The Theory of Dithered Quantization

by

Robert Alexander Wannamaker

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Applied Mathematics

Waterloo, Ontario, Canada, 2003

I hereby declare that I am the sole author of this thesis.

I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the University of Waterloo to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

A detailed mathematical model is presented for the analysis of multibit quantizing systems. Dithering is examined as a means for eliminating signal-dependent quantization errors, and subtractive and non-subtractive dithered systems are thoroughly explored within the established theoretical framework. Of primary interest are the statistical interdependences of signals in dithered systems and the spectral properties of the total error produced by such systems.

Regarding dithered systems, many topics of practical interest are explored. These include the use of spectrally shaped dithers, dither in noise-shaping systems, the efficient generation of multi-channel dithers, and the uses of discrete-valued dither signals.

Acknowledgements

I would like to thank my supervisor, Dr. Stanley P. Lipshitz, for giving me the opportunity to pursue this research and for the valuable guidance which he has provided throughout my time with the Audio Research Group. Also due many thanks is Dr. John Vanderkooy for much advice and encouragement. It has been an honour to work with individuals of such creativity and enthusiasm.

Dr. Lipshitz and Dr. Vanderkooy made it possible for me to present certain results of my research at conferences and workshops in various locales and I am grateful for having been given these excellent opportunities to travel and interact with other researchers in my field.

I wish to thank everyone with whom I have worked in the Audio Research Group for making my time in Waterloo stimulating and truly memorable. These illustrious individuals include: John Lalonde, Jeff Critten, Jeffrey Bamford, Scott Norcross, and Brad Gover. Finally, I would also like to thank my colleagues in the Department of Applied Mathematics who have provided timely encouragement and, sometimes, commiseration. Specifically, I have been fortunate in knowing Norm Corbett, Janet Grad, Debbie Maclean and Leslie Sayer.

This work is dedicated
to the memory of my father,
Clifford Isaac Wannamaker
(1924–1981).

Contents

List of Figures

Chapter 1

Introduction: Quantization

Dither and quantization are among the most frequently discussed topics in audio and other fields of signal processing. Dithering techniques are now commonplace in applications where it is necessary to reduce the precision of data prior to storage or transmission. In spite of widespread interest in dither and quantization, a comprehensive theory of their operation did not exist in print prior to the author's published investigations in this area, although certain unsubstantiated results could be found scattered among sundry journals. This thesis attempts to collect all of the significant known theory, to substantially extend it, and to provide rigorous justification for the various "rules of thumb" which have been adopted by the engineering community.

The author's interest in dithered quantization arose with an eye to its use in audio signal processing. Undithered quantization can produce audibly deleterious distortion and noise modulation in audio signals, indicating that the mean and variance of the quantization error signal are signal dependent. It will be seen

that the use of dither can eliminate such input dependences, yielding an audibly preferable error signal which is perceptually equivalent to a signal-independent random noise. Similar results are useful for grey-scale or colour quantization of images, in which at least the first two (and possibly the third) quantization error moments are perceptually meaningful and should be rendered signal independent. Data conversion and measurement instruments such as spectrum analyzers can also make profitable use of dithering when the statistical attributes of input signals need to be precisely deduced from quantized measurements.

1.1 Quantizers and Quantizing Systems

Analogue-to-digital conversion is customarily decomposed into two separate processes: *time sampling* of the input analogue waveform and *amplitude quantization* of the signal values in order that the samples may be represented by binary words of a prescribed length. The order of these two processes is immaterial in theory, although in practice quantization is usually second. The sampling operation incurs no loss of information as long as the input is bandlimited in accordance with the Sampling Theorem [1], but the approximating nature of the quantization operation *always* results in signal degradation. An operation with a similar problem is *requantization*, in which the wordlength of digital data is reduced after processing in order to meet specifications for its storage or transmission. An optimal (re)quantizer is one which minimizes the deleterious effects of the aforementioned signal degradation by converting the signal-dependent artifacts into benign signal-independent ones as far as possible.

Quantization and requantization possess similar "staircase" *transfer character-*

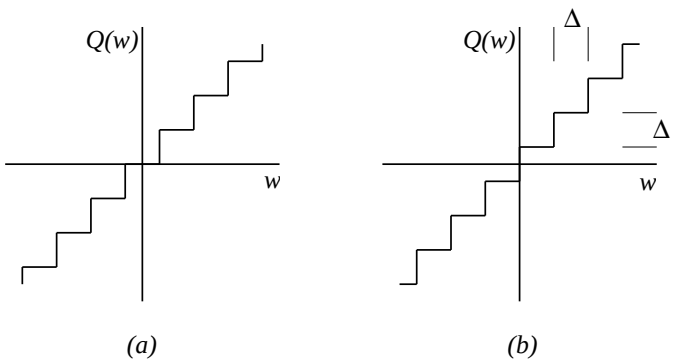


Figure 1.1: Quantizer transfer characteristics: (a) mid-tread, (b) mid-riser. The size of one LSB is denoted by Δ.

istics, which are generally of either the *mid-tread* or *mid-riser* variety illustrated in Fig. 1.1. We will only consider quantizers which are both *uniform,* meaning that all steps in the staircase are of an equal time-invariant size, and *infinite,* which, for practical purposes, means that the input signal is bounded such that it is never *clipped* by saturation of the quantizer. The step size, Δ, is commonly referred to as a *least significant bit (LSB),* since a change in input signal level of one step width corresponds to a change in the LSB of binary-coded output.

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

$$q(w) \stackrel{\triangle}{=} Q(w) - w, \tag{1.1}$$

where we use the symbol $\stackrel{\triangle}{=}$ to indicate equality by definition. This *quantization error* is shown as a function of w for a mid-tread quantizer in Fig. 1.2. It has a maximum magnitude of 0.5 LSB and is periodic in w with a period of 1 LSB.

We will refer to systems which restrict the accuracy of sample values using

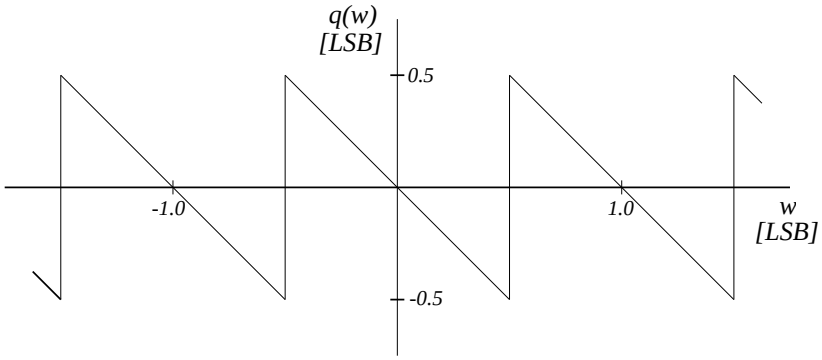


Figure 1.2: Quantization error, $q(w)$, as a function of quantizer input, w, for a mid-tread quantizer.

multi-bit quantization as *quantizing systems*, of which there exist three archetypes: *undithered (UD), subtractively dithered (SD),* and *non-subtractively dithered (NSD).* Schematics of these systems are shown in Fig. 1.3.

Throughout the sequel, we will refer to the *system input* as x, the *system output* as y, and the *total error* of the system as ε where

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

as distinguished from the quantization error, q, defined by Eq. (1.1). In an undithered quantizing system, the system input, x, is identical to the quantizer input, w, so that the total error equals the quantization error; i.e., $\varepsilon = q$. In the other two schemes, the quantizer input is comprised of the system input plus an additive random signal, ν, called *dither*, which is assumed to be stationary[1] and statistically independent of x. In such systems the quantizer input, $w = x + \nu$, is not a deterministic function of x and neither is the total error, ε. In the subtractively dithered

[1]A stationary random process is one whose statistical properties are time-invariant. Such notions from probability and statistics, which are crucial to the analysis of dithered systems, will be systematically introduced in Chapter 2.

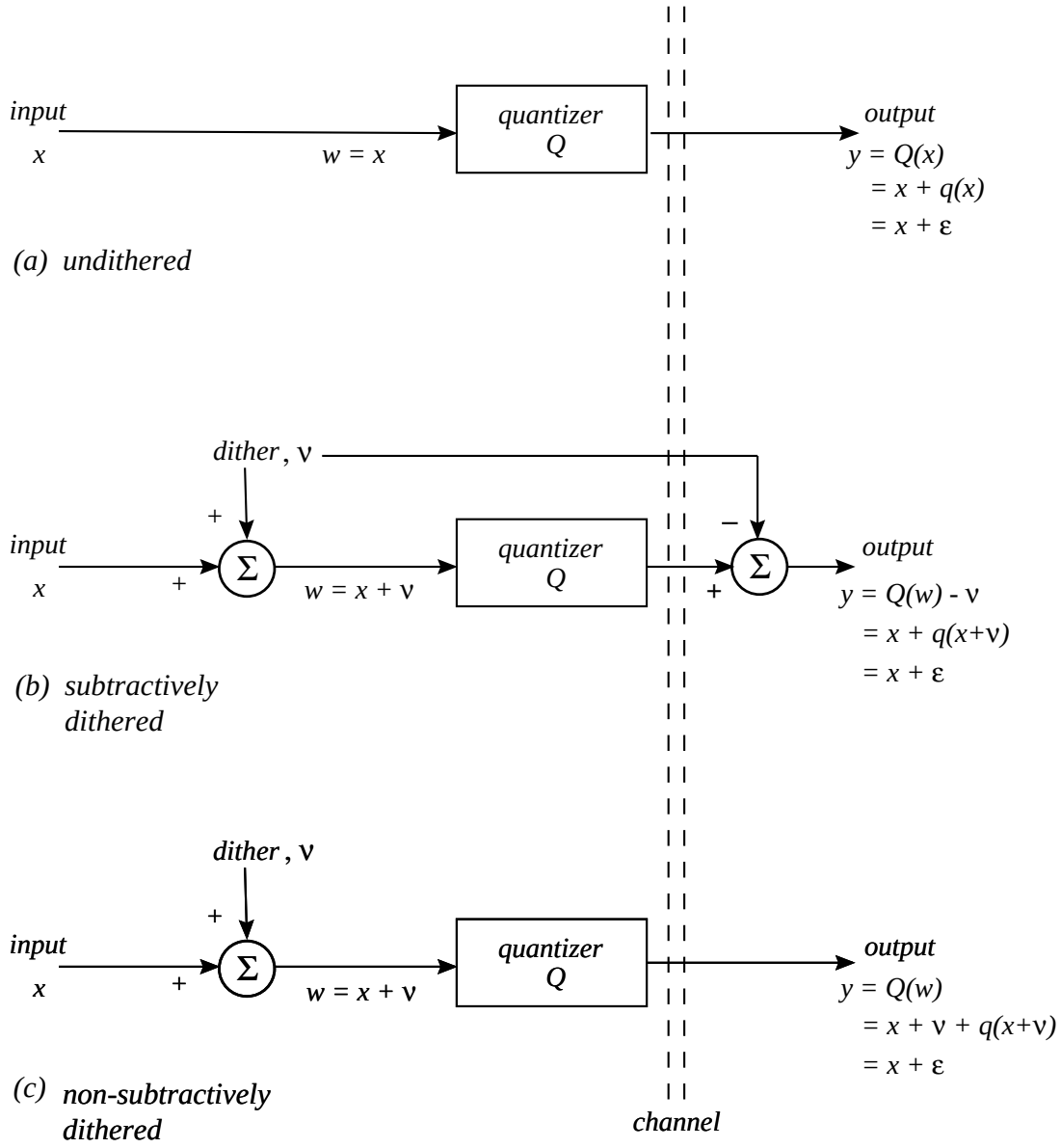


Figure 1.3: Archetypal quantizing systems: (a) undithered (UD), (b) subtractively dithered (SD), (c) non-subtractively dithered (NSD). Shown are the system input, x, the dither signal, ν, the quantizer input, w, and the system output, y.

topology, the dither signal is subtracted from the quantizer output, presumably after this output has been transmitted through some channel. This subtraction operation is omitted in a non-subtractively dithered system.

The objective of dithering is to control the statistical properties of the total error and its relationship to the system input. In undithered systems, we know that the error is a deterministic function of the input. If the input is simple and/or comparable in magnitude to the quantization step size, the total error signal is strongly input-dependent and audible as gross distortion. We shall see that use of dither with proper statistical properties can render the total error signal audibly equivalent to a steady noise floor.

1.2 A Brief History of Quantization Theory

Although citations will occur at appropriate points throughout the text, the formulation will be of a very general sort so that results will not appear in the order in which they were discovered. Hence a concise history of theoretical developments concerning quantization and dither is presented below to provide a contextual framework for the ensuing discussion.

It must be acknowledged that all mathematical treatments of quantization owe a substantial debt to the work of Widrow [2, 3, 4], who developed many of the essential mathematical tools while studying undithered quantizing systems. It was Widrow who first demonstrated the usefulness of characteristic functions in analyzing such systems.

Interest in SD systems arose long before that in non-subtractive schemes. The

original proponent of subtractive dither was Roberts [5], who experimented with it in video applications. It was later adapted for use in speech coding, where the first psychoacoustic evaluations of dithered systems were undertaken [6].

The unidimensional statistics of SD systems were first explored by Schuchman [7], who published conditions on the dither which would ensure uniform distribution of the error signal and its statistical independence of the system input. A more detailed analysis was undertaken by Sripad and Snyder [8], whose work was in turn extended with corrections by Sherwood [9]. Sherwood's paper represents a comprehensive treatment of SD systems (short of discussing noise-shaping error feedback, a technique not yet popular at the time of its writing).

SD systems have resisted widespread implementation due to the requirement that the dither sequence be available for subtraction at playback time, necessitating the storage/transmission of either the sequence itself or enough information to reliably reconstruct it. NSD systems, which avoid this drawback, were first investigated by Wright [10], but his findings were not published until recently [11]. Many of the principal results concerning moments of the error signal were discovered independently by Stockham and Brinton [12, 13], but again nothing was published until lately [14]. Vanderkooy and Lipshitz [15, 16, 17, 18, 19] were the first to make public the primary results regarding NSD systems, and published the first thoroughgoing mathematical treatments with the author [20, 21, 22, 11, 23, 24, 25, 26, 27]. These included the first explorations of the higher-order statistics, including power spectral densities, in such systems, as well as the first analyses of dithered systems with noise-shaping error feedback.

A thorough treatment of the first-order statistics of NSD systems (again short of addressing noise shaping) which uses a different approach has recently been

published by Stockham and Gray [14].

Although a handful of individuals in the engineering community are aware of certain results regarding dither, a number of misconceptions concerning the technique are widespread. In particular, the properties of SD and NSD quantizing systems are often confused. One objective of this thesis is to provide a consistent and rigorous account of the theory of dithered systems in order to promote a more universal understanding of dithering techniques.

The next chapter provides an overview of the mathematical tools to be used in the analysis of quantizing systems. Chapter 3 presents a short but intense development of the crucial theory underlying dithered quantizing systems, using a general approach with UD, SD, and NSD systems as special cases. Chapter 4 examines the distinctive characteristics of each of these systems in detail and makes recommendations for their implementation in specific applications. Chapter 5 examines the related topics of spectrally-shaped dither signals, dither in noise-shaping converters, and the efficient generation of multi-channel dither signals. Chapter 6 extends the theory to cover systems using discrete-valued (i.e., digital) dither signals. Chapter 7 makes some closing comments. Appendix A provides a brief discussion of generalized functions. Issues involving real-time estimation of statistical quantities in dithered quantizing systems are discussed in Appendix B.

Chapter 2

Mathematical Background

This chapter presents a brief introduction to the mathematical devices which will be used later, including stochastic processes and characteristic functions. The reader is assumed to be familiar with Fourier analysis of L_1 (i.e., absolutely integrable) functions. The definition of the Fourier transform maintained throughout the sequel is

$$\mathcal{F}[f](u) = \int_{-\infty}^{\infty} f(x) e^{-j2\pi ux} dx. \tag{2.1}$$

In some cases, ordinary functions will not suit our purposes and we will need to resort to *tempered generalized functions*. (These are sometimes called *tempered distributions* or *Schwartz distributions*, but in the body of the thesis we will eschew this usage in order to avoid confusion with the distinct notion of probability distributions.) In particular we will make frequent use of the Dirac delta function. Readers who are unfamiliar with the theory of generalized functions, but who have some working familiarity with delta functions, may proceed without trepidation. When references to such theory appear, they may be skipped without losing the

flow of the argument. Interested readers may consult Appendix A, which provides an outline of the theory and resolves certain mathematical issues associated with the generalized functions appearing in this thesis.

2.1 Stochastic Processes

Much confusion concerning dither and quantization arises from an unclear or incomplete understanding of the terms in the discussion. With this in mind, a succinct definition of the basic quantities to be discussed is in order. The discussion of probability will use Kolmogorov's axiomatics, as outlined below. For more details, the interested reader may consult [28].

Consider a random experiment with *outcomes* $\zeta \in S$, and a family $\mathcal{B}$ of subsets of S such that

1. $\emptyset \in \mathcal{B}$ and $S \in \mathcal{B}$,
2. $A \in \mathcal{B} \Rightarrow S - A \in \mathcal{B}$,
3. $\{A_n\}_{n=1}^{\infty} \subset \mathcal{B} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{B}$.

We assume that a probability measure P is defined on $\mathcal{B}$; i.e., a real, nonnegative set function P such that

1. $P(S) = 1$,
2. If the sets $A_1, A_2, \ldots$ in $\mathcal{B}$ are mutually disjoint then
$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

The triple $(S, \mathcal{B}, P)$ is called a probability space.

A *(real) random variable*, x, is any mapping

$$x : S \to \mathbf{R}$$

such that $\{\zeta \in S | x(\zeta) < \xi\} \in \mathcal{B}$ for any $\xi \in \mathbf{R}$. With a random variable x one associates a function $F_x : \mathbf{R} \to \mathbf{R}$ defined by

$$F_x(\xi) = P(\{\zeta \in S | x(\zeta) < \xi\}).$$

This function is called the *cumulative distribution function (cdf)* of x. We observe that $F_x(\xi)$ is non-decreasing and that

$$\lim_{\xi \to -\infty} F_x(\xi) = P(\emptyset) = 0, \qquad \lim_{\xi \to +\infty} F_x(\xi) = P(S) = 1.$$

When the cdf is everywhere differentiable, its derivative is called the *probability density function (pdf)* and is denoted by p_x:

$$p_x(x) = \frac{dF_x}{dx}(x).$$

Unfortunately, the cdf is often not differentiable everywhere. It is, however, locally integrable, and thus defines a generalized function (see Appendix A). Since the derivative of a generalized function is always well-defined, the pdf always exists as a generalized function. It can be shown, furthermore [29], that this distribution is defined by

$$\langle \phi, p_x \rangle = \int_{-\infty}^{\infty} \phi(x) dF_x(x) \qquad \forall \phi \in \mathcal{S}$$

where $\mathcal{S}$ is a space of *test functions*. Thus we may either treat pdf's as generalized functions, or eliminate them in favour of Stieltjes integrals.

The following theorem provides a useful characterization of cdf's [28]:

Theorem 2.1 (Lebesgue's Decomposition Theorem) *A cdf $F(x)$ can be written as*

$$F(x) = \alpha_1 F_d(x) + \alpha_2 F_c(x) + \alpha_3 F_s(x)$$

where α_1, α_2 and α_3 are nonnegative real numbers such that

$$\alpha_1 + \alpha_2 + \alpha_3 = 1$$

and $F_d(x)$, $F_c(x)$ and $F_s(x)$ are, respectively, a purely discontinuous cdf, an absolutely continuous cdf, and a singular cdf.

The singular function $F_s(x)$ is a continuous function whose derivative is zero almost everywhere (in Lebesgue measure) and which is not a constant. Such functions do not occur in practice and we will make the common assumption that $\alpha_3 = 0$ for the random quantities under consideration in the sequel. The function $F_c(x)$ possesses a density corresponding to an ordinary function. The purely discontinuous function $F_d(x)$ is constant except on at most a countable set of discontinuities. Thus $F_d(x)$ represents a countable sum of step functions so that the corresponding density is a countable sum of Dirac delta functions (see Appendix A).

We may also speak of the *joint cdf,* of a pair of random variables, x and y, as

$$F_{x,y}(\xi_x, \xi_y) = P(\{\zeta \in S | x(\zeta) < \xi_x \wedge y(\zeta) < \xi_y\}).$$

The corresponding *joint pdf* is

$$p_{x,y}(x, y) = \frac{\partial^2 F_{x,y}}{\partial x \partial y}(x, y)$$

where the derivatives are always meaningful in the sense of generalized functions. Corresponding definitions are possible in the case of more than two random variables.

We say that two random variables x and y are *statistically independent* if it is possible to write

$$F_{x,y}(x,y) = F_x(x)F_y(y)$$

or, equivalently,

$$p_{x,y}(x,y) = p_x(x)p_y(y).$$

The *marginal cdf's,* F_x and F_y are recoverable from $F_{x,y}$ as limits at infinity; for instance

$$\begin{aligned} F_x(\xi_x) &= P(\{\zeta \in S | x(\zeta) < \xi_x\}) \\ &= P(\{\zeta \in S | x(\zeta) < \xi_x \wedge y(\zeta) < \infty\}) \\ &= \lim_{\xi_y \to \infty} F_{x,y}(\xi_x, \xi_y) \end{aligned}$$

or, equivalently,

$$p_x(x) = \int_{-\infty}^{\infty} p_{x,y}(x,y)dy.$$

Also of interest are *conditional pdf's (cpdf's).* Any function $p_{x|y}$ such that

$$p_{x,y}(x,y) = p_{x|y}(x,y)p_y(y)$$

is referred to as a *version of the conditional pdf.* Clearly x and y are statistically independent if and only if $p_{x|y}(x,y) = p_x(x)$. We also observe that if $p_y(y) = \delta(y - y_0)$, $y_0 \in \mathbf{R}$, then

$$\begin{aligned} p_x(x) &= \int_{-\infty}^{\infty} p_{x,y}(x,y)dy \\ &= \int_{-\infty}^{\infty} p_{x|y}(x,y)\delta(y - y_0)dy \\ &= p_{x|y}(x,y_0). \end{aligned}$$

Thus $p_{x|y}(x,y_0)$ may be interpreted as the pdf of x given that y assumes a value $y_0 \in \mathbf{R}$ [30]. Note that if p_x in no way depends on the choice of p_y, then $p_{x|y}(x,y)$ is

a function of x alone so that x and y are statistically independent no matter how y is distributed. Thus the requirement that the pdf of x be unaffected by the choice of pdf for y ensures that these random variables are statistically independent for any choice of p_y.

Now consider a probability space $(S, \mathcal{B}, P)$ and any set $\mathbf{T}$, called a *parameter set*. A collection $\{x(\zeta, t); t \in \mathbf{T}\}$ of random variables on S is called a *stochastic* or *random process*. Usually we will simply refer to this random process as x and we will use the terms *signal* and random process interchangeably. For our purposes t represents a time parameter so that $\mathbf{T}$ is either $\mathbf{R}$ or $\mathbf{Z}$, in which case a random process represents a family of time functions (continuous or discrete, respectively), one for each $\zeta \in S$. Individually these are usually called *sample functions* and may correspond, for instance, to data records from single experimental trials. For a specific time value, t_i, the expression $x(\zeta, t_i)$ represents a quantity dependent on ζ (i.e., a random variable), which we will sometimes denote by x_i for convenience.

We define the pdf $p_x(x, t)$ of a random process x so that $p_x(x, t_i)$ is the pdf of the random variable $x(\zeta, t_i)$. We can also form the *joint pdf* $p_{x_1,x_2}(x_1, x_2, t_1, t_2)$ of the random variables x_1 and x_2 where $t_1 - t_2 \neq 0$. The explicit time dependence of these quantities will often be omitted where it may be understood from the context.

A random process x is said to be *(first-order) stationary in the strict sense* if its pdf is independent of time; i.e., if

$$p_x(x, t) = p_x(x, t + \tau) \qquad \forall \tau \in \mathbf{T}.$$

If

$$p_{x_1,x_2}(x_1, x_2, t_1, t_2) = p_{x_1,x_2}(x_1, x_2, t_1 + \tau, t_2 + \tau) \qquad \forall \tau \in \mathbf{T}$$

whenever $t_1 \neq t_2$ then the process is said to be *second-order stationary in the*

strict sense. If all of the random variables x_i and x_j are identically distributed and statistically independent of one another when $i \neq j$, then the random process is said to be *iid (independent and identically distributed).*

Given p_x, various statistical attributes of the stochastic process can be calculated, including *expected values* of functions of x, where the *expectation value operator* is defined by

$$E[f(x)](t) \stackrel{\triangle}{=} \int_{-\infty}^{\infty} f(w) dF_x(w,t) = \int_{-\infty}^{\infty} f(w) p_x(w,t) dw.$$

This definition extends in an obvious fashion to expectation values of multivariable functions. For these we observe that the expectation value operator is linear in the sense that

$$\begin{aligned} E[f(x)+g(y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [f(x)+g(y)] p_{x,y}(x,y) dx dy \\ &= \int_{-\infty}^{\infty} f(x) p_x(x) dx + \int_{-\infty}^{\infty} g(y) p_y(y) dy \\ &= E[f(x)] + E[g(y)]. \end{aligned}$$

When $E[|x|^k]$ exists, the *k-th moment* of x is defined as:

$$E[x^k](t) \stackrel{\triangle}{=} \int_{-\infty}^{\infty} w^k p_x(w,t) dw.$$

The zeroth moment of any random process is identically equal to unity; i.e.,

$$E[x^0](t) = E[1](t) = 1.$$

The first moment is usually referred to as the *mean* of the process. The term *variance* refers to the quantity

$$E\Big[(x - E[x])^2\Big](t) = E[x^2](t) - E^2[x](t),$$

so that if the mean of a random process is zero then its variance and second moment are equal. We emphasize that, in general, these quantities manifest an explicit time dependence, although hereafter it may be omitted unless explicitly required.

The quantity

$$E\Big[(x_1 - E[x_1])(x_2 - E[x_2])\Big](t_1, t_2) \\ = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1 - E[x_1])(x_2 - E[x_2])p_{x_1,x_2}(x_1, x_2, t_1, t_2)dx_1dx_2$$

is called the *autocovariance function* of the random process x. The *joint moment* $E[x_1x_2](t_1, t_2)$ is called the *autocorrelation function* of the random process, so that if the process has zero mean then its autocovariance and autocorrelation functions are equal. If

$$E[x_1x_2] = E[x_1]E[x_2]$$

then the random variables x_1 and x_2 are said to be *uncorrelated,* and if

$$E[x_1x_2] = 0$$

then they are said to be *orthogonal.* If x_1 and x_2 are statistically independent then they are uncorrelated, and if they are also zero mean then they are orthogonal, in which case

$$E[(x_1 + x_2)^2] = E[x_1^2] + 2E[x_1x_2] + E[x_2^2] = E[x_1^2] + E[x_2^2].$$

A random process is said to be *stationary in the wide sense* if

$$E[x](t) = E[x](0),$$

a constant for all t, and

$$E[x_1x_2](t_1, t_2) = E[x_1x_2](t_1 - t_2, 0)$$

for any t_1, t_2. That is, $E[x_1 x_2](t_1, t_2)$ depends only on the difference of t_1 and t_2. In this case we let $\tau = t_1 - t_2$ and use the notation

$$r_x(\tau) = E[x_1 x_2](\tau).$$

A random process is obviously wide-sense stationary if it is second-order strict-sense stationary, but the converse is not necessarily true. The *power spectral density (PSD)* of a wide-sense stationary random process is defined as the Fourier transform of its autocorrelation function:

$$\mathrm{PSD}_x(u) = \mathcal{F}[r_x](u).$$

When considering a random process in a sampled-data system we will for clarity write $r_x(k), k \in \mathbf{Z}$ instead of $r_x(\tau), \tau \in \mathbf{T}$. Its PSD may be calculated from $r_x(\tau)$ using delta functions at sampling intervals, or by using the *discrete-time Fourier transform (DTFT)* [31]:

$$\mathcal{F}_{\mathrm{DT}}[r_x](u) = 2T \sum_{k=-\infty}^{\infty} r_x(k) e^{-j2\pi kTu}$$

where T is the sampling period of the system. This definition is normalized such that

$$\int_0^{\frac{1}{2T}} \mathcal{F}_{\mathrm{DT}}[r_x](u) du = r_x(0),$$

which is the variance of the random process. The upper limit of integration, $\frac{1}{2T}$, is referred to as the *Nyquist frequency* of the system and is equal to half of the sampling frequency.

2.2 Characteristic Functions

An expectation of particular interest is the so-called *characteristic function* or *cf* of a random variable x:

$$P_x(u) \triangleq E[e^{-j2\pi ux}]$$

where u is a real variable.[1] Thus the cf of a random variable is precisely the Fourier transform of its pdf. We will denote cf's of random variables using upper case P's, while reserving lower case p's for their pdf's.

We observe that the cf always exists since

$$\left| \int_{-\infty}^{\infty} e^{-j2\pi ux} dF_x(x) \right| \leq \int_{-\infty}^{\infty} dF_x(x) = 1.$$

Furthermore we have the following.

Theorem 2.2 *The cf's of two random variables are identical if and only if their pdf's are identical.*

A proof may be found, for instance, in [28] and is simply a uniqueness proof for Fourier transforms. (Alternatively, viewing the quantities involved as generalized functions, we may appeal to the unicity results in Appendix A.) We conclude that the pdf and cf are equivalent descriptions of a random variable.

The characteristic function is a very useful tool in applications. The following theorems indicate some of the reasons why this is so. The first follows directly from the definition of the cf.

[1] Some authors use $P_x(u) = E[e^{j2\pi ux}]$. The choice of definition is a matter of preference since the results only differ by a complex conjugation.

Theorem 2.3 *Two random variables, x and y, are statistically independent if and only if their joint cf can be written as a product:*

$$P_{x,y}(u_x, u_y) = P_x(u_x)P_y(u_y).$$

Theorem 2.4 *Given two random variables x and y,*

$$P_x(u) = P_{x,y}(u, 0).$$

Proof:

$$\begin{aligned} E[e^{-j2\pi(xu_x+yu_y)}]\Big|_{u_y=0} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-j2\pi xu_x} dF_{x,y}(x,y) \\ &= \int_{-\infty}^{\infty} e^{-j2\pi xu_x} dF_x(x) \\ &= E[e^{-j2\pi xu_x}]. \end{aligned}$$

□

Theorem 2.5 *If x and y are two random variables, and $z = ax + by$ is a third where $a, b \in \mathbf{R}$, then*

$$P_{z,x,y}(u_z, u_x, u_y) = P_{x,y}(u_x + au_z, u_y + bu_z).$$

Proof:

$$p_{z|(x,y)}(z,x,y) = \delta(z - ax - by)$$

so

$$p_{z,x,y}(z,x,y) = \delta(z - ax - by)p_{x,y}(x,y).$$

This product of generalized functions (see Appendix A) is a composition of the tensor product $\delta(z)p_{x,y}(x,y)$ with a linear coordinate transformation

$$A \begin{bmatrix} z \\ x \\ y \end{bmatrix}$$

where

$$A = \begin{bmatrix} 1 & -a & -b \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The Fourier transform of the tensor product is $P_{x,y}(u_x, u_y)$, $\det(A) = 1$ and

$$\overline{A^{-1}} = \begin{bmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & 0 & 1 \end{bmatrix}$$

so by Theorem A.4(viii) we obtain the result.

□

We observe that a trivial generalization is allowed: if random variables other than z, x and y appear in the densities, these are unaffected; e.g.,

$$P_{z,x,y,w}(u_z, u_x, u_y, u_w) = P_{x,y,w}(u_x + au_z, u_y + bu_z, u_w).$$

Corollary 2.1 *If x and y are two random variables, and $z = ax + by$ is a third where $a, b \in \mathbf{R}$, then*

$$P_z(u) = P_{x,y}(au, bu).$$

Proof: Apply Theorem 2.4 to the result of Theorem 2.5.

□

If we define the *convolution* of two absolutely integrable functions f and g by

$$[f \star g](x) \stackrel{\triangle}{=} \int_{-\infty}^{\infty} f(x-w)g(w)dw$$

then we also have the following. (This definition can be extended to include appropriate pairs of generalized functions; see Appendix A.)

Corollary 2.2 *If x and y are two statistically independent random variables, and $z = x + y$ is a third, then*

$$P_z(u) = P_x(u)P_y(u)$$

and

$$p_z(x) = [p_x \star p_y](x).$$

Proof: The first equation follows from the previous corollary and Theorem 2.3. Taking the Fourier transform of the second equation and interchanging the order of integration yields the first.

□

Theorem 2.6 *If the first n moments of a random variable x exist, then $P_x(u)$ is n times differentiable and*

$$E[x^k] = \left(\frac{j}{2\pi}\right)^k P^{(k)}(0) \qquad k = 1, 2, \ldots, n. \tag{2.2}$$

Proof: Consider $n = 1$. If $E[|x|]$ exists, then

$$E[xe^{-j2\pi ux}] = \int_{-\infty}^{\infty} xe^{-j2\pi ux} dF(x)$$

converges uniformly in u. Thus

$$P^{(1)}(u) = \int_{-\infty}^{\infty} -j2\pi xe^{-j2\pi ux} dF(x).$$

In particular,

$$P^{(1)}(0) = -j2\pi E[x].$$

The result for higher n follows by iteration of the above procedure.

□

For our purposes we will consider only signals all of whose moments exist, so that the theorem holds for any n. The result is easily extended to yield

$$E[x^m y^n] = \left(\frac{j}{2\pi}\right)^{m+n} P_{x,y}^{(m,n)}(0,0),$$

where we take this opportunity to establish the useful convention

$$\begin{aligned} f^{\alpha}(x) &= f^{(\alpha_1,\alpha_2,\ldots,\alpha_N)}(x_1, x_2, \ldots, x_N) \\ &= \frac{\partial^{|\alpha|} f}{\partial^{\alpha_1} x_1 \partial^{\alpha_2} x_2 \ldots \partial^{\alpha_N} x_n}(x_1, x_2, \ldots, x_N), \end{aligned}$$

where $|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_N$. α is referred to as a *multi-index*.

We will now establish some general properties of characteristic functions which will prove useful in the sequel. The following is but the briefest of samplings, drawn from the extensive surveys in [33, 34].

Theorem 2.7 *The characteristic function, $P(u)$, of a random variable has the following properties for $u \in \mathbf{R}$:*

(i) *$P(u)$ is a uniformly continuous function of u;*

(ii) *$P(0) = 1$;*

(iii) *$|P(u)| \leq 1$;*

(iv) *if there exists $u_0 \neq 0$ such that $|P(u_0)| = 1$, then p is a* lattice distribution

$$p(x) = \sum_{k=-\infty}^{\infty} c_k \delta\left(x - \frac{k+\omega}{u_0}\right)$$

where $\omega \in \left[-\frac{1}{2}, \frac{1}{2}\right)$. If Re $P(u_0) = 1$ then $\omega = 0$.

Proof:

(i)

$$\begin{aligned} |P(u_1) - P(u_2)| &= |E[(e^{-j2\pi(u_1-u_2)x} - 1)e^{-j2\pi u_2 x}]| \\ &\leq E[|e^{-j2\pi(u_1-u_2)x} - 1|] \\ &\to 0 \qquad \text{as } |u_1 - u_2| \to 0. \end{aligned}$$

(ii)

$$E[e^{-j2\pi(0)x}] = E[1] = 1.$$

(iii)

$$|E[e^{-j2\pi ux}]| \leq E[|e^{-j2\pi ux}|] = 1.$$

(iv) There must exist $\omega \in \left[-\frac{1}{2}, \frac{1}{2}\right)$ such that $P(u_0)e^{j2\pi\omega} = 1$; i.e., such that

$$E[e^{-j2\pi u_0 x}]e^{j2\pi\omega} = E[1].$$

(If Re $P(u_0) = 1$ we may take $\omega = 0$.) Taking real parts, this implies that

$$E[1 - \cos(2\pi(u_0 x - \omega))] = 0.$$

The result follows since $1 - \cos(2\pi(u_0 x - \omega)) > 0$ unless $u_0 x - \omega = k \in \mathbf{Z}$.

