A Theory of Non-Subtractive Dither

Robert A. Wannamaker, Stanley P. Lipshitz*, and John Vanderkooy*

Audio Research Group, University of Waterloo, Waterloo, Ontario N2L 3G1, CANADA

and

J. Nelson Wright

Acuson, 1220 Charleston Rd., Mountain View, CA 94039-7393, USA

Abstract

A detailed mathematical investigation of multi-bit quantizing systems using non-subtractive dither is presented. It is shown that by the use of dither having a suitably-chosen probability density function, moments of the total error can be made independent of the system input signal, but that statistical independence of the error and the input signals is not achievable. Similarly, it is demonstrated that values of the total error signal cannot generally be rendered statistically independent of one another, but that their joint moments can be controlled and that, in particular, the error sequence can be rendered spectrally white. The properties of some practical dither signals are explored and recommendations are made for dithering in audio, video and measurement applications. Some of the results presented here are known to a handful of individuals in the engineering community, but many appear to be unpublished. In view of many widespread misunderstandings regarding non-subtractive dither, formal presentation of these results is long overdue.

*Members of the Guelph-Waterloo Program for Graduate Work in Physics.

1 Introduction

Analog-to-digital conversion is customarily decomposed into two separate processes: *sampling* of the input analog waveform and *quantization* of the sample values in order to represent them with binary words of a prescribed length. The sampling operation incurs no loss of information as long as the input is appropriately bandlimited, but the approximating nature of the quantization operation *always* results in signal degradation. Another common operation with a similar problem is *requantization,* in which the wordlength of digital data is reduced after arithmetical processing in order to meet specifications for its storage or transmission.

Dither, straightforwardly put, is a random "noise" process added to a signal prior to its (re)quantization in order to control the statistical properties of the quantization error. This is not a new idea. Subtractively dithered (SD) quantizing systems, in which the dither is subsequently subtracted from the output signal after quantization, have been discussed and used for over thirty years in speech and video processing applications [1, 2], and a satisfactory theory of their operation exists in print [3, 4]. Non-subtractively dithered (NSD) systems, in which the dither signal is not subtracted from the output, are a subject of more recent interest. The following provides a brief history of the theory of NSD quantization.

It must be acknowledged that all theoretical treatments of dithered quantization owe a substantial debt to the work of Widrow [5, 6, 7], who developed many of the essential mathematical tools while studying undithered quantization. Among Widrow's contributions was the "Quantizing Theorem," a counterpart in disrete-valued systems to the better known "Sampling Theorem" in discrete-time systems. Important extensions to Widrow's treatment were made by Sripad and Snyder [8], Sherwood [4] and Gray [9].

Early investigations into non-subtractive dither *per se* were conducted by one of the authors, Wright [10], in 1979, resulting in discovery of many of the important results which follow. This work remained unpublished until it was brought to the attention of the other authors of this manuscript [11].

The results concerning moments of the error signal were rediscovered independently by Stockham [12] in 1980, and documented in an unpublished Master's thesis by Brinton [13], a student of Stockham's, in 1984. Stockham, however, has remained silent on the matter until recently [14].

The properties of non-subtractive dither (and, in particular, triangular-pdf dither) were again discovered independently by two of the present authors, Lipshitz and Vanderkooy, in the mid-1980's. Vanderkooy and Lipshitz were the first researchers to publish their findings on non-subtractive dither [15, 16, 17, 18, 19], and this prompted collation and extension of the theoretical aspects by another of the authors, Wannamaker [20, 21, 22, 23, 24, 25]. Lipshitz, Wannamaker and Vanderkooy have published a broad theoretical survey of multi-bit quantization treating both undithered and dithered systems [26]. They have also extended the treatment to include an analysis of dithered quantizing systems using noise-shaping error feedback [27, 28, 24] and multichannel quantizing systems [29].

The results concerning error moments have also been independently discovered by Gray [30], using arguments employing Fourier series along with characteristic functions. Gray and Stockham have published a paper on the subject [14].

Although a small number of individuals in the engineering community are aware of the correct results regarding non-subtractive dither, a number of misconceptions concerning the technique are widespread. Particularly serious is a persistent confusion regarding the quite different properties of subtractive and non-subtractive dithering (see, for instance, [31, p. 170]). The aim of this paper is to provide a consistent

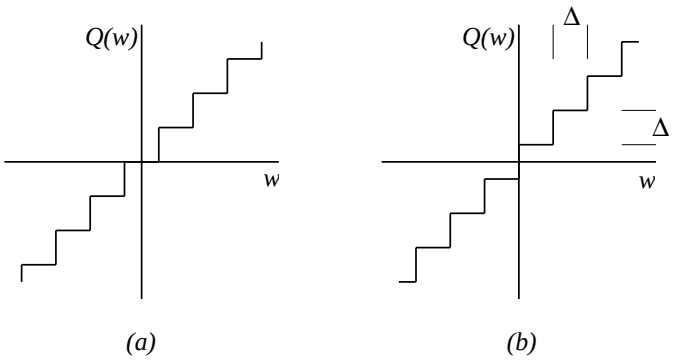


Figure 1: Quantizer transfer characteristics: (a) mid-tread, (b) mid-riser, with Δ *denoting the size of one LSB.*

and rigorous account of the theory of non-subtractively dithered systems in order to promote a more universal understanding of this dithering technique. As such, it greatly extends and elaborates the treatment provided by our earlier presentation [22].

1.1 The Classical Model of Quantization

Quantization and requantization processes possess similar transfer characteristics, which are generally of either the *mid-tread* or *mid-riser* variety illustrated in Fig. 1. We will assume that the quantizers involved are *infinite,* which, for practical purposes, means that the system input signal is never clipped by saturation of the quantizer. (Some comments regarding the application of dither to 1-bit and sigma-delta converters will be reserved for the Conclusions.) In this case the corresponding transfer functions relating the quantizer output to its input, w, can be expressed analytically in terms of the quantizer step size, Δ:

$$Q(w) = \Delta \left\lfloor \frac{w}{\Delta} + \frac{1}{2} \right\rfloor$$

for a mid-tread quantizer, or

$$Q(w) = \Delta \left\lfloor \frac{w}{\Delta} \right\rfloor + \frac{\Delta}{2}$$

for a mid-riser quantizer, where the "floor" operator, $\lfloor\ \rfloor$, returns the greatest integer less than or equal to its argument. The step size, Δ, is commonly referred to as a LSB ("least significant bit"), since a change in input signal level of one step width corresponds to a change in the LSB of binary coded output. Throughout the sequel, quantizers of the mid-tread variety will be assumed, but all derived results have obvious analogs for mid-riser quantizers and all stated theorems are valid for both types.

Quantization or requantization introduces an error signal, q, into the digital data stream, which is simply the difference between the output of the quantizer and its input:

$$q(w) \stackrel{\triangle}{=} Q(w) - w,$$

where we use $\stackrel{\triangle}{=}$ to indicate equality by definition. This *quantization error* is shown as a function of w for a mid-tread quantizer in Fig. 2. It has a maximum magnitude of 0.5 LSB and is periodic in w with a period of 1 LSB.

Although q is clearly a deterministic function of the input, the Classical Model of Quantization (CMQ) [31] holds that the quantization error can be modeled as an additive random process which is independent of the system input and iid (i.e., that distinct samples of the error are statistically independent of one another and identically distributed). The CMQ further postulates that the error is *uniformly distributed,* meaning that its values exhibit a probability density function (pdf) of the form

$$p_q(q) = \Pi_\Delta(q), \tag{1}$$

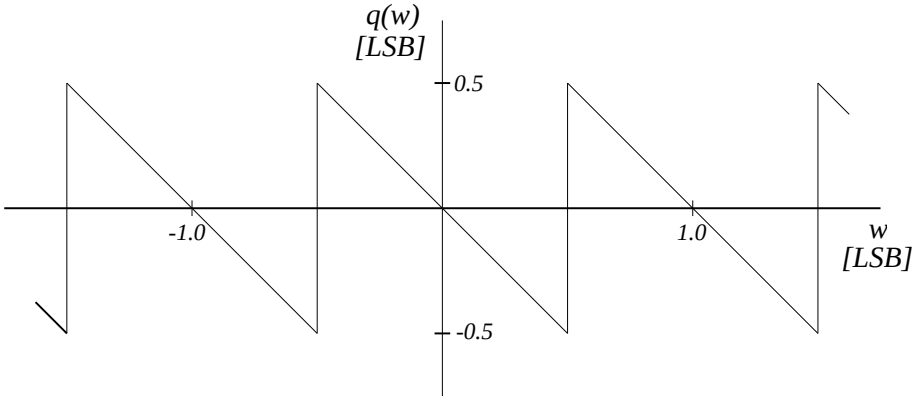


Figure 2: Quantization error, $q(w)$, as a function of quantizer input, w, for a mid-tread quantizer.

where the *rectangular window function* of width Γ, Π_Γ, is defined as

$$\Pi_\Gamma(q) \triangleq \begin{cases} \frac{1}{\Gamma}, & -\frac{\Gamma}{2} < q \leq \frac{\Gamma}{2}, \\ 0, & \text{otherwise.} \end{cases}$$

Such a pdf is referred to as a *uniform pdf* or *RPDF (rectangular pdf).*

The m-th *moment* of a random process q with pdf p_q is defined as the *expectation value* of $q^m, m \in \mathbf{Z}$:

$$E[q^m] \triangleq \int_{-\infty}^{\infty} q^m p_q(q) dq,$$

where $E[\,]$, the expectation value operator, is defined more generally by

$$E[f] \triangleq \int_{-\infty}^{\infty} f(q) p_q(q) dq.$$

The zeroth moment of any random process (i.e., $E[q^0]$) is identically equal to unity. The first moment is usually referred to as the *mean* of the process, whereas the term *variance* refers to the quantity $E[(q - E[q])^2] = E[q^2] - E^2[q]$. It is clear that if the mean of a random process is zero, its variance and second moment are equal.

If a quantization error signal is distributed according to Eq. (1), its moments are:

$$E[q] = 0 \tag{2}$$

$$E[q^2] = \frac{\Delta^2}{12} \tag{3}$$

$$E[q^m] = \begin{cases} \frac{1}{m+1}\left(\frac{\Delta}{2}\right)^m, & \text{for } m \text{ even,} \\ 0, & \text{for } m \text{ odd.} \end{cases} \tag{4}$$

Eq. (3) is the familiar expression for the variance of the quantization error in the classical model.

The CMQ is valid for input signals which exhibit smooth pdf's and which are large relative to an LSB [31, 25]. It fails catastrophically for small signals and many particularly simple (e.g., sinusoidal) signals, for which the quantization error retains the character of input-dependent distortion, rather than noise. The mid-tread quantization of a small signal of peak amplitude less than 0.5 LSB provides a simple example of this failure: the quantizer output is null, and the quantization error is just the input sign-inverted. Such an error is not uniformly distributed, iid or independent of the input. In such cases, application of an appropriate dither can be used to temper the statistical properties of the error signal.

1.2 Dither: Subtractive vs. Non-Subtractive

Schematics of subtractively dithered and non-subtractively dithered quantizing systems are shown in Fig. 3. In each case we denote the *system input* by x and the *system output* by y. We thus distinguish the system input from the *quantizer input,* which we continue to denote by w and which is given by $w = x + \nu$. ν represents the *dither* signal, a strict-sense stationary random process which is assumed to be statistically independent of x. Similarly, the *total error* of each quantizing system is defined as the difference between the system output and system input, and is denoted

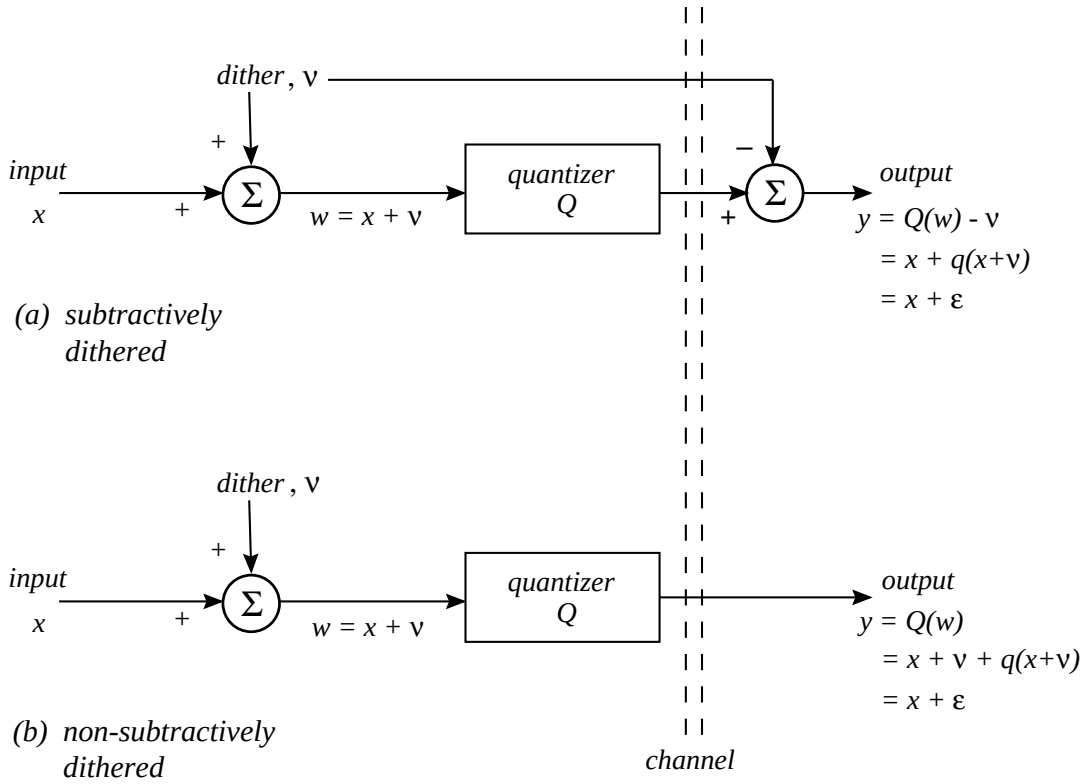


Figure 3: Dither quantizing systems: (a) subtractively dithered (SD), (b) non-subtractively dithered (NSD).

by

$$\varepsilon \stackrel{\triangle}{=} y - x$$

to distinguish it from the quantizer error, $q = Q(w) - w$.

