Granular Quantization Noise in a Class of Delta–Sigma Modulators

Ian Galton, *Member, IEEE*

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 input-referred circuit noise or dither is present, the second-order asymptotic statistics of the granular quantization noise sequences are determined and ergodic properties are derived.**

Index Terms—**Sigma–delta, delta–sigma, oversampling, analog-to-digital conversion, quantization.**

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 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. Without loss of generality, the output sequence from each quantizer can be taken to equal its input sequence plus a quantization noise sequence (equal to the output minus the input of the quantizer). In most $\Delta\Sigma$ modulators, the idea is to high-pass filter all the quantization noise sequences while simply delaying or low-pass filtering the $\Delta\Sigma$ modulator input sequence.

Manuscript received March 27, 1992; revised August 12, 1993.

The author is with the Department of Electrical and Computer Engineering, University of California, Irvine, CA 92717.

IEEE Log Number 9401816.

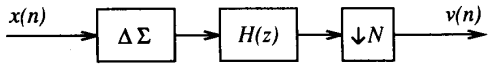

Fig. 1. A $\Delta\Sigma$ modulator-based oversampling A/D converter.

A $\Delta\Sigma$ modulator-based oversampling A/D converter consists of a $\Delta\Sigma$ modulator, a low-pass filter, and an N-sample decimator as shown in Fig. 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 $[-\pi/N, \pi/N]$. Since the $\Delta\Sigma$ modulator high-pass 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 $[-\pi/N, \pi/N]$ and can, in principle, be removed without distorting the input sequence by subsequent low-pass filtering. The decimator reduces the filtered output sequence to the Nyquist rate.

The filters applied to the input sequence and to the quantization noise sequences by a given $\Delta\Sigma$ modulator can generally be determined using linear system theory. However, quantization is a nonlinear process that is noninvertible, 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 distributed (i.i.d.) 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 have been limited to first-order $\Delta\Sigma$ modulators, cascades of first-order $\Delta\Sigma$ modulators, and certain higher order $\Delta\Sigma$ modulators operating on simple input sequences such as constants and sinusoids [4]–[6]. In the third category, various results have been developed for the first-order $\Delta\Sigma$ modulator [7], [8] and cascades of first-order $\Delta\Sigma$ modulators [7]. The approach has the benefit that it can be used to obtain rigorous results for a large class of input sequences [8]. The amplitude of the i.i.d. random component can be arbitrarily small, so the

assumption tends not to be restrictive in practice [8]. For example, thermal noise at the analog input of a ΔΣ modulator can be modeled as an i.i.d. random sequence.[1]

The current work extends the earlier results that characterize quantization noise for input sequences containing an i.i.d. random component. In particular, results are developed for a generic ΔΣ modulator of which many of the previously published ΔΣ modulators are special cases. The second-order asymptotic statistics of the quantization noise sequences are evaluated, and the various second-order statistical correlations are evaluated and shown to converge to the corresponding time-average correlations in probability. In addition to generalizing the earlier work to a larger class of ΔΣ modulators, the current work weakens the previous assumptions made about the i.i.d. random component of the input sequence, and does not assume that quantizer overload is avoided.

The remainder of the paper is divided into four main sections. The generic ΔΣ modulator is developed in Section II. In Section III, a granular quantization noise sequence expression is derived. In Section IV, the asymptotic statistics of the granular quantization noise sequences are determined, the second-order correlations are deduced, and the ergodic properties are derived. In Section V, the effect of quantizer overload is briefly discussed with respect to the results derived in the previous sections.

II. A Generic ΔΣ Modulator

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

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

In analyzing the generic ΔΣ modulator, we will make use of the *matrix transfer functions* of $\boldsymbol{T}(z)$ and $\boldsymbol{U}(z)$. The behavior of any multi-input multi-output LTI system can be represented mathematically by a matrix transfer function [9]. For example, $\boldsymbol{T}(z)$ 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 kth input node to the jth output node. If we define

$$\boldsymbol{r}(n) = \begin{bmatrix} r_1(n) \\ \vdots \\ r_K(n) \end{bmatrix} \text{ and } \boldsymbol{s}(n) = \begin{bmatrix} s_1(n) \\ \vdots \\ s_K(n) \end{bmatrix}$$

where $r_k(n)$ and $s_k(n)$ are the sequences denoted in Fig. 3, and define $\boldsymbol{R}(z)$ and $\boldsymbol{S}(z)$ to be the vectors obtained by z transforming the elements of $\boldsymbol{r}(n)$ and $\boldsymbol{s}(n)$, respectively, then $\boldsymbol{R}(z)$ is related to $\boldsymbol{S}(z)$ as

$$\boldsymbol{R}(z) = \boldsymbol{T}(z)\begin{bmatrix} X(z) \\ \boldsymbol{S}(z) \end{bmatrix}.$$

Similarly, $\boldsymbol{U}(z)$ can be represented by a $1 \times K$ vector of z transforms $U_k(z)$, $1 \le k \le K$. Thus, $Y(z) = \boldsymbol{U}(z)\boldsymbol{S}(z)$ where $Y(z)$ is the z transform of $y(n)$ in Fig. 3.

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

$$\boldsymbol{F}(z) = \begin{bmatrix} F_1(z) \\ \vdots \\ F_K(z) \end{bmatrix}$$

and

$$\boldsymbol{G}(z) = \begin{bmatrix} G_{1,1}(z) & \cdots & G_{1,K}(z) \\ \vdots & \ddots & \vdots \\ G_{K,1}(z) & \cdots & G_{K,K}(z) \end{bmatrix}$$

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

$$\boldsymbol{T}(z) = [\boldsymbol{F}(z) | \boldsymbol{G}(z)].$$

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 *product quantizer* of uniform scalar quantizers operating on the vector $\boldsymbol{r}(n)$ and producing the vector $\boldsymbol{s}(n)$. With $q_k(\cdot)$ denoting the functional operation of the kth quantizer in Fig. 3, define the vector-valued vector function $\boldsymbol{q}(\cdot)$ as

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

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

[1] A random sequence that is intentionally added to an input sequence is usually referred to as a *dither sequence*. However, random sequences such as those arising from thermal noise are not generally referred to as dither sequences, so we have avoided using this terminology. Nevertheless, from a mathematical point of view, it is irrelevant whether the i.i.d. random component of the input sequence arises from incidental causes or from dithering.