□

2.3 Definitions Regarding Dithered Quantizing Systems

Fig. 2.1 shows a quantizing system of a generalized sort, with SD and NSD systems representing specific instances of this generalized one. The *system input* is denoted by x, the *system output* by y, the *quantizer input* by w, and the *quantizer output* by w'. The signals q and ε represent the *quantization error* and the *total error* of the system, respectively. ν represents a strict-sense stationary *dither* process, which is usually chosen to be statistically independent of x although this assumption will not be made in the sequel except where it is stated explicitly. The signal ν' can assume one of two forms depending upon the specific type of system under consideration.

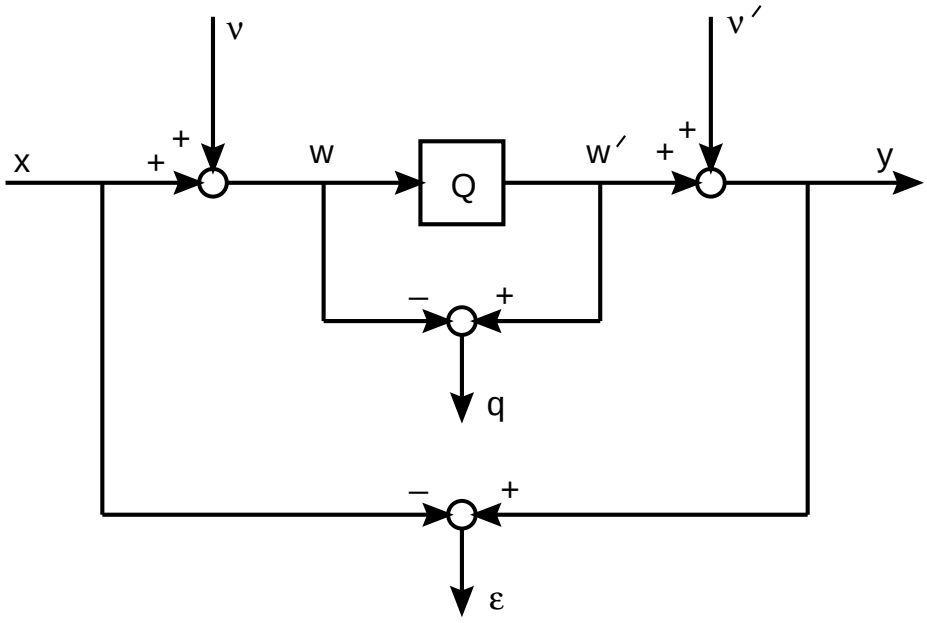


Figure 2.1: Schematic of a generalized dithered quantizing system.

If $\nu' \equiv -\nu$ then the system is SD, whereas if $\nu' \equiv 0$ then an NSD system is under consideration. If $\nu \equiv \nu' \equiv 0$ then the system is undithered.

We assume a uniform infinite quantizer with step size Δ. The corresponding transfer characteristics can be expressed analytically in terms of the input to the quantizer, w, and the quantizer step size, Δ, as

$$Q(w) = \Delta \left\lfloor \frac{w}{\Delta} + \frac{1}{2} \right\rfloor \tag{2.3}$$

for a mid-tread quantizer, or

$$Q(w) = \Delta \left\lfloor \frac{w}{\Delta} \right\rfloor + \frac{\Delta}{2} \tag{2.4}$$

for a mid-riser quantizer, where the "floor" operator, $\lfloor\ \rfloor$, returns the greatest integer less than or equal to its argument. These quantizers always round up at

step edges; i.e., $Q(k\Delta + \frac{1}{2}) = (k+1)\Delta$ for any $k \in \mathbf{Z}$. We could just as easily specify quantizers which round down at step edges, or *stochastic quantizers* which round either up or down at step edges with equal probability. Throughout the sequel, mid-tread stochastic quantizers will be assumed unless otherwise noted. All formulas will, however, possess analogues for mid-riser quantizers and all results stated as theorems are valid for either mid-tread or mid-riser types. We will see that the choice of a stochastic quantizer is the most convenient from a mathematical point of view, as it permits statistical modelling of the quantizer using certain products of generalized functions (see Appendix A). When appropriate, differences between stochastic and deterministic quantizers will be discussed, although these are usually not significant in practice since a dithered analogue signal will reside at the quantizer step edges with probability zero.

It is opportune to introduce a class of dither signals which we will show to have special useful properties. We begin by defining a *uniformly distributed* random process as one with a pdf of the form

$$p(x) = \Pi_\Delta(x), \tag{2.5}$$

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

$$\Pi_\Gamma(x) \stackrel{\triangle}{=} \begin{cases} \dfrac{1}{\Gamma}, & |x| < \dfrac{\Gamma}{2}, \\ \dfrac{1}{2\Gamma}, & |x| = \dfrac{\Gamma}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.6}$$

The pdf of Eq. (2.5) will be referred to as a *uniform* or *RPDF* (for *Rectangular Probability Density Function).* By direct calculation the moments of a uniformly

distributed random process ε are found to be

$$E[\varepsilon] = 0 \tag{2.7}$$

$$E[\varepsilon^2] = \frac{\Delta^2}{12} \tag{2.8}$$

$$E[\varepsilon^m] = \begin{cases} \dfrac{1}{m+1}\left(\dfrac{\Delta}{2}\right)^m, & \text{for } m \text{ even}, \\ 0, & \text{for } m \text{ odd}. \end{cases} \tag{2.9}$$

The cf of a uniformly distributed process is

$$P(u) = \frac{\sin(\pi\Delta u)}{\pi\Delta u}.$$

This function is commonly referred to as a "sinc" function[2], and we will often use the notation

$$\operatorname{sinc}(u) \triangleq \frac{\sin(\pi\Delta u)}{\pi\Delta u}.$$

Now denote by $\mathbf{Z}_0^N$ the space of all ordered N-tuples $(k_1, k_2, \ldots, k_N)$ with integer components with the exclusion of $0 = (0, 0, \ldots, 0)$. Thus, in particular, $\mathbf{Z}_0$ is the set of all integers except zero. Then we will refer to an iid dither whose cf, P_ν, obeys the condition

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

$$\text{for } m = 0, 1, 2, \ldots, n-1, \quad \text{and} \quad \forall k \in \mathbf{Z}_0$$

as a *dither of order n*. We shall see that dithers of this type are normally chosen for use in applications because of their desirable effects on the error signals.

[2]Actually, in much of the literature this function would be called $\operatorname{sinc}(\Delta u)$, but the stated definition is more convenient for our purposes and will be retained in the sequel.

The conditions for a dither to be of order n may also be expressed in terms of its pdf, although these are not as useful from a practical standpoint. Straightforward application of Poisson's summation formula (Theorem A.7) and the derivative property of Fourier transforms (Theorem A.4(v)) reveals that if and only if a dither is of order n then its pdf obeys

$$(x^m p_\nu(x)) \star W_\Delta(x) = E[\nu^m],$$

a constant, for $m = 0, 1, \ldots, n-1$, where have made use of the *impulse train*

$$W_\Delta(x) \stackrel{\triangle}{=} \sum_{k=-\infty}^{\infty} \delta\left(x - k\Delta\right).$$

In particular, for a dither of order greater than or equal to zero we have

$$\sum_{k=-\infty}^{\infty} p_\nu(x - k\Delta) = 1.$$

An example of a dither of order n is the so-called "nRPDF dither" produced by summing n statistically independent uniformly distributed random processes of peak-to-peak amplitude Δ. Summing statistically independent random processes convolves their pdf's, thus multiplying their cf's (see Theorem 2.2). Therefore the cf of an nRPDF dither is

$$P_\nu(u) = \left[\frac{\sin(\pi\Delta u)}{\pi\Delta u}\right]^n.$$

A general formula exists for the pdf of an nRPDF random process [35], and this may be integrated to find a general expression for the moments thereof, but these formulae are unwieldy and not very instructive. For our purposes two observations will suffice: first that all odd moments of nRPDF processes are zero since the pdf's are even, and second that for $n \geq 2$

$$\frac{d^2}{dx^2}[\mathrm{sinc}\,(x)]^n = n(n-1)[\mathrm{sinc}\,(x)]^{n-2}[\mathrm{sinc}^{(1)}\,(x)]^2 + n[\mathrm{sinc}\,(x)]^{n-1}\mathrm{sinc}^{(2)}\,(x)$$

so that

$$E[\nu^2] = \left(\frac{j}{2\pi}\right)^2 P_\nu^{(2)}(0) = n\frac{\Delta^2}{12}.$$

Of course, this is just the sum of the powers of n statistically independent uniformly distributed random processes, as expected.

2RPDF dither, being in common use, is frequently referred to as TPDF (for *Triangular Probability Density Function*), since the convolution of two uniform pdf's is triangular in shape:

$$[\Pi_\Delta \star \Pi_\Delta](\nu) = \begin{cases} \frac{1}{\Delta}\left(1 - \frac{|\nu|}{\Delta}\right), & 0 \le |\nu| < \Delta, \\ 0, & \text{otherwise.} \end{cases}$$

3RPDF dither is sometimes referred to as PPDF (for *Parabolic Probability Density Function)*, since this pdf is piecewise parabolic. We observe that an nRPDF random process has a maximum peak-to-peak amplitude of $n\Delta$ since its pdf is the convolution of n uniform pdf's.

Chapter 3

A General Theory of Dithered Quantization

This chapter presents a general theory of dithered quantization, with undithered (UD), subtractively dithered (SD) and non-subtractively dithered (NSD) systems as special cases to be elucidated later. Included is a thorough analysis of the statistical relationships between the signals indicated in Fig. 2.1. The approach used is to derive the joint cf of all random variables of interest so that the joint cf's of subsets of these variables are easily found by setting unwanted arguments to zero (see Theorem 2.4).

We define the vector

$$x \stackrel{\triangle}{=} (x_1, x_2, x_3, \ldots, x_N) \in \mathbf{R}^N$$

where the components represent N system input values occurring at distinct times. That is, x_1 and x_2, say, represent distinct but not necessarily successive values of

the system input. The following vectors in $\mathbf{R}^N$ are defined in an analogous fashion:

$$\nu, w, w', \nu', y, \varepsilon, q.$$

Corresponding entries in each vector are taken to be simultaneous.

Furthermore we define the vector

$$r \stackrel{\triangle}{=} (q, \varepsilon, y, \nu', w', w, \nu, x) \in \mathbf{R}^{8N}.$$

Taking $N = 1$ corresponds to considering the system at a single instant in time, and the reader should feel free to consider $N = 1$ upon a first reading if this aids in understanding. It turns out that taking $N \geq 1$ does not much complicate the analysis since each signal present in the system at any given time can be expressed algebraically in terms of the signals x and ν present at that time without reference to any later or earlier signal values. Initially we will make no assumptions regarding the statistical relationship between x and ν, since this may be complicated, with signal values at different instants in time affecting one another. (This is the case, for instance, when noise shaping error feedback is present in the system; see Section 5.2.)

Using the definition of conditional probability [30] we have

$$\begin{aligned} p_r(r) \;=\; & p_{q|\varepsilon,y,\nu',w',w,\nu,x}(r) \\ & \times p_{\varepsilon|y,\nu',w',w,\nu,x}(r) \\ & \times p_{y|\nu',w',w,\nu,x}(r) \\ & \times p_{\nu'|w',w,\nu,x}(r) \\ & \times p_{w'|w,\nu,x}(r) \\ & \times p_{w|\nu,x}(r) \\ & \times p_{\nu,x}(r). \end{aligned} \tag{3.1}$$

We will proceed to write down an expression for each quantity in this product.

Since $w \equiv x + \nu$ we have

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

where the delta function with a vector argument is defined as a tensor product of delta functions:

$$\delta(x) \stackrel{\triangle}{=} \prod_{i=1}^{N} \delta(x_i).$$

Consider the special case where $N = 1$ so that all function arguments are scalars and only one instant in time is involved. Quantizer output values are restricted to multiples of Δ, so we can write $p_{w'|w,\nu,x}(r)$ as a product of the impulse train

$$W_\Delta(w') \stackrel{\triangle}{=} \sum_{k=-\infty}^{\infty} \delta\left(w' - k\Delta\right).$$

with an appropriate window function. If a quantizer input value, w, satisfies

$$(2n-1)\Delta/2 < w < (2n+1)\Delta/2$$

for some $n \in \mathbf{Z}$, then the quantizer output value is $n\Delta$. Thus we can use a rectangular window function of width Δ to select the appropriate delta function from $W_\Delta(w')$. In particular, we can write[1]

$$p_{w'|w,\nu,x}(r) = \Delta\Pi_\Delta(w' - w)W_\Delta(w'). \tag{3.2}$$

[1]The astute reader may observe that the case where w falls at a quantizer step edge has been neglected. The indicated product of generalized functions in fact represents the cpdf of a *stochastic* quantizer, as is discussed in detail in Appendix A. For $w = (2n+1)\Delta/2$ the output of a stochastic quantizer is either $n\Delta$ or $(n+1)\Delta$ with equal probability.

Since quantizations occurring at different times have no effect on one another, the treatment is trivially extended to handle $N \geq 1$ by defining the following scalar functions of vector arguments:

$$\Pi_\Delta(x) \stackrel{\triangle}{=} \prod_{i=1}^{N} \Pi_\Delta(x_i)$$

and

$$\begin{aligned} W_\Delta(x) &\stackrel{\triangle}{=} \prod_{i=1}^{N} W_\Delta(x_i) \\ &= \sum_{k \in \mathbf{Z}^N} \delta(x - k\Delta) \end{aligned}$$

where

$$k \stackrel{\triangle}{=} (k_1, k_2, k_3, ..., k_N) \in \mathbf{Z}^N.$$

With these definitions, Eq. (3.2) applies when $N \geq 1$.

Now since

$$\begin{aligned} \nu' &= \begin{cases} 0, & \text{UD systems,} \\ 0, & \text{NSD systems,} \\ -\nu, & \text{SD systems,} \end{cases} \\ q &= w' - w \\ \varepsilon &= y - x \\ y &= \nu' + w' \end{aligned} \tag{3.3}$$

the other conditional pdf's are of the following obvious forms:

$$\begin{aligned}
p_{\nu'|w',w,\nu,x}(r) &= \begin{cases} \delta(\nu'), & \text{UD systems,} \\ \delta(\nu'), & \text{NSD systems,} \\ \delta(\nu'+\nu), & \text{SD systems.} \end{cases} \\
p_{q|\varepsilon,y,\nu',w',w,\nu,x}(r) &= \delta(q-w'+w), \\
p_{\varepsilon|y,\nu',w',w,\nu,x}(r) &= \delta(\varepsilon-y+x), \\
p_{y|\nu',w',w,\nu,x}(r) &= \delta(y-\nu'-w').
\end{aligned}$$

We now wish to form the product in Eq. (3.1) and to find its Fourier transform. We begin with

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

Using Theorem 2.5, the associated joint cf is given by

$$P_{w,\nu,x}(u_w,u_\nu,u_x) = P_{\nu,x}(u_\nu+u_w,u_x+u_w).$$

Then

$$\begin{aligned}
p_{w',w,\nu,x}(r) &= p_{w'|w,\nu,x}(w',w,\nu,x)p_{w,\nu,x}(w,\nu,x) \\
&= \Delta\Pi_\Delta(w'-w)W_\Delta(w')p_{w,\nu,x}(w,\nu,x) \\
&= \{A^*[\Delta\Pi_\Delta(w')p_{w,\nu,x}(w,\nu,x)]\}\,W_\Delta(w') \qquad (3.4)
\end{aligned}$$

where we use A^* to denote composition with a linear coordinate transformation of (w',w,ν,x) with transformation matrix

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Using Theorems A.4(viii) and A.5 from Appendix A we obtain the convolution

$$\begin{aligned} & P_{w',w,\nu,x}(u_{w'},u_w,u_\nu,u_x) \\ &= \{\mathrm{sinc}\,(u_{w'})P_{w,\nu,x}(u_{w'}+u_w,u_\nu,u_x)\} \star W_{\frac{1}{\Delta}}(u_{w'}) \qquad (3.5) \\ &= \sum_{k\in\mathbf{Z}^N} \mathrm{sinc}\left(u_{w'}-\frac{k}{\Delta}\right) P_{w,\nu,x}\left(u_{w'}+u_w-\frac{k}{\Delta},u_\nu,u_x\right) \\ &= \sum_{k\in\mathbf{Z}^N} \mathrm{sinc}\left(u_{w'}-\frac{k}{\Delta}\right) P_{\nu,x}\left(u_{w'}+u_w+u_\nu-\frac{k}{\Delta},u_{w'}+u_w+u_x-\frac{k}{\Delta}\right). \end{aligned}$$

The result is valid for $N \geq 1$ with the definition

$$\mathrm{sinc}\,(x) \stackrel{\triangle}{=} \prod_{i=1}^{N} \mathrm{sinc}\,(x_i).$$

The remaining factors in Eq. (3.1) are handled by repeated application of Theorem 2.5 using Eqs. (3.3). For the Fourier transform variables involved we will use the shorthand

$$u_r \stackrel{\triangle}{=} (u_q,u_\varepsilon,u_y,u_{\nu'},u_{w'},u_w,u_\nu,u_x) \in \mathbf{R}^{8N}.$$

In an NSD system we have

$$\begin{aligned} P_r(u_r) &= P_{\varepsilon,y,\nu',w',w,\nu,x}(u_\varepsilon,u_y,u_{\nu'},u_{w'}+u_q,u_w-u_q,u_\nu,u_x) \\ &= P_{y,\nu',w',w,\nu,x}(u_y+u_\varepsilon,u_{\nu'},u_{w'}+u_q,u_w-u_q,u_\nu,u_x-u_\varepsilon) \\ &= P_{\nu',w',w,\nu,x}(u_{\nu'}+u_y+u_\varepsilon,u_{w'}+u_y+u_\varepsilon+u_q,u_w-u_q,u_\nu,u_x-u_\varepsilon) \\ &= P_{w',w,\nu,x}(u_{w'}+u_y+u_\varepsilon+u_q,u_w-u_q,u_\nu,u_x-u_\varepsilon). \end{aligned}$$

In an SD system,

$$\begin{aligned} P_r(u_r) &= P_{\varepsilon,y,\nu',w',w,\nu,x}(u_\varepsilon,u_y,u_{\nu'},u_{w'}+u_q,u_w-u_q,u_\nu,u_x) \\ &= P_{y,\nu',w',w,\nu,x}(u_y+u_\varepsilon,u_{\nu'},u_{w'}+u_q,u_w-u_q,u_\nu,u_x-u_\varepsilon) \\ &= P_{\nu',w',w,\nu,x}(u_{\nu'}+u_y+u_\varepsilon,u_{w'}+u_y+u_\varepsilon+u_q,u_w-u_q,u_\nu,u_x-u_\varepsilon) \\ &= P_{w',w,\nu,x}(u_{w'}+u_y+u_\varepsilon+u_q,u_w-u_q,u_\nu-u_{\nu'}-u_y-u_\varepsilon,u_x-u_\varepsilon). \end{aligned}$$

For a UD system $\varepsilon = q$, $w' = y$, $w = x$ and $\nu' = 0$. We can treat such a system as a special case of SD (or NSD) systems by setting u_ε, $u_{w'}$, u_w and u_ν to zero and using $P_{\nu,x}(0, u_x) = P_x(u_x)$. The following relatively simple result is obtained:

Theorem 3.1 *In an undithered quantizing system*

$$P_{q,y,x}(u_q, u_y, u_x) = \sum_{k \in \mathbf{Z}^N} \operatorname{sinc}\left(u_q + u_y - \frac{k}{\Delta}\right) P_x\left(u_y + u_x - \frac{k}{\Delta}\right). \tag{3.6}$$

The results for SD and NSD systems are somewhat more complicated looking.

Theorem 3.2 *In an SD quantizing system*

$$\begin{aligned} P_r(u_r) = \sum_{k \in \mathbf{Z}^N} & \operatorname{sinc}\left(u_q + u_\varepsilon + u_y + u_{w'} - \frac{k}{\Delta}\right) \\ & \times P_{\nu,x}\left(-u_{\nu'} + u_{w'} + u_w + u_\nu - \frac{k}{\Delta}, u_y + u_{w'} + u_w + u_x - \frac{k}{\Delta}\right). \end{aligned} \tag{3.7}$$

Theorem 3.3 *In an NSD quantizing system*

$$\begin{aligned} P_r(u_r) = \sum_{k \in \mathbf{Z}^N} & \operatorname{sinc}\left(u_q + u_\varepsilon + u_y + u_{w'} - \frac{k}{\Delta}\right) \\ & \times P_{\nu,x}\left(u_\varepsilon + u_y + u_{w'} + u_w + u_\nu - \frac{k}{\Delta}, u_y + u_{w'} + u_w + u_x - \frac{k}{\Delta}\right). \end{aligned} \tag{3.8}$$

Now finding cf's or joint cf's of particular signals is straightforward since, according to Theorem 2.4, we only need to set the unwanted Fourier transform variables

to zero. For instance, for both SD and NSD systems with $N = 1$ we obtain the same expression for P_q by setting all variables except u_q to zero:

$$P_q(u_q) = \sum_{k=-\infty}^{\infty} \text{sinc}\left(u_q - \frac{k}{\Delta}\right) P_{\nu,x}\left(-\frac{k}{\Delta}, -\frac{k}{\Delta}\right).$$

If ν and x are assumed to be statistically independent, $P_{\nu,x}$ splits into a product yielding

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

Similarly, the cf of ε in an NSD system where ν and x are statistically independent is:

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

The corresponding expression for P_ε in an SD system is different. It is identical to the expression given above for P_q since, in an SD system, $q \equiv \varepsilon$.

Chapter 4

Practical Quantizing Systems

In this chapter we proceed from the general to the specific, interpreting the results obtained above with regard to particular realizations of quantizing systems. We begin, however, with a brief description of the classical model of undithered quantization in order that it may be contrasted with the more sophisticated treatment to follow.

4.1 The Classical Model of Undithered Quantization

We have seen that in an undithered quantizing system

$$\varepsilon = q(x).$$

Although this is a deterministic function of the input, the *classical model* of quantization treats this error as an additive iid random process which is independent of

the input and uniformly distributed. In particular, the quantization error variance (or "power") is taken to be $\Delta^2/12$ in the classical model [36].

This model of quantization error is suitable for complex (quasi-random) input signals which are large relative to an LSB. It fails catastrophically for small or simple signals where, in undithered systems, the quantization error retains the character of input-dependent distortion and/or noise modulation.

The non-random nature of the error can be demonstrated by using a computer to simulate the undithered quantization of a very simple signal: say, a 1 kHz sine wave of 4.0 LSB peak-to-peak amplitude. Fig. 4.1 shows the system input and output from such a simulation, as well as the resulting total error signal, and the estimated power spectrum of the system output. Evidence of the input signal is clearly visible in the total error waveform. In the power spectrum, many sharp peaks fall at multiples of the input sine wave frequency, indicating not only a high degree of non-random structure (i.e., harmonic distortion) in the error signal, but also a strong relationship between this signal and the system input.

The substantial discrepancies between the classical model of quantization and the observed behaviour of quantizing systems helped to spur the development of more sophisticated models of this process.

4.2 Widrow's Model of Undithered Quantization

A generalized statistical model of undithered quantization, valid for inputs with arbitrary statistical properties, was first developed by Widrow [2, 3, 4] in the 1950's. Widrow realized that quantizing a signal transforms its pdf into a train of weighted

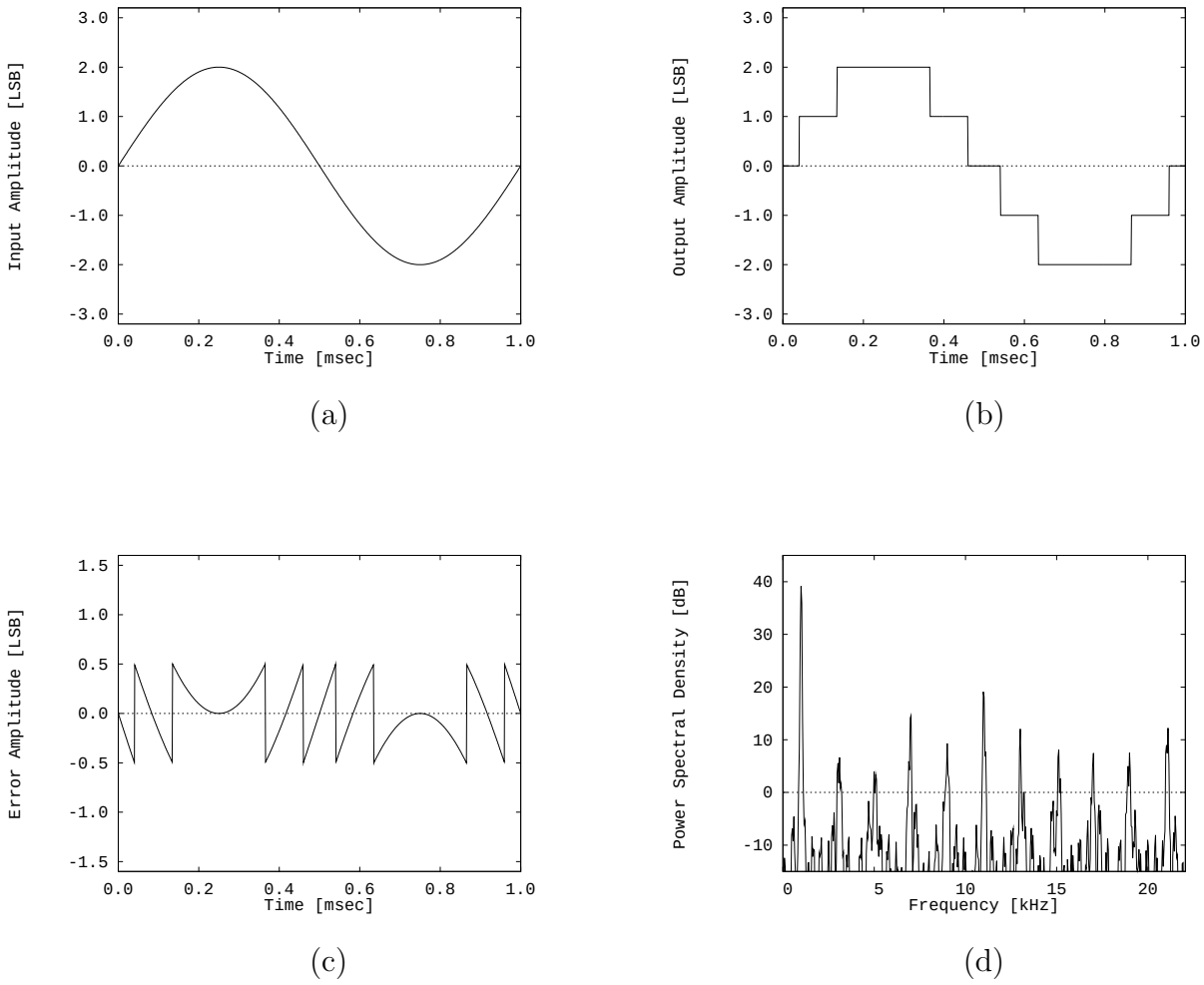


Figure 4.1: Results from the computer-simulated quantization of a 1 kHz sine wave of 4.0 LSB peak-to-peak amplitude without dither. Shown are (a) the system input signal, (b) the system output signal, (c) the resulting total error signal, and (d) the power spectrum of the system output signal (as estimated from sixty 50%-overlapping Hann-windowed 512-point time records with an assumed sampling frequency of 44.1 kHz; 0 dB represents a power spectral density of $\Delta^2 T/6$ where T is the sampling period).

impulse functions in a fashion reminiscent of time-sampling, so that recovery of the system input statistics from those of the system output must require conditions analogous to those of Sampling Theorem [1]. The development here differs from Widrow's in its details, the results are somewhat strengthened, and the proofs are new, but the essential nature of the approach owes much to his original.

4.2.1 UD Systems: Statistics of the Total Error

We begin by considering the statistical relationships between variables in the system at some given instant in time. (This corresponds to choosing $N = 1$, but in fact the argument is identical for $N > 1$.) Setting $u_y = 0$ in Eq. (3.6) we obtain

$$P_{q,x}(u_q, u_x) = \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_q - \frac{k}{\Delta}\right) P_x\left(u_x - \frac{k}{\Delta}\right). \tag{4.1}$$

If q and x are to be statistically independent, this must equal the product $P_q(u_q)P_x(u_x)$. Then, letting $u_q = \ell/\Delta$, we have

$$P_q\left(\frac{\ell}{\Delta}\right) P_x(u_x) = P_x\left(u_x - \frac{\ell}{\Delta}\right).$$

Now we must have

$$\left|P_q\left(\frac{\ell}{\Delta}\right)\right| = 1, \qquad \forall \ell \in \mathbf{Z}$$

otherwise $|P_x(u_x)| > 1$ for some value of u_x, which is impossible for a characteristic function (by Theorem 2.7(iii)). Then, letting $u_x = 0$ we have

$$\left|P_x\left(\frac{\ell}{\Delta}\right)\right| = 1, \qquad \forall \ell \in \mathbf{Z}.$$

Thus, by Theorem 2.7(iv), p_x and p_q are both lattice densities of delta functions separated by intervals of width Δ. That is

$$p_x(x) = \sum_{k=-\infty}^{\infty} c_k \delta\left(x - (k + \omega)\Delta\right)$$

for some $\omega \in \left[-\frac{1}{2}, \frac{1}{2}\right)$. Of course, this means that

$$p_q(q) = \delta(q + \omega\Delta)$$

so that the quantization error has a fixed value[1] of $-\omega\Delta$. Clearly this is statistical independence in only a purely formal sense and certainly does not imply that the error distribution is independent of the input distribution.

It is natural to wonder under what conditions q exhibits a uniform pdf of the sort assumed in the classical model.

Theorem 4.1 *The total error produced by an undithered quantizing system is uniformly distributed if and only if the cf of the system input, P_x, satisfies the condition that*

$$P_x\left(\frac{k}{\Delta}\right) = 0 \qquad \forall k \in \mathbf{Z}_0.$$

Proof: Setting $u_x = 0$ in Eq. (4.1) yields

$$P_q(u_q) = \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_q - \frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right). \tag{4.2}$$

If the error is to be uniformly distributed, Eq. (4.2) must reduce to a single sinc function centred at the origin. Thus the "if" direction is immediate. To prove "only if" suppose that

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

[1]Assuming a stochastic quantizer, $\omega = -\frac{1}{2}$ is a special case in which system outputs of $\pm\Delta/2$ are produced with equal probability.

Now let $u_q = \ell/\Delta$ where $\ell \in \mathbf{Z}_0$. This yields

$$\begin{aligned} 0 &= \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(\frac{\ell - k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right) \\ &= P_x\left(-\frac{\ell}{\Delta}\right). \end{aligned}$$

□

There are at least two other ways of showing this result. Firstly, we may write Eq. (4.2) as

$$P_q(u) = \operatorname{sinc}(u) \star \left[P_x \cdot W_{\frac{1}{\Delta}}\right](-u),$$

the inverse Fourier transform of which is (see Theorem A.5):

$$p_q(q) = \Delta \Pi_\Delta(q) \left[p_x \star W_\Delta\right](-q).$$

Using Poisson's summation formula (Theorem A.7) we have

$$\begin{aligned} \Delta \left[p_x \star W_\Delta\right](-q) &= \Delta \sum_{k=-\infty}^{\infty} p_x(-q - k\Delta) \\ &= \sum_{k=-\infty}^{\infty} P_x\left(-\frac{k}{\Delta}\right) e^{-j2\pi kq/\Delta}. \end{aligned}$$

If and only if the conditions of the theorem hold, the last summation reduces to $P_x(0) = 1$ so that $p_q = \Pi_\Delta$.

One may also reason as follows (after Gray and Stockham [14]). p_q can be non-zero only on $\left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right)$, so that we may expand it as a Fourier series on this interval:

$$p_q(q) = \frac{1}{\Delta} \sum_{k=-\infty}^{\infty} c_k e^{j2\pi kq/\Delta}$$

where

$$\begin{aligned} c_k &= \int_{-\Delta/2}^{\Delta/2} p_q(q) e^{-j2\pi kq/\Delta} dq \\ &= E\left[e^{-j2\pi kq/\Delta}\right] \\ &= E\left[e^{-j2\pi k\left(\lfloor \frac{x}{\Delta}+\frac{1}{2}\rfloor - \frac{x}{\Delta}\right)}\right] \\ &= E\left[e^{j2\pi kx/\Delta}\right] \\ &= P_x\left(-\frac{k}{\Delta}\right). \end{aligned}$$

Here we have used Eq. (2.3) and the fact that the floor operator returns an integer. We see that $p_q(q) = \frac{1}{\Delta}$ on $\left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right)$ if and only if the conditions of Theorem 4.1 hold.

The conditions in the theorem are not actually due to Widrow but to Sripad and Snyder [8]. Widrow [4] cites a different condition, which is sufficient but not necessary; viz., $P_x(u) = 0$ for $|u| \geq 1/\Delta$. Widrow calls this requirement "half-satisfaction" of the conditions of the Quantizing Theorem (cf. Theorem 4.3).

Note that if the requirements of Theorem 4.1 are satisfied, then the error is of the sort which is postulated by the classical model insofar as it is uniformly distributed with moments given by Eq. (2.9). Note also, however, that the error is not formally statistically independent of the input since

$$\left|P_x\left(\frac{k}{\Delta}\right)\right| \neq 1 \qquad \text{for } k \neq 0.$$

The statistical relationships between pairs of total error values separated in time are of particular interest since these determine the power spectral characteristics of the total error signal. Consider two system input values, x_1 and x_2, occurring at times t_1 and t_2, respectively, so that they are separated in time by $\tau = t_2 - t_1$ where

$\tau \neq 0$. Their statistical relationship is described by their joint pdf, $p_{x_1,x_2}(x_1, x_2)$. Taking $N = 2$ in Eq. (3.6) and letting $(u_{y_1}, u_{y_2}) = (0, 0)$ yields

$$\begin{aligned} P_{q_1,q_2,x_1,x_2}(u_{q_1}, u_{q_2}, u_{x_1}, u_{x_2}) &= \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \operatorname{sinc}\left(u_{q_1} - \frac{k_1}{\Delta}\right) \operatorname{sinc}\left(u_{q_2} - \frac{k_2}{\Delta}\right) \\ &\quad \times P_{x_1,x_2}\left(u_{x_1} - \frac{k_1}{\Delta}, u_{x_2} - \frac{k_2}{\Delta}\right). \end{aligned}$$

Proceeding as before, it is straightforward to show that this only splits into a product of $P_{q_1,q_2}(u_{q_1}, u_{q_2})$ with $P_{x_1,x_2}(u_{x_1}, u_{x_2})$ when the latter is a two-dimensional lattice distribution. Setting $(u_{x_1}, u_{x_2}) = (0, 0)$ yields

$$P_{q_1,q_2}(u_{q_1}, u_{q_2}) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \operatorname{sinc}\left(u_{q_1} - \frac{k_1}{\Delta}\right) \operatorname{sinc}\left(u_{q_2} - \frac{k_2}{\Delta}\right) P_{x_1,x_2}\left(-\frac{k_1}{\Delta}, -\frac{k_2}{\Delta}\right)$$

which leads to the following second-order version of Theorem 4.1:

Theorem 4.2 *In an undithered quantizing system, the joint cf, $P_{\varepsilon_1,\varepsilon_2}$, of total error values, ε_1 and ε_2, separated in time by $\tau \neq 0$ is given by*

$$p_{\varepsilon_1,\varepsilon_2}(\varepsilon_1, \varepsilon_2) = \Pi_\Delta(\varepsilon_1)\Pi_\Delta(\varepsilon_2) \tag{4.3}$$

if and only if the joint cf, P_{x_1,x_2}, of the corresponding system inputs, x_1 and x_2, satisfies the condition that

$$P_{x_1,x_2}\left(\frac{k_1}{\Delta}, \frac{k_2}{\Delta}\right) = 0 \qquad \forall (k_1, k_2) \in \mathbf{Z}_0^2.$$

Eq. (4.3) shows that, subject to the specified conditions, the joint pdf of ε_1 and ε_2 is a product of two rectangular window functions, one of which is a function of ε_1 alone and the other of ε_2 alone. Hence the two error values are statistically

independent of each other and each is uniformly distributed. Note that if the conditions of Theorem 4.2 are satisfied then so are those of Theorem 4.1.