The total errors introduced by subtractively dithered and non-subtractively dithered systems are not identical. In a subtractively dithered system, the dither is subtracted from the quantizer output to yield the system output. Hence, for such a system:

$$\begin{aligned} \varepsilon &= y - x \\ &= Q(x + \nu) - (x + \nu) \\ &= q(x + \nu). \end{aligned}$$

On the other hand, for the non-subtractively dithered system:

$$\begin{aligned} \varepsilon &= y - x \\ &= Q(x + \nu) - x \\ &= q(x + \nu) + \nu. \end{aligned}$$

In neither case is the total error equal to $q(x)$ as in an undithered system (i.e., one for which $\nu \equiv 0$), although in an SD system the total error does equal the quantization error associated with the *total* quantizer input w.

It has been shown by Schuchman [3] that the total error induced by an SD quantizing system can be rendered uniformly distributed for arbitrary input distributions if and only if the dither's *characteristic function* or *cf* (the Fourier transform of its pdf [32, 33]) obeys a certain condition. Defining the Fourier transform operator, $\mathcal{F}[\,]$, by

$$\mathcal{F}[f](u) \stackrel{\triangle}{=} F(u) \stackrel{\triangle}{=} \int_{-\infty}^{\infty} f(x) e^{-j2\pi ux} dx,$$

and denoting the dither pdf and cf as $p_\nu(\nu)$ and $P_\nu(u)$, respectively, Schuchman's condition is that

$$P_\nu\left(\frac{k}{\Delta}\right) = 0 \qquad \forall k \in \mathbf{Z}_0, \tag{5}$$

where we take this opportunity to define the set $\mathbf{Z}_0^n$ as the set of all n-vectors with integer components with the exception of the zero vector $\mathbf{0} = (0, 0, \ldots, 0)$; i.e., $\mathbf{Z}_0^n = \mathbf{Z}^n \setminus \mathbf{0}$.

Furthermore, it can be shown [8, 4, 26] that the total error in an SD quantizing system is statistically independent of the system input if and only if Eq. (5) holds. Thus, dither obeying Schuchman's condition renders the error statistically independent of the input and uniformly distributed. In particular, it exhibits a variance of $\Delta^2/12$. In these regards, then, it resembles the idealized quantization error of the CMQ. The simplest random process satisfying Schuchman's condition is one exhibiting a uniform pdf

$$p_\nu(\nu) = \Pi_\Delta(\nu),$$

whose associated characteristic function is a "sinc" function:

$$P_\nu(u) = \mathrm{sinc}\,(u) \triangleq \frac{\sin(\pi\Delta u)}{\pi\Delta u}.$$

(Different authors employ slightly different definitions of the sinc function. We will retain the above throughout the sequel.)

It can also be shown [8, 4, 26] that subtractive dither will render distinct samples of the total error signal statistically independent of one another for arbitrary input distributions if and only if

$$P_{\nu_1,\nu_2}\left(\frac{k_1}{\Delta}, \frac{k_2}{\Delta}\right) = 0 \qquad \forall (k_1, k_2) \in \mathbf{Z}_0^2, \tag{6}$$

where ν_1 and ν_2 represent dither values separated in time by $\tau \neq 0$, and where $P_{\nu_1,\nu_2}(u_1, u_2)$ represents their joint characteristic function (the two-dimensional Fourier

transform of their joint pdf, $p_{\nu_1,\nu_2}(\nu_1,\nu_2)$). This condition is satisfied by any dither which is iid, so that $P_{\nu_1,\nu_2}(u_1,u_2) = P_\nu(u_1)P_\nu(u_2)$, and which satisfies Eq. (5). For instance, this means that a subtractively dithered quantizing system employing iid dither of uniform distribution produces an iid total error signal whose values are uniformly distributed and statistically independent of the input. The error thus behaves like a purely additive independent noise process, as postulated by the CMQ. This beautiful result represents the ideal outcome for a quantization operation.

Unfortunately, subtractive dithering is difficult to use in many practical systems since the dither signal must be available at each end of the channel. This requires either the transmission of the dither values or the use of synchronized noise sources (pseudo-random number generators) separated, in general, by both time and distance. Furthermore, any digital processing of the dithered signal would necessitate processing of the dither prior to subtraction. For reasons such as these, the possibility of using dither without subsequently subtracting it is frequently of interest.

We will see that non-subtractively dithered systems, as distinct from subtractively dithered ones, *cannot* render the total error statistically independent of the input. Neither can they make temporally separated values of the total error statistically independent of one another. However, we shall prove that they *can* render any desired statistical moments of the error signal independent of the input and regulate the joint moments of errors which are separated in time. The theory underlying these features of non-subtractive dither is developed in Section 2, and is subsequently used to explore the properties of some practical dither signals in Section 3. Section 4 explores the important special case of quantizing systems in which the available dither is discrete valued, while Section 5 summarizes the most important observations and conclusions.

2 Non-Subtractive Dither Theory

We begin by describing the relationship between the total error and the input signal in probabilistic terms.

2.1 Total Error PDF's

The dependence of the total error on the system input can be analyzed in terms of its pdf as a function of a specified input value. This function is referred to as the *conditional pdf*, or *cpdf*, of the total error and is denoted $p_{\varepsilon|x}(\varepsilon, x)$ throughout the following discussion.

In order to derive an expression for $p_{\varepsilon|x}(\varepsilon, x)$, we consider a non-subtractively dithered quantizing system, as in Fig. 3(b), with a specified system input value, x. The input to the quantizer is $w = x + \nu$, the sum of the system input and the statistically independent dither process. This sum has a cpdf

$$p_{w|x}(w, x) = p_\nu(w - x).$$

Fig. 4 shows that total error depends not only on the system input value, but also on the value of the dither. In particular, if the input to the quantizer, w, is between $-\Delta/2$ and $+\Delta/2$, the output will be nil (for a mid-tread characteristic) so that the error is $\varepsilon = -x$. Similarly, if the input to the quantizer is between $+\Delta/2$ and $+3\Delta/2$ the output will be $+\Delta$, so that $\varepsilon = -x + \Delta$. Hence, the pdf of the error for a fixed input is a series of delta functions separated by intervals of Δ, each weighted by the probability that w falls upon the corresponding quantizer step:

$$p_{\varepsilon|x}(\varepsilon, x) = \sum_{k=-\infty}^{\infty} \delta(\varepsilon + x - k\Delta) \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} p_\nu(w - x)dw.$$

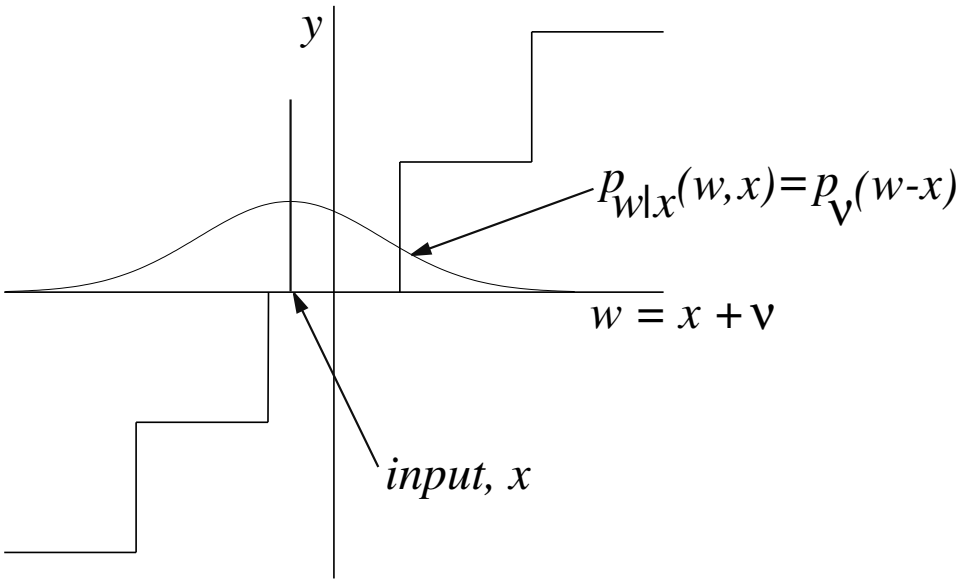


Figure 4: Cpdf of the quantizer input showing its justification relative to the quantizer transfer characteristic.

In the parlance of Widrow [7], the error cpdf is an *area sampled* version of the quantizer input cpdf.

Writing the integral in the last equation as a convolution (denoted by $\star$) of p_ν with a rectangular window function, $\Delta\Pi_\Delta$, it reduces to

$$p_{\varepsilon|x}(\varepsilon, x) = [\Delta\Pi_\Delta \star p_\nu](\varepsilon) W_\Delta(\varepsilon + x), \tag{7}$$

where

$$W_\Gamma(\varepsilon) \stackrel{\triangle}{=} \sum_{k=-\infty}^{\infty} \delta(\varepsilon - k\Gamma)$$

is a train of Dirac delta functions separated by intervals of width Γ.[1] Thus the pdf of ε is given by

$$\begin{aligned} p_\varepsilon(\varepsilon) &= \int_{-\infty}^{\infty} p_{\varepsilon|x}(\varepsilon, x) p_x(x) dx \\ &= [\Delta\Pi_\Delta \star p_\nu](\varepsilon)[W_\Delta \star p_x](-\varepsilon). \end{aligned} \tag{8}$$

As discussed in Section 1.2 in association with subtractively dithered systems, the quantization error, $q(w)$, *will* be statistically independent of x and uniformly distributed if the dither statistics obey Schuchman's condition, Eq. (5). Unfortunately, $q(w)$ is not the total error of a non-subtractively dithered system. Indeed, we will now show the following:

Theorem 1 *In an NSD quantizing system it is not possible to render the total error either statistically independent of the system input or uniformly distributed for system inputs of arbitrary distribution.*

[1] A problem arises in the formalism if the dither is null ($p_\nu(\nu) = \delta(\nu)$) and the system input occurs at a quantizer step edge, since the product of the generalized functions $W_\Delta(\nu - (2n+1)\Delta/2)$ and $\Pi_\Delta(\nu)$ is not conventionally defined. It is shown in [25] that an appropriate definition of this product for the purposes at hand is $\frac{1}{2}[\delta(\nu - n\Delta) + \delta(\nu - (n+1)\Delta)]$.

Proof: Eq. (7) makes it clear that $p_{\varepsilon|x}(\varepsilon, x)$ *cannot* be rendered independent of x by any choice of dither pdf, since the convolution of any dither pdf (which must be non-negative everywhere) with a rectangular window function yields a function at least as wide as the rectangular window. Hence, at least one delta function always makes a contribution to the sum, and the position of that delta function is dependent on the system input.

Taking the Fourier transform of Eq. (8) we find that the characteristic function of ε is given by

$$\begin{aligned} P_\varepsilon(u) &= [\operatorname{sinc}(u) P_\nu(u)] \star \left[W_{\frac{1}{\Delta}}(-u) P_x(-u) \right] \\ &= \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u - \frac{k}{\Delta}\right) P_\nu\left(u - \frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right), \end{aligned} \tag{9}$$

where P_x is the arbitrary cf of the input signal and P_ν is the cf of the dither. In order for ε to be uniformly distributed, this must reduce to $\operatorname{sinc}(u)$ for some choice of P_ν. Suppose that this is possible, in which case we obtain

$$\operatorname{sinc}(u) = \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u - \frac{k}{\Delta}\right) P_\nu\left(u - \frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right).$$

Now let $u = \ell/\Delta$ where $\ell \in \mathbf{Z}_0$. Then we have

$$\operatorname{sinc}\left(\frac{\ell}{\Delta}\right) = 0 = P_x\left(-\frac{\ell}{\Delta}\right),$$

which contradicts the assumption that P_x is arbitrary. Thus the total error cannot be made uniformly distributed in a non-subtractively dithered system for inputs of arbitrary distribution.