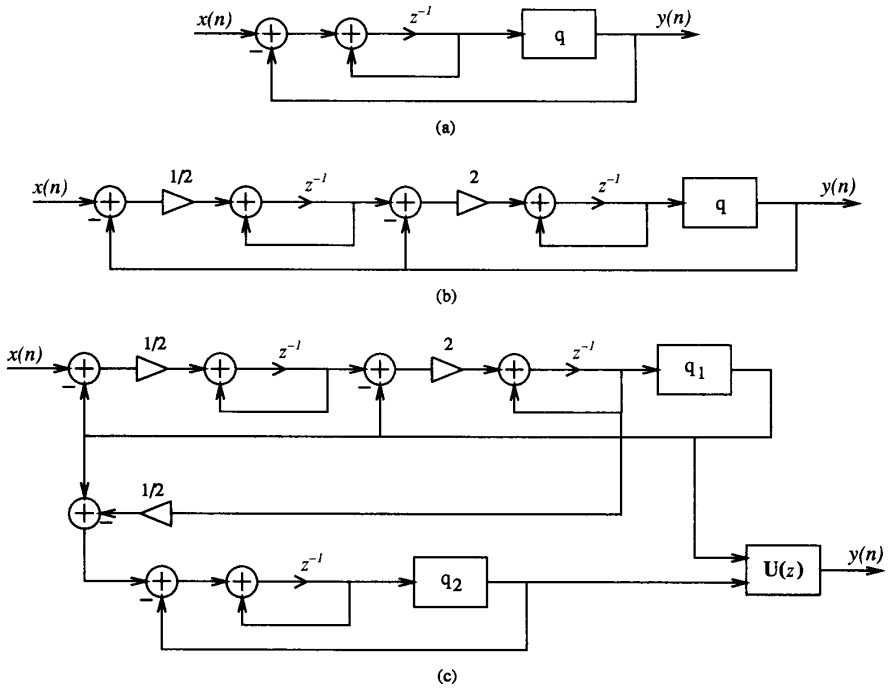

Fig. 2. (a) The first-order $\Delta\Sigma$ modulator. (b) The second-order double-loop $\Delta\Sigma$ modulator. (c) A cascade $\Delta\Sigma$ modulator that consists of a second-order double-loop $\Delta\Sigma$ modulator and a first-order $\Delta\Sigma$ modulator.

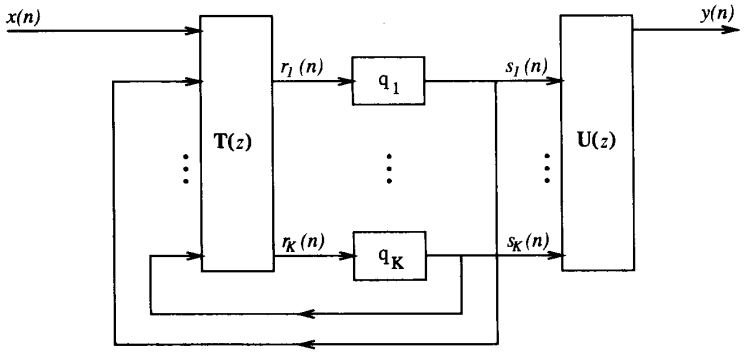

Fig. 3. A generic $\Delta\Sigma$ modulator architecture.

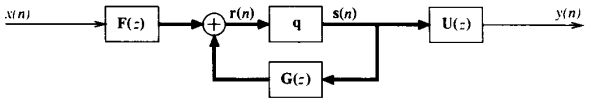

Fig. 4. A rearrangement of the generic $\Delta\Sigma$ modulator that is equivalent to the generic $\Delta\Sigma$ modulator in terms of its input–output relationship and the signals seen by its quantizers.

The results developed in this paper are applicable to any $\Delta\Sigma$ modulator that can be written in the form of Fig. 3 or, equivalently, Fig. 4, and that further satisfies the four additional conditions listed below. The four conditions make the analysis of the generic $\Delta\Sigma$ modulator tractable. Although the conditions place additional restrictions on the class of $\Delta\Sigma$ modulators to which the results of this paper apply, the restricted class is still quite general.

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

Condition 2: 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 all n.

Condition 3: For each k, the impulse response $f_k(n)$ does not converge to zero as $n \to \infty$, and for each p and $j \neq k$, the sequence $\{t_0 f_j(n) + t_1 f_k(n+p)\}$ does not converge to zero as $n \to \infty$ for any nonzero vector (t_0, t_1).

For some of the results, we will also assume that the ΔΣ modulator satisfies the following condition.

Condition 4: For each k and each $p \neq 0$, the sequence $\{t_0 f_k(n) + t_1 f_k(n+p)\}$ does not converge to zero as $n \to \infty$ for any nonzero vector (t_0, t_1).

By Condition 1, the quantizers are uniform midrise quantizers, each of arbitrary bit rate and number of output levels. The main restriction here is that the quantizers be uniform; although the results of the paper apply to uniform midrise quantizers, they can easily be derived for other uniform quantizers such as uniform midstep quantizers.

Condition 2 allows the separation of the granular and overload components of the quantization noise performed in the next section. If all the quantizers have the same step size, then the condition requires that the impulse responses of the feedback paths around the quantizers be integer-valued. Such is the case, for example, when the feedback paths contain combinations of discrete-time integrators and differencers.

Condition 3 rules out the possibility that the open-loop impulse responses between the input and any two quantizers converge to zero or to other linearly dependent sequences. The condition is essential to the derivation of Section IV where it is shown that the quantization noise sequences from any pair of quantizers are asymptotically independent. Similarly, Condition 4 rules out the possibility that the open-loop impulse response from the input to any quantizer and any shifted version of it converge to linearly dependent sequences. It is essential to the derivation of Section IV where it is shown that, provided the condition holds, the quantization noise sequence from each quantizer is asymptotically white.

Many of the common ΔΣ modulator variations satisfy Conditions 1–3. For example, the conditions are satisfied by typical first- and second-order ΔΣ modulators and most of the multistage ΔΣ modulators. With the notable exception of the first-order ΔΣ modulator, most of these also satisfy Condition 4. Throughout the paper, we will tacitly assume that the ΔΣ 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, applicable results that are analogous to those that assume Condition 4 in this paper have been developed in [8].

