, 57]. Here a single loop $\Delta\Sigma$ modulator is imbedded in a second loop with an integrator in the feedforward path. It can be interpreted as a first-order loop with the original input replaced by the integrated error between the input and the quantizer output. From this viewpoint, the second-order $\Delta\Sigma$ is equivalent to Cutler's quantizer with "two steps of error compensation" of Figure 3 of his 1960 patent application [37] except for the location of the delay.

The basic nonlinear difference equation for the quantizer input process u_n is given by

$$u_n = x_{n-1} - 2\epsilon_{n-1} + \epsilon_{n-2}, \tag{63}$$

where, as before, $\epsilon_n = q(u_n) - u_n$ is the quantizer error. Observe that the output of the second order sigma delta modulator is

$$q(u_n) = \epsilon_n + u_n = \epsilon_n - 2\epsilon_{n-1} + \epsilon_{n-2} + x_{n-1}, \tag{64}$$

a relation which bears a remarkable resemblance to (58) for the output of the two-stage $\Delta\Sigma$ modulator (and hence is capable of the same interpretation).

A well known difficulty with the second (and higher) order $\Delta\Sigma$ modulators is their potential for quantizer overload. In particular, if one uses a binary quantizer with levels $\pm b$ in a second order system, then it is easy to find an input within the range $[-b, b)$ which will overload the quantizer and hence will be capable of producing large errors. The potential overload also has the serious consequence for our purposes that it renders invalid a basic technique of the approach used here. No application of the techniques of this chapter seems possible for the case of a binary quantizer, but the techniques do apply if we permit a two bit (or higher) quantizer. It can be shown that the smallest value of Δ for which no overload occurs is given by [56] $\Delta = 2b/(M-3)$. Clearly this result is useful only if $M \geq 4$, that is, if the quantizer has at least two bits. For the present we make this assumption and we can then proceed as earlier.

As with the first order loop analysis we normalize the error and then sum the difference equation twice (since there is a second-order difference) to find

$$y_n = \frac{1}{2} - \frac{\epsilon_n}{\Delta} = \sum_{l=1}^{n} l(\frac{1}{2} + \frac{x_{n-l}}{\Delta}) - \sum_{l=1}^{n} l1(u_{n+1-l}). \qquad (65)$$

In the special case of a dc input, (65) can be written as

$$y_n = \sum_{l=1}^{n} l\beta - \sum_{l=1}^{n} l1(u_{n+1-l}) = \sum_{l=1}^{n} l(\beta - 1(u_{n+1-l})) = \sum_{l=1}^{n} (n+1-l)(\beta - 1(u_l)), \qquad (66)$$

where as before $\beta = (1/2 + x/\Delta)$.

As in the first-order case,

$$< y_n > = < \sum_{l=1}^{n} l(\frac{1}{2} + \frac{x_{n-l}}{\Delta}) > . \qquad (67)$$

If there is no overload, then $y_n =< y_n >$ and the solution is identical to that of He et al. (and to that for the two-stage $\Delta\Sigma$ modulator [58]). The problem is that if the quantizer has only one bit, then it is not in general true that $y_n =< y_n >$ and hence (67) does not provide a solution for the error. The analysis does characterize $< y_n >$ as being a uniform white noise sequence, but this is the fractional part of the quantizer error and not in general the quantizer error itself.

There is no simple solution to this problem. The failure of the analysis to apply to the one-bit second order system does not detract from the usefulness or popularity of the system, it only leaves open the issues of finding the properties of the quantizer error and of comparing those properties to those predicted by the white noise approximation.

As will be considered in Chapter 5, exact results for the one bit second-order case have been developed and reported by many researchers, including Wang [59], Hein and Zakhor [60], and by Pinault and Lopresti [61]. Both Wang and Pinault and Lopresti used ideas from dynamical systems techniques to show that the two integrator states must eventually lie in a compact set, demonstrating a form of stability for the system. The quantizer overloads, but an absolute bound on the integrator states (and hence the quantizer error) can be found as a function of the dc input. In particular their results provide a bound of the form

$$|\frac{1}{n}\sum_{k=1}^{n}(q(u_k)-x)| \leq \frac{C}{n}, \tag{68}$$

where $C = 5/4$ using the normalizations adopted here (it can be tightened to 1). It can further be shown that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\epsilon_n = 0,$$

but the evaluation of the variance, correlation, and spectra remains an open problem.