□

The counterintuitive nature of this result is the source of much confusion regarding NSD systems. For instance, it is tempting to accept the following line of reasoning:

suppose that a dither satisfying Schuchman's condition (Eq. (5)) is used so that q is independent of x. Then, since ν is also independent of x, the total error $\varepsilon = q+\nu$ is the sum of two random processes both of which are independent of x and thus should be independent of x as well. This conclusion is flatly false. In an NSD quantizing system, given the value of $q + \nu$ we know that the possible values of x satisfy the equation $x = -(q + \nu) + k\Delta$, $k \in \mathbf{Z}$, so that the distribution of x is highly dependent on $q + \nu$. To elucidate the source of the problem we may reason as follows: for arbitrary random variables q, ν, and x and a fourth $\varepsilon = q + \nu$ (none of these *necessarily* representing quantities in a quantizing system) it is clear that

$$p_{\varepsilon|q,\nu,x}(\varepsilon, q, \nu, x) = \delta(\varepsilon - q - \nu).$$

Then

$$\begin{aligned} p_{\varepsilon,x}(\varepsilon, x) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\varepsilon|q,\nu,x}(\varepsilon, q, \nu, x) p_{q,\nu,x}(q, \nu, x) dq d\nu \\ &= \int_{-\infty}^{\infty} p_{q,\nu,x}(\varepsilon - \nu, \nu, x) d\nu, \end{aligned}$$

the Fourier transform of which yields the joint cf of ε and x:

$$\begin{aligned} P_{\varepsilon,x}(u_\varepsilon, u_x) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{q,\nu,x}(\varepsilon - \nu, \nu, x) e^{-j2\pi(u_\varepsilon \varepsilon + u_x x)} d\nu d\varepsilon dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{q,\nu,x}(w, \nu, x) e^{-j2\pi[u_\varepsilon(w+\nu) + u_x x]} dw dx d\nu, \\ &\qquad\qquad \text{where } w = \varepsilon - \nu, \\ &= P_{q,\nu,x}(u_\varepsilon, u_\varepsilon, u_x). \end{aligned}$$

By definition, ε and x are statistically independent of one another if and only if $P_{\varepsilon,x}(u_\varepsilon, u_x)$ can be written as a product of two functions, one involving u_ε alone while the other involves u_x alone. From the above we see that this is the case if and only if

$$P_{q,\nu,x}(u_\varepsilon, u_\varepsilon, u_x) = P_{q,\nu}(u_\varepsilon, u_\varepsilon) P_x(u_x).$$

Unfortunately, knowing as we do that $P_{q,x}(u_q, u_x) = P_q(u_q)P_x(u_x)$ and $P_{\nu,x}(u_\nu, u_x) = P_\nu(u_\nu)P_x(u_x)$ is simply not sufficient to ensure satisfaction of the latter condition. Of course, the result would hold if $\{q, \nu, x\}$ formed a set of independent random variables; that is, if it were the case that

$$P_{q,\nu,x}(u_q, u_\nu, u_x) = P_q(u_q)P_\nu(u_\nu)P_x(u_x).$$

However, this even stronger condition is *certainly* not met in an NSD quantizing system with an arbitrarily distributed input.

Since we have shown that statistical independence of the total error from the system input is not achievable, we now turn our attention to the possibility of controlling moments of the error. For many applications, controlling relevant error moments is just as good as having full statistical independence of the input and error processes.

2.2 A Condition for the Independence of Total Error Moments

The m-th moment of the error signal is the expectation value of ε^m:

$$E[\varepsilon^m] = \int_{-\infty}^{\infty} \varepsilon^m p_\varepsilon(\varepsilon) d\varepsilon.$$

It can be shown that these moments may also be expressed in terms of the cf of the given random variable as [33]:

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m P_\varepsilon^{(m)}(0), \tag{10}$$

where $P_\varepsilon^{(m)}$ denotes the m-th derivative of P_ε. (This expression is easily derived by differenting with respect to u the definition of $P_\varepsilon(u)$ as the Fourier transform of p_ε.)

From Eq. (9) we obtain

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m \sum_{k=-\infty}^{\infty} G_\nu^{(m)}\left(\frac{k}{\Delta}\right) P_x\left(\frac{k}{\Delta}\right), \tag{11}$$

where

$$G_\nu(u) \stackrel{\triangle}{=} \mathrm{sinc}\,(u) P_\nu(u). \tag{12}$$

Since the cf, P_x, of the system input is arbitrary we obtain the following result [26]:

Theorem 2 *In an NSD quantizing system, $E[\varepsilon^m]$ is independent of the distribution of the system input, x, if and only if*

$$G_\nu^{(m)}\left(\frac{k}{\Delta}\right) = 0 \qquad \forall k \in \mathbf{Z}_0. \tag{13}$$

If the conditions of Theorem 2 are satisfied then, from Eq. (11),

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m G_\nu^{(m)}(0),$$

which is precisely the m-th moment of a notional random process with cf G_ν and pdf $\Delta\Pi_\Delta \star p_\nu$, although this is not, of course, the pdf of ε. We can derive the following expressions for the moments of the total error in terms of the moments of the dither signal by direct differentiation of $G_\nu(u)$:

$$E[\varepsilon] = E[\nu] \tag{14}$$

$$E[\varepsilon^2] = E[\nu^2] + \frac{\Delta^2}{12} \tag{15}$$

$$E[\varepsilon^m] = \sum_{\ell=0}^{\lfloor \frac{m}{2} \rfloor} \binom{m}{2\ell} \left(\frac{\Delta}{2}\right)^{2\ell} \frac{E[\nu^{m-2\ell}]}{2\ell+1}. \tag{16}$$

We emphasize that each of these equations for $E[\varepsilon^m]$ is only valid when Theorem 2 is satisfied for that particular value of m, and that the validity of one of these equations does not imply the validity of any others corresponding to different m values.

Eq. (15) merits special comment. It indicates that if the total error variance in an NSD quantizing system is input independent, then it always exceeds that of an SD

system (or a system described by the CMQ) by an amount equal to the variance of the dither. This characteristic increase in the error power is not problematic in most multi-bit applications, and the benefits of dithering typically far outweigh the slight noise penalty.

Two corollaries to Theorem 2 follow.

Corollary 1 *In an NSD quantizing system, if the condition of Eq. (13) is satisfied for any given m, then for any choice of n*

$$E[\varepsilon^m x^n] = E[\varepsilon^m]E[x^n];$$

i.e., ε^m and x^n are uncorrelated.

Proof: We observe that if $p_x(x) = \delta(x - x_0)$ then

$$\begin{aligned} p_\varepsilon(\varepsilon) &= \int_{-\infty}^{\infty} p_{\varepsilon,x}(\varepsilon, x)dx \\ &= \int_{-\infty}^{\infty} p_{\varepsilon|x}(\varepsilon, x)\delta(x - x_0)dx \\ &= p_{\varepsilon|x}(\varepsilon, x_0). \end{aligned} \tag{17}$$

By Theorem 2, $E[\varepsilon^m]$ is independent of the choice of p_x, and in particular it is independent of the choice of x_0 when $p_x(x) = \delta(x - x_0)$, as above. Thus

$$\int_{-\infty}^{\infty} \varepsilon^m p_{\varepsilon|x}(\varepsilon, x)d\varepsilon = E[\varepsilon^m]$$

for any x. In this case,

$$\begin{aligned} E[\varepsilon^m x^n] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varepsilon^m x^n p_{\varepsilon,x}(\varepsilon, x)d\varepsilon dx \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} \varepsilon^m p_{\varepsilon|x}(\varepsilon, x)d\varepsilon\right] x^n p_x(x)dx \\ &= \int_{-\infty}^{\infty} E[\varepsilon^m]x^n p_x(x)dx \\ &= E[\varepsilon^m]E[x^n]. \end{aligned}$$

□

In particular, if $E[\varepsilon]$ is independent of the distribution of x, then ε and x are uncorrelated in the usual mathematical sense:

$$E[\varepsilon x] = E[\varepsilon]E[x].$$

The second corollary is somewhat better known than Theorem 2 itself, but demands satisfaction of a stronger condition [10, 30].

Corollary 2 *In an NSD quantizing system, $E[\varepsilon^\ell]$ is independent of the distribution of the system input, x, for $\ell = 1, 2, \ldots, m$ if and only if*

$$P_\nu^{(i)}\left(\frac{k}{\Delta}\right) = 0$$

$$\forall k \in \mathbf{Z}_0 \qquad \textit{and} \qquad i = 0, 1, 2, \ldots, m-1.$$

Proof: Proof of the "if" direction follows immediately from repeated differentiation of Eq. (12):

$$G_\nu^{(\ell)}(u) = \sum_{i=0}^{\ell} \binom{\ell}{i} \mathrm{sinc}^{(\ell-i)}(u) P_\nu^{(i)}(u).$$

We see that the ℓ-th and all lower derivatives of G_ν will all go to zero at $u = k/\Delta$, $k \in \mathbf{Z}_0$ if the first $\ell - 1$ derivatives of P_ν do. The "only if" direction is easily proven using induction, but this requires more space than is justified here. The interested reader is referred to [25].

□

In most practical applications, we are interested in dither signals which satisfy the conditions of Corollary 2, and it turns out that the conditions of this corollary will be

of interest when we examine the statistics of the quantizer output (Section 2.4) and the special nature of digital dither signals (Section 4).

2.3 Second-Order Statistics of Total Error Values

We now begin an investigation into the joint statistics of temporally separated total error values, corresponding to input samples separated in time, in order to derive conclusions about the spectral characteristics of the total error sequence.

Consider two total error values, ε_1 and ε_2, which are separated in time by $\tau \neq 0$. (In the special case where $\tau = 0$, the analysis reduces to that of Section 2.2.) The corresponding system input values will be denoted as x_1 and x_2, respectively. Employing a derivation analogous to that of the Section 2.1 we find that

$$\begin{aligned} & p_{(\varepsilon_1,\varepsilon_2)\big|(x_1,x_2)}(\varepsilon_1,\varepsilon_2,x_1,x_2) \\ &\quad = \sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty} \delta(\varepsilon_1 + x_1 - k_1\Delta)\delta(\varepsilon_2 + x_2 - k_2\Delta) \\ &\qquad \times \int_{-\frac{\Delta}{2}+k_1\Delta}^{\frac{\Delta}{2}+k_1\Delta}\int_{-\frac{\Delta}{2}+k_2\Delta}^{\frac{\Delta}{2}+k_2\Delta} p_{\nu_1,\nu_2}(w_1 - x_1, w_2 - x_2)dw_1 dw_2 \\ &\quad = [\Delta^2\Pi_{\Delta\Delta} \star p_{\nu_1,\nu_2}](\varepsilon_1,\varepsilon_2)W_{\Delta\Delta}(\varepsilon_1 + x_1, \varepsilon_2 + x_2), \end{aligned}$$

where the convolution is two-dimensional, involving both ε_1 and ε_2, and where

$$\Pi_{\Gamma\Gamma}(\varepsilon_1,\varepsilon_2) \stackrel{\triangle}{=} \Pi_\Gamma(\varepsilon_1)\Pi_\Gamma(\varepsilon_2)$$

and

$$W_{\Gamma\Gamma}(\varepsilon_1,\varepsilon_2) \stackrel{\triangle}{=} W_\Gamma(\varepsilon_1)W_\Gamma(\varepsilon_2).$$

p_{ν_1,ν_2} represents the *joint pdf* of the dither values, ν_1 and ν_2, associated with the inputs x_1 and x_2, respectively.

Hence

$$\begin{aligned} p_{\varepsilon_1,\varepsilon_2}(\varepsilon_1,\varepsilon_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{(\varepsilon_1,\varepsilon_2)|(x_1,x_2)}(\varepsilon_1,\varepsilon_2,x_1,x_2)p_{x_1,x_2}(x_1,x_2)dx_1dx_2 \\ &= [\Delta^2\Pi_{\Delta\Delta} \star p_{\nu_1,\nu_2}](\varepsilon_1,\varepsilon_2)[W_{\Delta\Delta} \star p_{x_1,x_2}](-\varepsilon_1,-\varepsilon_2). \end{aligned} \tag{18}$$

The joint characteristic function of ε_1 and ε_2 is found by taking the two-dimensional Fourier transform of Eq. (18) with respect to ε_1 and ε_2, resulting in an expression in the corresponding frequency variables, u_1 and u_2:

$$\begin{aligned} P_{\varepsilon_1,\varepsilon_2}(u_1,u_2) &= \sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty} \operatorname{sinc}\left(u_1-\frac{k_1}{\Delta}\right)\operatorname{sinc}\left(u_2-\frac{k_2}{\Delta}\right) \\ &\quad \times P_{\nu_1,\nu_2}\left(u_1-\frac{k_1}{\Delta},u_2-\frac{k_2}{\Delta}\right) P_{x_1,x_2}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right). \end{aligned} \tag{19}$$

No choice of dither pdf will allow Eq. (19) to be expressed as a product of two characteristic functions, one involving u_1 alone and the other u_2 alone, for arbitrary choices of P_{x_1,x_2}. Thus ε_1 and ε_2 cannot be rendered statistically independent for arbitrary joint input distributions. Let us therefore proceed to investigate the joint moments of ε_1 and ε_2 in the hope that we can exercise some control over them by an appropriate choice of the dither statistics.