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

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

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

In the case of the second-order ΔΣ modulator shown in Fig. 1(b), we have

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

$$F(z) = \frac{z^{-2}}{(1-z^{-1})^2},$$

and

$$G(z) = -\frac{z^{-2}}{(1-z^{-1})^2} - \frac{2z^{-1}}{1-z^{-1}}. \quad (2)$$

Finally, in the case of the cascaded ΔΣ modulator shown in Fig. 2(c), we have

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

$$\boldsymbol{F}(z) = \begin{bmatrix} \dfrac{z^{-2}}{(1-z^{-1})^2} \\ \dfrac{1}{2}\dfrac{z^{-3}}{(1-z^{-1})^3} \end{bmatrix},$$

and

$$\boldsymbol{G}(z) = \begin{bmatrix} -\dfrac{z^{-2}}{(1-z^{-1})^2} - \dfrac{2z^{-1}}{1-z^{-1}} & 0 \\ \dfrac{1}{2}\dfrac{z^{-2}}{(1-z^{-1})^2} + \dfrac{2z^{-1}}{1-z^{-1}} & -\dfrac{z^{-1}}{1-z^{-1}} \end{bmatrix}. \quad (3)$$

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

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

Through straightforward matrix manipulations, the system can be rearranged as shown in Fig. 5(b) where

$$\boldsymbol{S}(z) = \boldsymbol{U}(z)(\boldsymbol{I} - \boldsymbol{G}(z))^{-1}\boldsymbol{F}(z) \quad (4)$$

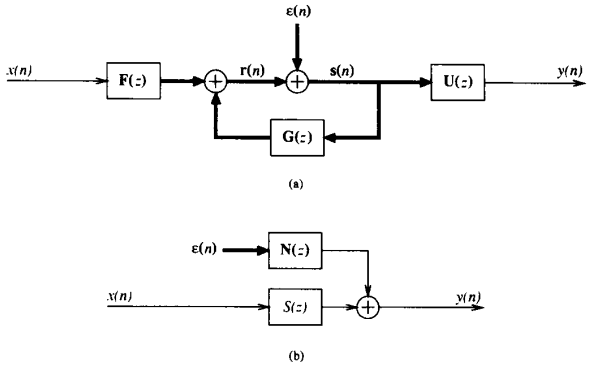

Fig. 5. (a) The system of Fig. 4 with the quantizer bank replaced with the equivalent additive noise source. (b) An equivalent form of the system showing the different filters that act on the input and quantization noise, respectively.

and

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

Note that $S(z)$ is a scalar transfer function, whereas $\boldsymbol{N}(z)$ 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 $\boldsymbol{N}(z)$. Therefore, we will refer to $S(z)$ as the *signal filter* and to $\boldsymbol{N}(z)$ as the *noise filter*. In many $\Delta\Sigma$ modulators, $\boldsymbol{T}(z)$ and $\boldsymbol{U}(z)$ are chosen such that $S(z)$ is a pure delay, while $\boldsymbol{N}(z)$ is a high-pass filter.

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

$$\boldsymbol{U}(z) = \begin{bmatrix} z^{-1}\left(1 + (1 - z^{-1})^2\right) & 2(1 - z^{-1})^2 \end{bmatrix}$$

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

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

respectively.

Up to this point, we have simply defined the generic $\Delta\Sigma$ modulator and have applied some basic linear system 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 Fig. 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 of the quantization noise introduced by the bank of quantizers and regarding ergodic properties of the system.

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 $n < a$, we will take the input sequence and all state variables of 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 always been in operation.

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 nonoverloadable 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

$$L_k(x) = \begin{cases} x & \text{if } x \in (-\gamma_k, \gamma_k] \\ -\gamma_k + \dfrac{\Delta_k}{2} & \text{if } x \le -\gamma_k \\ \gamma_k - \dfrac{\Delta_k}{2} & \text{if } x > \gamma_k. \end{cases}$$

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

Let us define the granular and overload quantization noise vectors as

$$\boldsymbol{\epsilon}_g(n) = \begin{bmatrix} \epsilon_{g_1}(n) \\ \vdots \\ \epsilon_{g_K}(n) \end{bmatrix} \text{ and } \boldsymbol{\epsilon}_o(n) = \begin{bmatrix} \epsilon_{o_1}(n) \\ \vdots \\ \epsilon_{o_K}(n) \end{bmatrix},$$

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 $\boldsymbol{\epsilon}(n) = \boldsymbol{\epsilon}_g(n) + \boldsymbol{\epsilon}_o(n)$.

Since Q_k is a non-overloadable uniform midrise quantizer with step size Δ_k, the granular quantization noise

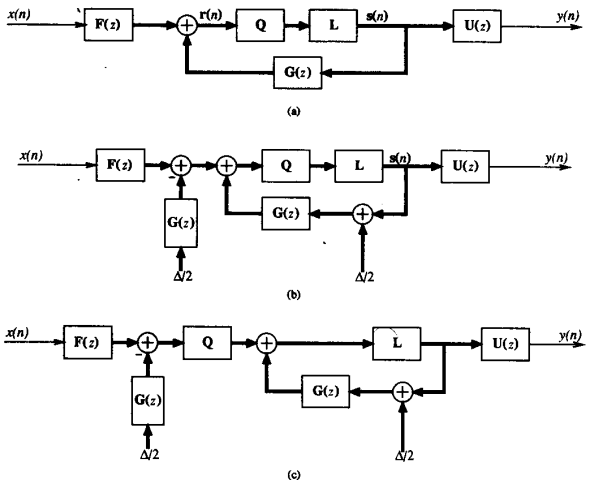

Fig. 6. (a) The system of Fig. 4 with the quantizer bank replaced by the equivalent cascade of nonoverloadable quantizers and amplitude limiters. (b) An equivalent form of the system in which $\Delta/2$ has been effectively added and subtracted from the system. (c) An equivalent form of the system in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

introduced by q_k can be written as

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

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 Fig. 6(a) 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 Fig. 6(b) is equivalent to that of Fig. 6(a) because $\Delta/2$ has been effectively added and subtracted from the system. Because the quantizers are midrise quantizers, $s(n) + \Delta/2$ is a vector whose kth element is an integer multiple of Δ_k for each index k and each time n. By Condition 2, the output of $G(z)$ 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 Q is independent of the feedback loop in Fig. 6(b). For this reason, we can move Q to the left of the feedback loop to obtain an equivalent system as shown in Fig. 6(c).