10 Some Extensions

Dithered Single-Loop $\Delta\Sigma$

The methods can be applied to a dithered the $\Delta\Sigma$ modulator with an input of the form

$$x_n = v_n + w_n, \tag{69}$$

where w_n is an i.i.d. process that is independent of the quasi-stationary process v_n and it is assumed that the input and dither are constrained within $[-b, b)$ to avoid quantizer overload. Again the characteristic functions can be found and use to show that [62] the error sequence has 0 mean and variance 1/12, but that the spectrum of the error is not flat. For a suitable dither sequence, however, it is smooth and tends to become increasingly white as the dither signal is increased (and the input is correspondingly decreased to avoid overload). In the limit of 0 signal the quantization error becomes white. As always, these results depend strongly on the underlying assumptions. Here, for example, the conclusions cannot be trusted if the sum of the dither and signal are allowed to cause quantizer overload.

Multistage and Higher Order $\Delta\Sigma$

The two stage $\Delta\Sigma$ results can be extended to multiple stages with binary quantizers, where it can be shown that dc inputs, sinusoidal inputs, and sums of sinusoidal inputs all yield white quantization noise if the integrators are assumed to be ideal and the dc or frequencies required to be irrational [63]. Similarly, dithering multistage (two stage or more) $\Delta\Sigma$ modulators with i.i.d. noise which does not cause overload also yields white quantization noise [62]. He, Buzo, and Kuhlman found the spectra for multi-bit higher order $\Delta\Sigma$ modulators [55, 57]. As with the second order case considered here, the quantizer must have sufficient bits to avoid overload (specifically, if the number of loops is k, then k bits are needed). They consider both dc and sinusoidal inputs. As with the two stage and second order systems considered here, the M-stage (one bit per stage) and Mth order (single k bit quantizer) $\Delta\Sigma$ modulators yield the same quantization noise spectra.

Leaky Integrating $\Delta\Sigma$

All of the systems considered here so far have a key aspect in common that permitted solution: the linear filtering within the loop consisted only of ideal discrete time integrators; more complicated filters such as leaky integrators or integrators with non-unity gain were not considered. While results for general filtering do not exist, the special case of leaky integration and non-unity gain has been considered. Kieffer [64, 65, 66] has extended the single-stage $\Delta\Sigma$ result to more general systems with dc inputs which include DPCM and leaky integrating $\Delta\Sigma$s, but his techniques differ from those considered here and have not been fully exploited for the $\Delta\Sigma$ application. Feely and Chua [50] used ideas from dynamical systems theory to describe various properties of a leaky $\Delta\Sigma$ modulator, including the input-output relation. The methods described here can be applied to the leaky integrator and the integrator with non-unit gain [67]. The analysis is complicated, but the results can be easily summarized.

For the special case of a dc input, the following properties hold:

- The quantizer sequence and error sequence are periodic, even for irrational dc inputs (unlike ideal integrator case).
- The error sequence is no longer uniform and the mean is not the input dc. This means traditional decoders (low pass filters) give biased reproduction (unlike ideal integrator

case). In fact, it can be shown that for a dc input, the input/output relation resulting when a long comb filter is used for digital-to-analog conversion is a form of Cantor function or a“devil's staircase” and it is a complicated function to describe analytically. Even when an arbitrarily large number of bits are used to reconstruct the input, the output is biased.

- The quantizer errors are not white.
- The SQNR is reduced from the ideal integrator case.

Multibit Quantizer, Single-bit Feedback

The methods described here do not work for all popular $\Delta\Sigma$ architectures, but they do work for a variety of systems. One such example of interest is the system introduced by Leslie and Singh [68, 69] Although the theoretical treatment of $\Delta\Sigma$ modulation given here permits multibit quantization, such systems have the practical shortcoming of requiring extremely accurate digital-to-analog conversion in the feedback loop (a problem which vanishes in the binary quantizer case). Leslie and Singh proposed combining multibit in the forward loop with single bit in the feedback loop. Analysis shows that the performance of this system is identical to that of an ordinary multibit single stage $\Delta\Sigma$ having an additional bit in the quantizer (and feedback loop). Hence previous analysis applies [70].