The (m_1,m_2)-*th joint moment* of the two signals of interest is given by:

$$\begin{aligned} E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] &\triangleq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varepsilon_1^{m_2}\varepsilon_2^{m_2} p_{\varepsilon_1,\varepsilon_2}(\varepsilon_1,\varepsilon_2)d\varepsilon_1 d\varepsilon_2 \\ &= \left(\frac{j}{2\pi}\right)^{m_1+m_2} P_{\varepsilon_1,\varepsilon_2}^{(m_1,m_2)}(0,0) \end{aligned} \tag{20}$$

where

$$P_{\varepsilon_1,\varepsilon_2}^{(m_1,m_2)}(u_1,u_2) \triangleq \frac{\partial^{(m_1+m_2)}P_{\varepsilon_1,\varepsilon_2}}{\partial u_1^{m_1}\partial u_2^{m_2}}(u_1,u_2).$$

Substituting Eq. (19) into Eq. (20), one finds that

$$E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] = \left(\frac{j}{2\pi}\right)^{m_1+m_2} \sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty} P_{x_1,x_2}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right) G_{\nu_1,\nu_2}^{(m_1,m_2)}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right) \tag{21}$$

where

$$G_{\nu_1,\nu_2}(u_1,u_2) \stackrel{\triangle}{=} \text{sinc}\,(u_1)\text{sinc}\,(u_2)P_{\nu_1,\nu_2}(u_1,u_2).$$

At this point we may deduce a theorem which represents a second-order analog of Theorem 2:

Theorem 3 *In an NSD quantizing system, the (m_1,m_2)-th joint moment, $E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}]$, of two total error values, ε_1 and ε_2, separated in time by $\tau \neq 0$, is independent of the system input for arbitrary input distributions if and only if*

$$G_{\nu_1,\nu_2}^{(m_1,m_2)}\left(\frac{k_1}{\Delta},\frac{k_2}{\Delta}\right) = 0 \qquad \forall (k_1,k_2) \in \mathbf{Z}_0^2. \tag{22}$$

The proof is completely analogous to that of Theorem 2. When Eq. (22) is satisfied, we have

$$E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] = \left(\frac{j}{2\pi}\right)^{m_1+m_2} G_{\nu_1,\nu_2}^{(m_1,m_2)}(0,0), \tag{23}$$

so that by explicitly performing the differentiation we can write an expression, analogous to Eq. (16), relating the joint moments of the total error to those of the dither:

$$E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] = \sum_{\ell_1=0}^{\lfloor \frac{m_1}{2} \rfloor} \sum_{\ell_2=0}^{\lfloor \frac{m_2}{2} \rfloor} \binom{m_1}{2\ell_1}\binom{m_2}{2\ell_2}\left(\frac{\Delta}{2}\right)^{2(\ell_1+\ell_2)} \frac{E[\nu_1^{m_1-2\ell_1}\nu_2^{m_2-2\ell_2}]}{(2\ell_1+1)(2\ell_2+1)}. \tag{24}$$

We attach the caveat that satisfaction of Eq. (24) for some particular m_1 and m_2 does not imply its satisfaction for any other values thereof.

If the dither process is iid so that ν_1 and ν_2 are statistically independent then

$$P_{\nu_1,\nu_2}(u_1,u_2) = P_\nu(u_1)P_\nu(u_2).$$

Then if the conditions of Corollary 2 are satisfied for $m = \max(m_1, m_2)$ we have

$$\begin{aligned} E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] &= \left(\frac{j}{2\pi}\right)^{m_1+m_2} G_\nu^{(m_1)}(0) G_\nu^{(m_2)}(0) \\ &= E[\varepsilon_1^{m_1}]E[\varepsilon_2^{m_2}] \end{aligned} \tag{25}$$

so that $\varepsilon_1^{m_1}$ and $\varepsilon_2^{m_2}$ are uncorrelated. In this case, of course, $E[\varepsilon_1^m] = E[\varepsilon_2^m] = E[\varepsilon^m]$. Hence:

Corollary 3 *Any iid non-subtractive dither signal which satisfies the conditions of Corollary 2 for $m = \max(m_1, m_2)$ will ensure that, for two error values, ε_1 and ε_2, separated in time by $\tau \neq 0$,*

$$E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] = E[\varepsilon^{m_1}]E[\varepsilon^{m_2}]. \tag{26}$$

In this case $E[\varepsilon^{m_2}]$ and $E[\varepsilon^{m_2}]$ will be given by Eq. (16). In particular, for an iid dither with zero mean we note that $E[\varepsilon_1\varepsilon_2] = 0$.

In a digital system, the total error is a discrete-time signal, so that $\tau = kT$ where T represents the sampling period and $k \in \mathbf{Z}$. The *autocorrelation function* of such a signal is defined to be $E[\varepsilon_1\varepsilon_2](k)$. The *power spectral density* (PSD) of a discrete-time random process is equal by definition to the discrete-time Fourier transform (DTFT) of its autocorrelation function, where we define the DTFT as

$$\mathcal{F}_{DT}[h](f) \stackrel{\triangle}{=} 2T \sum_{k=-\infty}^{\infty} h(k) e^{-j2\pi f k T}, \tag{27}$$

where the continuous frequency variable, f, is in hertz if T is in seconds. This definition is normalized such that the integral of the PSD from zero to the Nyquist frequency, $\frac{1}{2T}$, yields the variance of the signal.

Using Eqs. (15) and (26) we find that, for an NSD quantizing system using iid dither satisfying the conditions of Corollary 2 for $m = 2$, the autocorrelation function of the error is

$$E[\varepsilon_1\varepsilon_2](k) = \begin{cases} E[\nu^2] + \dfrac{\Delta^2}{12}, & k = 0, \\ E^2[\nu], & \text{otherwise.} \end{cases}$$

Comparing this with the autocorrelation function of the dither sequence,

$$E[\nu_1\nu_2](k) = \begin{cases} E[\nu^2], & k = 0, \\ E^2[\nu], & \text{otherwise,} \end{cases}$$

we conclude that

$$PSD_\varepsilon(f) = PSD_\nu(f) + \frac{\Delta^2 T}{6}$$

so that the total error signal must be spectrally white since the dither is spectrally white (apart from a dc component if the dither is not zero mean).

It is possible to derive conditions which ensure the satisfaction of Eq. (22) for the case where $m_1 = m_2 = 1$, but which do not require statistical independence of distinct dither values [24, 25]. This will allow the use of certain dither signals which are notspectrally white.

Theorem 4 *In an NSD system where all dither values are statistically independent of all system input values,*

$$E[\varepsilon_1\varepsilon_2] = E[\nu_1\nu_2] \tag{28}$$

for arbitrary input distributions if and only if the following three conditions are satisfied:

$$P_{\nu_1,\nu_2}\left(\frac{k_1}{\Delta}, \frac{k_2}{\Delta}\right) = 0 \quad \forall (k_1, k_2) \in \mathbf{Z}_0^2, \tag{29}$$

$$P^{(0,1)}_{\nu_1,\nu_2}\left(\frac{k_1}{\Delta}, 0\right) = 0 \quad \forall k_1 \in \mathbf{Z}_0, \tag{30}$$

$$P^{(1,0)}_{\nu_1,\nu_2}\left(0, \frac{k_2}{\Delta}\right) = 0 \quad \forall k_2 \in \mathbf{Z}_0. \tag{31}$$

This can be thought of as a second-order counterpart to Corollary 2 for the simple case $m_1 = m_2 = 1$. When the conditions of the theorem are satisfied, Eq. (28) follows immediately from direct differentiation of Eq. (23). It is possible to gain further insight into the meaning of the conditions involved by noting that

$$E[\varepsilon_1\varepsilon_2] = E[(q_1 + \nu_1)(q_2 + \nu_2)] = E[\nu_1\nu_2] + E[q_1\nu_2] + E[q_2\nu_1] + E[q_1q_2].$$

The last term is equal to zero as long as Eq. (29) is satisfied (see Eq. (6) or Theorem 2 of [26]) while it can be shown that the second and third terms vanish subject to the satisfaction of Eqs. (30) and (31), thus yielding Eq. (28). Necessity of the conditions follows from the arbitrariness of the input distribution.

Suppose that the conditions of both Theorem 4 and Corollary 2 with $m = 2$ are satisfied. Then the autocorrelation function of the error is given by

$$E[\varepsilon_1\varepsilon_2](k) = \begin{cases} E[\nu^2] + \dfrac{\Delta^2}{12}, & k = 0, \\ E[\nu_1\nu_2], & \text{otherwise.} \end{cases} \tag{32}$$

This indicates that the power spectrum of the error will be identical to the power spectrum of the dither, apart from a contribution due to the $k = 0$ case, manifested as an additive constant present at all frequencies (i.e., a white spectral component introduced by the properly dithered quantization operation). Hence, as before,

$$PSD_\varepsilon(f) = PSD_\nu(f) + \frac{\Delta^2 T}{6}, \tag{33}$$

except that now the dither PSD is not necessarily white. This will be illustrated by the discussion of high-pass dither in Section 3.5.

2.4 Statistics of the System Output

It is of interest to express the statistical attributes of the output, y, in terms of those of the input, x, since it is frequently required that one be deduced from the other. We apply the same brand of reasoning as used to determine the cpdf of the total error in Section 2.1. The values of the quantizer output are restricted to values of $k\Delta$, $k \in \mathbf{Z}$. Therefore $p_y(y)$ will consist of delta functions at these locations weighted by the probability that the quantizer input $w = x + \nu$ falls in the range of the corresponding quantizer step, $\frac{2k-1}{2}\Delta < w < \frac{2k+1}{2}\Delta$. This probability is just the integral of $p_w(w)$ over this range where, since x and ν are statistically independent, p_w is given by [34]

$$p_w(w) = [p_\nu \star p_x](w).$$

Thus we have

$$\begin{aligned} p_y(y) &= \sum_{k=-\infty}^{\infty} \delta(y - k\Delta) \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} [p_\nu \star p_x](w)dw \\ &= [\Delta\Pi_\Delta \star p_\nu \star p_x](y) W_\Delta(y). \end{aligned}$$

Taking the Fourier transform of this expression yields

$$\begin{aligned} P_y(u) &= [G_\nu(u)P_x(u)] \star W_{\frac{1}{\Delta}}(u) \\ &= \sum_{k=-\infty}^{\infty} G_\nu\left(u - \frac{k}{\Delta}\right) P_x\left(u - \frac{k}{\Delta}\right) \end{aligned} \tag{34}$$

and so

$$\begin{aligned} E[y^m] &= \left(\frac{j}{2\pi}\right)^m P_y^{(m)}(0) \\ &= \sum_{k=-\infty}^{\infty} \sum_{r=0}^{m} \binom{m}{r} \left[\left(\frac{j}{2\pi}\right)^r G_\nu^{(r)}\left(\frac{k}{\Delta}\right)\right] \left[\left(\frac{j}{2\pi}\right)^{m-r} P_x^{(m-r)}\left(\frac{k}{\Delta}\right)\right]. \end{aligned} \tag{35}$$

Now, if the first m derivatives of $G_\nu(u)$ are zero at all non-zero multiples of $1/\Delta$, then Eq. (35) reduces to

$$E[y^m] = \sum_{r=0}^{m} \binom{m}{r} E[\varepsilon^r] E[x^{m-r}], \tag{36}$$

where the expectation values of the total error are given in terms of the expectation values of the dither by Eq. (16). By direct differentiation of $G_\nu(u)$, the above condition is easily shown to be equivalent to the condition of Corollary 2. For the special cases $m = 1$ and $m = 2$ we note that

$$\begin{aligned} E[y] &= E[x] + E[\varepsilon] = E[x] + E[\nu] \\ E[y^2] &= E[x^2] + 2E[x]E[\nu] + E[\nu^2] + \frac{\Delta^2}{12} \end{aligned}$$

where Eqs. (14) and (15) have been substituted for the error moments.[2]

Proceeding similarly for the joint moments of output values y_1 and y_2, separated in time by $\tau \neq 0$, we find that

$$\begin{aligned} E[y_1^{m_1} y_2^{m_2}] &= \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \sum_{r_1=0}^{m_1} \sum_{r_2=0}^{m_2} \binom{m_1}{r_1} \binom{m_2}{r_2} \left[\left(\frac{j}{2\pi} \right)^{r_1+r_2} G_{\nu_1,\nu_2}^{(r_1,r_2)} \left(\frac{k_1}{\Delta}, \frac{k_2}{\Delta} \right) \right] \\ &\quad \times \left[\left(\frac{j}{2\pi} \right)^{(m_1-r_1)+(m_2-r_2)} P_{x_1,x_2}^{(m_1-r_1,m_2-r_2)} \left(\frac{k_1}{\Delta}, \frac{k_2}{\Delta} \right) \right]. \end{aligned} \tag{37}$$

If the indicated partial derivatives of G_{ν_1,ν_2} are zero for all $(k_1, k_2) \in \mathbf{Z}_0^2$, $r_i = 1, 2, \ldots, m_i$, $i \in \{1, 2\}$, then Eq. (37) reduces to

$$E[y_1^{m_1} y_2^{m_2}] = \sum_{r_1=0}^{m_1} \sum_{r_2=0}^{m_2} \binom{m_1}{r_1} \binom{m_2}{r_2} E[\varepsilon_1^{r_1} \varepsilon_2^{r_2}] E[x_1^{m_1-r_1} x_2^{m_2-r_2}], \tag{38}$$

[2]Note that the so-called *dither averaged transfer characteristic,* $E[y|x]$, is given by

$$E[y|x] = E[Q(x+\nu)|x] = \int_{-\infty}^{\infty} Q(x+\nu) p_\nu(\nu) d\nu = Q(x) * p_\nu(-x),$$

which is the convolution of the quantizer staircase with the dither pdf. For the $m = 1$ case this defines the line $y = x$ (see illustrations in [16, 17]).

where the joint moments of the total error are given in terms of those of the dither by Eq. (24).