For an undithered system satisfying the conditions of Theorem 4.2 at all times t_1 and t_2, the total error is wide-sense stationary with an autocorrelation function given by

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

Thus its PSD is given by

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

which is constant with respect to frequency so that the error signal is spectrally white and exhibits a total power of $\Delta^2/12$ up to the Nyquist frequency. In this respect the error is of the form assumed by the classical model of quantization.

4.2.2 UD Systems: Statistics of the System Output

We now proceed to investigate the statistical properties of the output of an undithered quantizing system. P_y can be obtained immediately from Eq. (3.6) but it is also instructive to consider Widrow's reasoning as follows [4].

The output can only assume values which are integer multiples of the quantization step size, Δ. Referring to Fig. 4.2, we see that the probability of an output having value $y = k\Delta$, for some specified integer k, is equal to the probability that

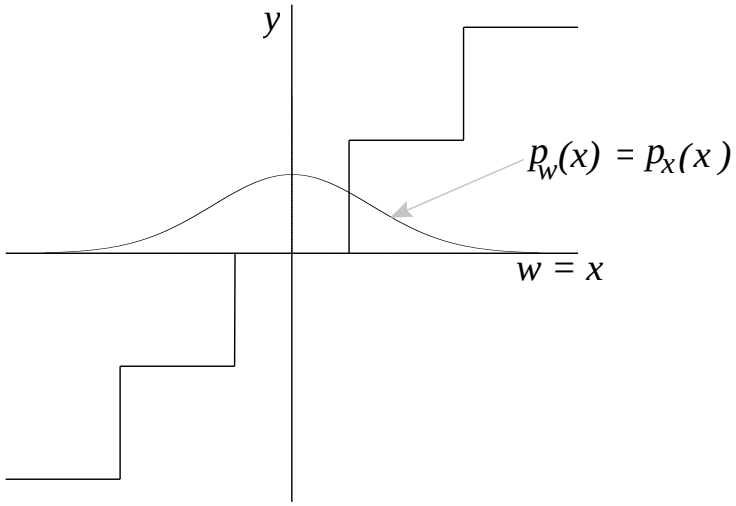


Figure 4.2: Pdf of the quantizer input in an undithered quantizing system, showing its justification relative to the quantizer characteristic.

the input lies between $-\frac{\Delta}{2} + k\Delta$ and $\frac{\Delta}{2} + k\Delta$. Hence,

$$p_y(y) = \sum_{k=-\infty}^{\infty} \delta(y - k\Delta) \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} p_x(x)dx. \tag{4.4}$$

Borrowing Widrow's terminology, we say that the quantization operation performs "area sampling" of the input distribution[2]. Writing the integral in Eq. (4.4) as a convolution of p_x with a rectangular window function, it reduces to

$$p_y(y) = [\Delta\Pi_\Delta \star p_x](y)W_\Delta(y). \tag{4.5}$$

Taking the Fourier transform of this expression yields (see Theorem A.5)

$$\begin{aligned} P_y(u) &= [\text{sinc}\,(u)P_x(u)] \star W_{\frac{1}{\Delta}}(u) \\ &= \sum_{k=-\infty}^{\infty} \text{sinc}\left(u - \frac{k}{\Delta}\right) P_x\left(u - \frac{k}{\Delta}\right), \end{aligned} \tag{4.6}$$

[2]Note that Eq. (4.4) loses its meaning when p_x contains delta functions at quantizer step edges, but that Eq. (3.6) does not.

which agrees with the expression obtained from Eq. (3.6).

Under what conditions are y and x identically distributed? Suppose that

$$P_x(u) = \sum_{k=-\infty}^{\infty} \text{sinc}\left(u - \frac{k}{\Delta}\right) P_x\left(u - \frac{k}{\Delta}\right),$$

and let $u = \ell/\Delta$, $\ell \in \mathbf{Z}$. We find that

$$P_x\left(\frac{\ell}{\Delta}\right) = P_x(0) = 1$$

so that, by Theorem 2.7(iv), we have

$$p_x(x) = \sum_{k=-\infty}^{\infty} c_k \delta(x - k\Delta).$$

Thus we obtain the intuitively satisfying result that $p_y \equiv p_x$ if and only if the input is restricted to integer multiples of Δ. We will see, however, that the *statistical properties* of the input can be recovered from the output subject to certain less restrictive conditions.

It is useful to rewrite Eq. (4.6) in the form

$$\begin{aligned} P_y(u) &= G_x(u) \star W_{\frac{1}{\Delta}}(u) \\ &= \sum_{k=-\infty}^{\infty} G_x\left(u - \frac{k}{\Delta}\right) \end{aligned} \tag{4.7}$$

where we have defined

$$G_x(u) \stackrel{\triangle}{=} \frac{\sin(\pi\Delta u)}{\pi\Delta u} P_x(u). \tag{4.8}$$

Hence, $P_y(u)$ consists of "aliases" of the function $G_x(u)$ separated by intervals of $1/\Delta$. Note, however, that if P_x is supported such that $P_x(u) = 0$ for $|u| \geq \frac{1}{2\Delta}$ (i.e., if, in the parlance of signal processing, p_x is "bandlimited"), then the aliased versions of $G_x(u)$ do not overlap, allowing recovery of the input cf (and hence the input pdf) from that of the output by bandlimiting. Indeed, this is [2, 3, 4]:

Theorem 4.3 (Widrow's Quantizing Theorem) *The pdf, $p_x(x)$, of the input, x, to an undithered quantizing system is recoverable from the pdf of its output if the cf of the input, P_x, is supported such that $P_x(u) = 0$ for $|u| \geq \frac{1}{2\Delta}$.*

Obviously, this theorem closely resembles the Sampling Theorem, which allows recovery of an appropriately bandlimited analogue signal from discrete-time samples thereof. The difference, of course, is that the Quantizing Theorem pertains not to time-sampling, but to amplitude quantizing of a signal (i.e., to area-sampling of the pdf of a signal).

It should be noted that the conditions of the Quantizing Theorem cannot be met unless $p_x(x)$ is *not* supported on a finite interval. This must be the case because if $P_x(u)$ is supported on a finite interval then its inverse Fourier transform cannot be [37]. Widrow [4] discusses signals, such as large amplitude processes with Gaussian distributions, which come close to satisfying the conditions. Here we will be satisfied with some qualitative observations. First, we have from Theorem 2.1 that

$$P_{ax}(u) = P_x(au), \qquad a \in \mathbf{R},$$

so that, roughly speaking, wide pdf's have narrow cf's. Also, it can be shown [38] that if $p_x, p_x^{(1)}, \ldots, p_x^{(n-1)}$ are continuous and tend to zero at infinity, and $p_x^{(n)}$ is absolutely integrable, then

$$\lim_{|u|\to\infty} u^n P_x(u) = 0,$$

so that the smoother the pdf of a random variable the more rapidly its cf tends to zero at infinity. Thus large amplitude signals with smooth pdf's will come closer to satisfying the Quantizing Theorem. In such cases relatively few terms will significantly contribute to Eq. (4.2) so that the quantization error will be more

uniformly distributed, and it is not difficult to show that correlations between samples of the error and between the error and the input diminish as well. Indeed, it is under these conditions that the CMQ has been found to be adequate for practical purposes.

In practice, recovering the *pdf* of the input is often unnecessary and it is sufficient to recover the *moments* of the input signal from the output. These are given by

$$E[y^m] = \left(\frac{j}{2\pi}\right)^m P_y^{(m)}(0).$$

If the Quantizing Theorem is satisfied then the aliased versions of $G_x(u)$ do not overlap, so that the m-th derivative of $P_y(u)$ at the origin is determined only by the "baseband" ($k = 0$) term in Eq. (4.7). This is also true, however, subject to the weaker condition that the Quantizing Theorem is only half-satisfied (see remarks following Theorem 4.1) or the still weaker condition that

$$G_x^{(m)}\left(\frac{k}{\Delta}\right) = 0 \qquad \forall k \in \mathbf{Z}_0. \tag{4.9}$$

If the input statistics obey this condition then

$$\begin{aligned} E[y^m] &= \left(\frac{j}{2\pi}\right)^m G_x^{(m)}(0) \\ &= \left(\frac{j}{2\pi}\right)^m \sum_{r=0}^{m} \binom{m}{r} \mathrm{sinc}^{(r)}(0) P_x^{(m-r)}(0) \\ &= \sum_{r=0}^{m} \binom{m}{r} E[\xi^r] E[x^{m-r}] \end{aligned}$$

where ξ is a notional uniformly distributed random variable which is sometimes thought of as a "quantization noise" but which, strictly speaking, is not physically meaningful. Thus we have succeeded in expressing the moments of y in terms of

the moments of x. Using Eq. (2.9) we obtain the following useful relationships:

$$E[y] = E[x] \tag{4.10}$$

$$E[y^2] = E[x^2] + \frac{\Delta^2}{12} \tag{4.11}$$

$$E[y^m] = \sum_{\ell=0}^{\lfloor \frac{m}{2} \rfloor} \binom{m}{2\ell} \left(\frac{\Delta}{2}\right)^{2\ell} \frac{E[x^{m-2\ell}]}{2\ell+1}. \tag{4.12}$$

Solving these equations to find the moments of x in terms of the moments of y yields the well-known Sheppard's corrections for grouping [39]. We emphasize that each of these equations for $E[y^m]$ is only valid when Eq. (4.9) 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. We observe, in particular, that if Eq. (4.9) is satisfied for $m = 2$, then the variance of $y = x + \varepsilon$ is the same as that of x plus a statistically independent additive random process with uniform pdf.

We note in passing that by repeated differentiation of Eq. (4.8) for $G_x(u)$ we can derive from Eq. (4.9) the following stronger, but perhaps more practical, condition in terms of the input cf, which ensures that $E[y^m]$ obeys Eq. (4.12) for $m = 1, 2, \ldots, M$:

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

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

From Eq. (3.6) with $N = 2$ we find that the joint pdf of two system output values, y_1 and y_2, separated in time by $\tau \neq 0$, is given by

$$P_{y_1,y_2}(u_1, u_2) = G_{x_1,x_2}(u_1, u_2) \star W_{\frac{1}{\Delta}}(u_1) W_{\frac{1}{\Delta}}(u_2),$$

where x_1 and x_2 are the corresponding system inputs and where we have defined

$$G_{x_1,x_2}(u_1,u_2) \triangleq \frac{\sin(\pi\Delta u_1)}{\pi\Delta u_1}\frac{\sin(\pi\Delta u_2)}{\pi\Delta u_2}P_{x_1,x_2}(u_1,u_2).$$

We can now write a second-order analogue of the Quantizing Theorem; namely, that the joint pdf of the input is recoverable from that of the output if

$$P_{x_1,x_2}(u_1,u_2) = 0 \quad \text{whenever } |u_1| \geq \frac{1}{2\Delta} \text{ or } |u_2| \geq \frac{1}{2\Delta}.$$

Of perhaps greater interest, however, is the second-order analogue of Eq. (4.9), which allows us to recover the joint moments of the system input from those of the output. That is, if

$$G_{x_1,x_2}^{(m_1,m_2)}\left(\frac{k_1}{\Delta},\frac{k_2}{\Delta}\right) = 0 \qquad \forall (k_1,k_2) \in \mathbf{Z}_0^2$$

then

$$\begin{aligned} E[y_1^{m_1}y_2^{m_2}] &= \left(\frac{j}{2\pi}\right)^{m_1+m_2} G_{x_1,x_2}^{(m_1,m_2)}(0,0) \\ &= \sum_{\ell_1=0}^{\lfloor \frac{m_1}{2} \rfloor}\sum_{\ell_2=0}^{\lfloor \frac{m_2}{2} \rfloor}\binom{m_1}{2\ell_1}\binom{m_2}{2\ell_2}\left(\frac{\Delta}{2}\right)^{2(\ell_1+\ell_2)}\frac{E[x_1^{m_1-2\ell_1}x_2^{m_2-2\ell_2}]}{(2\ell_1+1)(2\ell_2+1)}. \end{aligned} \tag{4.13}$$

Thus, assuming that x is wide-sense stationary,

$$r_y(k) = \begin{cases} E[x^2] + \dfrac{\Delta^2}{12}, & \text{for } k = 0, \\ E[x_1x_2](k), & \text{otherwise}, \end{cases} \tag{4.14}$$

so that the power spectral density of the output is identical to that of the input apart from an additive white-noise component arising from the quantization operation; that is:

$$\mathrm{PSD}_y(f) = \mathrm{PSD}_x(f) + \frac{\Delta^2 T}{6}. \tag{4.15}$$

4.2.3 Non-Stochastic Quantizers

UD quantization is the exceptional instance when the choice between a stochastic and deterministic quantizer would appear to make a difference to the statistical behaviour of signals in the system. One would expect this to be the case if the quantizer input pdf has the form

$$p_x(x) = \sum_{k=-\infty}^{\infty} c_k \delta\left(x - \frac{2k+1}{2}\Delta\right).$$

In this case, the input falls on a quantizer step edge with non-zero probability.

Suppose that a deterministic mid-tread quantizer is chosen such that inputs at step edges are consistently rounded up. In this case, we can deduce the system statistics by inspection:

$$p_{q,y,x}(q,y,x) = \delta\left(q + \frac{\Delta}{2}\right) \sum_{k=-\infty}^{\infty} c_k \delta(y - k\Delta)\delta\left(x - \frac{2k+1}{2}\Delta\right).$$

We find, as before, that the quantization error is formally statistically independent of the system input but certainly not uniformly distributed. This is similar to the result found when stochastic quantization was assumed.

4.2.4 Summary of Undithered Quantization

In a sense, the results of this section are primarily of theoretical, rather than practical, interest. All of the theorems given above impose conditions upon the statistics of the system input, and such restrictions are usually undesirable or impossible to meet in practice. Some not-uncommon system inputs satisfy the conditions of Theorem 4.1 (e.g., a 1RPDF random process) so that the associated error will be

uniformly distributed. On the other hand, however, the conditions of the Quantizing Theorem (Theorem 4.3) cannot be met by *any* system input whose pdf is supported on a finite interval so that, in practice, the distribution of the system input cannot be precisely recovered from the distribution of the system output.

There now becomes apparent, however, the possibility of dithering the system input with a suitably chosen dither signal, ν, so as to ensure that the quantizer input, $w = x + \nu$ in Fig. 2.1, satisfies some of the aforementioned conditions. In particular, if the dither is statistically independent of the system input, then the pdf, p_w, of w is the convolution $p_w = p_x \star p_\nu$, and hence its cf is the product $P_w = P_x \cdot P_\nu$. In this case the dither statistics can be chosen so as to cause P_w to vanish at required places, and so force the total quantizer input to meet the conditions of, say, Theorem 4.1. This accomplishment cannot then be subsequently undone by any system input which is statistically independent of the dither.

These tentative ideas will be developed in detail in the following sections.

4.3 Subtractive Dither

4.3.1 SD Systems: Statistics of the Total Error

In an SD system the quantizer input is $w = x + \nu$ so that the output of the system is (see Fig. 1.3(b))

$$y = Q(x + \nu) - \nu.$$

Hence the total error is given by

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

which is simply the quantization error, q, of the *total* quantizer input, w. We will assume that ν and x are statistically independent.

The following provides a new strengthening and proof of a result which was first reported by Schuchman [7].

Theorem 4.4 (Schuchman's Condition) *In an SD quantizing system, the total error will be statistically independent of the system input for arbitrary input distributions if and only if the cf of the dither, P_ν, satisfies the condition that*

$$P_\nu\left(\frac{k}{\Delta}\right) = 0 \qquad \forall k \in \mathbf{Z}_0. \tag{4.16}$$

Furthermore, the total error will be uniformly distributed for arbitrary input distributions if and only if this condition holds.

Proof: From Eq. (3.7) we obtain

$$P_{\varepsilon,x}(u_\varepsilon, u_x) = \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_\varepsilon - \frac{k}{\Delta}\right) P_\nu\left(-\frac{k}{\Delta}\right) P_x\left(u_x - \frac{k}{\Delta}\right). \tag{4.17}$$

If the condition of the theorem is met, this expression splits into a product of $P_x(u_x)$ with

$$P_\varepsilon(u_\varepsilon) = \operatorname{sinc}(u_\varepsilon)$$

so that the error is uniformly distributed and statistically independent of the input.

Now suppose that the input has some arbitrary distribution and that the error and input are statistically independent so that Eq. (4.17) can be written as a product $P_\varepsilon(u_\varepsilon)P_x(u_x)$. Then if $u_\varepsilon = \ell/\Delta$ for some $\ell \in \mathbf{Z}_0$ this yields

$$P_\varepsilon\left(\frac{\ell}{\Delta}\right) P_x(u_x) = P_\nu\left(-\frac{\ell}{\Delta}\right) P_x\left(u_x - \frac{\ell}{\Delta}\right).$$

Now if

$$\left|P_\nu\left(-\frac{\ell}{\Delta}\right)\right| \neq 0$$

then we must have

$$\left|P_\varepsilon\left(\frac{\ell}{\Delta}\right)\right| = \left|P_\nu\left(-\frac{\ell}{\Delta}\right)\right|$$

since otherwise $|P_x(u_x)| > 1$ for some u_x. Thus

$$|P_x(u_x)| = \left|P_x\left(u_x - \frac{\ell}{\Delta}\right)\right|.$$

Letting $u_x = 0$ shows that the input must have a lattice density, which contradicts the assumption that it is arbitrarily distributed. Thus we conclude that for any $\ell \in \mathbf{Z}_0$

$$\left|P_\nu\left(-\frac{\ell}{\Delta}\right)\right| = 0.$$

Finally suppose that the input has some arbitrary distribution and that ε is uniformly distributed. Eq. (4.17) then gives

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

Letting $u_\varepsilon = \ell/\Delta$ where $\ell \in \mathbf{Z}_0$ then gives

$$0 = P_\nu\left(-\frac{\ell}{\Delta}\right) P_x\left(-\frac{\ell}{\Delta}\right).$$

Since x is arbitrarily distributed this yields the desired result.

□

The above result regarding statistical independence was not explicitly mentioned by Schuchman [7]. It is found explicitly stated for the first time in [9].

Proceeding in similar fashion, we can use Eq. (3.7) with $N = 2$ to deduce that for two total error values, ε_1 and ε_2, separated in time by $\tau \neq 0$, and the corresponding input values x_1 and x_2 we have [23]:

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

where P_{ν_1,ν_2} represents the joint pdf of dither values ν_1 and ν_2, applied to input values x_1 and x_2, respectively. This leads, via the same brand of argument as used above for $N = 1$, to the following conclusion:

Theorem 4.5 *In an SD quantizing system, where ε_1 and ε_2 are two total error values separated in time by $\tau \neq 0$ with corresponding input values x_1 and x_2 and dither values ν_1 and ν_2, respectively, the random vector $(\varepsilon_1, \varepsilon_2)$ is statistically independent of the the random vector (x_1, x_2) for arbitrary input distributions if and only if*

$$P_{\nu_1,\nu_2}\left(\frac{k_1}{\Delta}, \frac{k_2}{\Delta}\right) = 0 \qquad \forall (k_1, k_2) \in \mathbf{Z}_0^2. \tag{4.18}$$

Furthermore, if and only if this condition holds then

$$p_{\varepsilon_1,\varepsilon_2}(\varepsilon_1, \varepsilon_2) = \Pi_\Delta(\varepsilon_1)\Pi_\Delta(\varepsilon_2), \tag{4.19}$$

so that ε_1 and ε_2 are both uniformly distributed and statistically independent of each other.

It should be noted that if ν_1 and ν_2 are statistically independent of each other, and the cf of each satisfies Eq. (4.16), then Eq. (4.18) will hold. This is the situation of interest in most practical applications using subtractive dither.

Subject to satisfaction of Eq. (4.18), the joint moments of ε_1 and ε_2 are given by

$$E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] = E[\varepsilon_1^{m_1}]E[\varepsilon_2^{m_2}],$$

so that $\varepsilon_1^{m_1}$ and $\varepsilon_2^{m_2}$ are, of course, uncorrelated. In particular, for $m_1 = m_2 = 1$

$$\begin{aligned} E[\varepsilon_1\varepsilon_2] &= E[\varepsilon_1]E[\varepsilon_2] \\ &= 0. \end{aligned}$$

Indeed, if the theorem is satisfied for all ν_1 and ν_2 separated in time by $\tau \neq 0$, and the conditions of Theorem 4.4 also hold, then

$$r_\varepsilon(k) = \begin{cases} \dfrac{\Delta^2}{12}, & k = 0, \\ 0, & \text{otherwise.} \end{cases}$$

so that

$$\mathrm{PSD}_\varepsilon(f) = \frac{\Delta^2 T}{6}. \tag{4.20}$$

This indicates that in a properly dithered SD quantizing system the total error signal will be spectrally white even if the dither signal is not.

4.3.2 SD Systems: Statistics of the System Output

From Eq. (3.7) we have

$$P_y(u) = \sum_{k=-\infty}^{\infty} \mathrm{sinc}\left(u - \frac{k}{\Delta}\right) P_\nu\left(-\frac{k}{\Delta}\right) P_x\left(u - \frac{k}{\Delta}\right). \tag{4.21}$$

Now suppose that the dither signal satisfies the conditions of Theorem 4.4. Then, since the total error is statistically independent of the input and uniformly distributed, and since the output is given by $y = x + \varepsilon$, the cf of the output should be

the product

$$P_y(u) = \frac{\sin(\pi\Delta u)}{\pi\Delta u} P_x(u). \tag{4.22}$$

Indeed, this is the expression to which Eq. (4.21) simplifies under the conditions of the theorem. This shows that

$$\begin{aligned} p_y(y) &= [p_\varepsilon \star p_x](y) \\ &= [\Delta\Pi_\Delta \star p_x](y). \end{aligned}$$

The output statistics assume this simple form for arbitrary input distributions only if the conditions of Theorem 4.4 are met, as may be verified by substituting Eq. (4.22) into Eq. (4.21) and letting $u = \ell/\Delta$, $\ell \in \mathbf{Z}_0$.

In this case the output is precisely the sum of the input plus a statistically independent uniformly distributed random process, and its cf and pdf exhibit the form expected of such a sum. The moments of the output in terms of the moments of the input are given by Eq. (4.12) above, which, in this case, is valid for all m.

Furthermore, if and only if P_{ν_1,ν_2} satisfies the conditions of Theorem 4.5, then

$$P_{y_1,y_2}(u_1, u_2) = \frac{\sin(\pi\Delta u_1)}{\pi\Delta u_1} \frac{\sin(\pi\Delta u_2)}{\pi\Delta u_2} P_{x_1,x_2}(u_1, u_2).$$

Hence, the joint moments of the output in terms of the moments of the input will be given by Eq. (4.13) above, and Eqs (4.14) and (4.15) will hold. That is to say that the quantization operation has merely added to the input signal a white noise process of variance $\Delta^2/12$.

4.3.3 SD Systems: Properties of Practical Dither Signals

It is naturally of interest to inquire as to which common random signals satisfy the criterion of Theorem 4.4. Perhaps the simplest imaginable candidate is dither with the uniform pdf

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

whose corresponding cf is the sinc function:

$$P_\nu(u) = \frac{\sin(\pi\Delta u)}{\pi\Delta u}.$$

This cf clearly satisfies the desired condition. We conclude that 1RPDF dither will render the total error statistically independent of the input and uniformly distributed in a subtractively dithered quantizing system. If we assume that values in the dither sequence are iid then the criterion of Theorem 4.5 is also satisfied and distinct values in the total error sequence are statistically independent of one another (thus ensuring that this sequence meets the weaker requirement of being spectrally white).

Of course, there are other cf's which meet the requirement of vanishing at all non-zero multiples of $1/\Delta$. For instance, nRPDF dithers with $n \geq 1$ all satisfy the criterion since their cf's are of the form

$$P_\nu(u) = \left[\frac{\sin(\pi\Delta u)}{\pi\Delta u}\right]^n.$$

However, in an SD system, such dithers usually have no inherent advantage over simple white 1RPDF dither. (An exceptional instance is discussed in Section 5.3.)

4.3.4 Summary of Subtractive Dither

The most practically important theoretical results concerning subtractively dithered quantizing systems are that:

1. the total error can be rendered uniformly distributed and statistically independent of the system input by choosing a dither which satisfies the conditions of Theorem 4.4, and

2. values of the total error separated in time can be rendered statistically independent of one another (so that the total error signal is spectrally white) by using a dither whose values, in addition to satisfying Theorem 4.4, are statistically independent of one another.

A familiar dither which satisfies all the required conditions is an iid 1RPDF process. Fig. 4.3 shows the results of a computer-simulated quantization operation performed upon a 1 kHz sine wave of 4.0 LSB peak-to-peak amplitude and using this type of subtractive dither. Shown are the system input and output, the total error, and the power spectrum of the system output. Note that the system output resembles a sine wave plus an independent additive noise without vestiges of the quantization staircase characteristic, and that no trace of the input signal is visible in the noise-like total error waveform. Furthermore, the power spectrum of the system output exhibits no distortion components whatsoever and shows that the total error is spectrally white. (The 0 dB noise floor in Fig. 4.3 represents a power spectral density of $\Delta^2 T/6$, with an integrated noise power of $\Delta^2/12$ up to the Nyquist frequency.) These results should be compared with those in Fig. 4.1, which illustrate the signal-dependent distortions produced by an undithered quantizing system with the same system input signal.

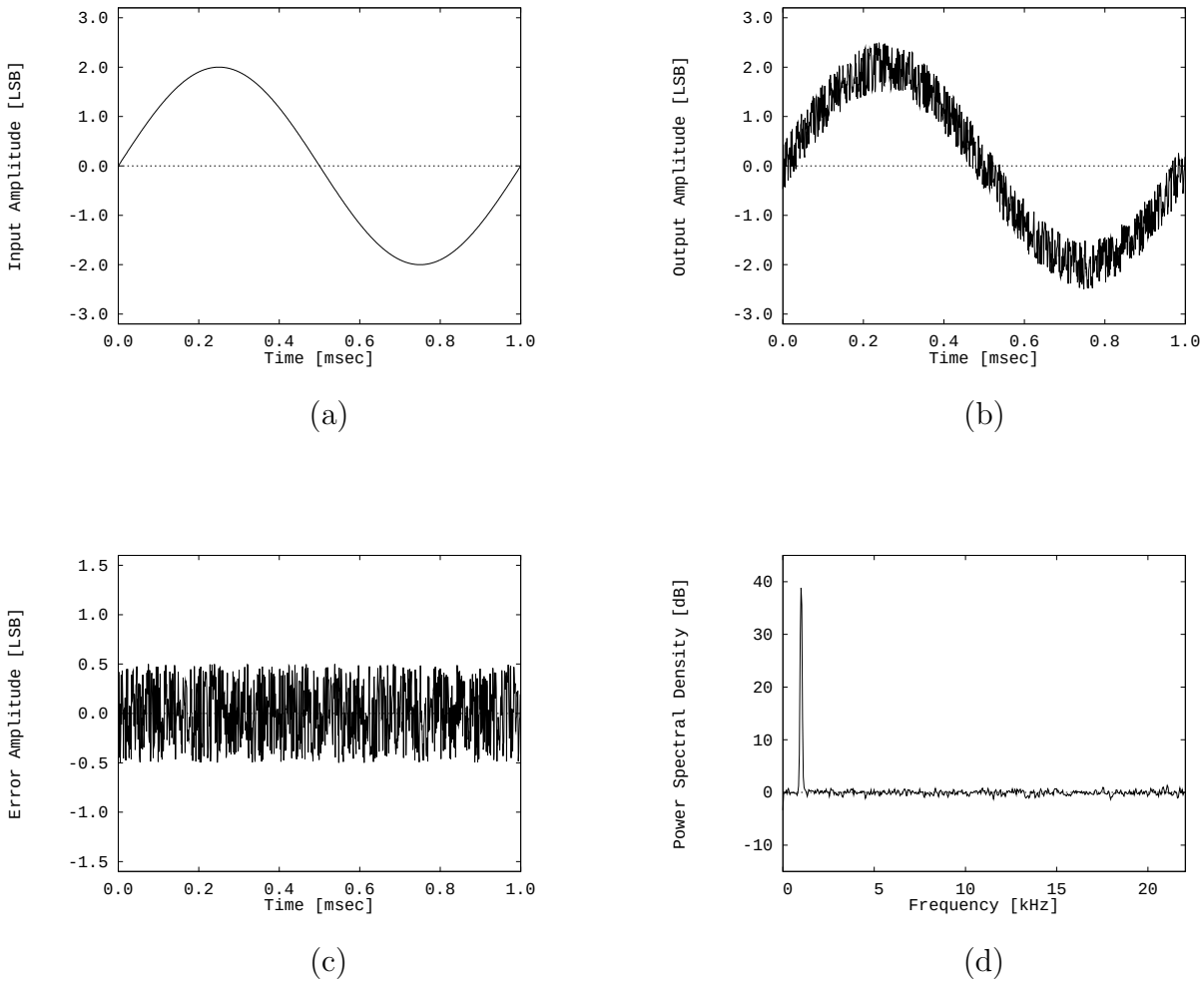


Figure 4.3: Results from the computer-simulated quantization of a 1 kHz sine wave of 4.0 LSB peak-to-peak amplitude using 1RPDF subtractive dither. Shown are (a) the system input signal, (b) the system output signal, (c) the resulting total error signal, and (d) the power spectrum of the system output signal (as estimated from sixty 50%-overlapping Hann-windowed 512-point time records with an assumed sampling frequency of 44.1 kHz; 0 dB represents a power spectral density of $\Delta^2 T/6$ where T is the sampling period).

Subtractively dithered quantizing systems are ideal in the sense that they render the total error an input-independent additive noise process. The requirement of dither subtraction at the system output, however, imposes restrictions which make it difficult to implement in practical applications. For one thing, the dither signal must be available at the output, and so either the dither must be transmitted along with the signal or synchronized dither generators must be present at both ends of the channel. Even more seriously, any signal editing or modification occurring between the original quantization and the subtraction of the dither necessitates a like operation on the dither sequence. It is for such reasons that subtractive dither is generally not a feasible option.

A proposed subtractive dithering scheme which may lead to practical implementations is due to Craven and Gerzon [40]. This scheme uses dither values determined from the input signal values by means of a suitably randomized look-up table. At this time, the proposed procedure awaits further testing and standardization. Even if these proceed in the future, non-subtractive dithering schemes are likely to remain preferable in many applications due to their relative simplicity.

Although many of the same benefits can be achieved with non-subtractive dither as with subtractive dither, the total error variance is inevitably greater in NSD systems, and the beautiful result regarding full statistical independence of the total error is unattainable, as we shall now see.

4.4 Non-Subtractive Dither

Although some 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 of subtractive and non-subtractive dithering, which have quite different properties (see, for instance, [36, p. 170]). We will see that non-subtractively dithered systems *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. They *can,* however, render certain statistical moments of the total error independent of the system input, and regulate the joint moments of total error values which are separated in time. For many applications, this is as good as full statistical independence.

4.4.1 NSD Systems: Statistics of the Total Error

The quantizer output in a non-subtractively dithered quantizing system is given by (see Fig. 1.3(c))

$$y = Q(x + \nu),$$

so that the total error is

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

Obviously, the total error is not simply the quantization error alone, but also involves the dither. This fact is responsible for the characteristics of NSD systems

which distinguish them from SD ones. Chief among these is the following [23]:

Theorem 4.6 *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.*

Proof: From Eq. (3.8) we obtain:

$$P_{\varepsilon,x}(u_\varepsilon, u_x) = \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_\varepsilon - \frac{k}{\Delta}\right) P_\nu\left(u_\varepsilon - \frac{k}{\Delta}\right) P_x\left(u_x - \frac{k}{\Delta}\right). \tag{4.23}$$

Now suppose that for arbitrarily distributed inputs we can write this as $P_\varepsilon(u_\varepsilon)P_x(u_x)$. Then for $u_\varepsilon = \ell/\Delta$, $\ell \in \mathbf{Z}_0$ we have

$$P_\varepsilon\left(\frac{\ell}{\Delta}\right) P_x(u_x) = P_x\left(u_x - \frac{\ell}{\Delta}\right).$$

Then

$$\left|P_\varepsilon\left(\frac{\ell}{\Delta}\right)\right| = 1$$

since otherwise $|P_x(u_x)| > 1$ for some u_x. Taking $u_x = 0$ we obtain

$$\left|P_x\left(\frac{\ell}{\Delta}\right)\right| = |P_x(0)| = 1$$

so that the input must have a lattice density, contradicting the assumption that it is arbitrarily distributed. We conclude that ε and x can never be made statistically independent in an NSD system.

Furthermore, setting $u_x = 0$ we have

$$P_\varepsilon(u_\varepsilon) = \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_\varepsilon - \frac{k}{\Delta}\right) P_\nu\left(u_\varepsilon - \frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right). \tag{4.24}$$

In order for ε to be uniformly distributed, this must reduce to $\mathrm{sinc}\,(u_\varepsilon)$ for some choice of P_ν; that is, we require

$$\mathrm{sinc}\,(u_\varepsilon) = \sum_{k=-\infty}^{\infty} \mathrm{sinc}\left(u_\varepsilon - \frac{k}{\Delta}\right) P_\nu\left(u_\varepsilon - \frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right).$$

Now suppose $u_\varepsilon = \ell/\Delta$ where $\ell \in \mathbf{Z}_0$. Then we have

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

so that P_x cannot be arbitrary. Thus the total error cannot in general be made uniformly distributed in an NSD system.

□

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 Theorem 4.4 is used so that q is independent of x. Then, since ν is also independent of x, the total error ε 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. The analytical approach of Chapter 2 can easily be used to show that for arbitrary random variables q, ν, and x and a third $\varepsilon = q + \nu$ (none of these necessarily representing quantities in a quantizing system) that

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

Obviously, ε and x are statistically independent of each other if and only if x is independent of the random vector (q, ν), since only in this instance does $P_{\varepsilon,x}(u_\varepsilon, u_x)$ split into a product of two functions one of which involves u_ε alone and the other u_x

alone. This is a stronger requirement than the one that x be independent of q and ν individually. In an NSD quantizing system, given q and ν, the possible values of x are restricted to $x = -(q+\nu) + k\Delta$, $k \in \mathbf{Z}$, so that the distribution of x is highly dependent on (q, ν). Of course, x would be independent of (q, ν) if $\{q, \nu, x\}$ formed a set of independent random variables, that is, if it were the case that

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

but this even stronger condition is *certainly* not met in an NSD quantizing system.