Related Work

During recent years many efforts have been made to find and apply exact analysis methods to $\Delta\Sigma$ modulators for the purpose of describing their behavior, predicting their performance, and developing improved systems. These works have in common with this chapter the goal of avoiding unjustified application of the white noise approximation, but the detailed methods and applications are not constrained to those described here. Of particular relevance to the issues considered here are the work of Delchamps on the behavior of control systems containing quantizers inside of feedback loops [71, 72], the work of Galton demonstrating the existence of stationary distributions for the error sequence for a general class of $\Delta\Sigma$ modulators with random inputs such as Gaussian processes [58, 73, 74, 75], the work of Kieffer [76, 64, 65] and Kieffer and Dunham [77] on the stability and convergence of one-bit feedback quantizers, and the work of Thao and Vetterli [79, 78] and Hein and Zakhor

[60, 80] on optimal nonlinear decoders for $\Delta\Sigma$ modulators.

11 Final Thoughts

It might be said that the theory has failed to keep up with the fast pace of practice, that the best $\Delta\Sigma$ modulators have been developed based on engineering insight and suspect approximations, and exact analysis has usually followed far behind, if at all. Nonetheless, this chapter argued for a proper appreciation of the common approximation techniques, their origins, and their limitations, and to demonstrate several important examples where exact analysis is possible. Many open problems remain, including the evaluation of the second order properties of the basic one-bit second order $\Delta\Sigma$ modulator as well as the general stability properties as well as first and second order properties of the wide variety of hybrid cascade and higher order systems that have been proposed. Perhaps with time some of these difficult problems may yet yield to solution or interesting new systems may be found by modifying successful systems to make them more amenable to exact analysis.

Acknowledgements

I would like to thank Ping-Wah Wong, Wu Chou, Sang Ju Park, and Rick VanderKam for their cooperation on much of the research reported herein. I would also like to thank A. H. Gray for first interesting me in quantization theory and pointing out the errors common in applications of Bennett's theory, Gabor Temes and Jim Candy for the pleasure of lecturing with them in several short courses on $\Delta\Sigma$ modulation, and István Kollár for helpful discussions on the differing properties of the approaches of Widrow, Sripad, and Snyder and of Bucklew and Gallagher. Much of the research leading to the results described here was funded by the National Science Foundation under Grant No. MIP-9014335 and by a Stanford University Center for Integrated Systems Seed Grant.

References

[1] W. R. Bennett, "Spectra of quantized signals," *Bell Systems Technical Journal*, vol. 27, pp. 446–472, July 1948.

[2] B. Widrow, "A study of rough amplitude quantization by means of Nyquist sampling theory," *IRE Transactions Circuit Theory*, vol. CT-3, pp. 266–276, 1956.

[3] 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 Process.*, vol. ASSP-25, pp. 442–448, October 1977.

[4] István Kollár, private communication.

[5] J.A. Bucklew and N.C. Gallagher, Jr., "Some properties of uniform step size quantizers," *IEEE Transactions on Information Theory*, Vol. IT-26, pp. 610–613, 1980.

[6] J.A. Bucklew and N.C. Gallagher, Jr., "A note on optimum quantization," *IEEE Transactions on Information Theory*, Vol. IT-25, pp. 365–366, 1979.

[7] J.A. Bucklew, "Two results on the asymptotic performance of quantizers," *IEEE Transactions on Information Theory*, Vol. IT-30, pp. 341–348, 1984.

[8] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Boston, 1992.

[9] T. Misawa, J. E. Iwersen, L. J. Loporcaro, and J. G. Rush, "A single-chip CODEC with filters utilizing $\Delta - \Sigma$ modulation," *IEEE J. Solid State Circuits*, vol. SC-16, pp. 333–341, August 1981.