Beginning from Eq. (37) with $m_1 = m_2 = 1$ it is straightforward to show that if the conditions of Theorem 4 are satisfied (i.e., Eqs. (29), (30) and (31)) then

$$E[y_1 y_2] = E[x_1 x_2] + E[\nu_1 \nu_2],$$

so that, with the aid of Eqs. (36) and (16), we find that the output has an autocorrelation function

$$E[y_1 y_2](k) = \begin{cases} E[x^2] + 2E[x]E[\nu] + E[\nu^2] + \dfrac{\Delta^2}{12}, & k = 0, \\ E[x_1 x_2] + E[\nu_1 \nu_2], & \text{otherwise.} \end{cases} \tag{39}$$

Then the spectrum of the output is the sum of the input and dither spectra apart from a white noise component, which is contributed by the $k = 0$ case of Eq. (39). The latter component is comparable to the white "quantization noise" posited in the CMQ. In particular, for a system using a zero-mean dither,

$$PSD_y(f) = PSD_x(f) + PSD_\nu(f) + \frac{\Delta^2 T}{6}.$$

3 Error Moments in Some Representative Systems

We proceed to apply the above results to realizable quantizing systems using a variety of dither signals.

3.1 Null Dither

We begin by considering an undithered system. The pdf of a "null dither" is

$$p_\nu(\nu) = \delta(\nu),$$

the Fourier transform of which is equal to unity everywhere. Hence, by Eq. (12),

$$G_\nu(u) = \text{sinc}\,(u).$$

No derivatives of this function vanish at non-zero multiples of $1/\Delta$, so no moments of the total error (excepting the zeroth) will be independent of the input distribution. Of course, it is not expected that they would be. We know that in the absence of dither, the error is a *deterministic function* of the input. Indeed, the mean value of the error as a function of the input is identical to the error function, $q(x)$, shown in Fig. 2.

3.2 Rectangular-PDF Dither

Now consider a system using dither with a simple rectangular (i.e., uniform) pdf of 1 LSB peak-to-peak amplitude:

$$p_\nu(\nu) = \Pi_\Delta(\nu),$$

with a corresponding cf

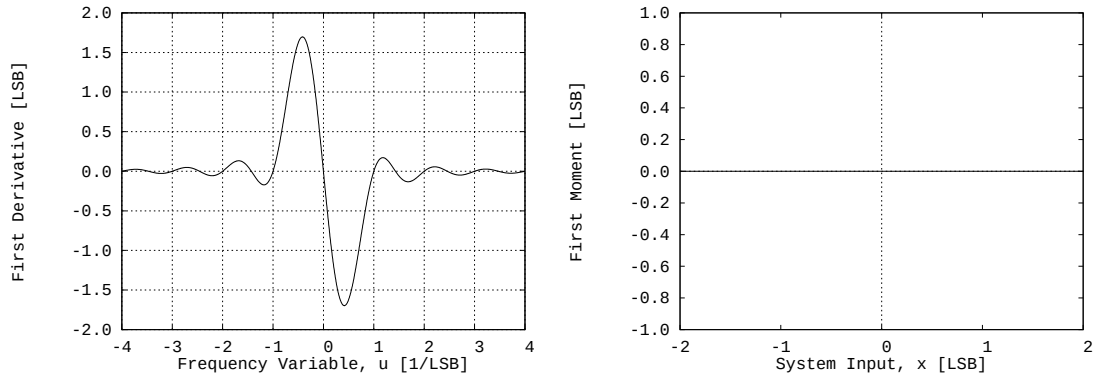
$$P_\nu(u) = \text{sinc}\,(u).$$

Hence, from Eq. (12):

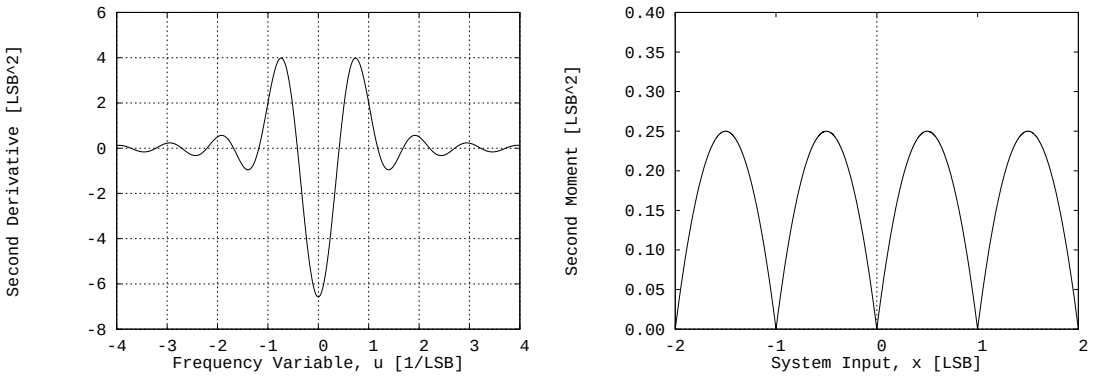
$$G_\nu(u) = \text{sinc}^2\,(u).$$

The first two derivatives of this function are plotted in Fig. 5. The first derivative clearly satisfies the condition of going to zero at the regularly spaced points stipulated by Eq. (13), while the second derivative does not (nor do higher derivatives). This indicates that the first moment of the error signal is independent of the input, but that its variance remains dependent. These conclusions are borne out by the accompanying plots in Fig. 5 of the *conditional moments*

$$E[\varepsilon^m|x] \stackrel{\triangle}{=} \int_{-\infty}^{\infty} \varepsilon^m p_{\varepsilon|x}(\varepsilon, x) d\varepsilon,$$



(a)



(b)

Figure 5: Derivatives of $G_\nu(u)$ (left) and conditional moments of the error (right) for a quantizer using RPDF dither of 1 LSB peak-to-peak amplitude: (a) $G_\nu^{(1)}(u)$ and $E[\varepsilon|x]$ (both in units of Δ), (b) $G_\nu^{(2)}(u)$ and $E[\varepsilon^2|x]$ (both in units of Δ^2). The frequency variable, u, is plotted in units of $1/\Delta$ and the system input, x, in units of Δ.

as computed using Eq. (7). The first moment, or mean error, is zero for all inputs, indicating that the quantizer has been *linearized* by the use of this dither. The error variance, on the other hand, is clearly signal dependent, so that the noise power in the signal varies with the input. This is sometimes referred to as *noise modulation* and is undesirable in audio or video signals.

If the dither is iid, then (as shown in Section 2.3), temporally separated error values will be uncorrelated. Thus, short-time error spectra will appear flat but their level will be input dependent.

3.3 Triangular-PDF Dither

The most straightforward means of generating dither signals with more complicated pdf's is to simply sum two or more statistically independent RPDF random processes. For instance, the sum of two such processes, ν_1 and ν_2, each of 1 LSB peak-to-peak amplitude, yields a dither with a triangular pdf (TPDF) of 2 LSB peak-to-peak amplitude, since the summation of statistically independent random processes convolves their pdf's (see Fig. 6):

$$\begin{aligned} p_\nu(\nu) &= [p_{\nu_1} \star p_{\nu_2}](\nu) \\ &= [\Pi_\Delta \star \Pi_\Delta](\nu). \end{aligned} \tag{40}$$

Convolution of pdf's corresponds to multiplication of the respective cf's [34], so that in a system employing this kind of dither P_ν is a squared sinc function and G_ν is given by

$$G_\nu(u) = \mathrm{sinc}^3(u).$$

The first *and* second derivatives of this function go to zero at the required places, so this dither renders both the first *and* second moments of the total error independent

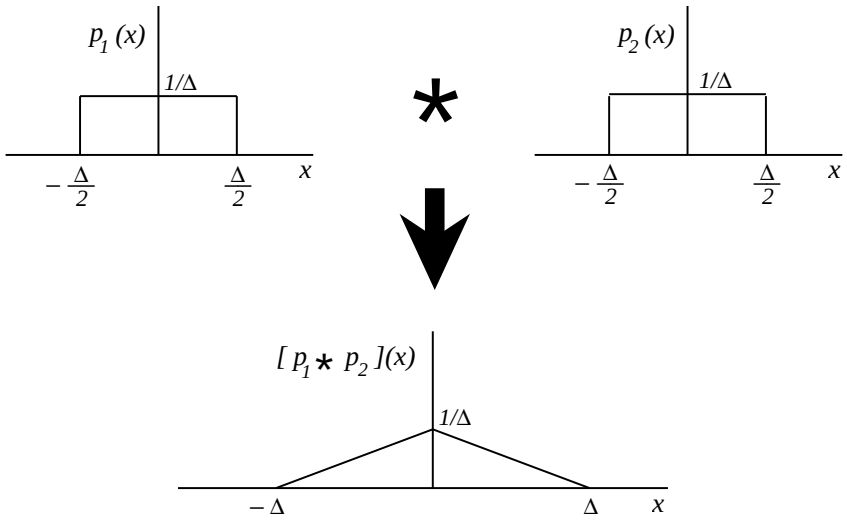


Figure 6: A triangular pdf, formed by the convolution of two rectangular pdf's.

of the system input.[3] The second derivative of G_ν is shown in Fig. 7, along with the second conditional moment of the total error, which is a constant $\Delta^2/4$ for all inputs, in agreement with Eq. (15). Higher derivatives of G_ν do not meet the required conditions, so that higher moments of the error remain dependent on the input.

We now proceed to show that a triangular pdf of 2 LSB peak-to-peak amplitude is the only choice of zero-mean dither pdf which renders the first two moments of the total error independent of the input while minimizing the second. We will begin by noting, from Corollary 2 and the stipulation of zero mean, that

$$\begin{aligned} P_\nu\left(\frac{k}{\Delta}\right) &= 0, \qquad \forall k \in \mathbf{Z}_0, \\ P_\nu^{(1)}\left(\frac{k}{\Delta}\right) &= 0, \qquad \forall k \in \mathbf{Z}. \end{aligned}$$

Also, $P_\nu(u)$ must be equal to unity at $u = 0$ if it is to be a valid characteristic function,

[3] A different proof of this result, using a direct method, was given in [16].

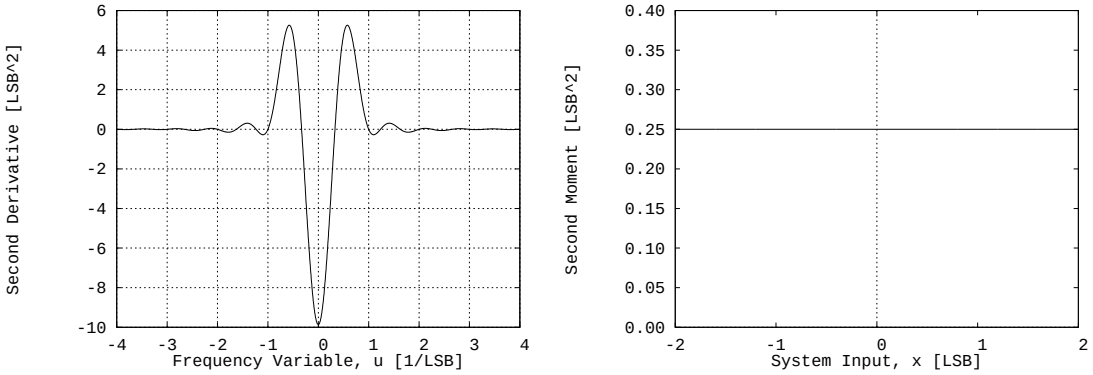


Figure 7: $G_\nu^{(2)}(u)$ (left) and $E[\varepsilon^2|x]$ (right) (both in units of Δ^2) for a quantizer using triangular-pdf dither of 2 LSB peak-to-peak amplitude. The frequency variable, u, is plotted in units of $1/\Delta$ and the system input, x, in units of Δ.

since

$$P_\nu(0) = \int_{-\infty}^{\infty} e^{-j2\pi(0)\nu} p_\nu(\nu) d\nu = 1.$$

We conclude that the dither cf and its first derivative are completely specified at *all* integer multiples of $1/\Delta$. According to the Generalized Sampling Theorem [34], this is sufficient to uniquely specify $P_\nu(u)$ for all u if $p_\nu(\nu)$ is Δ-bandlimited (i.e., if $p_\nu(\nu) = 0$ for $|\nu| \geq \Delta$). Since the pdf of Eq. (40) *is* Δ-bandlimited, and its corresponding cf satisfies all the given conditions, it must be the unique pdf in question.