From Fig. 6(c), it is straightforward to rewrite (6) as

$$\epsilon_{g_k}(n) = \frac{\Delta_k}{2} - \Delta_k \left\langle \beta_{k,n-a} + \frac{1}{\Delta_k} \sum_{m=0}^{n-a} f_k(m) x(n-m) \right\rangle \tag{7}$$

where

$$\beta_{k,p} = \sum_{l=1}^{K} \sum_{m=0}^{p} \frac{\Delta_l}{2\Delta_k} g_{k,l}(m).$$

As a consequence of Condition 2, $\beta_{k,p}$ only takes on values that are integer multiples of $\frac{1}{2}$.

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 this 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 (i.i.d) random variables that are independent of $x_d(n)$ and whose distribution function has an absolutely continuous component [10]. 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 assumed to be an unrelated sequence introduced by 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 i.i.d. random sequence.[2]

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.

Theorem 1: For each j, k, $1 \le j, k \le K$ and each set of integers n_1, n_2,

(i) $(\epsilon_{g_j}(n_1), x(n_2))$ converges in distribution to $(\epsilon'_{g_j}(n_1), x(n_2))$ as $a \to -\infty$ where $\epsilon'_{g_j}(n_1)$ and $x(n_2)$ are independent and $\epsilon'_{g_j}(n_1)$ is uniformly distributed on $(-\Delta_j/2, \Delta_j/2\,]$;

[2]Although other types of circuit noise that are not well approximated by i.i.d. sequences will also be present in $x(n)$, their presence does not affect our argument as they can be left as part of $x_d(n)$.

(ii) if $j \neq k$, or if $n_1 \neq n_2$ and Condition 4 is satisfied, then $(\epsilon_{g_j}(n_1), \epsilon_{g_k}(n_2))$ converges in distribution to $(\epsilon'_{g_j}(n_1), \epsilon'_{g_k}(n_2))$ as $a \to -\infty$ where $\epsilon'_{g_j}(n_1)$ and $\epsilon'_{g_k}(n_2)$ are independent.

Proof: Define $\alpha_m = \eta_{p+a-m}$, $c_m = (1/\Delta_j) f_j(m)$ and

$$\mu_p = \beta_{j,p} + \frac{1}{\Delta_j} \sum_{m=0}^{p} f_j(m) x_d(p + a - m).$$

Using (7) and (8), we can write $\epsilon_{g_j}(n_1) = \Delta_j/2 - \Delta_j U_{n_1 - a}$ where

$$U_p = \left\langle \mu_p + \sum_{m=0}^{p} c_m \alpha_m \right\rangle.$$

Define

$$V_p = \left\langle \nu_p + \sum_{m=0}^{p} d_m \alpha_m \right\rangle$$

where $\nu_p = 0$ and $d_m = c_m^2$. Since $f_j(m)$ does not converge to zero as $m \to \infty$, it follows that $\{t_0 c_m + t_1 d_m\}$ does not converge to zero unless $(t_0, t_1) = \mathbf{0}$. Therefore, U_p and V_p satisfy the hypotheses of Lemmas A1 and A2 (the implicit dependence of α_m on p does not cause problems because the η_m are i.i.d.).[3] It follows from Lemmas A1 and A2 that $(\epsilon_{g_j}(n_1), x(n_2))$ converges in distribution to $((\Delta_j/2) - \Delta_j U, x(n_2))$ where U and $x(n_2)$ are independent and U is uniformly distributed on $[0, 1)$. Therefore, $\epsilon'_{g_j}(n_1) = (\Delta_j/2) - \Delta_j U$ and is uniformly distributed on $(-\Delta_j/2, \Delta_j/2\,]$ and independent of $x(n_2)$. This proves part (i).

Part (ii) similarly follows from Lemma A1. By keeping α_m as above, but letting $d_m = (1/\Delta_k) f_k(m)$ and

$$\nu_p = \beta_{k,p} + \frac{1}{\Delta_k} \sum_{m=0}^{p} f_k(m) x_d(p + a - m),$$

we can write $\epsilon_{g_k}(n_2) = (\Delta_k/2) - \Delta_k V_{n_2 - a}$. Provided $j \neq k$, Condition 3 ensures that $\{t_0 c_m + t_1 d_m\}$ does not converge to zero unless $(t_0, t_1) = \mathbf{0}$. Similarly, provided $n_1 \neq n_2$, Condition 4 ensures that $\{t_0 c_m + t_1 d_m\}$ does not converge to zero. In either case, U_p and V_p again satisfy the hypothesis of Lemma A1. It follows from the lemma that $(\epsilon'_{g_j}(n_1), \epsilon'_{g_k}(n_2))$ is uniformly distributed on $(-\Delta_j/2, \Delta_j/2] \times (-\Delta_k/2, \Delta_k/2]$. Therefore $\epsilon'_{g_j}(n_1)$ and $\epsilon'_{g_k}(n_2)$ are independent. ■

Notice that it is valid to replace $x(n_2)$ with $x_d(n_2)$ in the statement and proof of Theorem 1. Hence, the granular component of the quantization noise is asymptotically independent of both the input sequence and the desired input sequence.

In accordance with the usual definitions, we will take the mean of the quantization noise from the kth quantizer to be

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

the autocorrelation of the quantization noise from the kth quantizer to be

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

the cross correlation of the quantization noise from the jth and kth quantizers to be

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

and the cross correlation of the quantization noise from the kth quantizer and the desired input sequence to be

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

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 $R_{\epsilon_{g_k} x_d}(n, p)$ are 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[4]

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

The main assertion of Corollary 2 is that if Conditions 1–4 are satisfied, then the granular component of the quantization noise sequence introduced by each quantizer is asymptotically white. In view of certain previous work, this result is somewhat surprising. For example, it has been shown in [6] that the quantization noise in a nonoverloading second-order double-loop $\Delta\Sigma$ modulator with a sinusoidal input is not asymptotically white. Nevertheless, Corollary 2 implies that any amount of input-referred i.i.d. noise per (8), no matter how small its amplitude, causes the quantization noise to be asymptotically white. Therefore, there is a discontinuity between the behavior of the system as the input noise amplitude approaches zero and its behavior when the input noise amplitude is zero. In this sense, for the case mentioned above, the purely deterministic result is not a physically stable solution.