[10] J. C. Candy and O. J. Benjamin, "The structure of quantization noise from sigma-delta modulation," *IEEE Trans. Comm.*, vol. COM-29, pp. 1316–1323, Sept. 1981.

[11] M. Vidyasagar, *Nonlinear Systems Analysis.* Englewood Cliffs,New Jersey: Prentice Hall, 1978.

[12] A. Gelbe and W. E. V. Velde, *Multiple-Input Describing Functions and Nonlinear Systems Design.* New York: McGraw-Hill, 1968.

[13] D. P. Atherton, *Stability of Nonlinear Systems.* Chichester: Research Studies Press: Wiley, 1981.

[14] D. P. Atherton, *Nonlinear Control Engineering.* New York: Van Nostrand Theinhold, 1982.

[15] A. R. Bergens and R. L. Franks, "Justification of the describing function method," *SIAM Journal of Control*, vol. 9, pp. 568–589, 1971.

[16] R. C. Booton, Jr., "The analysis nonlinear control systems with random inputs," in *Proceedings of the Symposium on Nonlinear Circuit Analysis*, (Polytechnic Institute of Brooklyn), April 1953.

[17] R. C. Booton, Jr., "Nonlinear control systems with statistical inputs," tech. rep., Massachusetts Institute of Technology, Cambridge, Mass., March 1952.

[18] S. H. Ardalan and J. J. Paulos, "An analysis of nonlinear behavior in delta-sigma modulators," *IEEE Trans. Circuits and Systems*, vol. CAS-34, pp. 593–603, June 1987.

[19] D. S. Arnstein, "Quantization error in predictive coders," *IEEE Trans. Comm.*, vol. COM-23, pp. 423–429, April 1975.

[20] D. Slepian, "On delta modulation," *Bell Syst. Tech. J.*, vol. 51, pp. 2101–2136, 1972.

[21] L. G. Roberts, "Picture coding using pseudo-random noise," *IRE Trans. on Information Theory*, vol. IT-8, pp. 145–154, February 1962.

[22] L. Schuchman, "Dither signals and their effects on quantization noise," *IEEE Transactions on Communication Technology*, vol. COM-12, pp. 162–165, December 1964.

[23] J. Vanderkooy and S. P. Lipshitz, "Dither in digital audio," *J. Audio Eng. Soc.*, vol. 35, pp. 966–975, December 1987.

[24] R. M. Gray and T. J. Stockham, Jr., "Dithered quantizers," *IEEE Trans. Inform. Theory*, vol. 38, pp. 805–812, May 1993.

[25] S. O. Rice, "Mathematical analysis of random noise," in *Selected papers on noise and stochastic processes* (N. Wax and N. Wax, eds.), pp. 133–294, New York, NY: Dover, 1954. Reprinted from Bell Systems Technical Journal,Vol. 23:282–332 (1944) and Vol. 24: 46–156 (1945).

[26] W. B. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise.* New York: McGraw-Hill, 1958.

[27] A. G. Clavier, P. F. Panter, and D. D. Grieg, "Distortion in a pulse count modulation system," *AIEE Transactions*, vol. 66, pp. 989–1005, 1947.

[28] A. G. Clavier, P. F. Panter, and D. D. Grieg, "PCM distortion analysis," *Electrical Engineering*, pp. 1110–1122, November 1947.

[29] B. Widrow, "Statistical analysis of amplitude quantized sampled data systems," *Transactions Amer. Inst. Elec. Eng.,Pt. II: Applications and Industry*, vol. 79, pp. 555–568, 1960.

[30] T. A. C. M. Claasen and A. Jongepier, "Model for the power spectral density of quantization noise," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-29, pp. 914–917, August 1981.

[31] J. E. Iwersen, "Calculated quantizing noise of single-integration delta-modulation coders," *Bell Syst. Tech. J.*, pp. 2359–2389, September 1969.

[32] L. Ljung, *System Identification.* Englewood Cliffs,NJ: Prentice-Hall, 1987.

[33] H. Bohr, *Almost Periodic Functions*, Chelsea, New York, 1947.