It remains to be shown that any dither pdf which is not thus bandlimited will produce a greater error variance. Since this variance is assumed to be constant with respect to the input, it is sufficient to show that this holds for a single input value. We will do so for $x = \Delta/2$.

$p_{\varepsilon|x}(\varepsilon, x)$ for $x = \Delta/2$ is obtained from Eq. (8) using $p_x(x) = \delta(x - \frac{\Delta}{2})$ (see Eq. (17)). As is shown in Fig. 8(a) it consists of two equally weighted delta functions at $\varepsilon = \pm\Delta/2$ when triangular-pdf dither of 2 LSB peak-to-peak amplitude is employed.

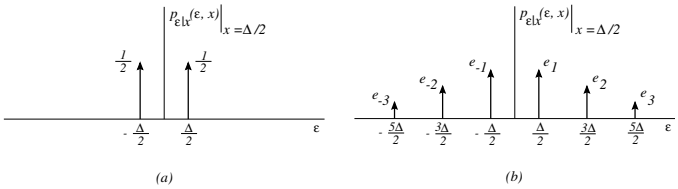


Figure 8: $p_{\varepsilon|x}(\varepsilon, x)$ evaluated at $x = \Delta/2$ for systems using (a) a triangular-pdf dither of 2 LSB peak-to-peak amplitude and (b) a wider dither pdf (the delta functions possess the indicated weightings).

Use of a wider dither pdf will result in the appearance of more delta functions in the error's cpdf, as shown in Fig. 8(b), where we denote the weighting of the delta function at $\varepsilon = \pm(2i-1)\Delta/2$, $i \geq 1$, by $e_{\pm i}$, so that

$$p_{\varepsilon|x}\left(\varepsilon, \frac{\Delta}{2}\right) = \sum_{i=1}^{\infty}\left[e_i\,\delta\!\left(\varepsilon - (2i-1)\frac{\Delta}{2}\right) + e_{-i}\,\delta\!\left(\varepsilon + (2i-1)\frac{\Delta}{2}\right)\right]. \tag{41}$$

We proceed by expressing the fundamental condition that the integral of this pdf must equal unity:

$$(e_1 + e_{-1}) + \sum_{i=2}^{\infty}(e_i + e_{-i}) = 1. \tag{42}$$

Now, by direct integration of Eq. (41), we have

$$\begin{aligned} E[\varepsilon^2|x = \Delta/2] &= \sum_{i=1}^{\infty}\left[(2i-1)\frac{\Delta}{2}\right]^2 (e_i + e_{-i}) \\ &= \frac{\Delta^2}{4}\left[(e_1 + e_{-1}) + \sum_{i=2}^{\infty}(2i-1)^2(e_i + e_{-i})\right]. \end{aligned} \tag{43}$$

Substituting Eq. (42) yields

$$E[\varepsilon^2|x = \Delta/2] = \frac{\Delta^2}{4}\left[1 + 4\sum_{i=2}^{\infty} i(i-1)(e_i + e_{-i})\right],$$

which is always greater than $\Delta^2/4$ since the e_i's must be positive. We have thus shown the following:

Theorem 5 *The choice of zero-mean dither pdf which renders the first and second moments of the total error independent of the input, such that the first moment is zero and the second is minimized, is unique and is a triangular pdf of 2 LSB peak-to-peak amplitude.*

3.4 The Sum of n Independent Rectangular-PDF Random Processes

Theorem 6 *A non-subtractive dither signal generated by the summation of n statistically independent RPDF random processes, renders $E[\varepsilon^\ell]$ independent of the system input distribution for $\ell = 0, 1, \ldots, n$, and results in a total error variance, for $n \geq 2$, of $(n+1)\Delta^2/12$.*

This must be the case since the use of n such dithers gives

$$G_\nu(u) = \text{sinc}^{n+1}(u),$$

the first n derivatives of which will consist entirely of terms containing non-zero powers of sinc (u). Since this function goes to zero at the required places, the first n moments of the error will always be independent of the input. Higher derivatives will not share this property [25]. Dithers of this form are sometimes referred to as nRPDF so that, for instance, TPDF dither may also be referred to as 2RPDF.

It is important to note that using uniformly distributed processes of peak-to-peak amplitude not equal to one LSB (or, rather, not equal to an integral number of LSB's) will not render error moments independent of the input since the zeros of the associated sinc functions will not fall at integral multiples of $1/\Delta$ (see illustrations in [16]).

Finally, it is easily shown from the Generalized Sampling Theorem that the $(n\Delta/2)$-bandlimited dither pdf which renders the first n moments of the total error independent of the input is unique, and must therefore be the pdf of Theorem 6.

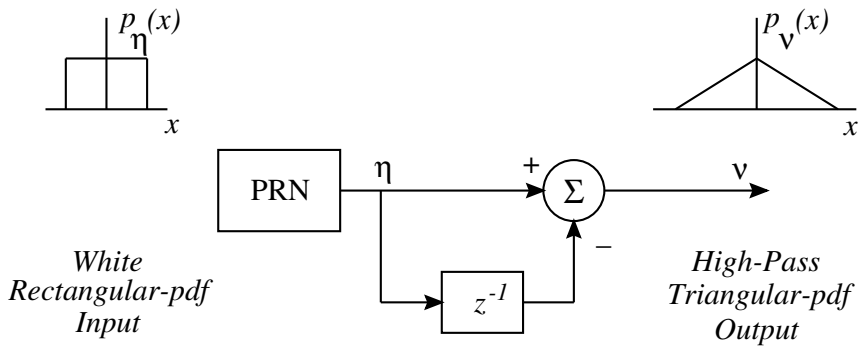


Figure 9: High-pass dither generator.

3.5 High-Pass Dither

A very simple discrete-time noise generator capable of producing dither with a high-pass spectrum is shown in Fig. 9. The system contains a pseudo-random number generator, marked PRN, producing iid, uniformly distributed random numbers, and a one-sample delay element marked z^{-1}. The output is the difference between the pseudo-random number most recently generated by the PRN, η_n, and the previous one, η_{n-1}; that is, $\nu_n = \eta_n - \eta_{n-1}$. The (first-order) pdf of the resulting dither sequence is triangular (TPDF), since it results from the summation of two statistically independent RPDF sequences, albeit one of these is simply a delayed version of the other. This means that all the beneficial effects of the TPDF dither discussed in Section 3.3 will also be associated with dither thus generated. Such high-pass TPDF dither may be preferable in some audio applications since it is less audible than spectrally white TPDF dither due to the ear's reduced sensitivity at high frequencies (although these dithers have equal variances of $\Delta^2/6$). Similar comments apply regarding reduced error visibility in imaging applications. Furthermore, the use of high-pass TPDF dither is more computationally efficient since it requires the calculation of only *one* new

RPDF random number per sample as compared to two when iid TPDF dither is used.

In order to investigate the spectral characteristics of the total error associated with this sort of dither, we must derive an expression for $p_{\nu_1,\nu_2}(\nu_1,\nu_2)$ as defined in Section 2.3. Suppose that the sampling period of the system is T. For time lags $|\tau| > T$, the dither values are statistically independent so that

$$\begin{aligned} p_{\nu_1,\nu_2}(\nu_1,\nu_2) &= p_{\nu_1}(\nu_1)p_{\nu_2}(\nu_2) \\ &= [\Pi_\Delta \star \Pi_\Delta](\nu_1)[\Pi_\Delta \star \Pi_\Delta](\nu_2) \end{aligned}$$

and

$$P_{\nu_1,\nu_2}(u_1,u_2) = \mathrm{sinc}^2(u_1)\mathrm{sinc}^2(u_2).$$

The non-trivial cases are those for $\tau = \pm T$. Consider two successive dither values (i.e., $\tau = T$):

$$\begin{aligned} \nu_1 &= \eta_1 - \eta_0 \\ \nu_2 &= \eta_2 - \eta_1. \end{aligned}$$

Then

$$\begin{aligned} & p_{\nu_1,\nu_2,\eta_0,\eta_1,\eta_2}(\nu_1,\nu_2,\eta_0,\eta_1,\eta_2) \\ &= p_{\nu_1|(\nu_2,\eta_0,\eta_1,\eta_2)}(\nu_1,\nu_2,\eta_0,\eta_1,\eta_2)p_{\nu_2|\eta_0,\eta_1,\eta_2}(\nu_2,\eta_0,\eta_1,\eta_2)p_{\eta_0,\eta_1,\eta_2}(\eta_0,\eta_1,\eta_2) \\ &= \delta(\nu_1-\eta_1+\eta_0)\delta(\nu_2-\eta_2+\eta_1)p_{\eta_0}(\eta_0)p_{\eta_1}(\eta_1)p_{\eta_2}(\eta_2). \end{aligned}$$

Taking the Fourier transform of this expression with respect to all variables present yields

$$\begin{aligned} & P_{\nu_1,\nu_2,\eta_0,\eta_1,\eta_2}(u_1,u_2,w_0,w_1,w_2) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-j2\pi u_1(\eta_1-\eta_0)}e^{-j2\pi u_2(\eta_2-\eta_1)}p_{\eta_0}(\eta_0)p_{\eta_1}(\eta_1)p_{\eta_2}(\eta_2) \\ &\qquad \times e^{-j2\pi(w_0\eta_0+w_1\eta_1+w_2\eta_2)}d\eta_0 d\eta_1 d\eta_2 \\ &= P_{\eta_0}(w_0-u_1)P_{\eta_1}(w_1+u_1-u_2)P_{\eta_2}(w_2+u_2). \end{aligned}$$

Desired marginal cf's can be obtained from a given joint cf by simply setting the unwanted variables to zero, since

$$P_{x,y}(u,0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-j2\pi(xu+y(0))}p_{x,y}(x,y)dxdy = P_x(u).$$

Thus we have

$$P_{\nu_1,\nu_2}(u_1,u_2) = P_{\eta_0}(-u_1)P_{\eta_1}(u_1-u_2)P_{\eta_2}(u_2).$$

Proceeding similarly for the case of $\tau = -T$, we find that

$$P_{\nu_1,\nu_2}(u_1,u_2) = P_{\eta_0}(-u_2)P_{\eta_1}(u_2-u_1)P_{\eta_2}(u_1).$$

For our purposes,

$$P_{\eta_0}(u) = P_{\eta_1}(u) = P_{\eta_2}(u) = \text{sinc}\,(u)$$

so that for both cases ($\tau = \pm T$):

$$P_{\nu_1,\nu_2}(u_1,u_2) = \text{sinc}\,(u_2-u_1)\text{sinc}\,(u_1)\text{sinc}\,(u_2).$$

Finally, using $\tau = kT$, $k \in \mathbf{Z}$, we can write that

$$P_{\nu_1,\nu_2}(u_1,u_2;k) = \begin{cases} \text{sinc}\,(u_2-u_1)\text{sinc}\,(u_1)\text{sinc}\,(u_2), & k = \pm 1, \\ \text{sinc}^2\,(u_1)\text{sinc}^2\,(u_2), & |k| > 1. \end{cases} \tag{44}$$

It is straightforward to check that this joint cf satisfies all three conditions of Theorem 4.

Using Eq. (44) and the knowledge that TPDF dither has a variance of $\Delta^2/6$, we find using Eq. (20) that the autocorrelation function of the dither under consideration is

$$E[\nu_1\nu_2](k) = \frac{\Delta^2}{6} \times \begin{cases} 1, & k = 0, \\ -\frac{1}{2}, & k = \pm 1, \\ 0, & \text{otherwise.} \end{cases}$$

This corresponds to a simple high-pass power spectral density

$$PSD_\nu(f) = \frac{\Delta^2 T}{3}[1 - \cos(2\pi f T)],$$

where the frequency variable, f, is in units of hertz if T is in seconds. Then, according to Eq. (32),

$$E[\varepsilon_1 \varepsilon_2](k) = \frac{\Delta^2}{4} \times \begin{cases} 1, & k = 0, \\ -\frac{1}{3}, & k = \pm 1, \\ 0, & \text{otherwise}, \end{cases}$$

which corresponds to a power spectral density of

$$PSD_\varepsilon(f) = \frac{\Delta^2 T}{6}[3 - 2\cos(2\pi f T)].$$

This is simply the high-pass spectrum of the dither, plus a white "quantization noise" component of $\Delta^2 T/6$ (which has a total power of $\Delta^2/12$ up to the Nyquist frequency, $\frac{1}{2T}$) in agreement with Eq (33).