We observe that the correct general expression for $p_\varepsilon(\varepsilon)$ in an NSD system may be obtained from Eq. (4.24) by writing it as

$$P_\varepsilon(u_\varepsilon) = [\text{sinc}\,(u_\varepsilon)P_\nu(u_\varepsilon)] \star [P_x(-u_\varepsilon)W_{\frac{1}{\Delta}}(u_\varepsilon)],$$

the inverse Fourier transform of which is:

$$p_\varepsilon(\varepsilon) = [\Delta\Pi_\Delta \star p_\nu](\varepsilon) \cdot [p_x \star W_\Delta](-\varepsilon). \tag{4.25}$$

Although the total error in an NSD system cannot be made statistically independent of the system input, it turns out that moments of the total error *can* be rendered independent of the input distribution. From Eq. (4.24) we have

$$\begin{aligned} E[\varepsilon^m] &\triangleq \left(\frac{j}{2\pi}\right)^m P_\varepsilon^{(m)}(0) \\ &= \left(\frac{j}{2\pi}\right)^m \sum_{k=-\infty}^{\infty} G_\nu^{(m)}\left(\frac{k}{\Delta}\right) P_x\left(\frac{k}{\Delta}\right), \end{aligned} \tag{4.26}$$

where

$$G_\nu(u) \triangleq \frac{\sin(\pi\Delta u)}{\pi\Delta u} P_\nu(u). \tag{4.27}$$

Since the cf, P_x, of the system input is arbitrary we obtain the following result [23]:

Theorem 4.7 *In an NSD quantizing system, $E[\varepsilon^m]$ is independent of the distribution of the system input, x, if and only if*

$$G_\nu^{(m)}\left(\frac{k}{\Delta}\right) = 0 \qquad \forall k \in \mathbf{Z}_0. \tag{4.28}$$

If the conditions of Theorem 4.7 are satisfied, then from Eq. (4.26),

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

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 $G_\nu(u)$:

$$E[\varepsilon] = E[\nu] \tag{4.29}$$

$$E[\varepsilon^2] = E[\nu^2] + \frac{\Delta^2}{12} \tag{4.30}$$

$$E[\varepsilon^m] = \sum_{\ell=0}^{\lfloor \frac{m}{2} \rfloor} \binom{m}{2\ell} \left(\frac{\Delta}{2}\right)^{2\ell} \frac{E[\nu^{m-2\ell}]}{2\ell+1}. \tag{4.31}$$

These exhibit the form of Sheppard's corrections (cf. Eq. (4.12)), but give expressions for the total error moments instead of the system input moments. We emphasize that each of these equations for $E[\varepsilon^m]$ is valid only when Theorem 4.7 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. (4.30) shows that with non-subtractive dither satisfying the conditions of Theorem 4.7, the total error variance is greater than that of classical UD or SD quantization by the dither variance.

We will prove a somewhat weaker theorem, which perhaps is really just a corollary to Theorem 4.7, but which is actually somewhat better known than that theorem itself [10, 14].

Theorem 4.8 *In an NSD quantizing system, $E[\varepsilon^m]$ is independent of the distribution of the system input, x, for $m = 1, 2, \ldots, M$ 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: The "if" direction follows immediately from repeated differentiation of Eq. (4.27), yielding

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m \sum_{k=-\infty}^{\infty} \sum_{r=0}^{m} \binom{m}{r} \mathrm{sinc}^{(r)}\left(-\frac{k}{\Delta}\right) P_\nu^{(m-r)}\left(-\frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right).$$

The "only if" direction employs an inductive argument. Consider first $M = 1$. By Theorem 4.7 we require

$$G^{(1)}\left(\frac{k}{\Delta}\right) = 0 \qquad \forall k \in \mathbf{Z}_0.$$

Direct computation yields

$$\begin{aligned} G_\nu^{(1)}\left(\frac{k}{\Delta}\right) &= \mathrm{sinc}^{(1)}\left(\frac{k}{\Delta}\right) P_\nu\left(\frac{k}{\Delta}\right) + \mathrm{sinc}\left(\frac{k}{\Delta}\right) P_\nu^{(1)}\left(\frac{k}{\Delta}\right) \\ &= \mathrm{sinc}^{(1)}\left(\frac{k}{\Delta}\right) P_\nu\left(\frac{k}{\Delta}\right) \qquad \text{for } k \in \mathbf{Z}_0. \end{aligned}$$

The derivative of the sinc function is (see Appendix C)

$$\mathrm{sinc}^{(1)}\left(\frac{k}{\Delta}\right) = \Delta \frac{(-1)^k}{k} \neq 0 \qquad \forall k \in \mathbf{Z}_0$$

so that the expression only vanishes $\forall k \in \mathbf{Z}_0$ if

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

Now consider

$$G_\nu^{(m+1)}\left(\frac{k}{\Delta}\right) = \sum_{r=0}^{m+1} \binom{m+1}{r} \mathrm{sinc}^{(r)}\left(\frac{k}{\Delta}\right) P_\nu^{(m-r+1)}\left(\frac{k}{\Delta}\right)$$

and suppose that the the theorem holds for $M = m$, in which case this expression reduces to

$$G_\nu^{(m+1)}\left(\frac{k}{\Delta}\right) = (m+1)\mathrm{sinc}^{(1)}\left(\frac{k}{\Delta}\right) P_\nu^{(m)}\left(\frac{k}{\Delta}\right).$$

Again the derivative of the sinc function does not vanish, so we must have

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

This proves the theorem.

□

We see that the dithers meeting the conditions of this theorem are those which were introduced as dithers of order M in Section 2.3.

Proceeding in the now accustomed fashion, we consider two total error values, ε_1 and ε_2, which are separated in time by $\tau \neq 0$, and the two corresponding input signal values, x_1 and x_2. We omit the demonstration that $(\varepsilon_1, \varepsilon_2)$ cannot be rendered statistically independent of (x_1, x_2), since this proceeds in a fashion analogous to that of the one-dimensional case discussed above, and instead directly use Eq. (3.8) with $N = 2$ to obtain

$$P_{\varepsilon_1,\varepsilon_2}(u_1,u_2) = \sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty} \frac{\sin[\pi\Delta(u_1-k_1/\Delta)]}{\pi\Delta(u_1-k_1/\Delta)} \frac{\sin[\pi\Delta(u_2-k_2/\Delta)]}{\pi\Delta(u_2-k_2/\Delta)} \times P_{\nu_1,\nu_2}\left(u_1-\frac{k_1}{\Delta}, u_2-\frac{k_2}{\Delta}\right) P_{x_1,x_2}\left(-\frac{k_1}{\Delta}, -\frac{k_2}{\Delta}\right). \qquad (4.32)$$

We 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.

From Eq. (4.32) we find that

$$E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] = \left(\frac{j}{2\pi}\right)^{m_1+m_2} \sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty} P_{x_1,x_2}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right) G_{\nu_1,\nu_2}^{(m_1,m_2)}\left(\frac{k_1}{\Delta},\frac{k_2}{\Delta}\right), \tag{4.33}$$

where

$$G_{\nu_1,\nu_2}(u_1,u_2) \triangleq \frac{\sin(\pi\Delta u_1)}{\pi\Delta u_1}\frac{\sin(\pi\Delta u_2)}{\pi\Delta u_2} P_{\nu_1,\nu_2}(u_1,u_2).$$

Since P_{x_1,x_2} is arbitrary, we find that $E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}]$ is independent of the joint pdf of the system input if and only if

$$G_{\nu_1,\nu_2}^{(m_1,m_2)}\left(\frac{k_1}{\Delta},\frac{k_2}{\Delta}\right) = 0 \qquad \forall (k_1,k_2) \in \mathbf{Z}_0^2, \tag{4.34}$$

in which case it is given by

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

In this case, we can write an expression analogous to Eq. (4.31), relating the joint moments of the total error to those of the dither:

$$E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] = \sum_{\ell_1=0}^{\lfloor \frac{m_1}{2} \rfloor}\sum_{\ell_2=0}^{\lfloor \frac{m_2}{2} \rfloor} \binom{m_1}{2\ell_1}\binom{m_2}{2\ell_2}\left(\frac{\Delta}{2}\right)^{2(\ell_1+\ell_2)} \frac{E[\nu_1^{m_1-2\ell_1}\nu_2^{m_2-2\ell_2}]}{(2\ell_1+1)(2\ell_2+1)}. \tag{4.35}$$

Note that if ν_1 and ν_2 are statistically independent of each other and satisfy Eq. (4.28) for $m = m_1$ and $m = m_2$, respectively, then Eq. (4.34) is automatically satisfied. In this case Eq. (4.35) factors such that $\varepsilon_1^{m_1}$ and $\varepsilon_2^{m_2}$ are uncorrelated (i.e., $E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] = E[\varepsilon_1^{m_1}]E[\varepsilon_2^{m_2}]$).

Let us now consider the special case where $m_1 = m_2 = 1$. Explicitly performing the differentiation in Eq. (4.33) we obtain

$$\begin{aligned}
E[\varepsilon_1\varepsilon_2] &\triangleq \left(\frac{j}{2\pi}\right)^2 P^{(1,1)}_{\varepsilon_1,\varepsilon_2}(0,0) \\
&= \left(\frac{j}{2\pi}\right)^2 \sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty} \Bigg\{ \mathrm{sinc}^{(1)}\left(-\frac{k_1}{\Delta}\right)\mathrm{sinc}^{(1)}\left(-\frac{k_2}{\Delta}\right) \\
&\qquad\qquad \times P_{\nu_1,\nu_2}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right) P_{x_1,x_2}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right) \\
&\quad + \mathrm{sinc}\left(-\frac{k_1}{\Delta}\right)\mathrm{sinc}^{(1)}\left(-\frac{k_2}{\Delta}\right) P^{(1,0)}_{\nu_1,\nu_2}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right) P_{x_1,x_2}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right) \\
&\quad + \mathrm{sinc}^{(1)}\left(-\frac{k_1}{\Delta}\right)\mathrm{sinc}\left(-\frac{k_2}{\Delta}\right) P^{(0,1)}_{\nu_1,\nu_2}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right) P_{x_1,x_2}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right) \\
&\quad + \mathrm{sinc}\left(-\frac{k_1}{\Delta}\right)\mathrm{sinc}\left(-\frac{k_2}{\Delta}\right) P^{(1,1)}_{\nu_1,\nu_2}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right) P_{x_1,x_2}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right) \Bigg\}.
\end{aligned}$$

Careful inspection of this expression, keeping in mind that the first derivatives of the sinc function vanish at the origin, shows that it reduces to

$$E[\nu_1\nu_2] = \left(\frac{j}{2\pi}\right)^2 P^{(1,1)}_{\nu_1,\nu_2}(0,0),$$

thereby becoming independent of the system input, only under the conditions of the following theorem:

Theorem 4.9 *In an NSD system where all dither values are statistically independent of all system input values,*

$$E[\varepsilon_1\varepsilon_2] = E[\nu_1\nu_2]$$

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

satisfied:

$$P_{\nu_1,\nu_2}\left(\frac{k_1}{\Delta},\frac{k_2}{\Delta}\right) = 0 \quad \forall (k_1,k_2) \in \mathbf{Z}_0^2, \tag{4.36}$$

$$P^{(0,1)}_{\nu_1,\nu_2}\left(\frac{k_1}{\Delta},0\right) = 0 \quad \forall k_1 \in \mathbf{Z}_0, \tag{4.37}$$

$$P^{(1,0)}_{\nu_1,\nu_2}\left(0,\frac{k_2}{\Delta}\right) = 0 \quad \forall k_2 \in \mathbf{Z}_0. \tag{4.38}$$

We may better understand the requirements of this theorem by writing

$$\begin{aligned} E[\varepsilon_1\varepsilon_2] &= E[(q_1+\nu_1)(q_2+\nu_2)] \\ &= E[q_1q_2]+E[q_1\nu_2]+E[q_2\nu_1]+E[\nu_1\nu_2]. \end{aligned}$$

We know from Theorem 4.5 that $E[q_1q_2] = 0$ in general if Eq. (4.36) holds. Furthermore, it is not difficult to show using Theorem 3.3 when Eqs. (4.37) and (4.38) hold then $E[q_1\nu_2] = 0$ and $E[q_2\nu_1] = 0$, respectively. Thus when all three equations hold we obtain $E[\varepsilon_1\varepsilon_2] = E[\nu_1\nu_2]$. (Necessity follows from the arbitrariness of P_{x_1,x_2}.)

We observe that if an iid dither is chosen so that $P_{\nu_1,\nu_2}(u_1,u_2) = P_\nu(u_1)P_\nu(u_2)$, and if the dither is of order at least one, then the conditions of the theorem will be satisfied. This is not sufficient to ensure that the error is wide-sense stationary, however, since a dither of at least second order is required to render the error variance independent of the input.

If the conditions of Theorem 4.9 hold for all ν_1 and ν_2 separated in time by $\tau \neq 0$, then assuming that a dither of at least second order is used so that the variance of ε is given by Eq. (4.30), we have

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

Discrete-time Fourier transforming this expression yields

$$\mathrm{PSD}_\varepsilon(f) = \mathrm{PSD}_\nu(f) + \frac{\Delta^2 T}{6}, \tag{4.39}$$

where PSD_ε represents the power spectral density of the total error and PSD_ν represents that of the dither. Eq. (4.39) indicates that the total error spectrum is the sum of the dither spectrum and a white noise component of total power $\Delta^2/12$. This white component is sometimes referred to as the "quantization noise."

The conditions of the theorem will certainly hold if an iid dither of second or higher order is chosen, in which case the total error spectrum will be white.

4.4.2 NSD Systems: Statistics of the System Output

We now turn our attention to the system output of an NSD system. Eq. (3.8) gives the cf of this process as

$$P_y(u) = \sum_{k=-\infty}^{\infty} G_\nu\left(u - \frac{k}{\Delta}\right) P_x\left(u - \frac{k}{\Delta}\right) \tag{4.40}$$

and hence

$$E[y^m] = \sum_{k=-\infty}^{\infty} \sum_{r=0}^{m} \binom{m}{r} \left[\left(\frac{j}{2\pi}\right)^r G_\nu^{(r)}\left(\frac{k}{\Delta}\right)\right] \left[\left(\frac{j}{2\pi}\right)^{m-r} P_x^{(m-r)}\left(\frac{k}{\Delta}\right)\right]. \tag{4.41}$$

We also observe, for completeness, that the inverse Fourier transform of Eq. (4.40) is

$$p_y(y) = \left[\Delta \Pi_\Delta \star p_\nu \star p_x\right](y) W_\Delta(y).$$

Now, if the first m derivatives of $G_\nu(u)$ vanish at all non-zero multiples of $1/\Delta$, then Eq. (4.41) reduces to

$$E[y^m] = \sum_{r=0}^{m} \binom{m}{r} E[\varepsilon^r] E[x^{m-r}], \tag{4.42}$$

where the expectation values of the total error are given in terms of the expectation values of the dither by Eq. (4.31). If P_x is arbitrary, then the converse must also hold. By direct differentiation of $G_\nu(u)$, the above condition is easily shown to be equivalent to the condition of Theorem 4.8 with $M = m$. Expanding Eq. (4.42) for the particularly interesting cases of $m = 1, 2$ under the assumption that $E[\nu] = 0$ we obtain

$$E[y] = E[x], \tag{4.43}$$

$$E[y^2] = E[x^2] + E[\nu^2] + \frac{\Delta^2}{12}. \tag{4.44}$$

Proceeding in the usual fashion, we find that the joint moments of output values y_1 and y_2, separated in time by $\tau \neq 0$, are given by

$$\begin{aligned} E[y_1^{m_1} y_2^{m_2}] = & \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \sum_{r_1=0}^{m_1} \sum_{r_2=0}^{m_2} \binom{m_1}{r_1} \binom{m_2}{r_2} \left[\left(\frac{j}{2\pi} \right)^{r_1+r_2} G_{\nu_1,\nu_2}^{(r_1,r_2)} \left(\frac{k_1}{\Delta}, \frac{k_2}{\Delta} \right) \right] \\ & \times \left[\left(\frac{j}{2\pi} \right)^{(m_1-r_1)+(m_2-r_2)} P_{x_1,x_2}^{(m_1-r_1,m_2-r_2)} \left(\frac{k_1}{\Delta}, \frac{k_2}{\Delta} \right) \right]. \end{aligned} \tag{4.45}$$

If the indicated partial derivatives of G_{ν_1,ν_2} are zero at all non-zero multiples of $1/\Delta$ for $r_i = 1, 2, \ldots, m_i$ where $i \in \{1, 2\}$ (this corresponding to a second-order analogue of the condition of Theorem 4.8), then Eq. (4.45) reduces to

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

where the joint moments of the total error are given in terms of those of the dither by Eq. (4.35). In particular, note that if these conditions are satisfied for $m_1 = m_2 = 1$, then

$$E[y_1 y_2] = E[x_1 x_2] + E[\varepsilon_1 \varepsilon_2].$$

Then substituting the moment formulae Eqs. (4.31), (4.35) and (4.42) and assuming the system input is wide-sense stationary, we have

$$E[y_1 y_2](k) = \begin{cases} E[x^2] + 2E[x]E[\nu] + E[\nu^2] + \dfrac{\Delta^2}{12}, & k = 0, \\ E[x_1 x_2](k) + E[\nu_1 \nu_2](k), & \text{otherwise.} \end{cases} \tag{4.46}$$

Hence, under these conditions, if the dither signal has zero mean then

$$\mathrm{PSD}_y(f) = \mathrm{PSD}_x(f) + \mathrm{PSD}_\nu(f) + \frac{\Delta^2 T}{6} \tag{4.47}$$

so that the spectrum of the output is the sum of the input and dither spectra, apart from a white noise component of variance $\Delta^2/12$ contributed by the $k = 0$ term in Eq. (4.46).

4.4.3 NSD Systems: Properties of Practical Dither Signals

Recall that an nRPDF random process is one generated by the summation of n statistically independent zero-mean uniformly distributed random processes, each of 1 LSB peak-to-peak amplitude. We will prove the following:

Theorem 4.10 *In an NSD quantizing system, an nRPDF dither renders the first n moments of the total error process, $E[\varepsilon^m]$, $m = 0, 1, \ldots, n$, independent of the distribution of the system input, and results, for a zero-mean dither with $n \geq 2$, in a total error variance of $(n+1)\Delta^2/12$. Higher moments of the error signal will, however, remain input dependent.*

Proof: The addition of n statistically independent RPDF random processes convolves their pdf's, hence multiplying their cf's and yielding

$$G_\nu(u) = \left[\frac{\sin(\pi\Delta u)}{\pi\Delta u}\right]^{n+1},$$

the first n derivatives of which will consist entirely of terms containing non-zero powers of $\sin(\pi\Delta u)/(\pi\Delta u)$. Since this function goes to zero at the places required by Theorem 4.8, the first n moments of the error will be independent of the input distribution. If the dither has a mean value of zero, then its variance is the sum of the variances of the n independent uniformly distributed random processes of which it is the sum, so that, according to Eq. (4.30), the variance of the total error is $(n+1)\Delta^2/12$ whenever $n \geq 2$. Lemma C.2 from Appendix C shows that higher derivatives of G_ν will not vanish at the required locations, so that, by Theorem 4.7, higher error moments will not be rendered input independent when such dither is in use.

□

Furthermore, it is important to note that using rectangular-pdf dithers 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 of input-dependent error moments in [16]).

We proceed to examine two important examples of non-subtractive dither pdf's. First, consider a system using dither with a simple RPDF (of 1 LSB peak-to-peak amplitude):

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

for which

$$G_\nu(u) = \left[\frac{\sin(\pi\Delta u)}{\pi\Delta u}\right]^2.$$

The first three derivatives of this function are plotted in Fig. 4.4. The first derivative satisfies the condition of going to zero at the regularly spaced intervals stipulated by Eq. (4.28), while the second derivative and higher derivatives do not. This indicates that the first moment of the error signal is independent of the input, but that its variance and higher moments remain dependent.

These conclusions are borne out by the accompanying plots of *conditional moments*, representing the error moments as functions of a given input:

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

The required cpdf may be found by substituting $p_x(x) = \delta(x - x_0)$ into Eq. (4.25), yielding

$$p_\varepsilon(\varepsilon) = p_{\varepsilon|x}(\varepsilon, x_0) = [\Delta\Pi_\Delta \star p_\nu](\varepsilon)W_\Delta(\varepsilon + x_0). \tag{4.48}$$

The first conditional moment, or mean error, in Fig. 4.4 is zero for all inputs, indicating that the quantizer has been *linearized* by the use of this dither thus eliminating distortion. The error variance, on the other hand, is clearly signal-dependent, so that the noise power in the signal varies with the system input. This is sometimes referred to as *noise modulation,* and is undesirable in many applications, such as in audio where audible time-dependent error signals are considered intolerable.

Now consider a 2RPDF (TPDF) dither resulting from the sum of two independent 1RPDF processes:

$$p_\nu(\nu) = [\Pi_\Delta \star \Pi_\Delta](\nu). \tag{4.49}$$

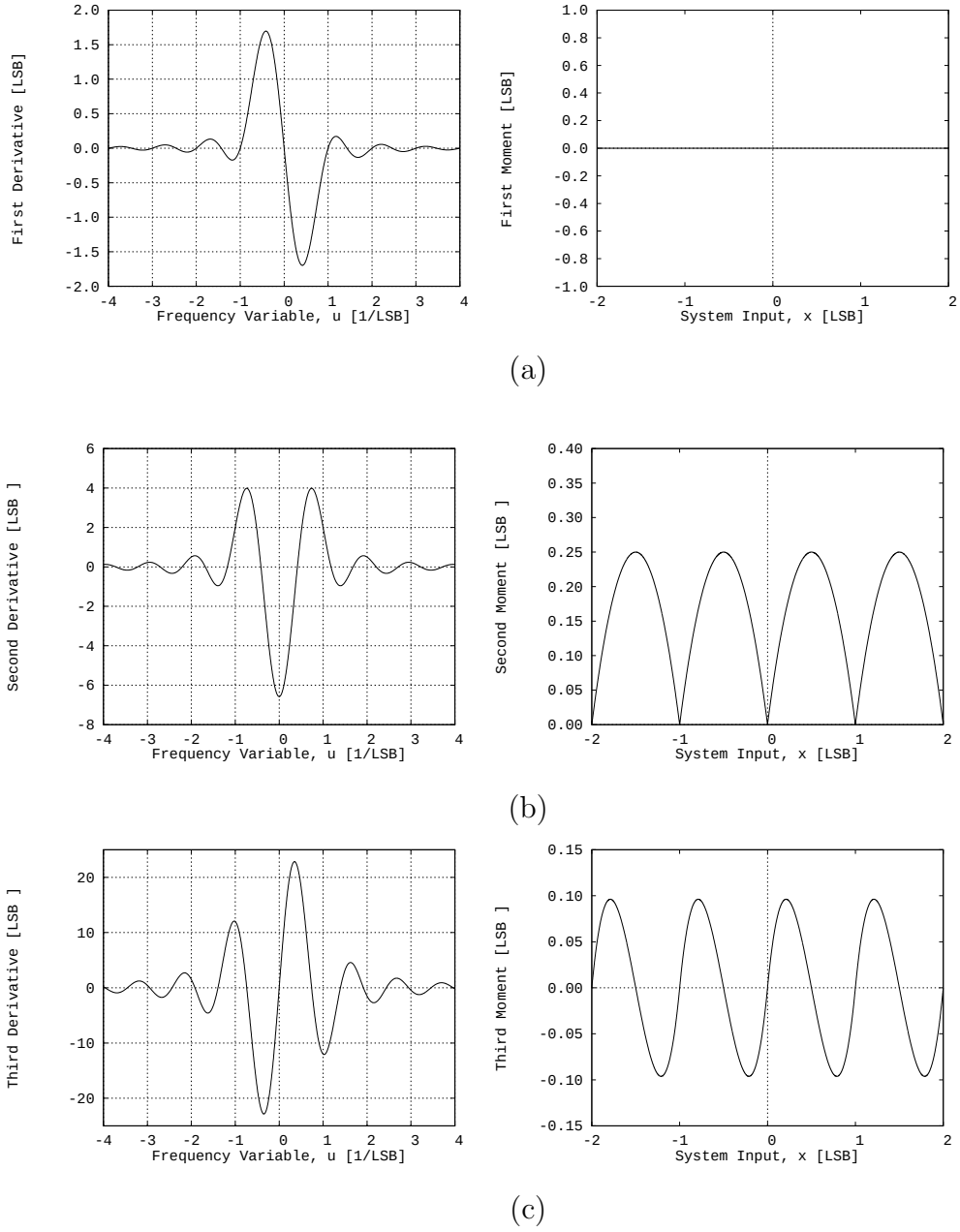


Figure 4.4: Derivatives of $G_\nu(u)$ (left) and conditional moments of the error (right) for a quantizer using 1RPDF dither: (a) $G_\nu^{(1)}(u)$ and $E[\varepsilon|x]$ (both in units of Δ), (b) $G_\nu^{(2)}(u)$ and $E[\varepsilon^2|x]$ (both in units of Δ^2), (c) $G_\nu^{(3)}(u)$ and $E[\varepsilon^3|x]$ (both in units of Δ^3). The frequency variable, u, is plotted in units of $1/\Delta$ and the input, x, in units of Δ.

In a system employing this kind of dither, $G_\nu(u)$ is given by

$$G_\nu(u) = \left[\frac{\sin(\pi\Delta u)}{\pi\Delta u}\right]^3 .$$

The first three derivatives of this function, and the corresponding moments as a function of the input, are plotted in Fig. 4.5. 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 of x. The second moment of the total error is a constant $\Delta^2/4$ for all inputs, in agreement with Eq. (4.30). In this case the use of an appropriate dither has eliminated both distortion and noise modulation. Higher derivatives of $G_\nu(u)$ do not meet the required conditions, so that higher moments of the error remain dependent on the input.

Using an argument derived from Wright [10, 11], we will now show that such 2RPDF dither is unique and optimal in the sense that it is the only zero-mean dither which renders the first and second moments of the total error input independent, while minimizing the second moment. That is, when used in an NSD quantizing system, this dither incurs the least possible increase in the total error variance of any dither which eliminates input-dependent distortion and noise modulation.

For 2RPDF dither with zero mean we know 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. 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 [30], this is sufficient to uniquely specify $P_\nu(u)$ for all u if $P_\nu(\nu)$ is Δ-bandlimited (i.e., if p_ν is supported such that $p_\nu(\nu) = 0$ for $|\nu| > \Delta$). Since

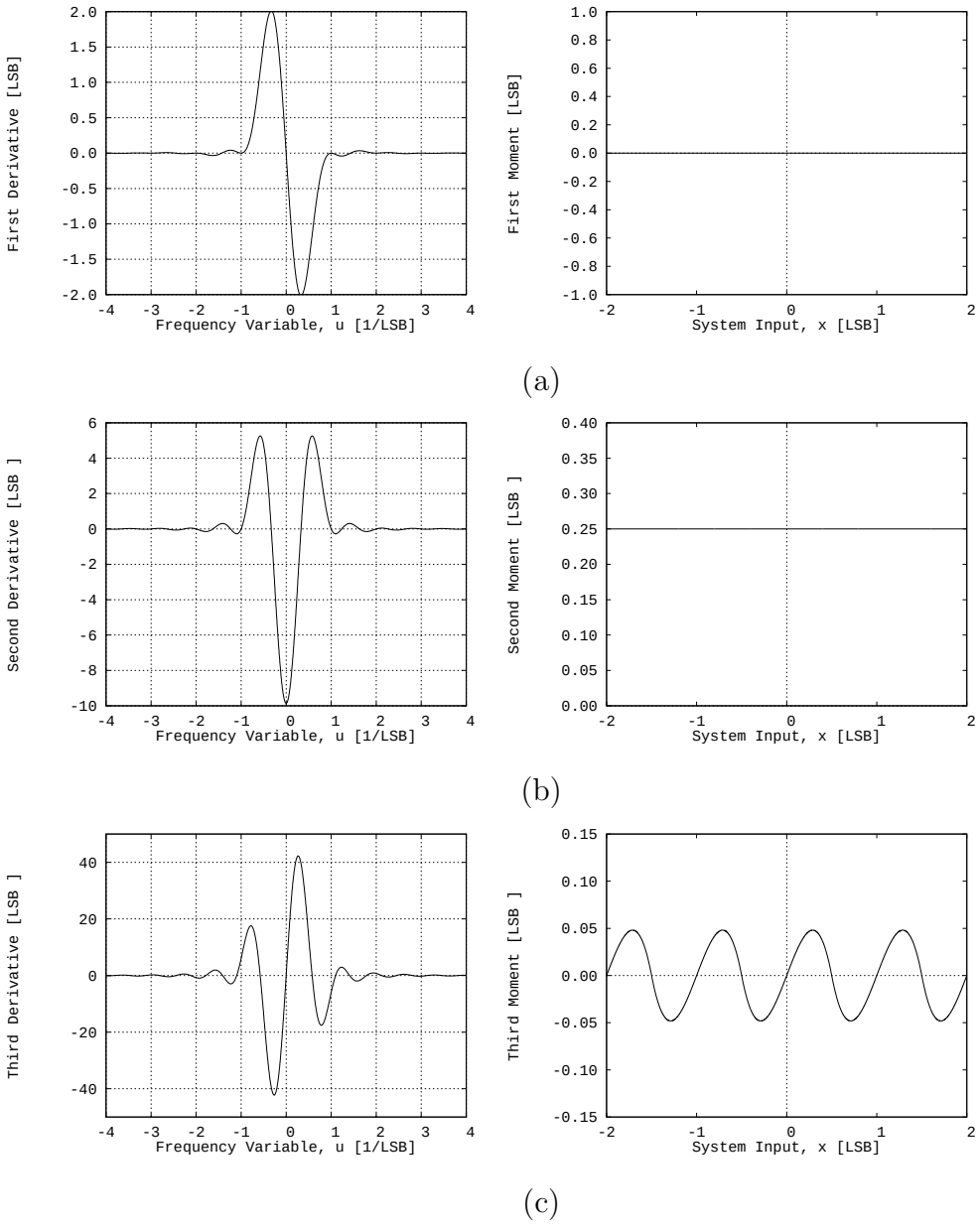


Figure 4.5: Derivatives of $G_\nu(u)$ (left) and conditional moments of the error (right) for a quantizer using 2RPDF dither: (a) $G_\nu^{(1)}(u)$ and $E[\varepsilon|x]$ (both in units of Δ), (b) $G_\nu^{(2)}(u)$ and $E[\varepsilon^2|x]$ (both in units of Δ^2), (c) $G_\nu^{(3)}(u)$ and $E[\varepsilon^3|x]$ (both in units of Δ^3). The frequency variable, u, is plotted in units of $1/\Delta$ and the input, x, in units of Δ.

the triangular dither pdf of Eq. (4.49) *is* thus supported, 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 non-zero outside of the interval $[-\Delta, \Delta]$ 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 an input value of $\Delta/2$; i.e., for $p_x(x) = \delta(x - \frac{\Delta}{2})$.

For $x = \Delta/2$, the cpdf of the total error, $p_{\varepsilon|x}(\varepsilon, x)$, is shown in Fig. 4.6(a). It consists of two equally weighted delta functions at $\varepsilon = \pm\Delta/2$ when 2RPDF dither is employed. Use of a wider dither pdf will result in the appearance of more delta functions in the error's cpdf, as shown in Fig. 4.6(b), where we denote the weighting of the delta function at $\varepsilon = \pm(2i-1)\Delta/2$, $i > 0$, by $e_{\pm i}$, so that

$$p_{\varepsilon|x}\left(\varepsilon, \frac{\Delta}{2}\right) = \sum_{i=1}^{\infty} \left[e_i \, \delta\Big(\varepsilon - (2i-1)\frac{\Delta}{2}\Big) + e_{-i} \, \delta\Big(\varepsilon + (2i-1)\frac{\Delta}{2}\Big) \right]. \tag{4.50}$$

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

$$(e_1 + e_{-1}) + \sum_{i=2}^{\infty} (e_i + e_{-i}) = 1. \tag{4.51}$$

Now, by direct integration of Eq. (4.50), we compute the conditional expectation

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

Substituting Eq. (4.51) yields

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

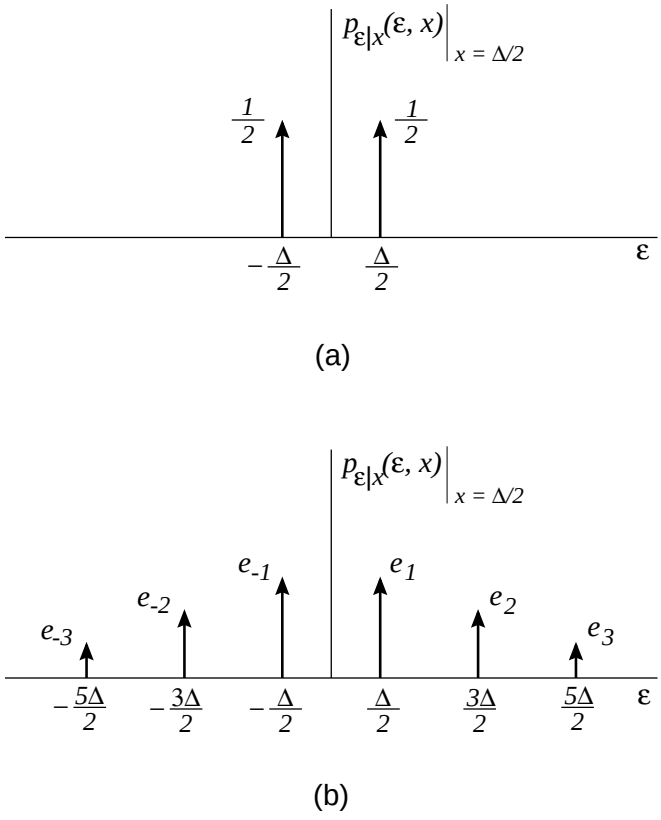


Figure 4.6: $p_{\varepsilon|x}(\varepsilon, x)$ evaluated at $x = \Delta/2$ for systems using (a) a triangular-pdf (2RPDF) dither of 2 LSB peak-to-peak amplitude and (b) a dither with wider pdf (the delta functions possess the indicated weightings).

which is always greater than $\Delta^2/4$ since the $e_{\pm i}$'s must be non-negative and some will be non-zero. We conclude the following:

Theorem 4.11 *The choice of zero-mean non-subtractive 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 2RPDF.*

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

The theorems of this chapter can also be applied to spectrally coloured dithers (i.e., ones for which $\mathrm{PSD}_\nu(f)$ is not a constant), but we will delay detailed investigation of such dithers until Chapter 5.

4.4.4 Summary of Non-Subtractive Dither

The results of greatest practical importance concerning NSD quantizing systems are reiterated below:

1. Non-subtractive dithering, unlike subtractive dithering, cannot render the total error statistically independent of the system input. It *can* render any desired moments of the total error independent of the input distribution provided that certain conditions on the cf of the dither are met (see Theorem 4.7). In particular, a dither of order n as defined in Section 2.3, such as 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. For instance, it can render the power spectrum of the total error signal white (see discussion following Eq. (4.39)).

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 4.4.2). 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. (4.30)).

5. 2RPDF (TPDF) dither incurs the least increase in the total error variance of any non-subtractive dither which eliminates input-dependent distortion and noise modulation.