[34] K. Petersen, *Ergodic Theory.* Cambridge: Cambridge University Press, 1983.

[35] N. S. Jayant and P. Noll, *Digital Coding of Waveforms.* Englewood Cliffs,New Jersey: Prentice-Hall, 1984.

[36] S. P. Lipshitz, R. A. Wannamaker, J. Vanderkooy, and J. N. Wright, "Non-subtractive dither," *IEEE Transactions on Signal Processing*, 1993. to appear.

[37] C. C. Cutler, "Transmission systems employing quantization," 1960. U. S. Patent No. 2,927,962.

[38] H. Inose and Y. Yasuda, "A unity bit coding method by negative feedback," *Proc. IEEE*, vol. 51, pp. 1524–1535, November 1963.

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

[40] J. C. Candy, "A use of double integration in sigma delta modulation," *IEEE Trans. Comm.*, vol. COM-33, pp. 249–258, March 1985.

[41] J. C. Candy, "Decimation for sigma delta modulation," *IEEE Trans. Comm.*, vol. COM-34, pp. 72–76, January 1986.

[42] J. C. Candy, Y. C. Ching, and D. S. Alexander, "Using triangularly weighted interpolation to get 13-bit PCM from a sigma delta modulator," *IEEE Trans. Comm.*, pp. 1268–1275, November 1976.

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

[44] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inform. Theory*, vol. IT-36, pp. 1220–1244, November 1990.

[45] V. Friedman, "Structure of the limit cycles in sigma delta modulation," *IEEE Trans. Commun.*, vol. 36, no. 8, , pp. 972–979, Aug 1988.

[46] J. E. Iwersen, "Comments on 'The structure of the limit cycles in sigma delta modulation'." *IEEE Transactions on Communications*, vol.38, no.8, p. 1117, Aug. 1990.

[47] R.M. Gray, "Spectral Analysis of Quantization noise in single-loop sigma-delta modulation with dc inputs," *IEEE Transactions on Communications,* pp. 588-599, June 1989.

[48] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals,Series,and Products.* New York: Academic Press, 1965.

[49] 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. IT-35, pp. 956–968, 1989.

[50] O. Feely and L. Chua, "The effect of integrator leak in $\Sigma - \Delta$ modulation," *IEEE Trans. Circuits and Systems*, vol. 38, pp. 1293–1305, November 1991.

[51] O. Feely and L. Chua, "Nonlinear dynamics of a class of analog-to-digital converters," *International Journal of Bifurcation and Chaos,* Vol. 2, pp. 325–340, June 1992.

[52] R. Schreier, "Destabilizing limit cycles in Delta-Sigma modulators with chaos," *Proceedings of the 1993 International Symposium on Circuits and Systems*, Chicago, Ill., pp. 1369–6, 1993.

[53] P. W. Wong and R. M. Gray, "Two stage sigma-delta modulation," *IEEE Trans. Acoust. Speech Signal Process.*, pp. 1937–1952, November 1989.

[54] N. He, A. Buzo, and F. Kuhlmann, "A frequency domain waveform speech compression system based on product vector quantizers," in *International Conference on Acoustics, Speech, and Signal Processing*, (Tokyo, Japan), April 1986.

[55] N. He, A. Buzo, and F. Kuhlmann, "Multi-loop sigma-delta quantization: Spectral analysis," in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 1870–1873, 1988.

[56] N. He, F. Kuhlmann, and A. Buzo, "Double-loop sigma-delta modulation with dc input," *IEEE Trans. Comm.*, vol. COM-38, pp. 487–495, 1990.

[57] N. He, F. Kuhlmann, and A. Buzo, "Multi-loop sigma-delta quantization with dc input," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1015–28, 1993.

[58] P. W. Wong and R. M. Gray, "Sigma-delta modulation with i.i.d. gaussian inputs," *IEEE Trans. Inform. Theory*, vol. IT-36, pp. 784–778, July 1990.

[59] H. Wang, "A geometric view of $\Sigma - \Delta$ modulation," *IEEE Transactions on Circuits and Systems-II*, vol. 39, pp. 402–405, June 1992.