Of course, it is possible to imagine many other spectrally shaped dither signals. The properties of such signals have now been investigated in detail. In particular, there is the following theorem, proven and extensively illustrated in [25, 27]:

Theorem 7 *In an NSD quantizing system using dither of the form*

$$\nu_n = \sum_{i=-\infty}^{\infty} c_i \eta_{n-i}$$

where η is an iid nRPDF random process, the total error will be wide-sense stationary and independent of the system input with a PSD given by

$$\mathrm{PSD}_\varepsilon(f) = \mathrm{PSD}_\nu(f) + \frac{\Delta^2 T}{6}$$

under the following conditions:

1. *for each* $\ell \in \mathbf{Z}_0$ *there exists an* i *such that of* c_i *and* $c_{i+\ell}$ *one is zero and the other is a non-zero integer,*

and

2. *either* η *is nRPDF with* $n \geq 1$ *and there exist at least two distinct values of* i *such that* c_i *is a non-zero integer, or* η *is nRPDF with* $n \geq 2$ *and there exists at least one value of* i *such that* c_i *is a non-zero integer.*

In particular, simple high-pass TPDF dither satisfies the above conditions.

4 Digital Dither

Some comment is required concerning the special nature of *re*quantization operations, in which the binary wordlength of data is reduced prior to its storage or transmission. This operation takes place entirely within the digital domain, so that both the input and dither signals are discrete valued due to the finite wordlengths available in practical digital systems. The continuous pdf's discussed thus far are unattainable in a purely digital scheme so that the properties of true digital dither signals require further investigation.

The following discussion represents a theoretical complement to empirical results presented in [16]. It is not intended to be exhaustive, but merely to demonstrate that there is no great difficulty in extending the results obtained for analog systems to digital ones, and to illustrate how this may be done. In particular, the discussion will be restricted to a treatment of first-order statistics with the extension to second-order being straightforward.

Consider a quantizing system which applies digital dither to digital data before removing its L least significant bits. We will use δ to denote the magnitude of an LSB

of the higher-precision signal which is to be requantized, and

$$\Delta = 2^L \delta$$

for an LSB of the requantized output.

Let us consider the following digital dither pdf

$$p_\nu(\nu) = \delta \tilde{p}_\nu(\nu) W_\delta(\nu), \tag{45}$$

where $\tilde{p}_\nu(\nu)$ represents an absolutely integrable function which serves as a "weighting" for the impulse train. $\tilde{p}_\nu$ is assumed to be normalized such that

$$\int_{-\infty}^{\infty} p_\nu(\nu) d\nu = \delta \sum_{\ell=-\infty}^{\infty} \tilde{p}_\nu(\ell\delta) = 1.$$

For instance, $\tilde{p}_\nu$ might be the pdf of a dither of order n, such as an nRPDF dither, in which case it is straightforward to show using Poisson's summation formula [35] that $\tilde{p}_\nu$ has the above normalization. In general, however, $\tilde{p}_\nu$ need not correspond to a pdf since it need not subtend unit area.

Taking the Fourier transform of Eq. (45) we find that

$$\begin{aligned} P_\nu(u) &= \left[\tilde{P}_\nu \star W_{\frac{1}{\delta}}\right](u) \\ &= \sum_{\ell=-\infty}^{\infty} \tilde{P}_\nu\left(u - \frac{\ell}{\delta}\right) \end{aligned} \tag{46}$$

where $\tilde{P}_\nu(u)$ is the Fourier transform of $\tilde{p}_\nu(\nu)$. Note that even if $\tilde{P}_\nu$ satisfies the conditions of Corollary 2 (for some m), P_ν will not, due to the modulation of $\tilde{P}_\nu(u)$ by the impulse train $W_{\frac{1}{\delta}}(u)$. Fortunately, we do not require that these conditions be satisfied in a digital system, since the requirement that $E[\varepsilon^m|x]$ be constant for *all* values of the system input is not of interest. Instead, we require only that the moments be constant for a subset of all conceivable x values, namely $\{x|x = n\delta, n \in \mathbf{Z}\}$, which

includes all values that are representable in the digital system. Thus we assume that the pdf of the system input can be expressed in the form

$$p_x(x) = \delta \tilde{p}_x(x) W_\delta(x) \tag{47}$$

where $\tilde{p}_x$ is a continuous function normalized such that the integral of Eq. (47) is unity. Then

$$\begin{aligned} P_x(u) &= [\tilde{P}_x \star W_{\frac{1}{\delta}}](u) \\ &= \sum_{\ell=-\infty}^{\infty} \tilde{P}_x \left(u - \frac{\ell}{\delta} \right). \end{aligned} \tag{48}$$

Now, from Eq. (12), we have

$$\begin{aligned} G_\nu(u) &\triangleq \frac{\sin(\pi \Delta u)}{\pi \Delta u} P_\nu(u) \\ &= \frac{\sin(\pi \Delta u)}{\pi \Delta u} \sum_{k=-\infty}^{\infty} \tilde{P}_\nu \left(u - \frac{k}{\delta} \right). \end{aligned} \tag{49}$$

Then, from Eq. (9),

$$P_\varepsilon(u) = \sum_{k=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} G_\nu \left(u - \frac{k}{\Delta} \right) \tilde{P}_x \left(-\frac{k + 2^L \ell}{\Delta} \right)$$

so that

$$\begin{aligned} E[\varepsilon^m] &= \left(\frac{j}{2\pi} \right)^m P_\varepsilon^{(m)}(0) \\ &= \left(\frac{j}{2\pi} \right)^m \sum_{k=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} G_\nu^{(m)} \left(-\frac{k}{\Delta} \right) \tilde{P}_x \left(-\frac{k + 2^L \ell}{\Delta} \right). \end{aligned} \tag{50}$$

The only way that this quantity can be independent of $\tilde{P}_x$ is if we require that

$$G_\nu^{(m)} \left(\frac{k}{\Delta} \right) = 0$$

$$\text{for all } k \in \mathbf{Z} \text{ except possibly for those values of } k \text{ such that } \frac{k}{2^L} \in \mathbf{Z}. \tag{51}$$

That is, the indicated derivative must vanish for all integral values of k except those which are integral multiples of 2^L, the value of this derivative being immaterial in the latter cases. In order to see that this is so, note that if a dither is chosen such that Eq. (51) holds then many terms vanish from Eq. (50), leaving

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m \sum_{k=-\infty}^{\infty} G_\nu^{(m)}\left(\frac{k}{\delta}\right) \sum_{\ell=-\infty}^{\infty} \tilde{P}_x\left(\frac{\ell}{\delta}\right).$$

Now, from Eq. (48) we know that

$$P_x(0) = \sum_{\ell=-\infty}^{\infty} \tilde{P}_x\left(\frac{\ell}{\delta}\right) = 1.$$

This leaves

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m \sum_{k=-\infty}^{\infty} G_\nu^{(m)}\left(\frac{k}{\delta}\right), \tag{52}$$

which does not depend on the input distribution. The necessity of Eq. (51) follows from the arbitrariness of $\tilde{P}_x$ (apart from its normalization). Furthermore, by inspection Eq. (52) is precisely the m-th moment of a notional random variable with pdf

$$\left[\frac{\Delta}{2^L}\Pi_\Delta \star p_\nu\right](\varepsilon) W_\delta(\varepsilon),$$

although this is not, of course, the pdf of ε. Some algebraic manipulation of this expression, exploiting the discrete-valued character of ν, reveals that it is equivalent to the following:

$$\left[\frac{\Delta}{2^L}\Pi_\Delta \cdot W_\delta\right](\varepsilon) \star p_\nu(\varepsilon). \tag{53}$$

This may be regarded as the pdf of a notional random variable which is the sum of the dither and an independent discrete-valued "quantization noise".

Note also that in the limit as $\delta \to 0$ (i.e., as $L \to \infty$) Eq. (51) becomes Eq. (13), the condition of Theorem 2 for analog systems.

Returning to Eq. (49) and differentiating, we have

$$\frac{d^m G_\nu}{du^m}(u) = \sum_{k=-\infty}^{\infty} \sum_{r=0}^{m} \binom{m}{r} \frac{d^r}{du^r} \left[\frac{\sin(\pi \Delta u)}{\pi \Delta u} \right] \tilde{P}_\nu^{(m-r)} \left(u - \frac{k}{\delta} \right). \tag{54}$$

If $\tilde{P}_\nu$ meets the conditions of Corollary 2, then all terms in Eq. (54) involving the derivatives of $\tilde{P}_\nu$ go to zero at the places required by Eq. (51) except for the single $(r = 0)$ term involving the m-th derivative. Fortunately, this term involves the zeroth derivative of the leading sinc function, which goes to zero at all the required places. This yields the following theorem:

Theorem 8 *For a digital NSD system in which requantization is used to remove the L least significant bits of binary data, $E[\varepsilon^\ell]$ is independent of the input distribution for $\ell = 1, 2, \ldots, m$, if a non-subtractive digital dither (with the same precision as the input data) is applied for which*

$$\tilde{P}_\nu^{(i)} \left(\frac{k}{\Delta} \right) = 0$$

$$\forall k \in \mathbf{Z}_0 \quad \textit{and} \quad i = 0, 1, 2, \ldots, m-1.$$

This theorem is the digital counterpart of Corollary 2. It is interesting to note that no such counterpart exists for Theorem 2 in terms of $\tilde{P}_\nu$.

We observe that using a dither of higher precision than the input signal is of no benefit. For instance, a dither cf which satisfies the conditions of Eq. (51) with $m = 1$ for $L = 8$ will also satisfy them for $L = 4$, but for a quantizing system in which the precision is reduced by only four bits there is no advantage associated with this cf over one which only satisfies the conditions for $L = 4$.

Frequently, dithers in digital systems will be given a 2's-complement [31] representation and thus will exhibit a mean which differs slightly from zero. This will be

reflected in the appearance of a small non-zero mean error which, of course, will be input independent if an appropriate dither pdf has been chosen.

To express the moments of the system output we impose the conditions of Theorem 8 upon Eq. (35), obtaining

$$\begin{aligned} E[y^m] &= \sum_{r=0}^{m} \binom{m}{r} \sum_{k=-\infty}^{\infty} \left[\left(\frac{j}{2\pi} \right)^r G_\nu^{(r)} \left(\frac{k}{\delta} \right) \right] \left[\left(\frac{j}{2\pi} \right)^{m-r} P_x^{(m-r)} \left(\frac{k}{\delta} \right) \right] \\ &= \sum_{r=0}^{m} \binom{m}{r} E[\varepsilon^r] E[x^{m-r}], \end{aligned}$$

where we have observed from Eq. (48) that $P_x(u)$ is periodic with period $1/\delta$ so that for any $k \in \mathbf{Z}$

$$\left(\frac{j}{2\pi} \right)^{m-r} P_x^{(m-r)} \left(\frac{k}{\delta} \right) = \left(\frac{j}{2\pi} \right)^{m-r} P_x^{(m-r)}(0) = E[x^{m-r}].$$

$E[\varepsilon^r]$ is given by Eq. (52).

The treatment presented above is most appropriate to dithers generated entirely in the digital domain using, for instance, pseudo-random number generation algorithms. In particular, we have shown that whenever the weighting function $\tilde{p}_\nu$ corresponds to the pdf of an analog nRPDF dither the associated digital dither with pdf given by Eq. (45) shares the beneficial properties of its analog counterpart.

In the case where a digital dither signal is generated by fine quantization of an analog dither signal, the details of the derivation change only slightly. The forms of the theorems, however, remain the same, with $\tilde{P}_\nu$ representing the cf of the analog signal. This can be seen directly using Eq. (34) (with null dither), for the pdf of the digital dither generated by quantization will be

$$p_\nu(\nu) = [\delta \Pi_\delta \star \tilde{p}_\nu](\nu) W_\delta(\nu)$$

with cf

$$P_\nu(u) = \left[\frac{\sin(\pi \delta u)}{\pi \delta u} \tilde{P}_\nu(u) \right] \star W_{\frac{1}{\delta}}(u).$$

This expression should be compared with Eq. (46). Note that if $\tilde{P}_\nu$ satisfies the conditions of the theorems, then so will the quantity

$$\frac{\sin(\pi\delta u)}{\pi\delta u}\tilde{P}_\nu(u).$$

Thus far, the behavior of the quantizer at step edges (i.e., when $w = \frac{2k-1}{2}\Delta$, $k \in \mathbf{Z}$) has not been explicitly considered. This is not a problem if the signals in question are continuous-valued. In this case the addition of dither will ensure that the quantizer input resides at a quantizer-step edge with zero probability. On the other hand, if digital signals are in use, the probability that the quantizer input resides at a step edge is always greater than zero. In this instance it makes a considerable difference to the quantizer output (and total error) whether the quantizer rounds up, down, or stochastically (up or down with equal probability) at these edges. Technically, it can be shown [25] that the above formalism yields correct predictions if a stochastic quantizer is used.