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 asymptotically white and independent of each other,

[3] Lemmas A1–A3 are presented in the Appendix.

[4] The function δ_p is the *Kronecker delta*, defined as

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

we can write the power spectral density of $e(n)$ as

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

where $N_k(z)$ is the kth element of $N(z)$ in (5). Suppose further that the input sequence is wide-sense stationary or, more generally, quasi-stationary [11]. Then the power spectral density of the output can be written as

$$S_{yy}(e^{j\omega}) = S_{xx}(e^{j\omega})|S(e^{j\omega})|^2 + S_{ee}(e^{j\omega})$$

where $S_{xx}(e^{j\omega})$ is the power spectral density of the input sequence.

Thus, Corollary 2 provides a simple means of evaluating the statistical performance of the ΔΣ 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 ΔΣ modulator satisfying Condition 4, do $\epsilon_{g_k}(n)$ and $\epsilon_{g_k}(n+1)$ for $n = a, a+1, \cdots$ take on values that are independent and uniformly distributed? Simulations such as that shown in Fig. 7 and theoretical results [12]–[14] 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 $\epsilon_{g_k}(n)$, $\epsilon_{g_k}(n)x_d(n+p)$ and $\epsilon_{g_j}(n)\epsilon_{g_k}(n+p)$ converge in probability to the corresponding statistical averages. Since each of the terms is bounded, convergence in probability implies convergence in the mean [15].

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)–(11) are similar, only the proof of (9) will be presented.

Without loss of generality, assume $a = 0$. Choose any integer $j \geq 0$, and for each $p > j$, define

$$S_p = \left\langle \mu'_p + \sum_{m=0}^{p-j-1} c'_m \alpha'_m \right\rangle$$

where $\alpha'_m = \eta_{p-m}$, $c'_m = (1/\Delta_k)f_k(m)$, and

$$\mu'_p = \beta_{k,p} + \frac{1}{\Delta_k}\sum_{m=0}^{p} f_k(m)x_d(p-m) + \frac{1}{\Delta_k}\sum_{m=p-j}^{p} f_k(m)\eta_{p-m}.$$

Then $\epsilon_{g_k}(p) = (\Delta_k/2) - \Delta_k S_p$. Applying Lemma A1 as in the proof of Theorem 1, except with S_{p+j} playing the role of U_p, it follows that as $n \to \infty$, S_p converges in distribution to a uniformly distributed random variable on $[0, 1)$ regardless of the statistics of the sequence $\{\mu'_k\}$. This implies that

$$E(S_p | \mu'_0, \mu'_1, \cdots) \to \tfrac{1}{2}$$

with probability 1 (and, consequently, in probability) as $p \to \infty$.

Because

$$E(\epsilon_{g_k}(p)|\eta_0, \cdots, \eta_j, x_d(0), x_d(1), \cdots) = \frac{\Delta_k}{2} - \Delta_k E(S_p | \mu'_0, \mu'_1, \cdots),$$

it follows that for each $j > 0$,

$$E(\epsilon_{g_k}(p)|\eta_0, \cdots, \eta_j, x_d(0), x_d(1), \cdots) \to 0$$

in probability as $p \to \infty$. From Lemma A3, this is a sufficient condition for (9) to hold 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 ΔΣ modulators [16]. In practical ΔΣ modulators, it is often best to choose the no-overload ranges of the quantizers so as to avoid overload altogether. However, in many practical systems, it is convenient to use coarse quantizers that sometimes overload.

For example, the second-order double-loop ΔΣ modulator of Fig. 2(b) is often used with a 1 b quantizer (i.e., a hard limiter). In this configuration, input sequences that overload the ΔΣ modulator can be found with arbitrarily small maximum amplitude. However, it has been observed that if the ratio of the maximum input amplitude to the quantizer step size is kept below approximately 0.1, the additional error introduced by the overload component of the quantization noise will not be overly severe [16].

Fig. 8 shows the simulated density of values taken on by the pair of quantization noise terms $(\epsilon(n), \epsilon(n+1))$, $n = 0, 1, \cdots$, for an overloading second-order double-loop ΔΣ modulator. The simulation parameters were the same as those of Fig. 7, except that a 1 b quantizer was used.

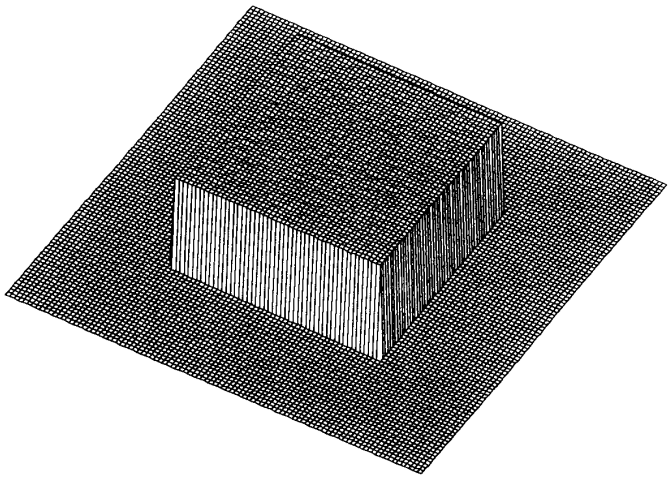

Fig. 7. The simulated density (i.e., the normalized histogram) of values taken on by $(\epsilon(n), \epsilon(n+1))$, $n = 0, 1, \cdots$, in a nonoverloading second-order double-loop $\Delta\Sigma$ modulator. The $\Delta\Sigma$ modulator used a 2 b quantizer with unity step size, the desired input sequence was a sinusoid of amplitude 0.1, and the input noise sequence had a variance of 10^{-8}. The density is plotted on the interval $[-2, 2]^2$.