[60] S. Hein and A. Zakhor, "Stability and scaling of double loop $\Sigma - \Delta$ modulators," in *Proceedings 1992 ISCAS*, pp. 1312–1315, IEEE, 1992.

[61] S. C. Pinault and P. V. Lopresti, "On the behavior of the double loop Sigma Delta modulator," *IEEE Trans. Circuits and Systems*, to appear.

[62] W. Chou and R. M. Gray, "Dithering and its effects on Sigma-Delta and multistage Sigma-Delta modulation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 500–513, May 1991.

[63] W. Chou, P. W. Wong, and R. M. Gray, "Multi-stage Sigma-Delta modulation," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 784–796, 1989.

[64] J. C. Kieffer, "Sturmian minimal systems associated with the iterates of certain functions on an interval," in *Proceedings of the Special Year on Dynamical Systems*, Lecture Notes in Mathematics, Springer-Verlag, 1988.

[65] J. C. Kieffer, "Analysis of DC input response for a class of one-bit feedback encoders," *IEEE Trans. Comm.*, vol. COM-38, pp. 337–340, 1990.

[66] J. C. Kieffer, "Note on 'Spectral analysis of quantization noise in a single-loop sigma-delta modulator with dc input'," *IEEE Trans. Comm.*, Vol. 38, pp. 337–340, March 1990.

[67] S. J. Park and R. M. Gray, "Sigma-Delta modulation with leaky integration and constant input," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1512–1533, September 1992.

[68] T. Leslie and B. Singh, "An improved sigma-delta modulator architecture," in *Proceedings 1990 IEEE International Symposium on Circuits and Systems*, vol. 1, (New Orleans, LA), pp. 372–5, IEEE, May 1990.

[69] T. Leslie and B. Singh, "Sigma-delta modulators with multibit quantising elements and single-bit feedback," *IEE Proceedings G (Circuits, Devices and Systems)*, vol. 139, pp. 356–62, June 1992.

[70] S. J. Park and R. M. Gray, "Sigma-Delta modulation with leaky integration and constant input," in *Abstracts of the 1991 IEEE International Symposium on Information Theory*, (Budapest, Hungary), p. 119, IEEE, June 1991.

[71] D. Delchamps, "Exact asymptotic statistics for sigma-delta quantization noise," in *Proceedings Twenty-Eighth Annual Allerton Conference on Communication, Control and Computing*, (Monticello, IL), Oct 1990.

[72] D. Delchamps, "Quantizer dynamics and their effect on the performance of digital feedback control systems," in *Proceedings of the 1992 American Control Conference*, vol. 3, (Chicago, IL), pp. 2498–503, American Autom. Control Council, June 1992.

[73] I. Galton, "Granular quantization noise in the first-order $\delta - \sigma$ modulator," *IEEE Trans. Inform. Theory*, Vol. 39, pp. 1944–1956, November 1993.

[74] I. Galton, "Granular quantization noise in a class of $\delta\sigma$ modulators," *IEEE Trans. Inform. Theory*, 1993. submitted.

[75] T. Koski, "Statistics of the binary quantizer error in sigma delta modulation with i.i.d. input," *IEEE Trans. Inform. Theory*, to appear 1994.

[76] J. C. Kieffer, "Stochastic stability for feedback quantization schemes," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 248–254, March 1982.

[77] J. C. Kieffer and J. G. Dunham, "On a type of stochastic stability for a class of encoding schemes," *IEEE Trans. Inform. Theory*, Vol. IT-29, pp. 703–797, November 1983.

[78] N. Thao and M. Vetterli, "Optimal MSE signal reconstruction in oversampled A/D conversion using convexity," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, IEEE, March 1992.

[79] N. Thao and M. Vetterli, "Oversampled A/D conversion using alternate projections," in *Proceedings of the Twenty-fifth Annual Conference on Information Sciences and Systems*, pp. 241–248, 1991.

[80] S. Hein and A. Zakhor, *Sigma Delta Modulators: Nonlinear Decoding Algorithms and Stability Analysis.* Boston: Kluwer Academic Publishers, 1993.