Fig. 4.7 shows the results of a computer-simulated quantization operation performed upon a 1 kHz sine wave of 4.0 LSB peak-to-peak amplitude and using iid dither with the aforementioned triangular pdf. Shown are the system input and output, the total error, and the estimated power spectrum of the system output. Note that vestiges of the input signal are clearly visible in the total error waveform, indicating that the two signals are *not* statistically independent. Also, the time-waveform of the system output in Fig. 4.7(b) does not visually resemble a sine wave plus an independent additive noise. Surprising as it may seem, listening experiments [21] show that the total error signal of Fig. 4.7(c) *sounds* like a

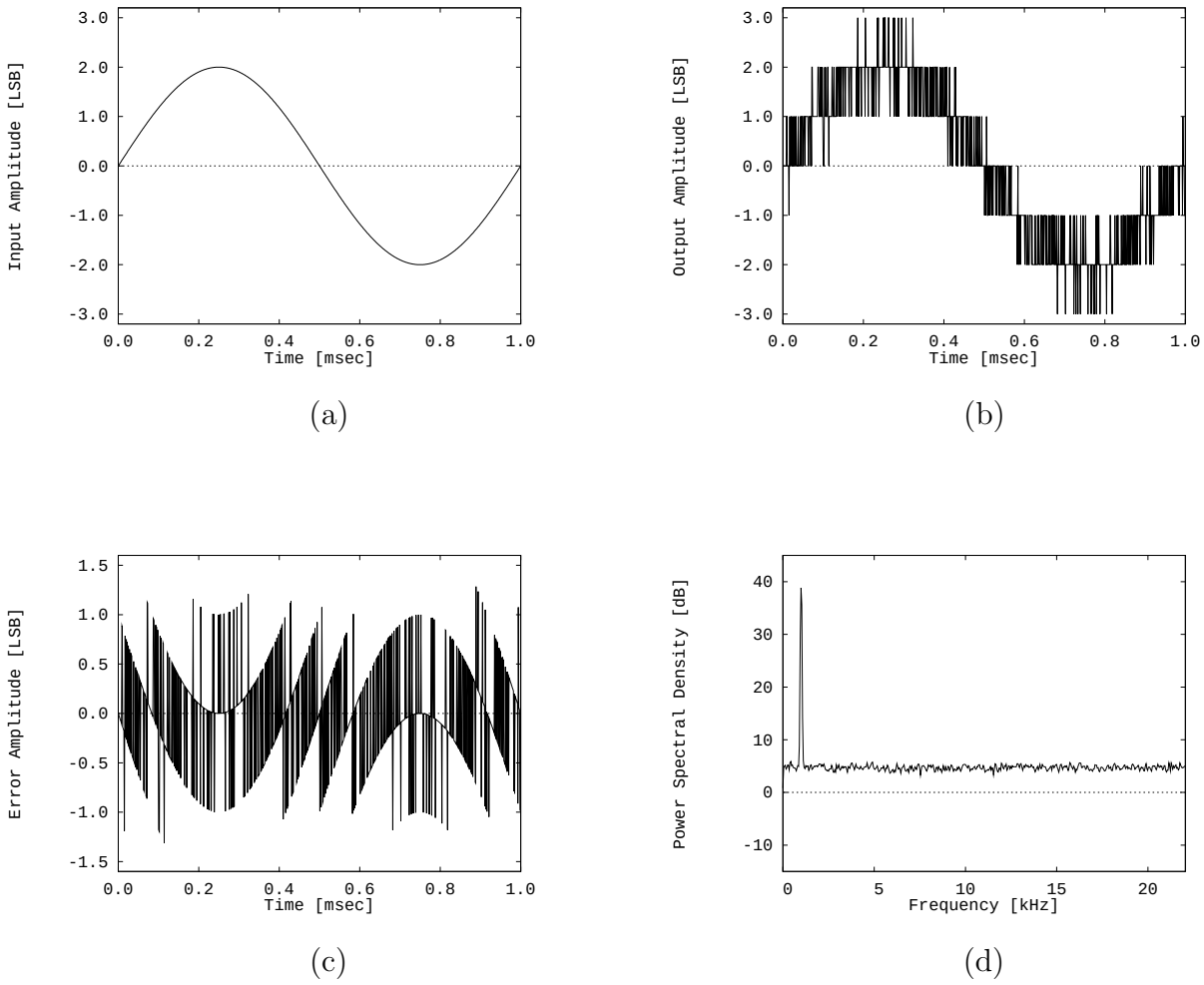


Figure 4.7: Results from the computer-simulated quantization of a 1 kHz sine wave of 4.0 LSB peak-to-peak amplitude using 2RPDF non-subtractive dither. Shown are (a) the system input signal, (b) the system output signal, (c) the resulting total error signal, and (d) the power spectrum of the system output signal (as estimated from sixty 50%-overlapping Hann-windowed 512-point time records with an assumed sampling frequency of 44.1 kHz; 0 dB represents a power spectral density of $\Delta^2 T/6$ where T is the sampling period).

constant white noise, independent of the nature of the input signal (with which it is indeed uncorrelated), and that the signal shown in Fig. 4.7(b) sounds identical to a noisy sine wave. Indeed, the estimated power spectrum of the system output in Fig. 4.7(d) exhibits no distortion components and indicates that the total error *is* spectrally white. These results should be compared with those in Figs. 4.1 and 4.3, which illustrate the results of quantizing a sine wave using undithered and SD systems, respectively. In particular, it should be noted that the noise floor in Fig. 4.7(d) is up by 4.8 dB relative to that of Fig. 4.3(d) due to the tripling of the noise spectral density in accordance with Eq. (4.47).

In audio applications, the PSD of the total error is perceptually meaningful and should be input independent. In particular the error should have zero mean, and *noise modulation* (i.e., variation in the second error moment) should be eliminated, so that a dither of at least second order should be used. In image processing, some evidence exists [13] that the third moment of the total error may be perceptually relevant and should perhaps be controlled by using third order dither. In instuments measuring parameters which depend on higher statistical moments, still higher order dithers may be appropriate.

Some specific comment is required concerning the special nature of *re*quantization. In a purely digital system, random processes exhibiting the continuous pdf's described in this section are not, strictly speaking, available since not all real numbers are representable using a finite number of binary digits. In fact, digital dither pdf's of necessity resemble discretized or "sampled" versions of the continuous pdf's (rectangular, triangular, etc.) described above. It is not immediately obvious that such dithers will retain the desirable properties of their analogue counterparts with respect to rendering total error moments independent of the system input. It is

rigorously proven in Chapter 6 that such dithers *do indeed* retain these properties, and empirical evidence corroborating this conclusion may be found in [16].

The question has been posed [41, 42] as to the extent to which real-time estimation of attributes of dithered quantizing systems proceeds in the same fashion as for signals with additive iid random noise processes. Readers interested in the similarity of the two cases are referred to the treatment of this question provided in Appendix B.

4.5 Summary of Statistical Relationships Between Signals

Fig. 4.8 indicates the statistical dependences between the signals indicated in Fig. 2.1 with and without the application of a first or higher order dither and under the assumption that that ν and x are statistically independent processes. Signals not rendered independent of one another by a first order dither are not so rendered by the use of higher order dithers. All entries in the charts were arrived at by inspecting the relevant joint cf's to determine whether a particular choice of dither cf would allow them to be written as a product[3]. For instance, let us consider NSD systems and take, by way of example, the pair of signals q and ε. Can these random variables ever be statistically independent in an NSD system?

[3]We point out that w' and y are identical in NSD systems (see Fig. 2.1), as are q and ε in SD systems, so that the corresponding entries in the charts are identical.

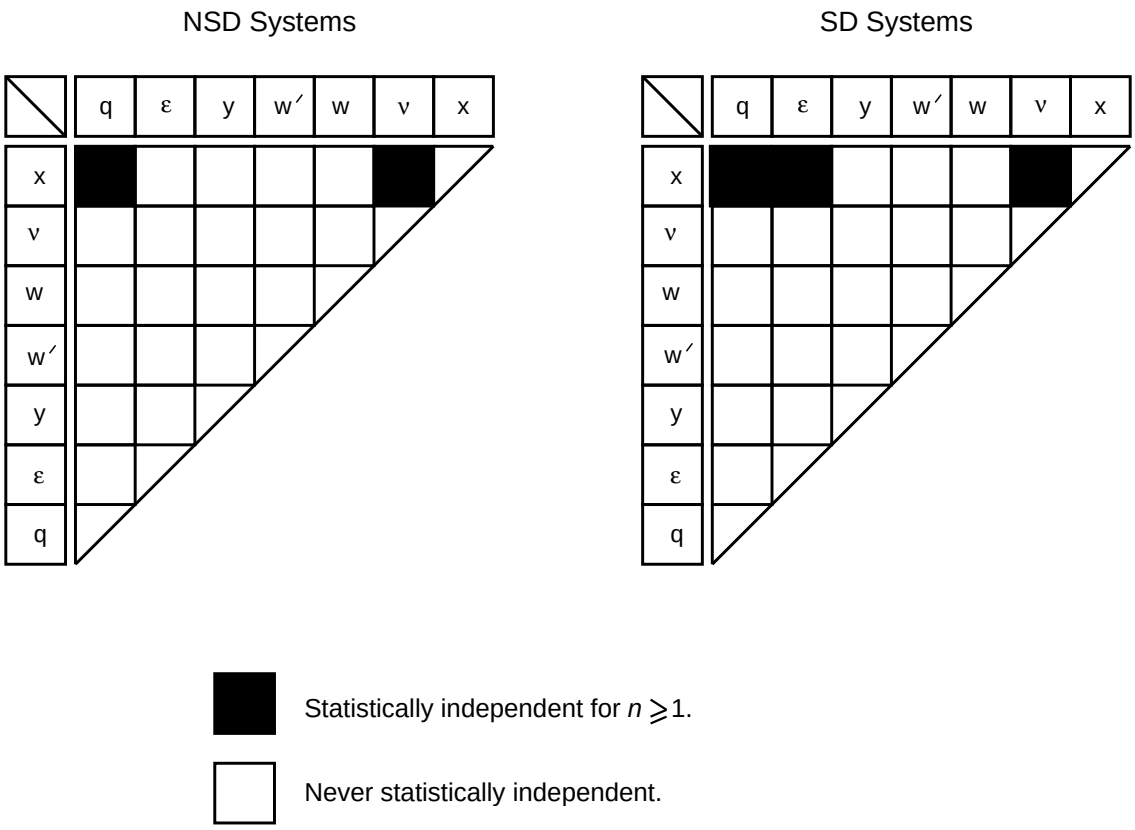


Figure 4.8: Statistical dependences between signals in SD and NSD quantizing systems where the dither and input signals are assumed to be statistically independent. (n refers to the order of the applied dither.)

Setting all unwanted variables to zero and simplifying Eq. (3.8), we see that

$$P_{q,\varepsilon}(u_q, u_\varepsilon) = \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_q + u_\varepsilon - \frac{k}{\Delta}\right) P_{\nu,x}\left(u_\varepsilon - \frac{k}{\Delta}, -\frac{k}{\Delta}\right).$$

Now, since the sum of u_q and u_ε occurs in the argument of the non-linear sinc function, it is clear that this expression can never be split into a product of a function involving u_q alone with one involving u_ε alone. Hence these random variables can never be statistically independent. This is obviously true for any other pair of random variables whose Fourier transform variables occur together in a function argument in Eq. (3.8).

Now consider the pair of random variables ε and x, and let us suppose that ν and x are statistically independent. The proof of Theorem 4.6 demonstrates that no choice of dither statistics can render ε and x statistically independent for arbitrarily distributed inputs. A similar conclusion is reached for the signal pairs (q, ν) and (q, w), although it can be shown that each of these pairs of random processes can be rendered *uncorrelated* by an appropriate choice of dither (see below).

We are left with only two signal pairs which might potentially be independent. These are (q, x) and (ν, x). ν is independent of x since we have specified that this is the case. Then Theorem 4.4 indicates that q and x are statistically independent if and only if

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

i.e., if a dither of order at least one is used. Combining all of the above considerations and conducting a similar analysis for SD systems allows construction of Fig. 4.8.

At the risk of belabouring the point, we observe that q is statistically independent of x in both SD and NSD systems if a first order dither is used. This is

especially good news in SD systems because the quantization error, q, and the total error of the system, ε, are identical. In NSD systems this is not true, however, and the total error is *never* statistically independent of the input for arbitrary input distributions. As we have seen, however, certain *moments* of the total error in an NSD system can be rendered independent of the system input distribution.

Fig. 4.9 indicates the correlation between various signals with different orders of dither assuming that ν and x are statistically independent. The charts were constructed by explicitly differentiating the relevant joint cf's and inspecting the results for conditions on the dither cf's which would render the corresponding random variables uncorrelated. As an example, consider ε and ν in an SD system. We are interested in conditions under which

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

From Eq. (3.7) we have

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

so that

$$\begin{aligned} E[\varepsilon\nu] &\triangleq \left(\frac{j}{2\pi}\right)^2 P_{\varepsilon,\nu}^{(1,1)}(0,0) \\ &= \left(\frac{j}{2\pi}\right)^2 \sum_{k=-\infty}^{\infty} \operatorname{sinc}^{(1)}\left(-\frac{k}{\Delta}\right) P_\nu^{(1)}\left(-\frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right). \end{aligned} \tag{4.52}$$

Furthermore,

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

so that

$$\begin{aligned} E[\varepsilon] &\triangleq \left(\frac{j}{2\pi}\right) P_\varepsilon^{(1)}(0) \\ &= \left(\frac{j}{2\pi}\right) \sum_{k=-\infty}^{\infty} \operatorname{sinc}^{(1)}\left(-\frac{k}{\Delta}\right) P_\nu\left(-\frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right). \end{aligned} \tag{4.53}$$

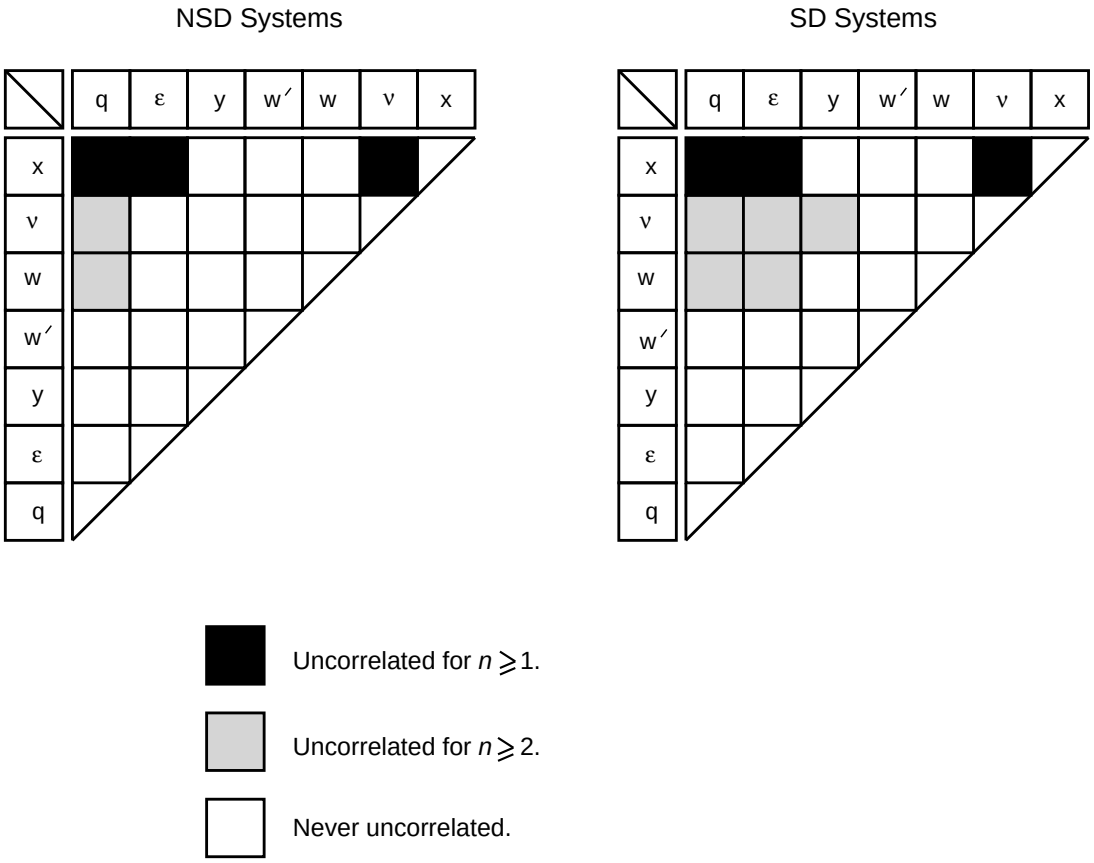


Figure 4.9: Statistical correlations between signals in SD and NSD quantizing systems where the dither and input signals are assumed to be statistically independent. (n refers to the order of the applied dither.)

Now, if

$$P_{\nu}^{(i)}\left(\frac{k}{\Delta}\right) = 0, \qquad \forall k \in \mathbf{Z}_0, \forall i \in \{0,1\}, \tag{4.54}$$

then Eqs. (4.52) and (4.53) both reduce to zero so that

$$E[\varepsilon\nu] = E[\varepsilon]E[\nu] = 0.$$

(The derivative of the sinc function vanishes at the origin, thereby taking care of the $k = 0$ case.) Thus, ε and ν are both uncorrelated and orthogonal. This analysis was repeated for all pairs of signals that were of interest in both SD and NSD systems in order to generate Fig. 4.9.

If the assumption is made that $E[\nu] = 0$ then this chart can be taken to indicate orthogonality as well as uncorrelatedness, in which case it can be used to deduce the variances of signals of interest. For instance, note from Fig. 2.1 that in an NSD system

$$\begin{aligned} E[\varepsilon^2] &= E[(\nu + q)^2] \\ &= E[\nu^2] + 2E[\nu q] + E[q^2]. \end{aligned}$$

From Fig. 4.9 we see that $E[\nu q] = 0$ if a zero mean dither of second or higher order is used. In this case, using $E[q^2] = \Delta^2/12$ we obtain

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

which is precisely Eq. (4.30). Furthermore, in such a system we can go on to write

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

where we have noted from Fig. 4.9 that $E[x\varepsilon] = 0$ for zero mean dithers of order $n \geq 1$. Note that in order to substitute for $E[\varepsilon^2]$, however, we require that the dither be of order $n \geq 2$. Indeed, $E[\varepsilon^2]$ is not independent of the system input otherwise. It should be observed that the expression obtained in this manner is identical to Eq. (4.44).

It turns out that this approach can be used to deduce the variance of any signal in an SD or NSD system in terms of the variances of x, ν and q provided an appropriate dither is in use.

Chapter 5

Coloured Errors and Multi-Channel Systems

This chapter will consider four topics related to discrete-time dithered quantizing systems: the use of spectrally coloured (i.e., non-white) dither signals, dither in systems using noise-shaping error feedback, the raw error of an SD system, and the efficient generation of multi-channel dither signals.

5.1 Spectrally Shaped Dithers

We now proceed to apply the analysis of the last chapter to a large family of wide-sense stationary but spectrally-shaped (i.e., non-white) dither signals of practical interest [25]. We will consider the family of dithers whose n-th sample can be

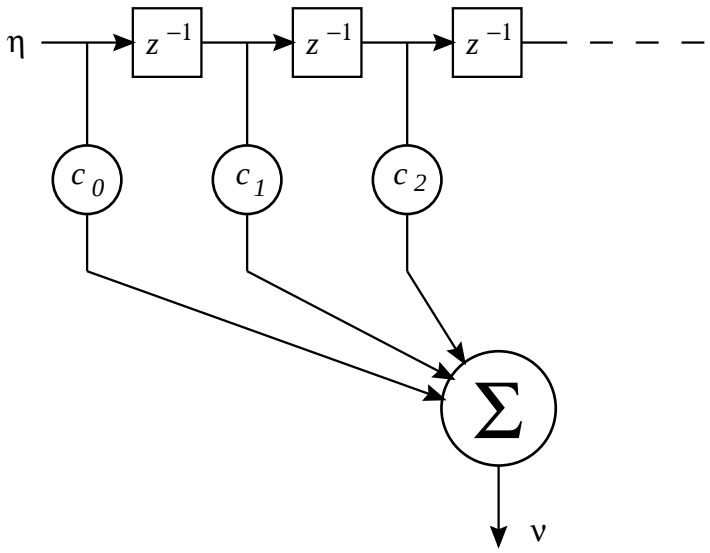


Figure 5.1: Schematic of a dither generator for producing spectrally shaped dithers.

written as

$$\nu_n = \sum_{i=-\infty}^{\infty} c_i \eta_{n-i} \tag{5.1}$$

where the η_i's are iid so that together they represent a strict-sense stationary random process, η. It will be tacitly assumed that $c_i = 0$ for $i < 0$, so that ν corresponds to the output of a causal non-recursive *dither filter*, G, of the form

$$G(z) = \sum_{i=0}^{\infty} c_i z^{-i}$$

with η as its input (see Fig. 5.1). η is also assumed to be statistically independent of the system input x, so that ν is as well. We will hereafter refer to such a dither as a *filtered dither.*

The objective is to find conditions such that dithers in this particular family will render the total error spectrum independent of the system input in a dithered

quantizing system. That is, we require that $E[\varepsilon^2]$ and $E[\varepsilon_1\varepsilon_2]$ are constant so that the autocorrelation function of the total error is input independent. We proceed by finding the characteristic functions required in order to use the theorems given in Chapter 4.

We begin by defining the vectors

$$\nu \stackrel{\triangle}{=} (\dots, \nu_{-1}, \nu_0, \nu_1, \dots)$$

and

$$\eta \stackrel{\triangle}{=} (\dots, \eta_{-1}, \eta_0, \eta_1, \dots).$$

Now we write the joint pdf

$$\begin{aligned} p_{\nu,\eta}(\nu,\eta) &\stackrel{\triangle}{=} p_{\nu|\eta}(\nu,\eta)p_\eta(\eta) \\ &= \prod_{j=-\infty}^{\infty} \delta\left(\nu_j - \sum_{i=-\infty}^{\infty} c_i \eta_{j-i}\right) p_\eta(\eta_j). \end{aligned}$$

Here we have used the facts that ν_j is completely determined by choosing the η_i's and that the η_i's are iid so that their joint pdf splits into a product of identical functions which we will simply denote by p_η; i.e.,

$$p_{\eta_i} \equiv p_\eta \quad \forall i.$$

To obtain the associated cf, we now Fourier transform all variables. The transform variable corresponding to ν_j will be u_j and that corresponding to η_i will be w_i, where, as above, we will form real vectors u and w from these components for

notational convenience.

$$\begin{aligned} P_{\nu,\eta}(u,w) &= \prod_{j=-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-j2\pi u_j \sum_{i=-\infty}^{\infty} c_i \eta_{j-i}\right) p_\eta(\eta_j) e^{-j2\pi\eta_j w_j} d\eta_j \\ &= \prod_{j=-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{i=-\infty}^{\infty} e^{-j2\pi u_j c_i \eta_{j-i}} p_\eta(\eta_j) e^{-j2\pi\eta_j w_j} d\eta_j \\ &= \prod_{i=-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{j=-\infty}^{\infty} e^{-j2\pi c_{j-i} u_j \eta_i} p_\eta(\eta_i) e^{-j2\pi\eta_i w_i} d\eta_i \\ &= \prod_{i=-\infty}^{\infty} P_\eta\left(w_i + \sum_{j=-\infty}^{\infty} c_{j-i} u_j\right). \end{aligned} \tag{5.2}$$

Setting $w_i = 0\ \forall i$ and $u_j = 0\ \forall j \neq n$ we directly obtain the cf we require:

$$P_{\nu_n}(u_n) = \prod_{i=-\infty}^{\infty} P_\eta(c_{n-i} u_n).$$

Since ν is strict-sense stationary, we will drop the unneeded time-index n and re-index the c's:

$$P_\nu(u) = \prod_{i=-\infty}^{\infty} P_\eta(c_i u). \tag{5.3}$$

Also, by setting to zero all of the w_i's and all of the u_j's except for u_n and $u_{n+\ell}$ (which we relabel u_1 and u_2), Eq. (5.2) yields

$$\begin{aligned} P_{\nu_n,\nu_{n+\ell}}(u_1,u_2) &= \prod_{i=-\infty}^{\infty} P_\eta(c_{n-i} u_1 + c_{n+\ell-i} u_2) \\ &= \prod_{i=-\infty}^{\infty} P_\eta(c_i u_1 + c_{i+\ell} u_2). \end{aligned} \tag{5.4}$$

Differentiation of Eqs. (5.3) and (5.4) and making the simplifying assumption that $E[\eta] = 0$ gives

$$r_\nu(\ell) = E[\eta^2] \sum_{j=-\infty}^{\infty} c_j c_{j+\ell}$$

and

$$\mathrm{PSD}_\nu(f) = 2TE[\eta^2] \left\{ \sum_{j=-\infty}^{\infty} c_j^2 + 2 \sum_{\ell=1}^{\infty} \sum_{j=-\infty}^{\infty} c_j c_{j+\ell} \cos(2\pi\ell T f) \right\}. \tag{5.5}$$

5.1.1 Filtered Dithers in NSD Systems

We return to Theorem 4.8 in order to see what demands it places upon the cf's derived above. We begin with the case of the error mean ($m = 1$), which entails the requirement that

$$P_\nu\left(\frac{k}{\Delta}\right) = 0 \qquad \forall k \in \mathbf{Z}_0 \tag{5.6}$$

in order that this quantity be independent of the input and given by Eq. (4.29). Clearly, this condition will be satisfied by the dither of Eq. (5.3) if and only if for each $k \in \mathbf{Z}_0$ there exists at least one value of i such that:

$$P_\eta\left(c_i\frac{k}{\Delta}\right) = 0.$$

Requiring that the error variance be input independent introduces an additional constraint:

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

From Eq. (5.3) we have

$$P_\nu^{(1)}\left(\frac{k}{\Delta}\right) = \sum_{j=-\infty}^{\infty} c_j P_\eta^{(1)}\left(c_j\frac{k}{\Delta}\right) \prod_{\substack{i=-\infty \\ i\neq j}}^{\infty} P_\eta\left(c_i\frac{k}{\Delta}\right). \tag{5.7}$$

This expression will go to zero at the required locations if for each $k \in \mathbf{Z}_0$ either

1. there exists an i such that

$$P_\eta^{(1)}\left(c_i\frac{k}{\Delta}\right) = 0$$

and

$$P_\eta\left(c_i\frac{k}{\Delta}\right) = 0,$$

or

2. there exist two distinct values of i such that

$$P_\eta\left(c_i\frac{k}{\Delta}\right) = 0$$

so that, although terms occur in Eq. (5.7) in which either one of these two functions alone will be differentiated, in any given term one will be undifferentiated and will cause the respective term to vanish in the required places.

We now proceed to address the question of what conditions ensure that a spectrally shaped dither will render the total error spectrum input independent. As usual, we approach the question by investigating correlations between errors separated in time; e.g., $E[\varepsilon_{n_1}\varepsilon_{n_2}]$. To apply Theorem 4.9 we use $P_{\nu_n,\nu_{n+\ell}}$ as given by Eq. (5.4). We proceed by treating separately the three conditions required by the theorem.

Condition I (Eq. (4.36)) is satisfied for all lags $\ell \in \mathbf{Z}_0$ if and only if $\forall(k_1,k_2) \in \mathbf{Z}_0^2$ and $\forall \ell \in \mathbf{Z}_0$ there exists an i such that

$$P_\eta\left(c_i\frac{k_1}{\Delta} + c_{i+\ell}\frac{k_2}{\Delta}\right) = 0.$$

Note that if this equation holds, then Eq. (5.6) necessarily holds as well.

Proceeding to Condition II (Eq. (4.37)) we have:

$$P^{(0,1)}_{\nu_n,\nu_{n+\ell}}\left(\frac{k_1}{\Delta},0\right) = \sum_{j=-\infty}^{\infty} c_{j+\ell}P^{(1)}_\eta\left(c_j\frac{k_1}{\Delta}\right)\prod_{\substack{i=-\infty\\ i\neq j}}^{\infty} P_\eta\left(c_i\frac{k_1}{\Delta}\right).$$

All terms in this sum will go to zero at the required locations $\forall \ell \in \mathbf{Z}_0$ under the same condition that we found for constancy of the error variance above; that is, we

require $\forall k \in \mathbf{Z}_0$ that either $P_\eta(c_i k/\Delta) = 0$ and $P_\eta^{(1)}(c_i k/\Delta) = 0$ for some value of i, or $P_\eta(c_i k/\Delta) = 0$ for any two values of i.

Condition III (Eq. (4.38)) is symmetric with Condition II and yields the same conditions on the cf of η.

Collecting the above conditions yields the following set of sufficient conditions for the error spectrum to be constant and input independent:

Theorem 5.1 *In an NSD quantizing system using filtered dither the total error will be wide-sense stationary and independent of the system input under the following conditions:*

1. $\forall (k_1, k_2) \in \mathbf{Z}_0^2$ *and for each* $\ell \in \mathbf{Z}_0$ *there exists an* i *such that*

$$P_\eta\left(c_i \frac{k_1}{\Delta} + c_{i+\ell} \frac{k_2}{\Delta}\right) = 0, \tag{5.8}$$

and

2. *for each* $k \in \mathbf{Z}_0$, *either there exists a value of* i *such that*

$$P_\eta\left(c_i \frac{k}{\Delta}\right) = 0 \tag{5.9}$$

and

$$P_\eta^{(1)}\left(c_i \frac{k}{\Delta}\right) = 0, \tag{5.10}$$

or there exist two distinct values of i *such that*

$$P_\eta\left(c_i \frac{k}{\Delta}\right) = 0. \tag{5.11}$$

Subject to the conditions of the theorem, we have

$$E[\varepsilon_n \varepsilon_{n+\ell}] = \begin{cases} E[\varepsilon_n^2], & \ell = 0, \\ E[\nu_n \nu_{n+\ell}], & \text{otherwise.} \end{cases}$$

so that

$$\mathrm{PSD}_\varepsilon(f) = \mathrm{PSD}_\nu(f) + \frac{\Delta^2 T}{6}. \tag{5.12}$$

The conditions in the Theorem 5.1 are sufficient but not necessary, with more complicated and general conditions probably existing. In spite of this, the conditions of this theorem are so general as to be difficult to use, but they are the form required for certain η pdf's (see [43]). Here, let us interpret them in the common case where η represents a strict-sense stationary mRPDF random process.

If the η's are iid and mRPDF, then Condition 1 of Theorem 5.1 will be satisfied $\forall (k_1, k_2) \in \mathbf{Z}_0^2$ if for each $\ell \in \mathbf{Z}_0$ there exists an i, call it i_0, such that of c_{i_0} and $c_{i_0+\ell}$ one is zero and the other is a non-zero integer. To see why this is, note that for an η of this sort Eq. (5.8) involves

$$P_\eta\left(c_i \frac{k_1}{\Delta} + c_{i+\ell} \frac{k_2}{\Delta}\right) = \mathrm{sinc}^m\left(c_{i_0} \frac{k_1}{\Delta} + c_{i_0+\ell} \frac{k_2}{\Delta}\right).$$

This equation must hold if both $k_1 \neq 0$ and $k_2 \neq 0$ since the argument of the sinc function will then be a non-zero integer multiple of $1/\Delta$ under the above condition. What happens in the case where $c_{i_0} = 0$ and $k_2 = 0$ ($k_1 \neq 0$)? Then there exists $i_1 = i_0 + \ell$ such that Eq. (5.8) holds and becomes

$$\mathrm{sinc}^m\left(c_{i_1} \frac{k_1}{\Delta}\right) = 0.$$

A similar factor exists if $c_{i_0+\ell} = 0$ and $k_1 = 0$ ($k_2 \neq 0$). Hence for each pair $(k_1, k_2) \in \mathbf{Z}_0^2$ there exists, under the stated condition, an i such that Eq. (5.8) holds.

What does Condition 2 of Theorem 5.1 entail when η is mRPDF with $m \geq 1$? In such a case, we see that the existence of two distinct c_i's with values which are non-zero integers is sufficient to satisfy the requirement of Eq. (5.11). If, on the other hand, η is mRPDF with $m \geq 2$ then it is sufficient that one non-zero integral c_i exist to satisfy the requirements of both Eq. (5.9) and Eq. (5.10). For instance, the cf of a 2RPDF process,

$$P_\eta(u) = \mathrm{sinc}^2(u),$$

goes to zero at $u = c_i k/\Delta, \forall k \in \mathbf{Z}_0$ if $c_i \in \mathbf{Z}_0$, and so does its first derivative.

We collect these conclusions into the following useful corollary to Theorem 5.1.

Corollary 5.1 *In an NSD quantizing system using filtered dither with η being an iid mRPDF random process, the total error will be wide-sense stationary and independent of the system input with a PSD given by Eq. (5.12) 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 mRPDF with $m \geq 1$ and there exist at least two distinct values of i such that c_i is a non-zero integer, or η is mRPDF with $m \geq 2$ and there exists at least one value of i such that c_i is a non-zero integer.*

Note that the above conditions are sufficient but not necessary. On the other hand, Eq. (5.4) reveals that a necessary (but *not* sufficient) condition is that there must exist at least one value of i such that c_i is a non-zero integer.

Consider a system with a stationary RPDF η signal. What sets of dither filter coefficients satisfy the above conditions? Obviously, the requirements are met by the dither filter coefficients

$$\{1, -1\},$$

(where we have omitted the infinite sequences of zeros preceding and following the coefficients shown). This coefficient set is associated with a dither whose spectrum has a simple highpass form, as given by Eq. (5.5):

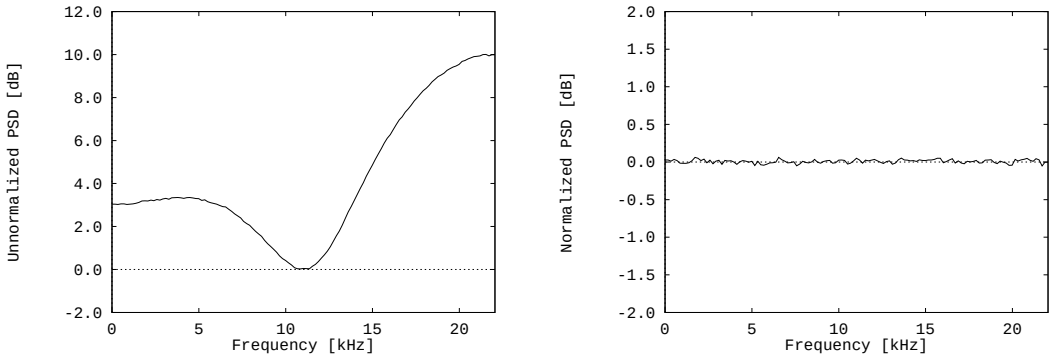
$$\mathrm{PSD}_\nu(f) = \frac{\Delta^2 T}{3} \left\{1 - \cos(2\pi T f)\right\}.$$

Also, the coefficient sequences

$$\left\{1, -1, \frac{1}{2}, -\frac{1}{2}\right\},$$
$$\left\{\frac{1}{2}, -\frac{1}{2}, 1, -1\right\},$$
$$\left\{1, -\frac{1}{2}, 1, -\frac{1}{2}\right\},$$
$$\left\{\frac{1}{2}, -1, \frac{1}{2}, -1\right\},$$
$$\left\{1, -\frac{1}{2}, \frac{1}{2}, -1\right\},$$
$$\left\{1, -\frac{1}{2}, 0, \frac{1}{2}, -1\right\},$$

all satisfy the requirements. Fig. 5.2 is in agreement with this conclusion. It shows the output error spectrum from a system using the fourth in this list of dithers with a null system input, as well as that error spectrum normalized by the error PSD as predicted by Eq. (5.12) for a properly dithered system[1]. The result of the

[1]All power spectra shown in this chapter represent the average of 12000 256-point FFT's of 50%-overlapping Hann-windowed data generated by computer-simulated quantization. 0 dB represents the PSD of a random process whose values are RPDF and iid; i.e., 0 dB represents $\Delta^2 T/6$.