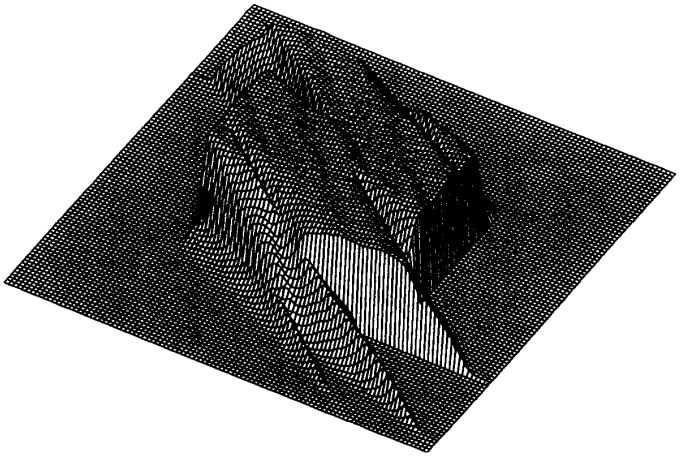

Fig. 8. The simulated density of values taken on by $(\epsilon(n), \epsilon(n+1))$, $n = 0, 1, \cdots$, in an overloading second-order double-loop $\Delta\Sigma$ modulator. The simulation parameters were the same as in the case of Fig. 7, except that the $\Delta\Sigma$ modulator used a 1 b quantizer.

Specifically, the quantizer had unity step size, the desired input sequence was a sinusoid of amplitude 0.1, and the input noise sequence had a variance of 10^{-8}. In comparing the two figures, we see that overload significantly complicates the structure of the joint distribution. In particular, the quantization noise terms are no longer uniformly distributed or independent.

In such cases, the results of the previous section still apply to the granular component of the quantization noise, but do not provide insight into the structure of the overload component of the quantization noise. In general, the problem of rigorously characterizing the overload component of the quantization noise is still open.

Nevertheless, one conclusion regarding the effect of overload can be deduced. It follows from Fig. 6(c) that the overload component of the quantization noise $\epsilon_o(n)$ originates in the bank of limiters $\boldsymbol{L}$ and is subjected to the noise filter $\boldsymbol{N}(z)$. In many systems, including all those in Fig. 2, the noise filter has a finite impulse response. In such cases, if the quantizers were to overload infrequently, the overall quantization error $e(n)$ would only infrequently be a function of overload quantization noise; at all other times, the quantization error would be equivalent to the output of the noise filter operating on the granular component of the quantization noise given by (7). In this sense, the effect of the overload component of the quantization noise will decay with the frequency of occurrence of the overload condition. This argument is

similar to one used by Wong and Gray [17] in considering the effect of infrequently occurring overload for the special case of a first-order ΔΣ modulator with an i.i.d. Gaussian input sequence.

VI. Conclusion

We have defined and analyzed a generic ΔΣ modulator architecture. The results provide a unified framework for analyzing a large class of ΔΣ modulators because many of the known ΔΣ 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 ΔΣ modulator, we have performed a statistical analysis of the granular quantization noise for arbitrary input sequences. In particular, we have shown that ΔΣ modulators in the class with orders greater than one each have quantization noise sequences that converge in distribution to sequences of random variables that are uniformly distributed, white, 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 ΔΣ modulator as developed in [8]. We have also derived ergodic properties relating the various second-order statistical averages of the quantization noise to the corresponding time averages. Unlike most other theoretical treatments, we do not require that the quantizers never overload.

Appendix

Supporting Lemmas

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

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

where $\{\alpha_k\}$ is a sequence of independent, identically distributed random variables whose distribution contains an absolutely continuous component, $\{c_k\}$ and $\{d_k\}$ are any real sequences such that $\{t_0 c_k + t_1 d_k\}$ does not converge to zero unless $(t_0, t_1) = \mathbf{0}$, and $\{\mu_p\}$ and $\{\nu_p\}$ are any two sequences of random variables, each of which is independent of every α_k. Then as $p \to \infty$, (U_p, V_p) converges in distribution to a random vector (U, V) that is uniformly distributed on $[0, 1)^2$.

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

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

and

$$Y_p = \nu_p + \sum_{k=0}^{p} d_k \alpha_k.$$

Then $U_p = \langle X_p \rangle$ and $V_p = \langle Y_p \rangle$.

Define the following random vectors:

$$\boldsymbol{r}_p = \begin{bmatrix} \mu_p \\ \nu_p \\ \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} \text{ and } \boldsymbol{s}_p = \begin{bmatrix} X_p \\ Y_p \\ \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix}.$$

It is easy to verify that $\boldsymbol{s}_p = \boldsymbol{A}_p \boldsymbol{r}_p$ where

$$\boldsymbol{A}_p = \begin{bmatrix} 1 & 0 & c_0 & c_1 & \cdots & c_p \\ 0 & 1 & d_0 & d_1 & \cdots & d_p \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & & \\ 0 & \cdots & & & 0 & 1 \end{bmatrix}. \tag{12}$$

Moreover, $\det \boldsymbol{A}_p = 1$ so $\boldsymbol{A}_p$ is nonsingular.

Because the α_k are mutually independent and are independent of μ_p and ν_p, the probability measure of $\boldsymbol{r}_p$ can be written as

$$P_{r_p} = P_{\mu_p, \nu_p} \times \underbrace{P_\alpha \times \cdots \times P_\alpha}_{p+1} \tag{13}$$

where P_{μ_p, ν_p} is the joint probability measure of μ_p and ν_p and P_α is the probability measure of each α_k. If $T_p : \boldsymbol{R}^{p+3} \to \boldsymbol{R}^{p+3}$ is the mapping associated with the matrix $\boldsymbol{A}_p$, it follows that the probability measure of $\boldsymbol{s}_p$ is

$$P_{s_p} = P_{r_p} T_p^{-1}.$$

The characteristic function of $\boldsymbol{s}_p$ is

$$\Phi_{s_p}(t_0, \cdots, t_{p+2}) = \int_{\boldsymbol{R}^{p+3}} e^{j\boldsymbol{t}\cdot\boldsymbol{s}_p} P_{s_p}(d\boldsymbol{s}_p)$$

where

$$\boldsymbol{t} = \begin{bmatrix} t_0 \\ \vdots \\ t_{p+2} \end{bmatrix}.$$

Performing a change of variables gives[5]

$$\Phi_{s_p}(t_0, \cdots, t_{p+2}) = \int_{\boldsymbol{R}^{p+3}} e^{j\boldsymbol{t}\cdot T\boldsymbol{r}_p} P_{r_p}(d\boldsymbol{r}_p). \tag{14}$$