The extension of the results to deterministic (i.e., non-stochastic) quantizers employs a simple trick. Consider, for instance, the consequences of choosing a quantizer which always rounds up at step edges (a similar argument applies to quantizers which round down). We note that if a (dc) *virtual offset* τ such that $0 < \tau < \delta$ is introduced into the dither signal, the quantizer output is unaffected except that quantizer inputs residing at step edges are consistently rounded up. We can thus analyze digitally dithered systems with deterministic requantizers using such a notional dc offset, which is a purely mathematical device without physical counterpart. It can be shown [25] that Theorem 8 holds precisely as before. Eq. (52) holds if the virtually offset dither pdf

$$p_\nu(\nu) = \delta\tilde{p}_\nu(\nu - \tau)W_\delta(\nu - \tau)$$

is used in the calculations. In this case, the expression (53) becomes

$$\left[\frac{\Delta}{2^L}\Pi_\Delta(\varepsilon-\tau)W_\delta(\varepsilon)\right]\star p_\nu(\varepsilon+\tau).$$

5 Conclusions

The following conclusions bear repeating:

1. Non-subtractive dithering, unlike subtractive dithering, cannot render the total error statistically independent of the system input, but it *can* render any desired conditional moments of the total error independent of the input distribution provided that certain conditions on the cf of the dither are met (see Theorems 1 and 2). In particular, a nRPDF dither will render the first n moments of the total error input independent.

2. Non-subtractive dithering, unlike subtractive dithering, cannot render total error values separated in time statistically independent of one another. It can, however, regulate the joint moments of such errors and, in particular, it can render the power spectrum of the total error signal equal to the power spectrum of the dither signal plus a white "quantization noise" component (see Theorem 4 and Eq. (33)).

3. Non-subtractive dithering can render any desired moments of the system input recoverable from those of the system output, provided that the statistical attributes of the dither are properly chosen (see Section 2.4). This includes joint moments of system inputs separated in time, so that the spectrum of the input can be recovered from the spectrum of the output.

4. Proper non-subtractive dithering always results in a total error variance greater than $\Delta^2/12$ (see Eq. (15) and Theorem 6).

It is also worth noting that, since the dither is simply an additive signal which is independent of the system input, we are free to add it at *any* time prior to (re)quantization. In particular, once a signal is properly dithered, other signals (which are statistically independent of the dither) may be added to it and the resulting total signal will still be properly dithered for (re)quantization purposes.

For audio signal processing purposes, there seems to be little point in rendering any error moments other than the first and second independent of the input. Variations in higher moments are believed to be inaudible and this has been corroborated by a large number of psycho-acoustic tests conducted by the authors and others [13, 21]. These tests involved listening to a large variety of signals (sinusoids, sinusoidal chirps, slow ramps, various periodically switched inputs, piano and orchestral music, etc.) which had been very coarsely requantized (from 16 bits to 8 bits) in order to render the requantization error essentially independent of low-level non-linearities in the digital-to-analog conversion system used for listening purposes. In addition, the corresponding error signals (output minus input) were used in listening tests in order to check for any vestiges of audible dependence on the input. Using undithered quantizers resulted in clearly audible distortion and noise modulation in the output and error signals. Rectangular-pdf dither of 1 LSB peak-to-peak amplitude eliminated all distortion, but the residual noise level was found to vary audibly in an input-dependent fashion. When triangular-pdf dither of 2 LSB peak-to-peak amplitude (either white or high-pass) was employed, no instance was found in which the error was audibly distinguishable from a steady random noise entirely unrelated to the input. Admittedly, these tests were informal, and there remains a need for formal psycho-acoustic tests of this sort involving many participants under carefully controlled conditions.

We recommend the use of spectrally-white triangular-pdf (TPDF) dither of 2 LSB peak-to-peak amplitude for most audio applications requiring non-subtractively dithered

multi-bit quantization or requantization operations, since this type of dither renders the first and second moments of the total error signal constant with respect to the system input while incurring the minimum increase in error variance. This kind of dither is easy to generate for digital requantization by simply summing two independent rectangular-pdf (RPDF) pseudo-random processes, each of 1 LSB peak-to-peak amplitude, which may easily be generated using a linear congruential algorithm [21, 36]. The resulting digital dither can be used to feed a digital-to-analog converter for analog dithering applications. It should be noted, however, that many analog signals and digital conversion systems exhibit a Gaussian noise component which is of large enough amplitude to act as a satisfactory dither without the requirement of an explicit dithering operation [7, 17].

High-pass TPDF dither is of interest for audio processing since it yields a total error which is *audibly* quieter than that associated with iid TPDF dither. It also can be generated with greater computational efficiency, since only one new pseudo-random number needs to be calculated per sampling period instead of two. Other spectrally-shaped dithers can also be used [24, 25]. The use of spectrally-shaped dither will usually be superseded, however, by the powerful technique of *noise shaping* in applications where the total audibility of the error signal needs to be reduced [37, 38]. This technique employs error feedback in order to spectrally shape the total error of a quantizing system, including the white component arising from a properly dithered quantization. The necessity of and criteria for proper dithering of such systems have now been explored in some detail [27, 24].

With regard to video signals, at least the first two moments of the total error should be rendered independent of the input by the use of an appropriate dither. Some evidence exists that input-dependent variations in the third moment of the total error can be perceptually significant in some video signals [13], but the effects of such

variations are probably not noticeable in most cases.

For signal measurement and statistical signal analysis applications in which signal moments are being measured, appropriate dither (possibly of a quite high order) must be used to render the input signal statistics correctly determinable from the statistics of the quantized output, in accordance with Eqs. (36) and (38).

Some of the results obtained above for multi-bit systems may be applied to 1-bit quantizers with certain caveats. For instance, a 1-bit quantizing system using a full-scale RPDF dither signal will exhibit a total error signal which is zero-mean and spectrally white whenever its peak input value always remains less than the peak dither amplitude. In this instance, the 1-bit quantizer may be regarded as a RPDF-dithered multi-bit mid-riser quantizer whose peak input amplitude is restricted to less than $\Delta/2$, so that the analysis presented above applies without modification. Granted, the amount of dither required is large, but this may be acceptable when no distortion is tolerable and the oversampling ratio is high. A sigma-delta converter topology may be employed to move the noise power out-of-band so long as the dither is iid (in order to avoid statistical dependences between the dither and input which would otherwise be introduced by the error feedback [24, 27]). The potential benefits of using lower dither amplitudes in such systems have been explored in [39].

We maintain that the use of appropriate dither prior to (re)quantization is as fundamental as the use of an appropriate anti-aliasing filter prior to sampling—both serve to eliminate classes of signal-dependent errors. In each case this is accomplished by protecting the system from inputs which, if left unmodified, may introduce such errors.

6 Acknowledgments

Stanley P. Lipshitz and John Vanderkooy have been supported by operating grants from the Natural Sciences and Engineering Research Council of Canada.

References

[1] Roberts, L.G., "Picture Coding Using Pseudo-Random Noise," *IRE Trans. Inform. Theory*, vol. IT-8, pp. 145–154, (1962 Feb.).

[2] Jayant, N.S. and L.R. Rabiner, "The Application of Dither to the Quantization of Speech Signals," *Bell Syst. Tech. J.*, vol. 51, pp. 1293–1304 (1972 July–Aug.).

[3] Schuchman, L., "Dither Signals and Their Effect on Quantization Noise," *IEEE Trans. Commun. Technol.*, vol. COM-12, pp. 162–165 (1964 Dec.).

[4] Sherwood, D.T., "Some Theorems on Quantization and An Example Using Dither," *Conference Record, 19th Asilomar Conference on Circuits, Systems, and Computers*, Pacific Grove, CA (1985 Nov).

[5] Widrow, B., "A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory," Ph.D Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA, (1956 Jun.).

[6] Widrow, B., "A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory," *IRE Trans. Circuit Theory*, vol. PGCT-3, no. 4, pp. 266–276, (1956 Dec.).

[7] Widrow, B., "Statistical Analysis of Amplitude-Quantized Sampled-Data Systems," *Trans. Amer. Inst. Elec. Eng.*, Pt. II, Applications and Industry, vol. 79, pp.555–568 (1961 Jan.).

[8] Sripad, A.B., 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 (1977 Oct.).

[9] Gray, R.N., "Quantization Noise Spectra," *IEEE Trans. Info. Theory.,* pp. 1220–1244 (1990 Nov.).

[10] Wright, J.N., unpublished manuscripts (1979 Jun.–Aug.).

[11] Wright, J.N, private communication, (1991 Apr.).

[12] Stockham, T.G., private communication (1988).

[13] Brinton, L.K., "Nonsubtractive Dither," M.Sc. Thesis, Dept. of Elec. Eng., Univ. of Utah, Salt Lake City, UT (1984 Aug.).

[14] Gray, R.M., and T.G. Stockham, "Dithered Quantizers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 805–811 (1993 May).

[15] Vanderkooy, J., and S.P. Lipshitz, "Resolution Below the Least Significant Bit in Digital Systems with Dither," *J. Audio Eng. Soc.*, vol. 32, pp. 106–113 (1984 Mar.); correction ibid., p. 889 (1984 Nov.).

[16] Lipshitz, S.P., and J. Vanderkooy, "Digital Dither," presented at the 81st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 34, p. 1030 (1986 Dec.), preprint 2412.

[17] Vanderkooy, J., and S.P. Lipshitz, "Dither in Digital Audio," *J. Audio Eng. Soc.*, vol. 35, pp. 966–975 (1987 Dec.).

[18] Vanderkooy, J., and S.P. Lipshitz, "Digital Dither: Signal Processing with Resolution Far Below the Least Significant Bit," *Proc. of the AES 7th International Conference: Audio in Digital Times*, Toronto, Canada, pp. 87–96 (1989 May).

[19] Lipshitz, S.P., and J. Vanderkooy, "High-Pass Dither," presented at the 4th Regional Convention of the Audio Engineering Society, Tokyo (1989 Jun.); in *Collected Preprints* (AES Japan Section, Tokyo, 1989), pp. 72–75.

[20] Wannamaker, R.A., S.P. Lipshitz and J. Vanderkooy, "Dithering to Eliminate Quantization Distortion," *Proc. Annual Meeting Can. Acoustical Assoc.*, Halifax, NS, Canada, pp. 78–86 (1989 Oct.).

[21] Wannamaker, R.A., "Dither and Noise Shaping in Audio Applications," M.Sc. Thesis, Dept. of Physics, Univ. of Waterloo, Waterloo, ON, Canada, (1990 Dec.).

[22] Lipshitz, S.P., R.A. Wannamaker, J. Vanderkooy, and J.N. Wright, "Non-Subtractive Dither," *Proc. of the 1991 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* New Paltz, NY (1991 Oct.), Paper No. 6.2.

[23] Wannamaker, R.A., and S.P. Lipshitz, "Time Domain Behavior of Dithered Quantizers," presented at the 93rd Convention of the Audio Engineering Society, San Francisco, CA (1992 Oct.), preprint 3418.

[24] Wannamaker, R.A., "Subtractive and Non-Subtractive Dithering: A Comparative Analysis," presented at the 97th Convention of the Audio Engineering Society, San Francisco, CA (1994 Nov.), preprint 3920.

[25] Wannamaker, R.A., "The Mathematical Theory of Dithered Quantization," Ph.D. Thesis, Dept. of Applied Mathematics, Univ. of Waterloo, Waterloo, ON, Canada, (1997 Jun.).

[26] Lipshitz, S.P., R.A. Wannamaker and J. Vanderkooy, "Quantization and Dither: A Theoretical Survey," *J. Audio Eng. Soc.,* vol. 40, pp. 355–375 (1992 May).

[27] Lipshitz, S.P., R.A. Wannamaker, and J. Vanderkooy, "Dithered Noise Shapers and Recursive Digital Filters," presented at the 94th Convention of the Audio Engineering Society, Berlin, Germany (1993 Mar.), preprint 3515.

[28] Wannamaker, R.A., and S.P. Lipshitz, "Dithered Quantizers With and Without Feedback," *Proc. of the 1993 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* New Paltz, NY (1993 Oct.).

[29] Wannamaker, R.A., "Efficient Generation of Multichannel Dither Signals," presented at the 103rd Convention of the Audio Engineering Society, New York, NY (1997 Sept.), preprint 4533.

[30] Gray, R.M., private communication, (1991 Apr.).

[31] Jayant, N.S., and P. Noll, *Digital Coding of Waveforms,* Prentice Hall, Englewood Cliffs, NJ, (1984).

[32] Lukacs, E., *Characteristic Functions,* Charles Griffin & Co., London, UK (1960).

[33] Kawata, T., *Fourier Analysis in Probability Theory,* Academic Press, NY, NY (1972).

[34] Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, 2nd ed., McGraw-Hill, New York, NY, (1984).

[35] Papoulis, A., *The Fourier Integral and Its Applications,* McGraw-Hill, NY, (1962).

[36] Knuth, D., *The Art of Computer Programming,* vol. 2, 2nd ed. Addison-Wesley, Reading, MA (1981).

[37] Lipshitz, S.P., J. Vanderkooy and R.A. Wannamaker, "Minimally Audible Noise Shaping," *J. Audio Eng. Soc.,* vol. 39, pp. 836–852 (1991 Nov.)

[38] Wannamaker, R.A., "Psychoacoustically Optimal Noise Shaping," *J. Audio Eng. Soc.,* vol. 40, pp. 611–620 (1992 July/Aug.).

[39] Norsworthy, S.R., R. Schreier and G.C. Temes, *Delta-Sigma Data Converters: Theory, Design and Simulation,* IEEE Press, Piscataway, NJ (1997).