(a) (b)

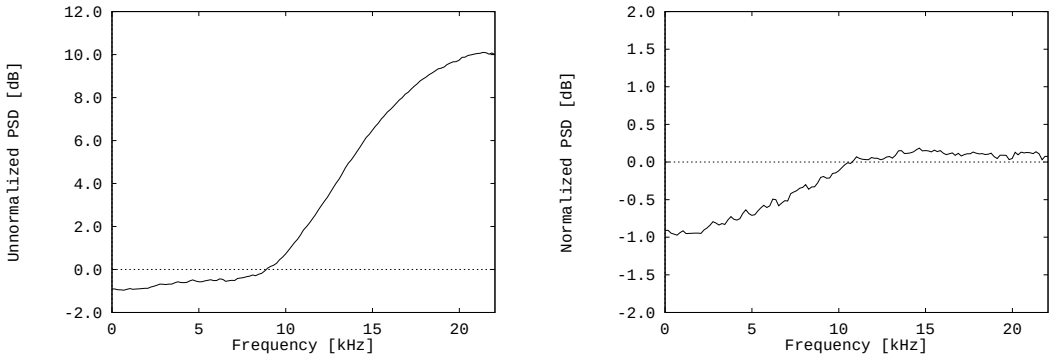
Figure 5.2: $PSD_\varepsilon(f)$ for a NSD quantizing system and using a dither filter with RPDF input, η, and coefficients {0.5,-1.0,0.5,-1.0}. The system was presented with a static null input (0.0 LSB). (a) Observed PSD, (b) observed PSD normalized by expected PSD.

normalization is flat, indicating that the spectrum is of the expected shape. On the other hand,

$$\left\{\frac{1}{2}, -1, 1, -\frac{1}{2}\right\}$$

does not meet Part 1 of the condition for $\ell = \pm 1$. Fig. 5.3 shows the error spectrum from a system using this sort of dither with a null system input, along with that spectrum normalized by Eq. (5.12). The results of the normalization are not flat, indicating that the error spectrum is not of the sort predicted.

As a final note, we observe that in NSD systems we cannot generate arbitrary total error spectra by varying the dither spectrum, since Eq. (5.12) indicates that an additive white noise component will always be present. The are many applications where more complete control of the error spectrum is desirable, and this may be achieved using noise-shaping error feedback (see Section 5.2). Spectrally shaped dithers remain of interest in certain applications, however (see Section 5.3). Furthermore, they are useful in high speed applications where it is prohibitively



(a) (b)

Figure 5.3: $PSD_\varepsilon(f)$ for an NSD quantizing system using a dither filter with RPDF input, η, and coefficients $\{0.5,-1.0,1.0,-0.5\}$. The system was presented with a static null input (0.0 LSB). (a) Observed PSD, (b) observed PSD normalized by expected PSD.

time-consuming to generate nRPDF dither using n newly calculated random numbers per data sample. In such cases, a single new η may be generated per sample and placed in a delay line to generate spectrally shaped dither of the sort described by Eq. (5.1). A commonly used example is the simple highpass dither mentioned above, which may be generated using dither filter coefficients

$$\{1, -1\}.$$

Such dither is 2RPDF, but only one new random number is calculated per sampling period.

5.1.2 Filtered Dithers in SD Systems

Let us compare the above results for NSD systems with analogous ones for SD systems. We require only that Eq. (4.18) be satisfied. This is the same requirement as imposed by Eq. (4.36) and so leads directly to the following theorem:

Theorem 5.2 *In an SD quantizing system using filtered dither, the total error will be strict-sense stationary with its PSD given by*

$$PSD_\varepsilon = \frac{\Delta^2 T}{6} \tag{5.13}$$

if and only if for each pair $(k_1, k_2) \in \mathbf{Z}_0^2$ *and for each* $\ell \in \mathbf{Z}_0$ *there exists an* i *such that*

$$P_\eta\left(c_i \frac{k_1}{\Delta} + c_{i+\ell}\frac{k_2}{\Delta}\right) = 0.$$

The condition here is, of course, just the first condition of Theorem 5.1.

Again, conditions specifically for nRPDF η's can be derived. Note that the condition of the following corollary to Theorem 5.2 is precisely the first condition of Corollary 5.1.

Corollary 5.2 *In an SD quantizing system using filtered dither with* η *being an iid* n*RPDF random process, the total error will be wide-sense stationary and independent of the system input with a PSD given by Eq. (5.13) if 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.*

Of course, there exist filter coefficient sequences, $\{c_i\}$, which satisfy the conditions of Corollary 5.2 without satisfying those of Corollary 5.1. That is to say that just because filtered dither is suitable for an SD system does not imply that it is suitable for an NSD system. One example is

$$\left\{\frac{1}{2}, 1\right\}$$

which is certainly not a suitable dither for an NSD system. Fig. 5.4, however,

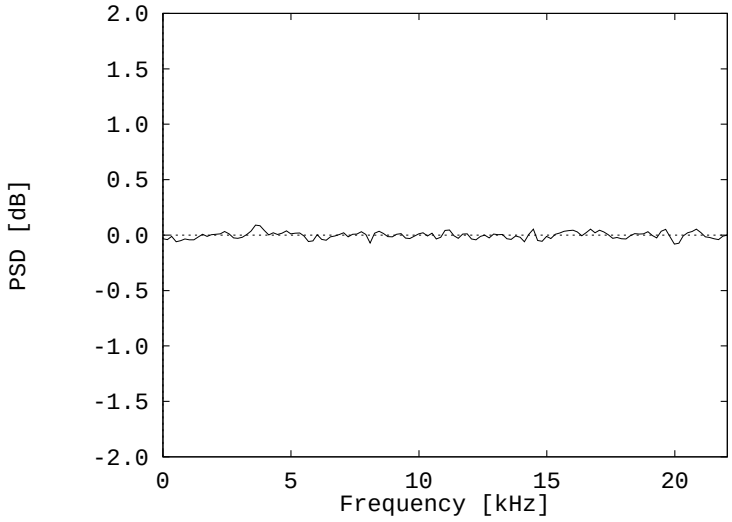


Figure 5.4: $PSD_\varepsilon(f)$ for an SD quantizing system using a dither filter with RPDF input, η, and coefficients $\{0.5, 1\}$. The system had a nominal sampling rate of 44.1 kHz and was presented with a static null input.

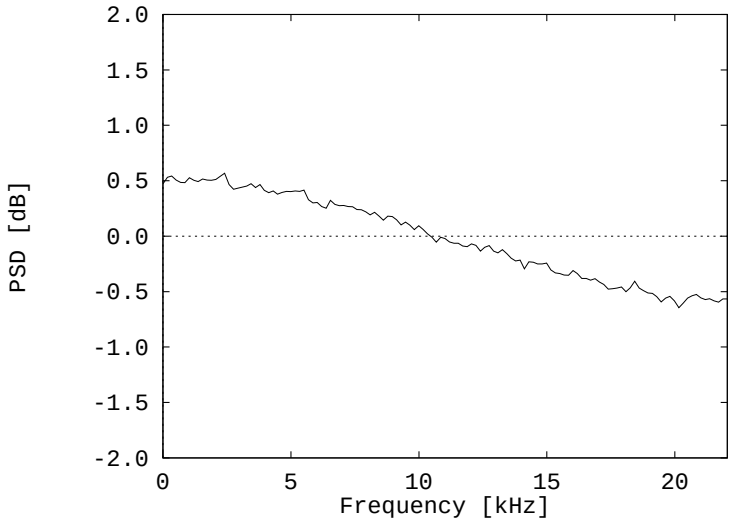


Figure 5.5: $PSD_\varepsilon(f)$ for an SD quantizing system using a dither filter with RPDF input, η, and coefficients $\{0.5, 1, 0.5\}$. The system had a nominal sampling rate of 44.1 kHz and was presented with a static null input.

shows $\mathrm{PSD}_\varepsilon(f)$ as calculated from a computer simulation of an SD system using this dither. It is flat as expected. On the other hand, Fig. 5.5 shows the error spectrum from a simulated SD system using a dither with filter coefficient sequence

$$\left\{\frac{1}{2}, 1, \frac{1}{2}\right\}.$$

This sequence does not satisfy the conditions of Corollary 5.2 and the corresponding error spectrum is not flat.

In light of the fact that the total error $\varepsilon = q$ of an SD system is spectrally flat irrespective of the spectral shape of the dither, the reader may wonder why one would ever bother using a spectrally shaped dither, or indeed any dither other than simple iid (white) RPDF, in such a system. We will see in Section 5.3 that this may be desirable if the output of an SD system with noise-shaping error feedback will sometimes be played back without subtraction of the dither, in which case the resulting error signal *will* be spectrally shaped if a non-white dither is used.

5.2 Dithered Noise-Shaping Quantizing Systems

The use of noise-shaping error feedback in quantizing systems is a powerful technique which allows the total error alone to be spectrally shaped in a fashion determined by the system designer without affecting the signal. For instance, in an audio system it may be preferable to shape the quantization error such that most of its power resides in high frequency bands where the human ear is relatively insensitive. (A considerable decrease in the perceived noise level is possible even in systems operating at commercial audio sampling rates [44, 45].)

Fig. 5.6 shows a schematic for a dithered quantizing system with noise-shaping

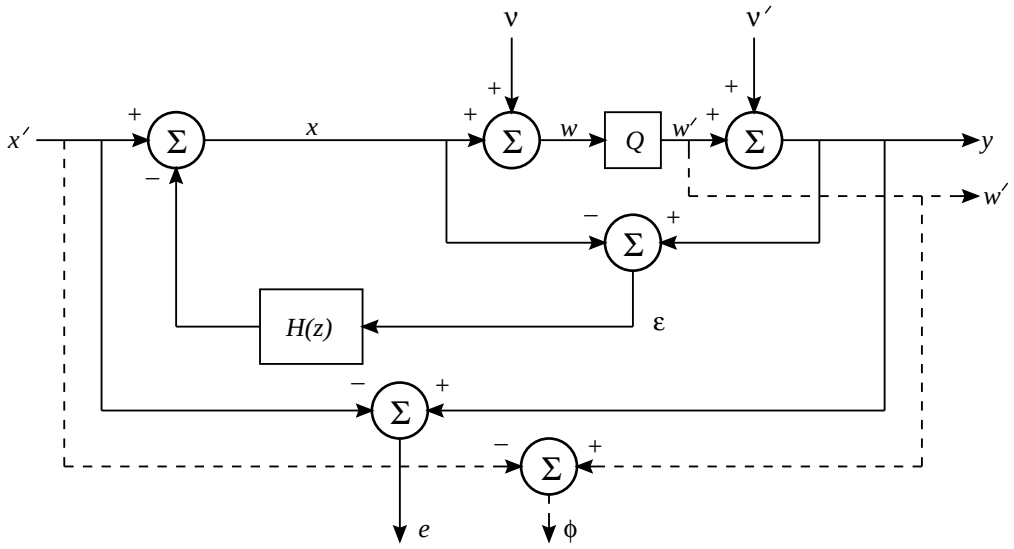


Figure 5.6: Schematic of a generalized dithered quantizing system using noise-shaping error feedback. Shown are the shaped total error, *e, of the system and also its* raw error, *ϕ (discussed in Section 5.3).*

error feedback. Note that only the total error ε of the quantization operation is fed back. The effect of the feedback filter $H(z)$ on the *shaped total error*, e, can be assessed by expressing the z-transform of the system output, $y(z)$, in two different ways [45]:

$$\begin{aligned} y(z) &= x'(z) + e(z) \\ y(z) &= x'(z) - H(z)\varepsilon(z) + \varepsilon(z). \end{aligned}$$

Subtraction yields

$$e(z) = [1 - H(z)]\varepsilon(z)$$

where $e(z)$ and $\varepsilon(z)$ are the z-transforms of the signals e and ε, respectively, and where $H(z)$ is the transfer function of the noise-shaping filter. Hence, the power

spectrum of e is given by [31]

$$\mathrm{PSD}_e(f) = \left|1 - H(e^{j2\pi fT})\right|^2 \mathrm{PSD}_\varepsilon(f), \tag{5.14}$$

where $H(e^{j2\pi fT})$ represents the frequency response of the noise-shaping filter, $H(z)$. This filter always includes one *implicit* delay element which prevents the current error from being subtracted from the current input. We note that $\mathrm{PSD}_\varepsilon(f)$ may itself be shaped, its form being determined by whether the system is NSD (see Eq. (4.39)) or SD (see Eq. (4.20)) and whether or not ν is spectrally shaped.

The use of noise shaping complicates the analysis of the error statistics. The reason for this is that x and ν will not be statistically independent if a filtered dither generator of the sort shown in Fig. 5.1 is used. Consider for instance the case where H(z) is a simple delay element

$$H(z) = z^{-1}.$$

Using subscripts to denote the temporal order of the quantities involved, we note that the input sample x_n contains vestiges of η_{n-1} arriving via the feedback path, and that in general this signal is also present in ν_n. The theorems given above cannot be applied in this situation because they all assume independence of x and ν. Under these circumstances, Eq. (5.14) holds but we cannot be certain of the form or even the constancy of $\mathrm{PSD}_\varepsilon(f)$. Fortunately, the theorems can be generalized to handle the case at hand.

5.2.1 NSD Noise Shaping Systems

We begin with NSD systems and the derivation of results analogous to Theorems 4.8 and 4.9. Eq. (3.8) yields

$$P_\varepsilon(u_\varepsilon) = \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_\varepsilon - \frac{k}{\Delta}\right) P_{\nu,x}\left(u_\varepsilon - \frac{k}{\Delta}, -\frac{k}{\Delta}\right).$$

Thus

$$\begin{aligned} E[\varepsilon^m] &\triangleq \left(\frac{j}{2\pi}\right)^m P_\varepsilon^{(m)}(0) \\ &= \left(\frac{j}{2\pi}\right)^m \sum_{k=-\infty}^{\infty} \sum_{r=0}^{m} \binom{m}{r} \operatorname{sinc}^{(m)}\left(\frac{k}{\Delta}\right) P_{\nu,x}^{(m-r,0)}\left(\frac{k}{\Delta}, \frac{k}{\Delta}\right). \end{aligned}$$

In this case we have by analogy with Theorem 4.8:

Lemma 5.1 *In an NSD quantizing system in which the dither, ν, and system input signal, x, are not necessarily statistically independent, $E[\varepsilon^\ell]$ is independent of the distribution of the input x for $\ell = 1, 2, \ldots, N$ if and only if the joint characteristic function of the dither and the input, $P_{\nu,x}(u, v)$, obeys the condition that*

$$\begin{gathered} P_{\nu,x}^{(i,0)}\left(\frac{k}{\Delta}, \frac{k}{\Delta}\right) = 0 \\ \forall k \in \mathbf{Z}_0 \quad \textit{and} \quad i = 0, 1, 2, \ldots, N-1. \end{gathered} \tag{5.15}$$

Subject to the conditions of Lemma 5.1, $E[\varepsilon^m]$ for $0 \le m \le N$ is given by Eq. (4.31), as before.

The derivation of $P_{\nu,x}$ in terms of the η_i's proceeds precisely as for the case where x is not involved, and we simply state the result:

$$P_{\nu,\eta,x}(u, w, v) = P_{\eta,x}(\gamma, v), \tag{5.16}$$

where

$$x = (\dots, x_{-1}, x_0, x_1, \dots),$$

and where

$$v = (\dots, v_{-1}, v_0, v_1, \dots)$$

is the corresponding vector of Fourier transformed variables. γ is a similar vector with components

$$\gamma_i = w_i + \sum_{j=-\infty}^{\infty} c_{j-i} u_j.$$

By setting all the unwanted variables in Eq. (5.16) to zero we obtain:

$$P_{\nu_n,x_n}(u_n, v_n) = P_{\eta,x_n}(\mu, v_n), \tag{5.17}$$

where the components of μ are

$$\mu_i = c_{n-i} u_n$$

and where we will retain the time indices since the relative times of η_i and x_n must be taken into account. (Note that if the η's are all mutually independent and we let $v_n = 0$, then Eq. (5.17) reduces to Eq. (5.3).)

In order for the mean and variance of the error to be input independent, Lemma 5.1 requires that:

$$P_{\nu_n,x_n}\left(\frac{k}{\Delta}, \frac{k}{\Delta}\right) = 0 \quad \forall k \in \mathbf{Z}_0 \tag{5.18}$$

and

$$P^{(1,0)}_{\nu_n,x_n}\left(\frac{k}{\Delta}, \frac{k}{\Delta}\right) = 0 \quad \forall k \in \mathbf{Z}_0. \tag{5.19}$$

At first glance, interpretation of these conditions in terms of Eq. (5.17) appears to be frustrated by the fact that we know nothing about the quantity P_{η,x_n}. However,

we can assume that (a) the dither filter is causal so that $c_i = 0 \; \forall i < 0$, and that (b) η_i is statistically independent of the random vector $(\ldots, x_{n-2}, x_{n-1}, x_n)$ for $i \geq n$, where we recall that the dither filter, $H(z)$, must contain an implicit single-sample delay. Thus there exists exactly one value of i such that $c_i \neq 0$ and for which η_i is statistically independent of x_n. This is $i = n$, so that Eq. (5.17) can be written as the product

$$P_{\nu_n, x_n}(c_0 u_n, v_n) = P_{\eta_n}(c_0 u_n) P_{\eta, x_n}(\mu', v_n).$$

where

$$\mu'_i = \begin{cases} \mu_i, & i < n, \\ 0, & i \geq n. \end{cases}$$

We conclude that Eq. (5.18) holds if

$$P_\eta \left(c_0 \frac{k}{\Delta} \right) = 0 \quad \forall k \in \mathbf{Z}_0, \tag{5.20}$$

and similarly that Eq. (5.19) holds if

$$P_\eta^{(1)} \left(c_0 \frac{k}{\Delta} \right) = 0 \quad \forall k \in \mathbf{Z}_0. \tag{5.21}$$

The analysis of the 2-D statistics proceeds in the usual fashion. We state without proof the obvious generalization of Theorem 4.9.

Lemma 5.2 *Consider two values, ε_n and $\varepsilon_{n+\ell}$, of the total error produced by an NSD quantizing system in which the dither and the input to the quantizing system are not necessarily statistically independent. Let these error samples be separated in time by $\tau = \ell T$ where T is the sampling period of the system and $\ell \neq 0$. Denote by $P_{(\nu_n, \nu_{n+\ell}), (x_n, x_{n+\ell})}$ the joint cf of the dither and input values, ν_n, $\nu_{n+\ell}$, x_n, and*

$x_{n+\ell}$, corresponding to ε_n and $\varepsilon_{n+\ell}$, respectively. If and only if

$$P_{(\nu_n,\nu_{n+\ell}),(x_n,x_{n+\ell})}\left(\frac{k_1}{\Delta},\frac{k_2}{\Delta},\frac{k_1}{\Delta},\frac{k_2}{\Delta}\right) = 0 \quad \forall (k_1,k_2) \in \mathbf{Z}_0^2 \tag{5.22}$$

$$P^{(0,1,0,0)}_{(\nu_n,\nu_{n+\ell}),(x_n,x_{n+\ell})}\left(\frac{k_1}{\Delta},0,\frac{k_1}{\Delta},0\right) = 0 \quad \forall k_1 \in \mathbf{Z}_0 \tag{5.23}$$

$$P^{(1,0,0,0)}_{(\nu_n,\nu_{n+\ell}),(x_n,x_{n+\ell})}\left(0,\frac{k_2}{\Delta},0,\frac{k_2}{\Delta}\right) = 0 \quad \forall k_2 \in \mathbf{Z}_0 \tag{5.24}$$

then

$$E[\varepsilon_n\varepsilon_{n+\ell}] = E[\nu_n\nu_{n+\ell}].$$

From Eq. (5.16) we have

$$P_{(\nu_n,\nu_{n+\ell}),(x_n,x_{n+\ell})}(u_1,u_2,v_1,v_2) = P_{\eta,(x_n,x_{n+\ell})}(\mu,v_1,v_2) \tag{5.25}$$

where

$$\mu_i = c_{n-i}u_1 + c_{n+\ell-i}u_2.$$

We first consider the case where $\ell > 0$. Using the same brand of reasoning that we used in the 1-D case, we note that there exists exactly one value of i for which $(c_{n-i}, c_{n+\ell-i}) \neq (0,0)$ and for which η_i is statistically independent of $(x_n, x_{n+\ell})$. This is $i = n+\ell$, so that Eq. (5.25) can be written

$$P_{(\nu_n,\nu_{n+\ell}),(x_n,x_{n+\ell})}(u_1,u_2,v_1,v_2) = P_{\eta_{n+\ell}}(c_0u_2)P_{\eta,(x_n,x_{n+\ell})}(\mu',v_1,v_2) \tag{5.26}$$

where only c_0 remains since the other coefficient is zero, and where

$$\mu_i' = \begin{cases} \mu_i, & i < n+\ell, \\ 0, & i \geq n+\ell. \end{cases}$$

According to Eq. (5.26), Condition I (Eq. (5.22)) of Lemma 5.2 will be satisfied for $\ell > 0$ and $k_2 \in \mathbf{Z}_0$ if

$$P_\eta\left(c_0\frac{k}{\Delta}\right) = 0 \quad \forall k \in \mathbf{Z}_0.$$

On the other hand, if $k_2 = 0$, then Eqs. (5.22) and (5.25) yield

$$P_{(\nu_n,\nu_{n+\ell}),(x_n,x_{n+\ell})}\left(\frac{k_1}{\Delta}, 0, \frac{k_1}{\Delta}, 0\right) = P_{\eta,x_n}\left(\mu'', \frac{k_1}{\Delta}\right)\Bigg|_{u_1=k_1/\Delta} \tag{5.27}$$

where

$$\mu_i'' = c_{n-i}u_1.$$

Then there exists exactly one i such that $c_{n-i} \neq 0$ and for which η_i is independent of x_n. This is $i = n$. Thus the right-hand side of Eq. (5.27) splits into a product which goes to zero if

$$P_\eta\left(c_0\frac{k}{\Delta}\right) = 0 \quad \forall k \in \mathbf{Z}_0.$$

Thus Condition I is satisfied for all $(k_1, k_2) \in \mathbf{Z}_0^2$ subject to this requirement. By symmetry, the $\ell < 0$ case produces identical conditions.

Conditions II and III (Eqs. (5.23) and (5.24)) are handled by application of the product rule as before. We omit the details, but it can be shown that these conditions are satisfied if Eqs. (5.20) and (5.21) hold. All three conditions being satisfied, Eq. (4.39) gives the total error spectrum in terms of the dither spectrum.

We collect below the conclusions from the above analysis.

Theorem 5.3 *In an NSD quantizing system with arbitrary noise-shaping error feedback and using filtered dither of the form described by Eq. (5.1), the total error will be wide-sense stationary and independent of the system input with a PSD given*

by

$$PSD_e(f) = \left|1 - H(e^{j2\pi fT})\right|^2 \left[PSD_\nu(f) + \frac{\Delta^2 T}{6}\right] \tag{5.28}$$

under the following conditions:

$$P_\eta\left(c_0\frac{k}{\Delta}\right) = 0 \quad \forall k \in \mathbf{Z}_0$$

and

$$P_\eta^{(1)}\left(c_0\frac{k}{\Delta}\right) = 0 \quad \forall k \in \mathbf{Z}_0.$$

If η is mRPDF we reach the simple but quite restrictive conclusion that:

Corollary 5.3 *In an NSD quantizing system with arbitrary noise-shaping error feedback and using filtered dither with η being an iid mRPDF random process, the total error will be wide-sense stationary and independent of the system input with a PSD given by Eq (5.28) if c_0 is a non-zero integer and $m \geq 2$.*

To appreciate just how restrictive this condition really is, it should be noted that it is not satisfied by simple highpass dither formed from the difference of two successive samples of an RPDF random process. This is confirmed by Fig. 5.7 which shows the spectrum of ε from a noise shaper using this kind of dither and a one tap feedback filter with coefficient -0.5. (Of course, the $\mathrm{PSD}_e(f)$ will have the expected form given by Eq. (5.28) if and only if $\mathrm{PSD}_\varepsilon(f)$ has the form given by Eq. (5.12); i.e., the sum of the dither spectrum and that of a white noise process.) Also shown is the spectrum normalized by the predicted spectrum of Eq. (5.12). Two static inputs ($x = 0.0$ and 0.5 LSB, respectively) were used. The normalized

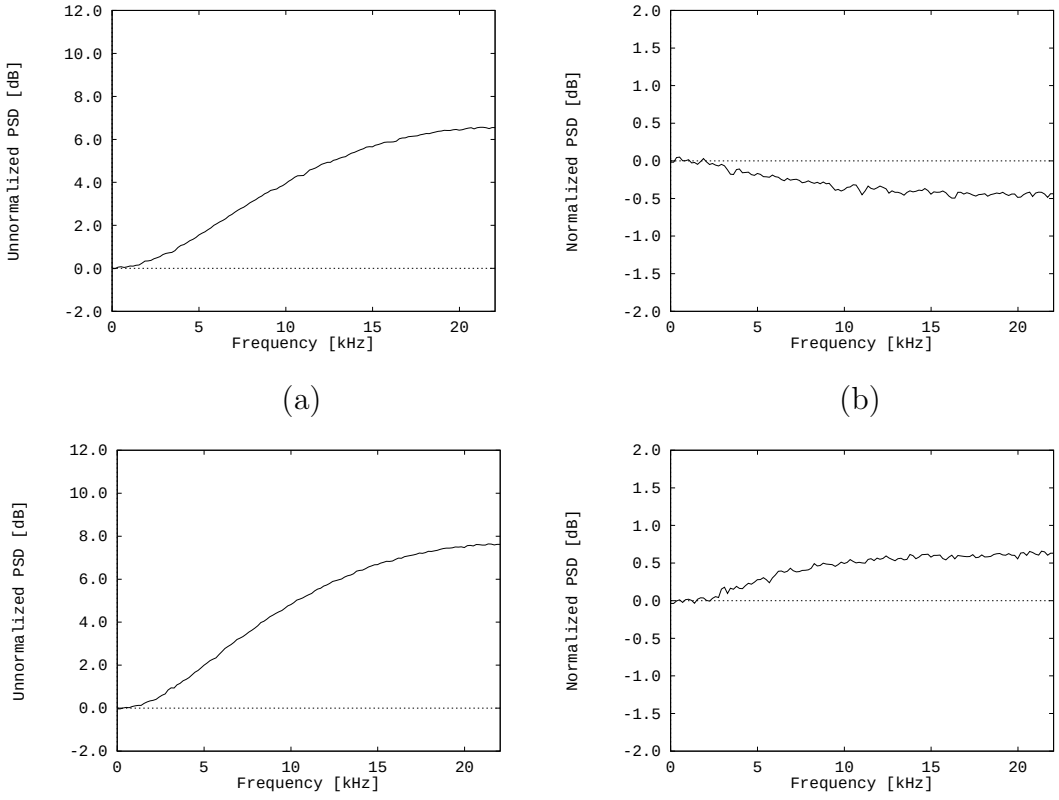


Figure 5.7: $PSD_\varepsilon(f)$ for an NSD quantizing system with error feedback and using a dither filter with RPDF input and coefficients {1.0,-1.0}. A single-tap noise-shaping filter with coefficient -0.5 was used. (a) Observed PSD for 0.0 LSB input, (b) observed PSD normalized by expected PSD for 0.0 LSB input, (c) observed PSD for 0.5 LSB input, (d) observed PSD normalized by expected PSD for 0.5 LSB input.

spectra are not flat, indicating that the error spectra are not of the expected shape. Furthermore, the two spectra are different, indicating that the error spectrum is input dependent.

These effects decrease in size with increasing gain and complexity of the noise-shaping filter, since the quantizer input then begins to resemble the sum of the system input with a large weakly-correlated Gaussian noise which will act as a suitable dither signal. For instance, the plots in Fig. 5.8 correspond to those in Fig. 5.7 with the sole difference being the use of a 3-coefficient noise-shaping filter with psychoacoustically optimized coefficients (refer to [45]). Although some variation of the spectrum with input is probably still present, it is apparently negligible.

5.2.2 SD Noise Shaping Systems

The analysis of SD systems with noise-shaping feedback is next. The straightforward generalization of Eq. (4.18), offered without proof, is

$$P_{\nu_1,\nu_2,x_1,x_2}\left(\frac{k_1}{\Delta},\frac{k_2}{\Delta},\frac{k_1}{\Delta},\frac{k_2}{\Delta}\right)=0 \qquad \forall(k_1,k_2)\in\mathbf{Z}_0^2. \tag{5.29}$$

This, however, is the same condition as Eq. (5.22), and thus leads immediately to the following theorem:

Theorem 5.4 *In an SD system with arbitrary noise-shaping error feedback and using filtered dither of the form described by Eq. (5.1), the total error will be wide-sense stationary and independent of the system input with a PSD given by*

$$PSD_e(f)=\left|1-H(e^{j2\pi fT})\right|^2\frac{\Delta^2T}{6} \tag{5.30}$$

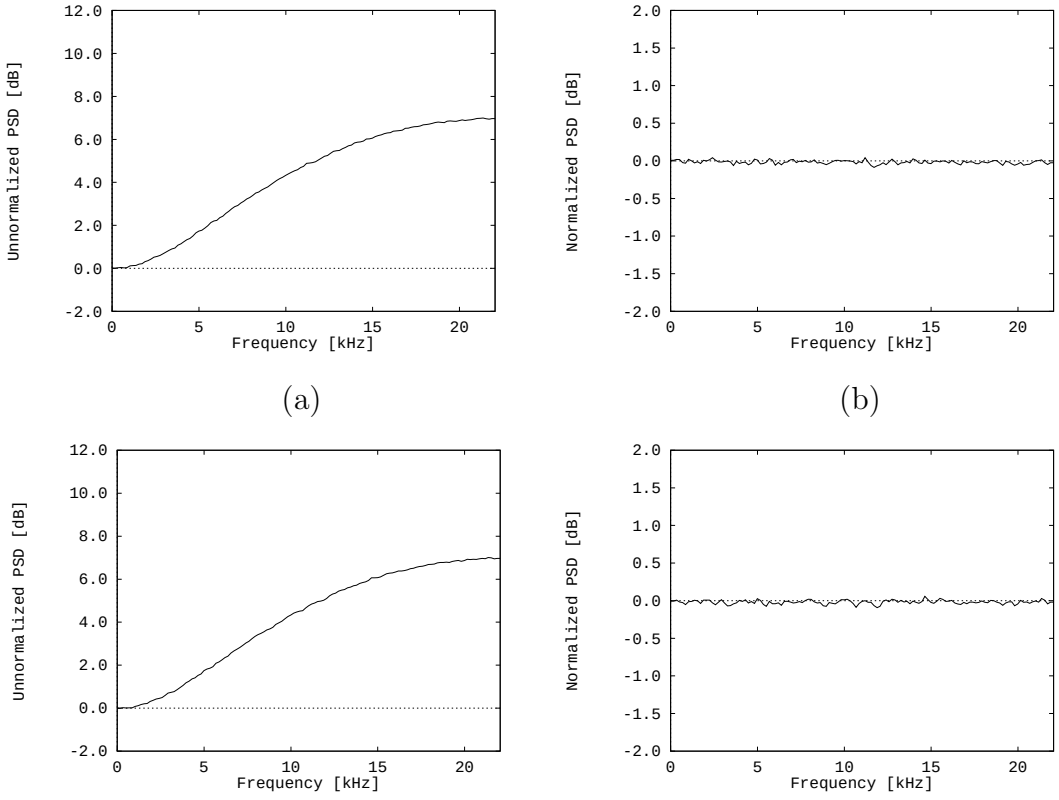


Figure 5.8: $PSD_\varepsilon(f)$ for an NSD quantizing system with error feedback and using a dither filter with RPDF input and coefficients {1.0,-1.0}. A 3-tap FIR noise-shaping filter with coefficients $\{1.33, -0.73, 0.065\}$ was used. (a) Observed PSD for 0.0 LSB input, (b) observed PSD normalized by expected PSD for 0.0 LSB input, (c) observed PSD for 0.5 LSB input, (d) observed PSD normalized by expected PSD for 0.5 LSB input.

under the following condition:

$$P_\eta\left(c_0\frac{k}{\Delta}\right) = 0 \quad \forall k \in \mathbf{Z}_0.$$

Corollary 5.4 *In an SD quantizing system with arbitrary noise-shaping error feedback and using filtered dither with η being an iid mRPDF random process, the total error will be wide-sense stationary and independent of the system input with a PSD given by Eq (5.30) if c_0 is a non-zero integer and $m \geq 1$.*

This latter is a weaker restriction than for NSD systems, insofar as $m \geq 2$ is required for the satisfaction of Theorem 5.3 (see Corollary 5.3).

A practical point regarding the implementation of SD systems with noise shaping should be made. Subtraction of the dither obviously must occur when the signal is replayed, because the point of quantizing is to restrict the resolution of transmitted/stored data to Δ, and the dither signal will have finer resolution than this. Hence, the signal transmitted or stored is not y but w' (see Fig. 5.6). The dither must be either transmitted/stored along with the signal or regenerated so that it can be subtracted at playback, but the dither must also be subtracted from w' before transmission/storage in order to calculate the total error ε to be fed back through $H(z)$.

5.2.3 Results For Special Classes of Shapers

Although we have so far been unable to find weaker sufficient conditions than those given in the theorems above, which guarantee input independence of the

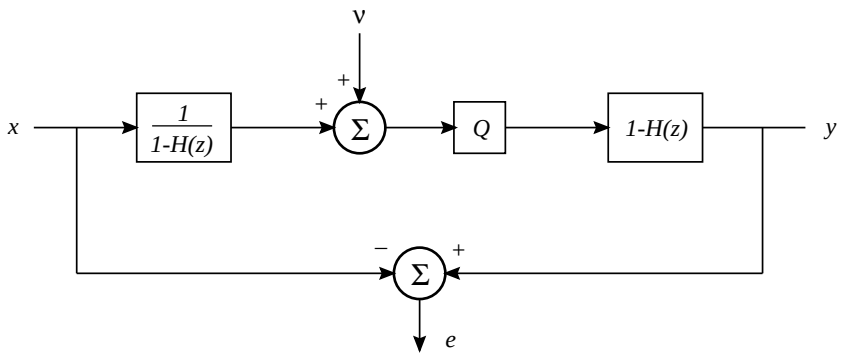
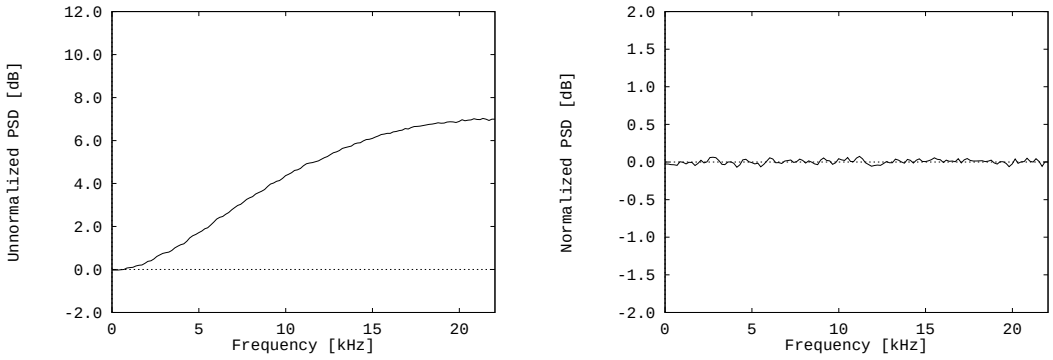


Figure 5.9: A system equivalent to that of Fig. 5.6 in the NSD case where all the coefficients of the error-feedback filter, $H(z)$, are integers.

error spectrum for an arbitrary noise shaper, some interesting results are known for certain special classes of shapers. Consider for instance an NSD system in which the feedback filter $H(z)$ is FIR and its first ℓ coefficients are all zero. Then the total error spectrum is wide-sense stationary and given by Eq. (5.28) if the conditions of Theorem 5.1 are satisfied and the dither filter, $G(z)$, is FIR with $c_i = 0$ for $i > \ell$. This ensures that x_i contains no vestiges of any η_j's which will also be present in the current dither sample, ν_i, so that x_i and ν_i will be independent. An analogous result exists for SD systems.