From (12), we can write

$$\boldsymbol{t} \cdot T\boldsymbol{r}_p = t_0 \left[\mu_p + \sum_{k=0}^{p} c_k \alpha_k \right] + t_1 \left[\nu_p + \sum_{k=0}^{p} d_k \alpha_k \right] + \sum_{k=0}^{p} t_{k+2} \alpha_k. \tag{15}$$

Substituting (13) and (15) into (14) gives

$$\Phi_{s_p}(t_0, \cdots, t_{p+2}) = \int_{\boldsymbol{R}^2} e^{j(t_0 \mu_p + t_1 \nu_p)} P_{\mu_p, \nu_p}(d\mu_p, d\nu_p) \times \prod_{k=0}^{p} \int_{\boldsymbol{R}^1} e^{j(t_0 c_k + t_1 d_k + t_{k+2})\alpha_k} P_\alpha(d\alpha_k),$$

which can be written more compactly as

$$\Phi_{s_p}(t_0, \cdots, t_{p+2}) = \Phi_{\mu_p, \nu_p}(t_0, t_1) \prod_{k=0}^{p} \Phi_\alpha(t_0 c_k + t_1 d_k + t_{k+2}) \tag{16}$$

where $\Phi_\alpha(t)$ is the characteristic function common to each α_k, and $\Phi_{\mu_p, \nu_p}(t_0, t_1)$ is the joint characteristic function of μ_p and ν_p.

[5] See, for example, [15, Theorem 16.12].

By definition, the joint characteristic function of X_p and Y_p can be written as

$$\Phi_{X_p,Y_p}(t_0,t_1) = \Phi_{s_p}(t_0,t_1,0,\cdots,0).$$

Using (16), this becomes

$$\Phi_{X_p,Y_p}(t_0,t_1) = \Phi_{\mu_p,\nu_p}(t_0,t_1)\prod_{k=0}^{p}\Phi_\alpha(t_0c_k + t_1d_k).$$

All characteristic functions equal one at the origin and have absolute value less than or equal to one elsewhere. Since the distribution of the α_k is not a purely discrete distribution, we are assured that $|\Phi_\alpha(t)|$ is strictly less than one for all $t \neq 0$.[6] Moreover, since the distribution function of the α_k contains an absolutely continuous component, it follows from the Riemann–Lebesgue Lemma that $|\Phi_\alpha(t)|$ is bounded by a number less than one outside each neighborhood of the origin. Because $\{t_0c_k + t_1d_k\}$ does not converge to zero when $(t_0,t_1) \neq \mathbf{0}$, we have

$$\lim_{p\to\infty}\Phi_{X_p,Y_p}(t_0,t_1) = \begin{cases}1, & \text{if } (t_0,t_1) = \mathbf{0}\\ 0, & \text{otherwise.}\end{cases} \tag{17}$$

Let P_{X_p,Y_p} be the joint probability measure of X_p and Y_p. Given $A \subset \boldsymbol{R}^2$, let

$$P_{\langle X_p\rangle,\langle Y_p\rangle}(A) = \sum_{i=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}P_{X_p,Y_p}((A\cap[0,1)^2) + (i,k)).$$

By the definition of the fractional part operator, it follows that $P_{\langle X_p\rangle,\langle Y_p\rangle}$ must equal the joint probability measure of $\langle X_p\rangle$ and $\langle Y_p\rangle$ for every set on which the sum converges. But by the countable additivity of P_{X_p,Y_p} and the fact that if A is measurable so is $\bigcup_{i,k=-\infty}^{\infty}((A\cap[0,1)^2) + (i,k))$, the sum converges for all sets on which P_{X_p,Y_p} is defined. Therefore, $P_{\langle X_p\rangle,\langle Y_p\rangle}$ is the joint probability measure of $\langle X_p\rangle$ and $\langle Y_p\rangle$.

The joint characteristic function of $\langle X_p\rangle$ and $\langle Y_p\rangle$ is

$$\begin{aligned}\Phi_{\langle X_p\rangle,\langle Y_p\rangle}(t_0,t_1) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{j(t_0x+t_1y)}P_{\langle X_p\rangle,\langle Y_p\rangle}(dx,dy)\\ &= \int_0^1\int_0^1 e^{j(t_0x+t_1y)}\sum_{i=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\\ &\quad\cdot P_{X_p,Y_p}(dx+i,dy+k).\end{aligned}$$

Hence, for all integers m and n,

$$\begin{aligned}&\Phi_{\langle X_p\rangle,\langle Y_p\rangle}(2\pi m,2\pi n)\\ &\quad= \int_0^1\int_0^1\sum_{i=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}e^{j2\pi(m(x+i)+n(y+k))}\\ &\qquad\cdot P_{X_p,Y_p}(dx+i,dy+k)\\ &\quad= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{j2\pi(mx+ny)}P_{X_p,Y_p}(dx,dy)\\ &\quad= \Phi_{X_p,Y_p}(2\pi m,2\pi n).\end{aligned}$$

[6] See, for example, [10, Theorem 2.1.4].

Since $U_p = \langle X_p\rangle$ and $V_p = \langle Y_p\rangle$, it follows from (17) that the joint characteristic function of U_p and V_p satisfies

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

Define

$$\Phi_{U,V}(t_0,t_1) = e^{-j(t_0+t_1)/2}\frac{\sin(t_0/2)}{t_0/2}\frac{\sin(t_1/2)}{t_1/2}.$$

Then

$$\lim_{p\to\infty}\Phi_{U_p,V_p}(2\pi m,2\pi n) = \Phi_{U,V}(2\pi m,2\pi n). \tag{18}$$

Taking the Fourier transform of $\Phi_{U,V}(t_0,t_1)$ shows it to be the characteristic function of a random vector (U,V) that is uniformly distributed on $[0,1)^2$.