A remarkable result has been obtained for one important special class of NSD noise shaper designs by Craven [46]. These shapers employ feedback filters, $H(z)$, whose filter coefficients are all integers. Craven has shown that any such system produces *precisely* the same output as the system of Fig. 5.9, which employs no feedback. (The effective dither filter, $1-H(z)$, must be minimum phase for Fig. 5.9 to be realizable; i.e., it must be invertible.) This means that for such noise shapers, the broad class of shaped dithers satisfying only the conditions of Theorem 5.1 *must* produce the expected, input-independent error spectra. This is confirmed



(a) (b)

Figure 5.10: $PSD_\varepsilon(f)$, for an NSD quantizing system with error feedback and using a dither filter with RPDF input and coefficients {1.0,-1.0}. The system was presented with a null static input (0.0 LSB) and a single-tap noise-shaping filter with coefficient 1.0 was used. (a) Observed PSD, (b) observed PSD normalized by expected PSD.

by Fig. 5.10 which shows error spectra, unnormalized and normalized, for such a system using the simple highpass dither which failed when a feedback filter with non-integer coefficients was used.

5.3 The Raw Error of SD Systems

Consider an SD quantizing system with noise-shaping error feedback. It would be nice to be able to play back the quantizer output w' without subtraction of the dither if, for instance, the playback system did not have facilities for subtraction. We will let ϕ denote the *raw error* associated with the signal w', where (see Fig. 5.6)

$$\begin{aligned} \phi &\stackrel{\triangle}{=} w' - x' \\ &= e + \nu. \end{aligned}$$

Using the z-transforms of the signals involved, we obtain

$$\phi(z) = [1 - H(z)]\varepsilon(z) + \nu(z),$$

where the possible presence of a noise-shaping feedback filter, $H(z)$, has been assumed.

Now if all samples of ε were in general uncorrelated with all samples of ν, we could conclude that the power spectrum associated with $\phi(z)$ was the sum of those associated with $e(z) = [1 - H(z)]\varepsilon(z)$ and $\nu(z)$, but this is not the case. We anticipate that the dither signal will have to satisfy certain additional conditions for this to be true. If $H(z) \equiv 0$ we see that w' is just the output of an NSD system, so we require that the dither satisfy the conditions appropriate to such a system (see Theorem 5.3). If $H(z) \neq 0$, however, the properties of the error ϕ are not apparent from the analysis conducted thus far.

We make the following observations:

$$\begin{aligned} E[\phi^2] &= E[(e+\nu)^2] \\ &= E[e^2] + 2E[e\nu] + E[\nu^2] \qquad (5.31) \\ E[\phi_1\phi_2] &= E[(e_1+\nu_1)(e_2+\nu_2)] \\ &= E[e_1e_2] + E[e_1\nu_2] + E[e_2\nu_1] + E[\nu_1\nu_2]. \qquad (5.32) \end{aligned}$$

Let us write

$$1 - H(z) = \sum_{n=-\infty}^{\infty} h_n z^{-n}$$

so that the i-th sample of e can be expressed as

$$e_i = \sum_{n=-\infty}^{\infty} h_n \varepsilon_{i-n}.$$

We will consider the correlation between the i-th sample of e and the j-th sample of ν. When $i = j$ this quantity may be denoted by $E[\varepsilon_i \nu_j] = E[\varepsilon\nu]$. When $i \neq j$ we will denote $E[\varepsilon_i \nu_j]$ as $E[\varepsilon_1 \nu_2]$ (or, alternatively, as $E[\varepsilon_2 \nu_1]$). Now

$$\begin{aligned} E[e_i \nu_j] &= E\left[\sum_{n=-\infty}^{\infty} h_n \varepsilon_{i-n} \nu_j\right] \\ &= \sum_{n=-\infty}^{\infty} h_n E[\varepsilon_{i-n} \nu_j] \\ &= \sum_{m=-\infty}^{\infty} h_{i-m} E[\varepsilon_m \nu_j]. \end{aligned} \tag{5.33}$$

Let us consider the terms in this last sum without assuming, for now, that ν and x are statistically independent. We will use Eq. (3.7) which gives, for $N = 2$,

$$\begin{aligned} P_{\varepsilon_1,\varepsilon_2,\nu_1,\nu_2}(u_{\varepsilon_1}, u_{\varepsilon_2}, u_{\nu_1}, u_{\nu_2}) &= \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \operatorname{sinc}\left(u_{\varepsilon_1} - \frac{k_1}{\Delta}\right) \operatorname{sinc}\left(u_{\varepsilon_2} - \frac{k_2}{\Delta}\right) \\ &\quad \times P_{\nu_1,\nu_2,x_1,x_2}\left(u_{\nu_1} - \frac{k_1}{\Delta}, u_{\nu_2} - \frac{k_2}{\Delta}, -\frac{k_1}{\Delta}, -\frac{k_2}{\Delta}\right). \end{aligned} \tag{5.34}$$

Consider first the $m = j$ term in Eq. (5.33), and the following reduced form of Eq. (5.34) where we have written $P_{\varepsilon_m,\nu_m} = P_{\varepsilon,\nu}$:

$$P_{\varepsilon,\nu}(u_\varepsilon, u_\nu) = \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_\varepsilon - \frac{k}{\Delta}\right) P_{\nu,x}\left(u_\nu - \frac{k}{\Delta}, -\frac{k}{\Delta}\right).$$

We see that

$$\begin{aligned} E[\varepsilon\nu] &= \left(\frac{j}{2\pi}\right)^2 P^{(1,1)}_{\varepsilon,\nu}(0,0) \\ &= \left(\frac{j}{2\pi}\right)^2 \sum_{k=-\infty}^{\infty} \operatorname{sinc}^{(1)}\left(-\frac{k}{\Delta}\right) P^{(1,0)}_{\nu,x}\left(-\frac{k}{\Delta}, -\frac{k}{\Delta}\right). \end{aligned}$$

Thus $E[\varepsilon\nu] = 0$ if

$$P^{(1,0)}_{\nu,x}\left(\frac{k}{\Delta}, \frac{k}{\Delta}\right) = 0 \qquad \forall k \in \mathbf{Z}_0. \tag{5.35}$$

On the other hand, if $m \neq j$ then from Eq. (5.34) we obtain

$$P_{\varepsilon_1,\nu_2}(u_{\varepsilon_1}, u_{\nu_2}) = \sum_{k_1=-\infty}^{\infty} \operatorname{sinc}\left(u_{\varepsilon_1} - \frac{k_1}{\Delta}\right) P_{\nu_1,\nu_2,x_1,x_2}\left(-\frac{k_1}{\Delta}, u_{\nu_2}, -\frac{k_1}{\Delta}, 0\right). \tag{5.36}$$

Now

$$\begin{aligned} E[\varepsilon_1\nu_2] &= \left(\frac{j}{2\pi}\right)^2 P^{(1,1)}_{\varepsilon_1,\nu_2}(0,0) \\ &= \left(\frac{j}{2\pi}\right)^2 \sum_{k_1=-\infty}^{\infty} \operatorname{sinc}^{(1)}\left(-\frac{k_1}{\Delta}\right) P^{(0,1,0,0)}_{\nu_1,\nu_2,x_1,x_2}\left(-\frac{k_1}{\Delta}, 0, -\frac{k_1}{\Delta}, 0\right) \end{aligned} \tag{5.37}$$

so that $E[\varepsilon_1\nu_2] = 0$ if

$$P^{(0,1,0,0)}_{\nu_1,\nu_2,x_1,x_2}\left(\frac{k_1}{\Delta}, 0, \frac{k_1}{\Delta}, 0\right) = 0 \qquad \forall k_1 \in \mathbf{Z}_0. \tag{5.38}$$

A similar analysis reveals that $E[\varepsilon_2\nu_1] = 0$ if

$$P^{(1,0,0,0)}_{\nu_1,\nu_2,x_1,x_2}\left(0, \frac{k_2}{\Delta}, 0, \frac{k_2}{\Delta}\right) = 0 \qquad \forall k_2 \in \mathbf{Z}_0. \tag{5.39}$$

Hence if Eqs. (5.35)–(5.39) are satisfied for all time lags between ε_1 and ν_2 then we may state that

$$E[\varepsilon_m\nu_j] = 0 \qquad \forall (m, j).$$

In this case Eq. (5.33) indicates that

$$E[e_i\nu_j] = 0 \qquad \forall (i, j).$$

Then Eqs. (5.31) and (5.32) become

$$E[\phi^2] = E[e^2] + E[\nu^2] \tag{5.40}$$

$$E[\phi_1\phi_2] = E[e_1e_2] + E[\nu_1\nu_2]. \tag{5.41}$$

Now, if Eq. (5.29) is also satisfied, then ε has a well-defined autocorrelation function and

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

so that

$$\mathrm{PSD}_e(f) = \left|1 - H(e^{j2\pi fT})\right|^2 \frac{\Delta^2 T}{6}.$$

In this case the input-independent autocorrelation function of e is:

$$r_e(\ell) \stackrel{\triangle}{=} \begin{cases} E[e^2], & \ell = 0, \\ E[e_1 e_2](\ell), & \text{otherwise.} \end{cases}$$

Comparing this with Eqs. (5.40) and (5.41), we conclude subject to satisfaction of Eqs. (5.29), (5.35), (5.38) and (5.39) that

$$r_\phi(\ell) = r_e(\ell) + r_\nu(\ell),$$

so that

$$\mathrm{PSD}_\phi(f) = \mathrm{PSD}_e(f) + \mathrm{PSD}_\nu(f).$$

Let us compare the additional requirements imposed above on an SD system with the requirements typically imposed in an NSD system. Suppose that Eq. (5.29) holds. If Eq. (5.38) and Eq. (5.39) also hold then all the conditions of Lemma 5.2 (i.e., Eqs. (5.22)–(5.24)) are satisfied. Furthermore, if Eq. (5.29) holds, then it necessarily follows that

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

If Eq. (5.35) also holds, then the conditions of Lemma 5.1 (i.e., Eqs. (5.15)) are satisfied for $N = 2$. We know, however, that *all* of these conditions are satisfied under the conditions of Theorem 5.3. Thus we can state the following:

Theorem 5.5 *If the dither signal in an SD system, possibly using noise-shaping error feedback, satisfies the conditions of Theorem 5.3, then the raw error will be wide-sense stationary and independent of the system input with a PSD given by*

$$PSD_\phi(f) = \left|1 - H(e^{j2\pi fT})\right|^2 \frac{\Delta^2 T}{6} + PSD_\nu(f). \tag{5.42}$$

If we assume that ν and x *are* statistically independent (i.e., $H(z) \equiv 0$ so that no error feedback is present) then the conditions weaken to those imposed to yield Theorem 5.1, and Eq. (5.42) simplifies to yield Eq. (4.39). This is not surprising, since, in the absence of feedback, w' is the output of an ordinary NSD system.

These results allow for spectral shaping of the raw error of an SD system. Say for instance that a highpass error spectrum is desired in a noise shaping SD system whether or not the dither is subtracted at playback. By using a simple highpass 4RPDF dither, generated using a 2RPDF η and a dither filter with coefficients

$$\{1, -1\},$$

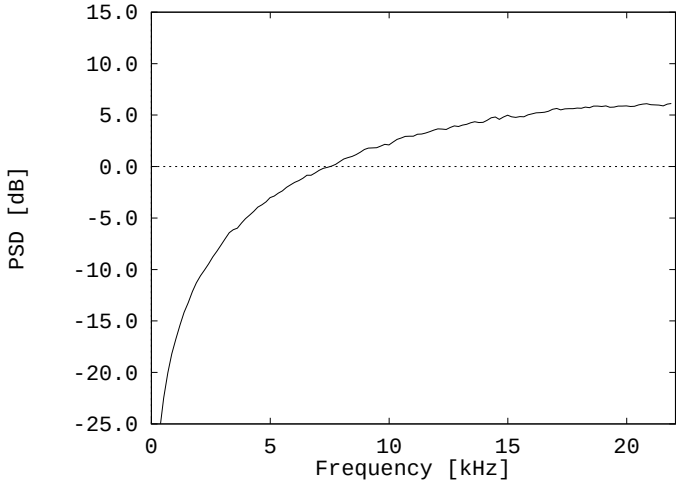
Theorem 5.3 will be satisfied. If a simple highpass noise-shaping feedback filter, $H(z) = z^{-1}$, is used, then $\mathrm{PSD}_e(f)$ and $\mathrm{PSD}_\nu(f)$ will both be highpass so that $\mathrm{PSD}_\phi(f)$ will be as well. This is confirmed by the spectra in Fig. 5.11 which are taken from a computer simulation of the described system. Note that a lower total noise power is still achieved by subtracting the dither at playback. In units of $\Delta^2/12$ the variance of e is 2 (the power gain of $1 - H(z)$) while that of ϕ is 6 (the power gain of $1 - H(z)$, plus the power of 2RPDF dither multiplied by the power gain of the dither filter). It should be noted that in accordance with the conditions of Theorem 5.3, 1RPDF noise η is not sufficient to eliminate spectral modulation

(indeed, if such dither is used in the system described above and $x' \equiv 0$, then $\phi \equiv 0$).

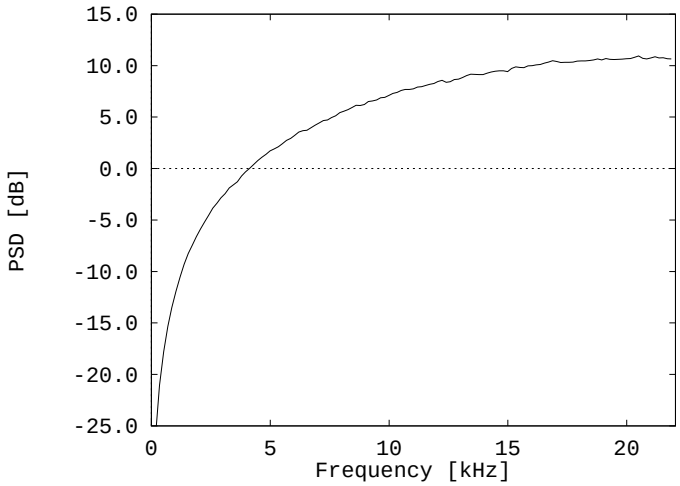
5.4 Multi-Channel Dither Generation

When multiple channels of nRPDF dither are to be generated, the generation of n new RPDF pseudo-random values per channel per sample may become computationally burdensome. It is tempting to try to reuse computed random numbers in different channels. For instance, Gerzon et al. [46] have proposed an efficient non-subtractive dither generation scheme for stereo signals which they call "diamond dither". A schematic of the proposed generator is shown in Fig. 5.12. Here η_1 and η_2 are iid, statistically independent of each other, and 1RPDF. Thus ν_1 and ν_2 are iid and 2RPDF, but not statistically independent of each other. In this design, only two new 1RPDF pseudorandom numbers need to be generated each sampling period, as opposed to four if statistically independent 2RPDF dithers were to be generated for each channel.

In general, interchannel sharing of random numbers for the purposes of dither generation will introduce interchannel correlations between error signals. This interchannel error correlation may be undesirable in certain applications. For instance, such correlations may affect the spatial image of the noise in multi-channel audio signals. The remainder of this section is dedicated to the assessment of such correlations and to methods of eliminating them. Generalizations of the Gerzon scheme to efficiently produce multi-channel dithers with other pdf's will be explored. (Only NSD quantizing systems will be considered, since SD systems only require one new RPDF dither value per sample per channel anyway.)



(a)



(b)

Figure 5.11: $PSD_e(f)$ and $PSD_\phi(f)$ for an SD quantizing system with error feedback and using a dither filter with 2RPDF input and coefficients $\{1, -1\}$. A simple highpass noise-shaping filter $H(z) = z^{-1}$ was used. The system had a nominal sampling rate of 44.1 kHz and was presented with a static null input. (a) $PSD_e(f)$, (b) $PSD_\phi(f)$.

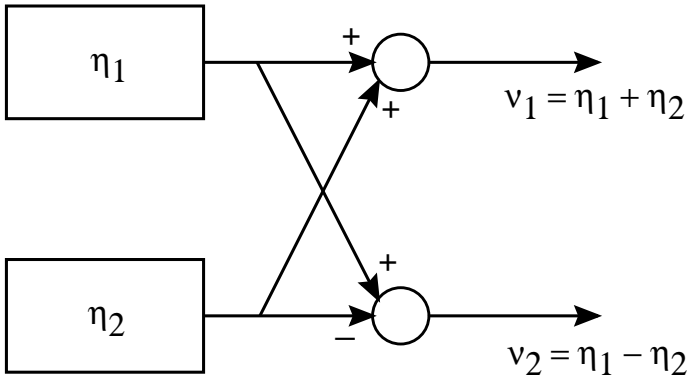


Figure 5.12: Efficient generation scheme for stereo non-subtractive dither. The η_i's are iid and uniformly distributed.

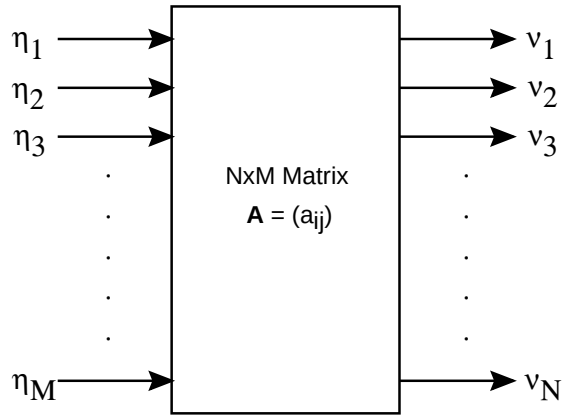


Figure 5.13: Efficient generation scheme for multi-channel non-subtractive dither. The η_i's are assumed to be iid.

Fig. 5.13 illustrates a generalized multi-channel dither generation scheme. We denote the output of the system by the random vector

$$\nu = (\nu_1, \nu_2, \ldots, \nu_N)^T$$

where the superscript T denotes matrix transposition. ν is assumed to be given by the equation

$$\nu = \boldsymbol{A}\eta$$

where

$$\eta = (\eta_1, \eta_2, \ldots, \eta_M)^T$$

is a random vector with iid components and

$$\boldsymbol{A} = (a_{ij})$$

is a constant real $N \times M$ matrix. (It will be shown below that it is not possible to generate more than N uncorrelated random processes from combinations of only N random processes, and thus we will assume that $M \geq N$.) The dither values obtained are

$$\nu_i = \sum_{j=1}^{M} a_{ij}\eta_j, \qquad i = 1, 2, \ldots, N.$$

We will assume that the η_j's are each iid random processes of the easily generated 1RPDF variety, and furthermore that they are statistically independent of one another at any given instant in time.

We are interested in correlations between total errors in different channels. Theorem 4.9 has thus far been applied to errors separated in time in a single channel system, but also applies directly to simultaneous errors in different channels (or to any other pair of errors generated by identical NSD quantizers). In order to

use this theorem, we must first find P_{ν_1,ν_2}. We begin by considering the statistical relationship between two typical ν_i's, say ν_1 and ν_2. Now

$$\begin{aligned} p_{\nu_1,\nu_2,\eta}(\nu_1,\nu_2,\eta) &= p_{\nu_1,\nu_2|\eta}(\nu_1,\nu_2,\eta)\prod_{j=1}^{M} p_{\eta_j}(\eta_j) \\ &= \delta\left(\nu_1 - \sum_{j=1}^{M} a_{1j}\eta_j\right)\delta\left(\nu_2 - \sum_{j=1}^{M} a_{2j}\eta_j\right)\prod_{j=1}^{M} p_{\eta_j}(\eta_j). \end{aligned}$$

Performing the necessary Fourier transforms yields

$$P_{\nu_1,\nu_2,\eta}(u_{\nu_1},u_{\nu_2},u_\eta) = \prod_{j=1}^{M} P_{\eta_j}(u_{\eta_j} + a_{1j}u_{\nu_1} + a_{2j}u_{\nu_2})$$

where $u_\eta = (u_{\eta_1}, u_{\eta_2}, \ldots, u_{\eta_M})$. Then, setting $u_\eta = 0$, we have

$$P_{\nu_1,\nu_2}(u_{\nu_1},u_{\nu_2}) = \prod_{j=1}^{M} P_{\eta_j}(a_{1j}u_{\nu_1} + a_{2j}u_{\nu_2}). \tag{5.43}$$

By way of example, we consider the stereo dither scheme discovered by Gerzon, with its associated matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Here we have

$$P_{\nu_1,\nu_2}(u_{\nu_1},u_{\nu_2}) = \operatorname{sinc}(u_{\nu_1} + u_{\nu_2})\ \operatorname{sinc}(u_{\nu_1} - u_{\nu_2}),$$

the inverse Fourier transform of which is

$$p_{\nu_1,\nu_2}(\nu_1,\nu_2) = \frac{1}{2}\Pi_\Delta\left(\frac{\nu_1+\nu_2}{2}\right)\Pi_\Delta\left(\frac{\nu_1-\nu_2}{2}\right).$$

As illustrated in Fig. 5.14, this pdf is supported on a diamond-shaped region in the $\nu_1\nu_2$-plane, giving rise to the denotation "diamond dither."

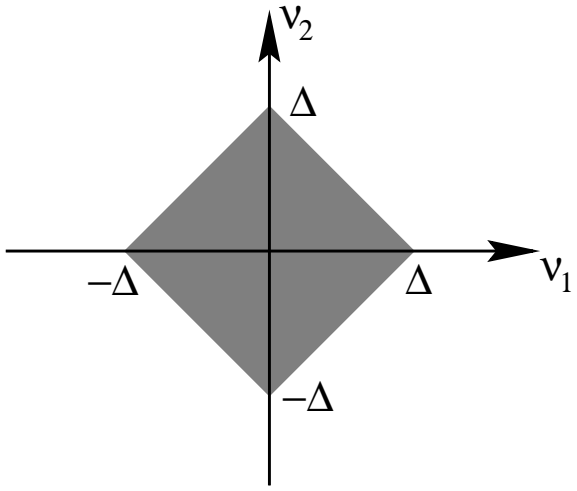


Figure 5.14: The support of the "diamond dither" joint pdf, $p_{\nu_1,\nu_2}(\nu_1,\nu_2)$.

The interchannel dither correlation can now be calculated in the usual fashion:

$$\begin{aligned} E[\nu_1\nu_2] &= \left(\frac{j}{2\pi}\right)^2 P^{(1,1)}_{\nu_1,\nu_2}(0,0) \\ &= \sum_{j=1}^{M} a_{1j}a_{2j}E[\eta_j^2] + \sum_{j=1}^{M}\sum_{\substack{i=1\\ i\neq j}}^{M} a_{1i}a_{2j}E[\eta_i]E[\eta_j]. \end{aligned}$$

Since η is assumed to be iid 1RPDF, it has zero mean and a variance of $\Delta^2/12$. In this case the above equation simplifies to give

$$E[\nu_1\nu_2] = \frac{\Delta^2}{12}\sum_{j=1}^{M} a_{1j}a_{2j}. \tag{5.44}$$

If we require that interchannel error correlations be independent of the input signal distribution, then we must ensure that the conditions of Theorem 4.9 are satisfied. We will briefly defer discussion of the requirements placed upon the matrix $\boldsymbol{A}$ by the conditions of this theorem, and proceed under the assumption that they are satisfied. In this case

$$E[\varepsilon_1\varepsilon_2] = E[\nu_1\nu_2].$$

Thus, in order to eliminate all interchannel error correlations, we require that $E[\nu_i \nu_j] = 0$ for all i, j such that $i \neq j$. Eq. (5.44) indicates that this requirement is simply that the coefficient vectors

$$\{(a_{i1}, a_{i2}, a_{i3}, \ldots, a_{iM}), \quad i = 1, 2, \ldots, N\}$$

(i.e., the row vectors of $\boldsymbol{A}$) form a mutually orthogonal set[2]. Since we can only have N orthogonal M-vectors if $M \geq N$ this implies that we cannot generate more orthogonal dither processes than we employ independent η's. While matrices meeting the orthogonality requirements are abundant, the additional requirement that the resulting dither be nRPDF for some given n complicates matters. This requires that the coefficient vectors each contain precisely n entries equal to either 1 or -1 and that the remaining entries be zeros.

It turns out that if the the row vectors of $\boldsymbol{A}$ *are* mutually orthogonal, then the conditions of Theorem 4.9 will be satisfied whenever the desired order of dither is $n \geq 2$. In order to demonstrate this we consider a typical pair of dither values, ν_1 and ν_2, and refer to Eq. (5.43). The first condition of the theorem (Eq. (4.36)) is that

$$P_{\nu_1,\nu_2}\left(\frac{k_1}{\Delta}, \frac{k_2}{\Delta}\right) = \prod_{j=1}^{M} P_{\eta_j}\left(a_{1j}\frac{k_1}{\Delta} + a_{2j}\frac{k_2}{\Delta}\right) = 0 \qquad \forall (k_1, k_2) \in \mathbf{Z}_0^2.$$

Let us assume, for purposes of contradiction, that η is RPDF and that the rows of $\boldsymbol{A}$ are mutually orthogonal vectors consisting of elements $a_{ij} \in \{0, 1, -1\}$, but that the above condition does not hold. That is, there exists $(k_1, k_2) = (k_1^*, k_2^*) \neq (0, 0)$ such that no term in the given product vanishes. Since the terms are sinc functions,

[2]This is not quite the same as saying that $\boldsymbol{A}$ is an *orthogonal matrix,* which requires furthermore that the matrix be square and that each of its rows has unit magnitude.

this means that

$$\begin{bmatrix} k_1^* & k_2^* \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1M} \\ a_{21} & a_{22} & \dots & a_{2M} \end{bmatrix} = \begin{bmatrix} 0 & 0 & \dots & 0 \end{bmatrix}$$

which would imply, in particular, that for all $j_1, j_2 \in \{1, 2, \dots, M\}$, $j_1 \neq j_2$,

$$\begin{bmatrix} k_1^* & k_2^* \end{bmatrix} \begin{bmatrix} a_{1j_1} & a_{1j_2} \\ a_{2j_1} & a_{2j_2} \end{bmatrix} = \begin{bmatrix} 0 & 0 \end{bmatrix}. \tag{5.45}$$

However, since the rows of $\boldsymbol{A}$ are mutually orthogonal, there must exist at least one pair (j_1, j_2) such that Eq. (5.45) has only the trivial solution $(k_1^*, k_2^*) = (0, 0)$, which provides a contradiction.

Again from Eq. (5.44) we have

$$P^{(0,1)}_{\nu_1,\nu_2}\left(\frac{k_1}{\Delta}, 0\right) = \sum_{j=1}^{M} a_{2j} P_\eta^{(1)}\left(a_{1j}\frac{k_1}{\Delta}\right) \prod_{\substack{i=1 \\ i\neq j}}^{M} P_\eta\left(a_{1i}\frac{k_1}{\Delta}\right).$$

This expression goes to zero for all $k_1 \in \mathbf{Z}_0$, so that the second condition of Theorem 4.9 (Eq. (4.37)) is satisfied, whenever η is RPDF, $a_{i,j} \in \{0, 1, -1\}$ and $n \geq 2$. In this case, the final condition of the theorem (Eq. (4.38)) is similarly satisfied.

A multi-channel dither generator may be considered optimal if it yields uncorrelated dither values and requires the generation of just one new random number per sample per channel. The latter will be the case if the matrix $\boldsymbol{A}$ is square (i.e., $N = M$). We will call such schemes and their associated matrices (N, n)-*optimal,* where, again, N is the number of channels of dither produced and n is the order of the dither. We have seen that a $N \times N$ matrix $\boldsymbol{A} = (a_{ij})$ is (N, n)-optimal if:

1. $a_{ij} \in \{0, 1, -1\}$ $\forall (i, j)$,

2. each row of the matrix contains precisely n entries of absolute value one, and

3. the rows of the matrix form a set of mutually orthogonal vectors.

The following simple (N, n)-optimal matrices, the first of which corresponds to a stereo "diamond dither" generator, can serve as building blocks for the construction of many others:

$$(N, n) = (2, 2) : \qquad \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

$$(N, n) = (4, 3) : \qquad \begin{bmatrix} 0 & -1 & 1 & 1 \\ 1 & 0 & -1 & 1 \\ -1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix},$$

$$(N, n) = (6, 5) : \qquad \begin{bmatrix} 0 & 1 & 1 & -1 & -1 & 1 \\ 1 & 0 & -1 & 1 & -1 & 1 \\ 1 & -1 & 0 & -1 & 1 & 1 \\ -1 & 1 & -1 & 0 & 1 & 1 \\ -1 & -1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

The following rules then allow construction of other optimal schemes (the proofs are by inspection):

Rule 1. Interchanging two rows or two columns in a (N, n)-optimal matrix yields a (N, n)-optimal matrix.

Rule 2. Multiplying a row or a column of a (N, n)-optimal matrix by -1 yields a (N, n)-optimal matrix.

Rule 3. If $\boldsymbol{A}$ is a (N_1, n)-optimal matrix, $\boldsymbol{B}$ is a (N_2, n)-optimal matrix, and $\mathbf{0}$ is a $N_1 \times N_2$ matrix of zeros, then the *direct sum*

$$\boldsymbol{A} \oplus \boldsymbol{B} \triangleq \begin{bmatrix} \boldsymbol{A} & \mathbf{0} \\ \mathbf{0}^T & \boldsymbol{B} \end{bmatrix}$$

is a $(N_1 + N_2, n)$-optimal matrix.

Rule 4. If $\boldsymbol{A} = (a_{ij})$ is a (N_1, n_1)-optimal matrix and $\boldsymbol{B}$ is a (N_2, n_2)-optimal matrix then the *Kronecker* or *direct product* [47]

$$\boldsymbol{A} \otimes \boldsymbol{B} \triangleq \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \dots & a_{1N_1}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \dots & a_{2N_1}\boldsymbol{B} \\ \vdots & \vdots & & \vdots \\ a_{N_1 1}\boldsymbol{B} & a_{N_1 2}\boldsymbol{B} & \dots & a_{N_1 N_1}\boldsymbol{B} \end{bmatrix}$$

is a $(N_1 N_2, n_1 n_2)$-optimal matrix.

For example, combining two Gerzon-type (2,2)-optimal matrices of the form

$$A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

using Rule 3 yields the (4,2)-optimal matrix

$$\boldsymbol{A} \oplus \boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

This corresponds to two Gerzon-type schemes operating independently in parallel. On the other hand, combining the same two (2,2)-optimal matrices using Rule 4

we obtain the (4,4)-optimal matrix

$$\boldsymbol{A} \otimes \boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

This corresponds to two pairs of Gerzon-type schemes, each member of the second pair receiving one of its inputs from each member of the first pair.

For arbitrary N and n, such optimal matrices do not generally exist. For instance, it can be checked by trial and error that no (N, n)-optimal scheme exists for $(N, n) \in \{(3, 2), (3, 3), (5, 2), (5, 3), (5, 4), (5, 5)\}$. In such cases, extra dither values can be generated using an optimal scheme and some then discarded. This reduces the computational efficiency of the scheme, but using the above rules a matrix with roughly the desired number of channels and order of dither can be found.

For most multi-channel audio applications, Gerzon-type optimal generators operating independently in parallel are appropriate, since these will produce the 2RPDF dither required to render the first and second moments of the total error input independent. For image processing or measurement applications, optimal schemes generating higher order dithers may be of interest in order to render higher error moments input independent.

Chapter 6

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 analogue systems to digital ones, and to illustrate how this may be done.

6.1 Digital Dither pdf's

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 to be requantized, and

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

for an LSB of the requantized output.

Let us consider the following digital dither pdf

$$p_\nu(\nu) = \delta \tilde{p}_\nu(\nu) W_\delta(\nu), \tag{6.1}$$

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 (Theorem A.7) 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. (6.1) we find that

$$\begin{aligned} P_\nu(u) &= \left[\tilde{P}_\nu \star W_{\frac{1}{\delta}}\right](u) \\ &= \sum_{\ell=-\infty}^{\infty} \tilde{P}_\nu\left(u - \frac{\ell}{\delta}\right) \end{aligned} \tag{6.2}$$

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 Theorem 4.8 (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

$$p_x(x) = \delta\tilde{p}_x(x)W_\delta(x) \tag{6.3}$$

where $\tilde{p}_x$ is a continuous function normalized such that the integral of Eq. (6.3) is unity. Then

$$\begin{aligned} P_x(u) &= [\tilde{P}_x \star W_{\frac{1}{\delta}}](u) \\ &= \sum_{\ell=-\infty}^{\infty} \tilde{P}_x\left(u - \frac{\ell}{\delta}\right). \end{aligned} \tag{6.4}$$

We will make similar assumptions regarding joint pdf's of interest. Thus we will consider

$$p_{\nu_1,\nu_2}(\nu_1,\nu_2) = \delta^2\tilde{p}_{\nu_1,\nu_2}(\nu_1,\nu_2)W_\delta(\nu_1,\nu_2)$$

with

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\nu_1,\nu_2}(\nu_1,\nu_2)d\nu_1 d\nu_2 = \delta^2 \sum_{\ell_1=-\infty}^{\infty}\sum_{\ell_2=-\infty}^{\infty} \tilde{p}_{\nu_1,\nu_2}(\ell_1\delta,\ell_2\delta) = 1.$$

Then

$$\begin{aligned} P_{\nu_1,\nu_2}(u_1,u_2) &= \left[\tilde{P}_{\nu_1,\nu_2} \star W_{\frac{1}{\delta}}\right](u_1,u_2) \\ &= \sum_{\ell_1=-\infty}^{\infty}\sum_{\ell_2=-\infty}^{\infty} \tilde{P}_{\nu_1,\nu_2}\left(u_1 - \frac{\ell_1}{\delta}, u_2 - \frac{\ell_2}{\delta}\right), \end{aligned}$$

where $\tilde{P}_{\nu_1,\nu_2}(u_1,u_2)$ is the two-dimensional Fourier transform of $\tilde{p}_{\nu_1,\nu_2}(\nu_1,\nu_2)$. Similarly, we assume that we can write

$$P_{x_1,x_2}(u_1,u_2) = \sum_{\ell_1=-\infty}^{\infty}\sum_{\ell_2=-\infty}^{\infty} \tilde{P}_{x_1,x_2}\left(u_1 - \frac{\ell_1}{\delta}, u_2 - \frac{\ell_2}{\delta}\right).$$

6.2 Digital SD Systems

Now,

$$\begin{aligned} P_x\left(u - \frac{k}{\Delta}\right) &= \sum_{\ell=-\infty}^{\infty} \tilde{P}_x\left(u - \frac{k}{\Delta} - \frac{\ell}{\delta}\right) \\ &= \sum_{\ell=-\infty}^{\infty} \tilde{P}_x\left(u - \frac{k+2^L\ell}{\Delta}\right) \end{aligned}$$

so that, from Eq. (3.7), we have

$$P_{q,x}(u_q,u_x) = \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_q - \frac{k}{\Delta}\right) P_\nu\left(-\frac{k}{\Delta}\right) \sum_{\ell=-\infty}^{\infty} \tilde{P}_x\left(u_x - \frac{k+2^L\ell}{\Delta}\right).$$

Thus for q and x to be statistically independent for arbitrary $\tilde{P}_x$ we require that

$$P_\nu\left(\frac{k}{\Delta}\right) = 0 \qquad \text{for all } k \in \mathbf{Z} \text{ except, possibly, when } \frac{k}{2^L} \in \mathbf{Z}. \tag{6.5}$$

In this case

$$\begin{aligned} P_{q,x}(u_q,u_x) &= \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_q - \frac{2^L k}{\Delta}\right) P_\nu\left(-\frac{2^L k}{\Delta}\right) \sum_{\ell=-\infty}^{\infty} \tilde{P}_x\left(u_x - \frac{2^L(k+\ell)}{\Delta}\right) \\ &= \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_q - \frac{k}{\delta}\right) P_\nu\left(-\frac{k}{\delta}\right) \sum_{\ell=-\infty}^{\infty} \tilde{P}_x\left(u_x - \frac{\ell}{\delta}\right) \\ &= P_x(u_x) \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_q - \frac{k}{\delta}\right) P_\nu\left(-\frac{k}{\delta}\right) \end{aligned}$$

where Eq. (6.4) has been used in the last step. Note that in the limit as $\delta \to 0$ (i.e., as $L \to \infty$) Eq. (6.5) becomes Eq. (4.16), the condition of Theorem 4.4 for analogue systems. This reflects the conception of an analogue system as a digital system with infinite precision (i.e., an infinite number of bits).