To finish the proof, it is necessary to show that the vector (U_p,V_p) converges in distribution to (U,V). A necessary and sufficient condition for this to occur is that $\Phi_{U_p,V_p}(t_0,t_1)$ converge to $\Phi_{U,V}(t_0,t_1)$ for each $(t_0,t_1) \in \boldsymbol{R}^2$. However, so far, we have only shown that $\Phi_{U_p,V_p}(t_0,t_1)$ converges to $\Phi_{U,V}(t_0,t_1)$ at the points $\{(2\pi m, 2\pi n): n, m = 0, \pm1, \pm2, \cdots\}$.

Because the distribution of (U,V) and the distribution of (U_p,V_p) both have support restricted to $[0,1)^2$, it follows that $\Phi_{U,V}(t_0,t_1)$ and $\Phi_{U_p,V_p}(t_0,t_1)$ are each uniquely determined by their values at the $\{(2\pi m,2\pi n): n,m = 0, \pm1, \cdots\}$. Therefore, any subsequence of (U_p,V_p) that converges in distribution at all must converge in distribution to (U,V). Moreover, the distribution of (U_p,V_p) is tight (a consequence of the sequence having bounded support). A sufficient condition for a tight sequence of probability measures to converge weakly to a particular probability measure is that all subsequences that converge weakly at all converge weakly to the particular probability measure.[7] It follows that (U_p,V_p) converges in distribution to (U,V). ■

Part of the previous lemma is an extension of a result proven by Chou and Gray [7]. They proved that U_p converges in distribution to U under the restrictions that $c_k = 1$ for all k, that the α_n are i.i.d. with a distribution that has a density, and that the μ_k are deterministic.

Lemma A2: Let U_p and V_p be defined as in Lemma A1, and let X be any random variable that is independent of all but a finite subset of $\{\alpha_k\}$. Then as $p \to \infty$, (U_p,V_p,X) converges in distribution to (U,V,X) where X is independent of U and V.

Proof: By the definition of independence, it is sufficient to show that for any number x, (S_p,T_p) defined as (U_p,V_p) under the constraint that $X = x$ (i.e., $(S_p,T_p) = (U_p,V_p)|_{X=x}$) converges in distribution to (U,V).

Because X only depends on a finite number of the α_k, there exists some K such that X is independent of the set $\{\alpha'_k = \alpha_{k+K} : k = 0,1,\cdots\}$. Therefore, we can write

$$S_p = \left\langle \mu'_p + \sum_{k=0}^{p-K}c'_k\alpha'_k\right\rangle$$

[7] See, for example, the corollary in [15, p. 395].

and

$$T_p = \left\langle \nu'_p + \sum_{k=0}^{p-K} d'_k \alpha'_k \right\rangle$$

where $c'_k = c_{k+K}$, $d'_k = d_{k+K}$,

$$\mu'_p = \left(\mu_p + \sum_{k=0}^{K-1} c_k \alpha_k \right) \Bigg|_{X=x},$$

and

$$\nu'_p = \left(\nu_p + \sum_{k=0}^{K-1} d_k \alpha_k \right) \Bigg|_{X=x}.$$

Since μ'_p and ν'_p are independent of every α'_k, it follows from Lemma A1 that (S_p, T_p) converges in distribution to (U, V). ■

Lemma A3: For each $k = 1, 2, \cdots$, let $f_k : \boldsymbol{R}^{2k+2} \to \boldsymbol{R}$ be a measurable function that has absolute value less than β. Let $\{\eta_0, \cdots, \eta_k\}$ and $\{\mu_0, \cdots, \mu_k\}$, be two sequences of random variables where the η_n are independent of each other and independent of the μ_n, and let $X_k = f_k(\eta_0, \cdots, \eta_k, \mu_0, \cdots, \mu_k)$. Suppose that

$$E(X_k | \eta_0, \cdots, \eta_j, \mu_0, \mu_1, \cdots) \to 0 \tag{19}$$

in probability as $k - j \to \infty$ with $k > j \geq 0$. Then

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

in probability as $N \to \infty$.

Proof: This lemma corresponds to [8, Lemma A2]. See [8] for the proof. ■

Acknowledgment

The author would like to thank Professor Howard G. Tucker for his careful review of Lemma A1 and his suggestions leading to improvements of its proof. As with previous work on this subject, the author is also grateful for the support, encouragement, and friendship provided by the late Professor Edward C. Posner.

References

[1] F. Goodenough, "High-resolution ADCs up dynamic range in more applications," *Electron. Design*, pp. 65–79, Apr. 11, 1991.

[2] J. C. Candy and G. C. Temes, Eds., *Oversampling Delta-Sigma Data Converters Theory, Design and Simulation.* New York: IEEE Press, 1992.

[3] ——, "Oversampling methods for A/D and D/A conversion," in *Oversampling Delta-Sigma Data Converters Theory, Design and Simulation*, New York: IEEE Press, 1992, pp. 1–25.

[4] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1220–1244, Nov. 1990.

[5] N. He, F. Kuhlmann, and A. Buzo, "Multiloop sigma delta quantization," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1015–1028, May 1992.

[6] N. He, A. Buzo, and F. Kuhlmann, "Double-loop sigma-delta modulation with DC inputs," *IEEE Trans. Commun.*, vol. 38, pp. 487–495, Apr. 1990.

[7] 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.

[8] I. Galton, "Granular Quantization Noise in the First-Order Delta-Sigma Modulator," *IEEE Transactions on Information Theory*, vol. 39, no. 6, pp. 1944–1956, November, 1993.

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

[10] E. Lukacs, *Characteristic Functions.* New York: Hafner, 1970.

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

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

[13] D. F. Delchamps, "Exact asymptotic statistics for sigma-delta quantization noise," in *Proc. 28th Annu. Allerton Conf. Commun., Contr., Computing*, Oct. 1990.

[14] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences.* New York: Wiley, 1974.

[15] P. Billingsley, *Probability and Measure.* New York: Wiley, 1986.

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

[17] P. W. Wong and R. M. Gray, "Sigma-delta modulation with iid Gaussian inputs," *IEEE Trans. Inform. Theory*, vol. 36, pp. 784–798, May 1990.