Now from Eq. (6.2) we see that if $\tilde{P}_\nu$ meets the conditions of Theorem 4.4, i.e. that

$$\tilde{P}_\nu\left(\frac{k}{\Delta}\right) = 0 \qquad \forall k \in \mathbf{Z}_0,$$

then P_ν will go to zero at the places required by Eq. (6.5). Since Eq. (6.2) shows that P_ν is periodic such that

$$P_\nu\left(\frac{k}{\delta}\right) = P_\nu(0) = 1 \qquad \forall k \in \mathbf{Z},$$

we then obtain

$$\begin{aligned} P_q(u_q) &= \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_q - \frac{k}{\delta}\right) \\ &= \left[\operatorname{sinc} \star W_{\frac{1}{\delta}}\right](u_q). \end{aligned}$$

Thus (using Theorem A.5)

$$p_q(q) = \frac{\Delta}{2^L}\Pi_\Delta(q)W_\delta(q)$$

and in this sense the total error is uniformly distributed.

Thus we have the following theorem:

Theorem 6.1 *For a digital SD system in which requantization is used to remove the L least significant bits of binary data, the total error is statistically independent of the system input and uniformly distributed if a digital dither (with the same precision as the input data) is applied for which*

$$\tilde{P}_\nu\left(\frac{k}{\Delta}\right) = 0 \qquad \forall k \in \mathbf{Z}_0.$$

It is worth noting 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 Theorem 6.1 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 $L = 4$.

By the usual means the analysis may be extended to the joint statistics of errors separated in time. It is straightforward to show that for two such errors, q_1 and q_2,

$$\begin{aligned} & P_{q_1,q_2,x_1,x_2}\left(u_{q_1}, u_{q_2}, u_{x_1}, u_{x_2}\right) \\ &= \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \sum_{\ell_1=-\infty}^{\infty} \sum_{\ell_2=-\infty}^{\infty} \operatorname{sinc}\left(u_{q_1} - \frac{k_1}{\Delta}\right) \operatorname{sinc}\left(u_{q_2} - \frac{k_2}{\Delta}\right) \\ & \quad \times P_{\nu_1,\nu_2}\left(-\frac{k_1}{\Delta}, -\frac{k_2}{\Delta}\right) \tilde{P}_{x_1,x_2}\left(u_{x_1} - \frac{k_1 + 2^L\ell_1}{\Delta}, u_{x_2} - \frac{k_2 + 2^L\ell_2}{\Delta}\right) \end{aligned}$$

so that if

$$P_{\nu_1,\nu_2}\left(\frac{k_1}{\Delta}, \frac{k_2}{\Delta}\right) = 0$$

$$\text{for all } (k_1, k_2) \in \mathbf{Z}^2 \text{ except, possibly, when } \left(\frac{k_1}{2^L}, \frac{k_2}{2^L}\right) \in \mathbf{Z}^2$$

then

$$p_{q_1,q_2}(u_{q_1}, u_{q_2}) = \frac{\Delta}{2^L}\Pi_\Delta(q_1)W_\delta(q_1) \cdot \frac{\Delta}{2^L}\Pi_\Delta(q_2)W_\delta(q_2).$$

Hence q_1 and q_2 are statistically independent so that the total error will be spectrally white.

Subject to the satisfaction of the conditions of Theorem 6.1, q and x are statistically independent so we may immediately write down an expression for the cf of the system output:

$$P_y(u) = P_q(u)P_x(u).$$

If, in addition, the dither is iid, then

$$P_{y_1,y_2}(u_1,u_2) = P_{q_1}(u_1)P_{q_2}(u_2)P_{x_1,x_2}(u_1,u_2).$$

6.3 Digital NSD Systems

From Eq. (4.27) we have

$$\begin{aligned} G_\nu(u) &= \frac{\sin(\pi\Delta u)}{\pi\Delta u} P_\nu(u) \\ &= \frac{\sin(\pi\Delta u)}{\pi\Delta u} \sum_{k=-\infty}^{\infty} \tilde{P}_\nu\left(u - \frac{k}{\delta}\right). \end{aligned} \tag{6.6}$$

Then from Eq. (3.8) we have

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

so that

$$\begin{aligned} E[\varepsilon^m] &= \left(\frac{j}{2\pi}\right)^m P_\varepsilon^{(m)}(0) \\ &= \left(\frac{j}{2\pi}\right)^m \sum_{k=-\infty}^{\infty}\sum_{\ell=-\infty}^{\infty} G_\nu^{(m)}\left(-\frac{k}{\Delta}\right)\tilde{P}_x\left(-\frac{k+2^L\ell}{\Delta}\right). \end{aligned} \tag{6.7}$$

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

$$\begin{gathered} G_\nu^{(m)}\left(\frac{k}{\Delta}\right) = 0 \\ \text{for all } k \in \mathbf{Z} \text{ except, possibly, when } \frac{k}{2^L} \in \mathbf{Z}. \end{gathered} \tag{6.8}$$

Note that in the limit as $\delta \to 0$ (i.e., as $L \to \infty$) Eq. (6.8) becomes Eq. (4.28), the condition of Theorem 4.7 for analogue systems.

Returning to Eq. (6.6) and differentiating, we have

$$\frac{d^m G_\nu}{du^m}(u) = \sum_{k=-\infty}^{\infty} \sum_{r=0}^{m} \binom{m}{r} \frac{d^r}{du^r} \left[\frac{\sin(\pi \Delta u)}{\pi \Delta u} \right] \frac{d^{m-r} \tilde{P}_\nu}{du^{m-r}} \left(u - \frac{k}{\delta} \right). \tag{6.9}$$

If $\tilde{P}_\nu$ meets the conditions of Theorem 4.8 (for $M = m$), then all terms in Eq. (6.9) involving the derivatives of $\tilde{P}_\nu$ go to zero at the places required by Eq. (6.8) 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 6.2 *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 a digital counterpart of Theorem 4.8. It is interesting to note that no such analogue exists for Theorem 4.7 in terms of $\tilde{P}_\nu$.

As before, 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. (6.8) 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 $L = 4$.

We would like to write down expressions for the moments of the total error. If we choose a dither such that Eq. (6.8) holds, many terms vanish from Eq. (6.7), 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. (6.4) we know that

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

Thus

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m \sum_{k=-\infty}^{\infty} G_\nu^{(m)}\left(\frac{k}{\delta}\right) \tag{6.10}$$

which 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.

Frequently, dithers in digital systems will be given a 2's-complement [36] 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 6.2 upon Eq. (4.41), 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. (6.4) 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. (6.10).

6.4 Quantized Dithers

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 analogue dither of order n (as defined in Section 2.3), the associated digital dither with pdf given by Eq. (6.1) shares the beneficial properties of its analogue counterpart.

In the case where a digital dither signal is generated by fine quantization of an analogue 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 analogue signal. This can be seen directly using Eq. (4.5), for the pdf of the digital dither 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. (6.2). Note that if $\tilde{P}_\nu$ satisfies the conditions of the Theorems, then so will the quantity

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

6.5 Non-Stochastic Quantizers

In some cases, stochastic quantizers may not be practical to implement. 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 at these edges.

We will now explore 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 digital dithered systems with deterministic requantizers using such a notional dc offset, which is a purely mathematical device without physical counterpart. Proceeding otherwise as we did before, Eq. (6.2) becomes

$$P_\nu(u) = e^{-j2\pi\tau u} \sum_{\ell=-\infty}^{\infty} \tilde{P}_\nu\left(u - \frac{\ell}{\delta}\right).$$

First consider an SD system. Eq. (6.5) holds under the same assumptions as before; i.e., that

$$\tilde{P}_\nu\left(\frac{k}{\Delta}\right) = 0 \qquad \forall k \in \mathbf{Z}_0.$$

In this case we obtain

$$\begin{aligned} P_q(u_q) &= \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(u_q - \frac{k}{\delta}\right) e^{j2\pi\tau k/\delta} \\ &= \operatorname{sinc}(u_q) \star \left[W_{\frac{1}{\delta}}(u_q) e^{j2\pi\tau u_q}\right] \end{aligned}$$

so that

$$p_q(q) = \frac{\Delta}{2^L}\Pi_\Delta(q)W_\delta(q+\tau).$$

This equation is not quite right. It is offset by τ because the dither subtracted after quantization contained the virtual offset. Removing this offset yields the correct expression:

$$p_q(q) = \frac{\Delta}{2^L}\Pi_\Delta(q-\tau)W_\delta(q).$$

In other words, the input and quantization error are statistically independent of each other under the same conditions as before and the error pdf is precisely what one would expect.

Now consider an NSD system with virtually offset digital dither. Eq. (6.9) becomes

$$\frac{d^m G_\nu}{du^m}(u) = \sum_{k=-\infty}^{\infty}\sum_{r=0}^{m}\binom{m}{r}\frac{d^r}{du^r}\left[\frac{\sin(\pi\Delta u)}{\pi\Delta u}e^{-j2\pi\tau u}\right]\frac{d^{m-r}\tilde{P}_\nu}{du^{m-r}}\left(u-\frac{k}{\delta}\right) \tag{6.11}$$

so that Theorem 6.2 holds precisely as before. Eq. (6.10) holds if the offset dither pdf is used in the calculations since, in this case, no dither subtraction takes place to introduce spurious offsets.

Chapter 7

Conclusions

7.1 SD and NSD Quantizing Systems

We will take this final opportunity to summarize the principal differences between SD and NSD systems.

First, the dither signal must be available for subtraction at playback in SD systems, and so either the dither sequence or information sufficient to reconstruct it must be stored or transmitted with the signal. That NSD systems do not require this added information at playback is their primary advantage over SD systems.

On the other hand, SD systems can render the total error signal statistically independent of the input signal as well as rendering error samples separated in time statistically independent of one another. This ensures that the power spectrum of the total error is independent of the system input, and that it is spectrally flat (white) even if the dither signal is not. A dither capable of doing all this is simple

iid RPDF dither. The total error variance in SD systems is always $\Delta^2/12$.

NSD systems, on the other hand, cannot render the total error statistically independent of the input, but can only render specified moments of the error input independent. Furthermore, dithers of successively higher order are required for each moment to be so rendered. For instance, to make the mean and variance of the total error independent of the input, a second-order dither is required—say 2RPDF (TPDF) dither with twice the variance of simple RPDF dither. The increased dither variance is reflected in increased total error variance, which is $\Delta^2/4$ for 2RPDF dither, and it has been shown (see Theorem 4.11) that this is the lowest possible total error variance achievable if the first two error moments are to be successfully rendered input independent. Note that the resulting error variance is three times as great as that of an SD system, which renders the error statistically independent of the system input, thereby ensuring the constancy of *all* the error moments. This difference in the resulting total error variance is the principal advantage of SD systems over NSD systems.

Another difference between the two types of systems is that in an SD system the total error spectrum is flat irrespective of the dither spectrum, whereas spectrally shaped non-subtractive dither will result in a non-flat error spectrum which, if the system is properly dithered, will be the sum of the dither spectrum and a white "quantization noise" component. Some interest has been expressed in tailoring the shape of the dither to result in total error spectra which are perceptually quieter than flat spectra. Unfortunately the aforementioned white component is unaffected by altering the dither spectrum. Thus, for such purposes, it is usually preferable to use noise-shaping error feedback, which can shape the entire error spectrum as desired. Conditions have been given above (see Section 5.2) which will ensure

that the resulting error spectrum is of a fixed predictable form. Spectrally shaped dithers may still be of interest in high-speed applications, however, since non-white dithers of any order can be generated using only one new pseudo-random number per input sample.

It has been shown that if the quantizer output from an SD system, with or without noise-shaping error feedback, is to be replayed without subtraction of the dither signal, then, to avoid input-dependent spectral modulation of the error, the dither used should satisfy the conditions necessary to ensure absence of error spectral modulation in an NSD system.

7.2 Audio Applications

Much of the present investigation was originally motivated by questions which arose in audio signal processing. Some comments regarding such applications seems appropriate.

For audio signal processing purposes, there seems to be little point in rendering any moments of the total error 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 requantized very coarsely (to 8 bits from 16) in order to render the requantization error essentially independent of low-level non-linearities in the digital-to-analogue conversion system through which the lis-

tening took place. In addition, the corresponding total error signals (output minus input) were used in listening tests in order to check for any audible dependences on the input. Using undithered quantizers resulted in clearly audible distortion and noise modulation in the output and error signals. A subtractively dithered quantizing system using iid 1RPDF dither was found to eliminate all audible input dependences in the error signal, which was confirmed to be audibly equivalent to a steady white noise. A non-subtractively dithered quantizing system using the same dither eliminated all distortion, but the residual noise level was found to vary audibly in an input-dependent fashion. When 2RPDF dither was employed, no instance was found in which the error was audibly distinguishable from a steady white noise entirely unrelated with the input, although the level of this noise was, of course, somewhat higher than that observed in the subtractively-dithered system. Admittedly, these tests were informal, and there remains a need for formal psychoacoustic tests of this sort involving many participants under carefully controlled conditions.

The use of of non-subtractive, iid 2RPDF dither is recommended for most audio applications requiring multi-bit quantization or requantization operations, since this type of dither renders the power spectrum of the total error independent of the input, while incurring the minimum increase in error variance. This kind of dither is easy to produce for digital requantization purposes by simply summing two independent 1RPDF pseudo-random processes, which may be rapidly generated using linear congruential algorithms [48, 21]. The resulting digital dither can be used to feed a digital-to-analogue converter for analogue dithering applications.

Important extensions of the work reported herein would include the analysis of systems of interest incorporating non-linearities other than infinite, uniform quan-

tizers. In particular, a complete statistical description of non-linear systems with feedback, such as sigma-delta converters, awaits development.

In closing, it is proposed that appropriate dithering prior to (re)quantization is as fitting as appropriate anti-alias filtering prior to sampling—both serve to eliminate classes of signal-dependent errors.

Bibliography

[1] Whittaker, E.T., "On the Functions which are Represented by the Expansions of the Interpolation-Theory," *Proc. Roy. Soc. Edinburgh*, vol. 35, pp. 181–194, (1915).

[2] 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.).

[3] 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.).

[4] 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.).

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

[6] 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.).

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

[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] 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).

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

[11] Wannamaker, R.A., S.P. Lipshitz, J. Vanderkooy, and J.N. Wright, "A Theory of Non-Subtractive Dither," *IEEE Trans. Signal Processing,* accepted for publication.

[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] 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).

[24] 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.

[25] 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.

[26] 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.).

[27] 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.

[28] Kingman, J.F.C., and S.J. Taylor, *Introduction to Measure and Probability,* Cambridge University Press, Cambridge, UK (1966).

[29] Cristescu, R., and G. Marinescu, *Applications of the Theory of Distributions,* John Wiley & Sons, NY, NY (1973).

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

[31] Oppenheim, A.V., and R.W. Schafer, *Digital Signal Processing,* Prentice-Hall, Englewood Cliffs, NJ (1983).

[32] Davenport, W.B. Jr., and W.L. Root, *An Introduction to the Theory of Random Signals and Noise,* McGraw-Hill, New York, NY, (1958).

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

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

[35] Rényi, A., *Probability Theory,* North-Holland Publ. Co., Amsterdam, the Netherlands (1970).

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

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

[38] Titchmarsh, E.C., *Introduction to the Theory of Fourier Integrals,* Clarendon Press, Oxford, UK (1950).

[39] Sheppard, W.F., "On the Calculation of the most Probable Values of Frequency-Constants, for Data arranged according to Equidistant Divisions of a Scale," *Proc. London Math. Soc.*, Series 1, vol. 29, pp. 353–380 (1897-98).

[40] Craven, P.G., and M.A. Gerzon, "Compatible Improvement of 16-Bit Systems Using Subtractive Dither," presented at the 93rd Convention of the Audio Engineering Society, San Francisco, CA (1992 Oct.), preprint 3356.

[41] Lagadec, R., "New Frontiers in Digital Audio," presented at the 89th Convention of the Audio Engineering Society, Los Angeles, CA (1990 Sept.), preprint 3002.

[42] Lagadec, R., private communications (1990 Sept., 1991 Feb.).

[43] Lipshitz, S.P., and C. Travis, "The Generation of Non-White Dithers of Specified Probability Density Function," presented at the 94th Convention of the Audio Engineering Society, Berlin, Germany, (1993 Mar.).

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

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

[46] Gerzon, M.A., P.G. Craven, J.R. Stuart and R.J. Wilson, "Psychoacoustic Noise Shaped Improvements in CD and Other Linear Digital Media," presented at the 94th Convention of the Audio Engineering Society, Berlin, Germany (1993 Mar.), preprint 3501.

[47] *CRC Standard Mathematical Tables and Formulae, 30th Ed.,* CRC Press, Boca Raton, FL, 1995.

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

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

[50] Friedlander, F.G., *Introduction to the Theory of Distributions,* Cambridge University Press, Cambridge, UK (1982).

[51] Richards, I., and H. Youn, *Theory of Distributions,* Cambridge University Press, Cambridge, UK (1990).

[52] Jones, D.S., *The Theory of Generalised Functions,* Cambridge University Press, Cambridge, UK (1982).

[53] Lighthill, M.J., *Introduction to Fourier Analysis and Generalised Functions,* Cambridge University Press, Cambridge, UK (1970).

[54] Helson, H., *Harmonic Analysis,* Addison-Wesley Publ. Co., Reading, MA (1983).

Appendix A

Generalized Functions

This appendix provides a brief outline of the theory of *tempered generalized functions*, also known as *tempered distributions.* Few results will be proven in detail, but appropriate references will be given and some theoretical issues arising in the body of the thesis will be resolved. It will be assumed that the reader is familiar with the L_1 Fourier transform as defined by Eq. (2.1).

Definition A.1 *A function $\phi \in C^\infty(\mathbf{R}^n)$ is said to be a* **rapidly decreasing test function** *if*

$$\sup_{x \in \mathbf{R}^n} |x^\alpha \phi^{(\beta)}(x)| < \infty$$

for all pairs of multi-indices α, β. The vector space of such functions is denoted by $\mathcal{S}$.

This space contains, for instance, Gaussians and even functions compactly supported on any given interval such as $\phi(ax+b)$ where $a, b \in \mathbf{R}$ and

$$\phi(x) = \begin{cases} e^{1/(x^2-1)}, & |x| < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Theorem A.1 *$\mathcal{S}$ is stable under the following operations: differentiation, multiplication by polynomials, affine transformations and the (L_1) Fourier transform.*

Proof: The assertion is that each of the indicated operations maps $\mathcal{S}$ into $\mathcal{S}$. This is obvious from the definition of a rapidly decreasing test function in each case except for the last, which we prove for $\mathbf{R}^1$ (the extension to $\mathbf{R}^n$ is straightforward). We wish to show that if $\phi \in \mathcal{S}$ then its L_1 Fourier transform $\hat{\phi} \in \mathcal{S}$; i.e., that $\hat{\psi}(t) = t^N \hat{\phi}^{(k)}(t)$ is bounded for any given integers $N, k > 0$. Now, $\hat{\psi}$ is the (L_1) Fourier transform [49] of

$$\begin{aligned} \psi(x) &= \left(\frac{1}{2\pi j}\right)^N \frac{d^N}{dx^N}(x^k \phi(x)) \\ &= \left(\frac{1}{2\pi j}\right)^N \sum_{i=0}^{k} \binom{N}{i} \frac{k!}{(k-i)!} x^{k-i} \phi^{(N-i)}(x). \end{aligned}$$

Each term in this sum is a rapidly decreasing test function and thus so is $\psi(x)$. Thus $\psi(x)$ is absolutely integrable and

$$|\hat{\psi}(t)| = \left| \int_{-\infty}^{\infty} \psi(x) e^{-j2\pi xt} dx \right| \leq \int_{-\infty}^{\infty} |\psi(x)| dx < \infty.$$

□

Definition A.2 *A linear functional* $u : \mathcal{S} \to \mathbf{R}$ *is called a* **tempered generalized function** *or* **tempered distribution** *if there exist a real number* $C \geq 0$ *and a nonnegative integer* N *such that*

$$|\langle u, \phi \rangle| \leq C \sum_{|\alpha| \leq N} \sup |\phi^{\alpha}|$$

for all $\phi \in \mathcal{S}$*. The generalized function is then said to be of order* N*. The vector space of tempered generalized functions is denoted by* $\mathcal{S}'$*.*

Inequalities of this sort are known as *semi-norm estimates* [50]. The use of the inner product notation $\langle u, \phi \rangle$ to denote the operation of the functional u on the test function ϕ is conventional. We will now show how this operation in fact corresponds to the formation of an inner product in the usual sense in many cases of interest.

The generalized functions appearing in this thesis are all of order $N = 0$. An example of such is the so-called Dirac delta function, δ, defined by

$$\langle \delta, \phi \rangle = \phi(0).$$

This is a special case of the general result that any finite Borel measure μ determines a generalized function of order zero by

$$\langle \mu, \phi \rangle = \int \phi d\mu.$$

(The converse is also true; see [50].) Another example is the tempered distribution associated with an ordinary locally integrable function, f, of polynomial growth, which is determined by

$$\langle u, \phi \rangle = \int_{-\infty}^{\infty} f \phi dx.$$

Since ϕ can be chosen with support on any given interval, this determination is unique up to an equivalence class of functions equal almost everywhere (i.e., differing only on a set of measure zero).

Theorem A.2 *The following operations on tempered distributions $u, v \in \mathcal{S}'$ produce tempered distributions:*

(i) $\langle u+v, \phi\rangle = \langle u, \phi\rangle + \langle v, \phi\rangle$;

(ii) $\langle cu, \phi\rangle = c\langle u, \phi\rangle, \quad c \in \mathbf{C}$;

(iii) $\langle u^\alpha, \phi\rangle = (-1)^{|\alpha|}\langle u, \phi^\alpha\rangle, \quad$ *for multi-indices* α;

(iv) $\langle A^* u, \phi\rangle = \dfrac{1}{|\det A|}\Big\langle u, \phi(A^{-1}x)\Big\rangle, \quad$ *for a real non-singular* $n \times n$ *matrix* A;

(v) $\langle u(x-a), \phi\rangle = \langle u, \phi(x+a)\rangle, \quad a \in \mathbf{R}^n$;

(vi) $\langle gu, \phi\rangle = \langle u, g\phi\rangle, \quad g \in C^\infty(\mathbf{R}^n)$ *and of polynomial growth.*

Furthermore, when u and v correspond to ordinary functions, the results of these operations are consistent with those for ordinary functions.

The proofs are straightforward [50, 51]. As an example we prove Part (v). u is a tempered distribution and $\phi(x+a) \in \mathcal{S}$, so $u(x-a)$ has a finite semi-norm estimate and is a tempered distribution. If u corresponds to an ordinary function f then

$$\langle u(x-a), \phi\rangle = \int f(x)\phi(x+a)dx = \int f(x-a)\phi(x)dx$$

which is the definition of the generalized function associated with $f(x-a)$.

Writing $u(x-a)$ is an abuse of notation, although the meaning should be clear. Some authors also denote the composition of a distribution with a coordinate transformation by $u(Ax)$ instead of A^*u.

As an application of Theorem A.2, consider the generalized function associated with the Heaviside step function

$$H(x) = \begin{cases} 1, & x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

To compute its derivative we write

$$\left\langle H^{(1)}, \phi \right\rangle = -\left\langle H, \phi^{(1)} \right\rangle = -\int_0^\infty \phi^{(1)}(x)dx = \phi(0) = \langle \delta, \phi \rangle, \qquad \forall \phi \in \mathcal{S}.$$

Thus

$$H^{(1)} = \delta.$$

Arbitrary products of distributions are not defined. Theorem A.2(vi) shows how one can straightforwardly define a product when one distribution corresponds to an infinitely differentiable function of polynomial growth. The problem is that $\mathcal{S}$ is not stable under products with arbitrary functions, although some special cases can be handled. Particularly useful is the following [52]:

Definition A.3 *If g is a continuous function in some neighbourhood of the origin, then*

$$g\delta = g(0)\delta.$$

A product of generalized functions which is always well-defined is the so-called *tensor product* of two distributions in distinct spaces:

Theorem A.3 *Suppose that $u \in \mathcal{S}'(\mathbf{R}^n)$ and $v \in \mathcal{S}'(\mathbf{R}^m)$. Then there is a unique element of $\mathcal{S}'(\mathbf{R}^{m+n})$ called the* tensor product *of u and v, written $u \otimes v$, such that*

$$\langle u \otimes v, \phi\psi \rangle = \langle u, \phi \rangle \langle v, \psi \rangle, \qquad \phi \in \mathcal{S}(\mathbf{R}^m), \quad \psi \in \mathcal{S}(\mathbf{R}^n).$$

For a proof, see [50]. We will freely abuse notation and write down such tensor products as

$$\delta(x, y) = \delta(x)\delta(y).$$

Partial derivatives are defined in the obvious fashion. The tensor product of countable distributions is definable in the same manner.

We now introduce the Fourier transform of a tempered generalized function.

Definition A.4 *The (forward) Fourier transform $\hat{u}$ and inverse Fourier transform $\check{u}$ of a tempered distribution are defined by*

$$\begin{aligned} \langle \hat{u}, \phi \rangle &= \left\langle u, \hat{\phi} \right\rangle \\ \langle \check{u}, \phi \rangle &= \left\langle u, \check{\phi} \right\rangle, \end{aligned}$$

where $\hat{\phi}$ and $\check{\phi}$ are the ordinary (L_1) forward and inverse Fourier transforms, respectively, of test functions $\phi \in \mathcal{S}$.

Note that $\hat{u}$ and $\check{u}$ are tempered distributions since $\mathcal{S}$ is stable under Fourier transforms.

The following identities hold.

Theorem A.4 *Let u, v and the constants be the same as in Theorem A.2 and let $\overline{A}$ denote the transpose of A. Then*

(i) $\hat{u}(x) = \check{u}(-x)$;

(ii) $[u + v]\hat{} = \hat{u} + \hat{v}$;

(iii) $[cu]\hat{} = c\hat{u}$;

(iv) $[u^\alpha]\hat{} = (j2\pi)^{|\alpha|} x^\alpha \hat{u}, \qquad$ *where* $x^\alpha \triangleq \prod_{j=1}^n x_j^{\alpha_j}$;

(v) $[x^\alpha u]\hat{} = \left(\frac{j}{2\pi}\right)^{|\alpha|} \hat{u}^\alpha$;

(vi) $[u(x-a)]\hat{} = e^{-j2\pi a\cdot x}\hat{u}(x)$;

(vii) $[e^{j2\pi a\cdot x}u(x)]\hat{} = \hat{u}(x-a)$;

(viii) $[A^*u]\hat{} = \frac{1}{|\det A|}\overline{A^{-1}}^*\hat{u}$;

(ix) $u\hat{}\check{} = u\check{}\hat{} = u$.

Furthermore, where u and v correspond to ordinary functions, the results of these operations are consistent with those for ordinary functions.

Again the proofs are not difficult (see, for instance, [51]). As an example we will prove Part (vi):

$$\begin{aligned}
[u(x-a)]\hat{} &= \left\langle u(x-a), \hat{\phi}(x)\right\rangle \\
&= \left\langle u(x-a), \int \phi(t)e^{-j2\pi x\cdot t}dt\right\rangle \\
&= \left\langle u, \int \phi(t)e^{-j2\pi (x+a)\cdot t}dt\right\rangle \\
&= \left\langle u, \int [\phi(t)e^{-j2\pi a\cdot t}]e^{-j2\pi x\cdot t}dt\right\rangle \\
&= \left\langle \hat{u}, \phi(x)e^{-j2\pi a\cdot x}\right\rangle \\
&= \left\langle \hat{u}e^{-j2\pi a\cdot x}, \phi(x)\right\rangle.
\end{aligned}$$

Note that Part (i) of the theorem can be used to rewrite each of the subsequent parts in terms of inverse Fourier transforms. Furthermore we observe that the Fourier transform of a distribution is unique since the Fourier transform operation has an inverse; i.e., the Fourier transform is a bijective mapping between $\mathcal{S}'$ and $\mathcal{S}'$.

As a simple example, consider the Fourier transform of the Dirac delta:

$$\left\langle \hat{\delta}, \phi \right\rangle = \left\langle \delta, \hat{\phi} \right\rangle = \hat{\phi}(0) = \langle 1, \phi \rangle, \qquad \forall \phi \in \mathcal{S}.$$

Thus $\hat{\delta} = 1$.

Now let us consider a more complicated example: the Fourier transform of the tempered generalized function

$$W_\Delta(x) = \sum_{k=-\infty}^{\infty} \delta(x - k\Delta).$$

We should first check that this is in fact a tempered distribution, for which we require the following:

Definition A.5 *Consider a sequence* $\{u_n\} \subset \mathcal{S}'$ *and* $u \in \mathcal{S}'$. *We say that* u_n **converges** *to* u, *written* $u_n \to u$, *if*

$$\lim_{n\to\infty} \langle u_n, \phi \rangle = \langle u, \phi \rangle$$

for each $\phi \in \mathcal{S}$.

We can show that the partial sums

$$\left\langle \sum_{k=-n}^{n} \delta(x - k\Delta), \phi(x) \right\rangle = \sum_{k=-n}^{n} \phi(k\Delta)$$

converge as $n \to \infty$ and that the limit is in $\mathcal{S}'$. In fact, this is trivial since $\phi(x)$ decreases faster than any power of $|x|$. (How we index the summands is also clearly irrelevant.) Now we can state the following important theorem [51, 53]:

Theorem A.5 *If*

$$W_\Delta(x) = \sum_{k=-\infty}^{\infty} \delta(x - k\Delta)$$

then

$$\hat{W}_\Delta = \frac{1}{|\Delta|} W_{\frac{1}{\Delta}}$$

Proof:[Outline.] W_Δ is a periodic generalized function; i.e., $W_\Delta(x+\Delta) = W_\Delta(x)$. Using Theorem A.4(vi) we observe that

$$(e^{-j2\pi\Delta t} - 1)\hat{W}_\Delta(t) = 0.$$

$(e^{-j2\pi\Delta t} - 1)$ vanishes at $t = k/\Delta$, $\forall k \in \mathbf{Z}$. We consider only the origin, $k = 0$, since the situation for other values of k is similar. $(e^{-j2\pi\Delta t} - 1)$ is $O(t)$ at the origin and it can be shown [51] that $tu(t) = 0$ if and only if $u = C\delta$ for some $C \in \mathbf{R}$. Thus

$$\hat{W}_\Delta = \sum_{k=-\infty}^{\infty} c_k \delta\left(t - \frac{k}{\Delta}\right).$$

Now W_Δ is itself a sum of delta functions, so by the same brand of reasoning $\hat{W}_\Delta$ is periodic with period $1/\Delta$; i.e., we can write

$$(e^{-j2\pi x/\Delta} - 1)W_\Delta(x) = 0$$

which implies that $\hat{W}_\Delta(t + \frac{1}{\Delta}) = \hat{W}_\Delta(t)$. Thus

$$\hat{W}_\Delta(t) = C \sum_{k=-\infty}^{\infty} \delta\left(t - \frac{k}{\Delta}\right) = CW_\Delta(\Delta^2 t)$$

for some real constant C. Then Theorems A.4(viii) and (ix) give

$$W_\Delta = W_\Delta{}^{\check{}\check{}} = \frac{C^2}{\Delta^2} W_\Delta$$

whence $C^2 = 1/\Delta^2$. Finally we observe that $\phi(x) = e^{-\pi x^2} = \hat{\phi}(x) \in \mathcal{S}$ is everywhere greater than zero, so $C = 1/|\Delta|$.

□

Definition A.6 *When $\hat{u}$ corresponds to an ordinary function continuous in some neighbourhood of the origin, we may define the* **definite integral** *of u by*

$$\int_{-\infty}^{\infty} u(x)dx = \hat{u}(0).$$

Thus we obtain, for instance, the intuitively satisfying results that $\int \delta(x)dx = 1$ but that $\int W_\Delta(x)dx$ is undefined.

A popular operation on generalized functions which requires some care is that of convolution. We introduce the notion of a compactly supported generalized function:

Definition A.7 *A distribution u is said to have* **compact support** $\text{support}(u) \subseteq [a,b]$ *if $\langle u, \phi\rangle = 0$ for all test functions whose support lies outside $[a,b]$. The vector space of compactly supported distributions is denoted by $\mathcal{E}'$.*

The following elegant and useful theorem may be found in [50]:

Theorem A.6 *Suppose that $u \in \mathcal{S}'(\mathbf{R}^n)$ and $v \in \mathcal{E}'(\mathbf{R}^n)$. Then*

$$\begin{aligned} u \star v &= \langle v(y), \langle u(x), \phi(x+y)\rangle\rangle \\ &= \langle u(x), \langle v(y), \phi(x+y)\rangle\rangle, \quad \phi \in \mathcal{S}, \end{aligned}$$

is an element of $\mathcal{S}'(\mathbf{R}^n)$, called the **convolution** *of u with v, and furthermore*

$$(u \star v)\hat{} = \hat{u}\hat{v}.$$

Since $v \in \mathcal{E}'(\mathbf{R}^n)$ it turns out that $\hat{v}$ corresponds to an ordinary function of polynomial growth in $C^\infty(\mathbf{R}^n)$. Thus the product $\hat{u}\hat{v}$ is well-defined by Theorem A.2(vi).

Unfortunately this result is not quite as general as we might like it to be. That one of the distributions must be compactly supported is a severe restriction, and one which is not always warranted. Of particular interest are convolutions involving W_Δ. We will prove some useful results concerning such convolutions, but first we require the following notions from the Fourier theory of ordinary functions.

Definition A.8 *For a given function f we say that $f \in L_1(\mathbf{R})$ if*

$$\int_{-\infty}^{\infty} |f(x)|dx < \infty.$$

Definition A.9 *A function f is said to have* **bounded variation** *on $\mathbf{R}$ if*

$$\sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|$$

is bounded above for all ordered finite sequences $x_0 < x_1 < \ldots < x_n$ in $\mathbf{R}$.

Any function displaying only a finite number of finite discontinuities in any closed interval will have bounded variation.

Definition A.10 *A function f is said to be* **normalized** *if for each $x \in \mathbf{R}$*

$$f(x) = \frac{f(x^+) + f(x^-)}{2}.$$

Theorem A.7 (Poisson's Summation Formula) *Suppose $f \in L_1(\mathbf{R})$, is of bounded variation and normalized. Then*

$$\sum_{k=-\infty}^{\infty} f(x+k) = \sum_{k=-\infty}^{\infty} \hat{f}(k) e^{-j2\pi kx}.$$

In particular,

$$\sum_{k=-\infty}^{\infty} f(k) = \sum_{k=-\infty}^{\infty} \hat{f}(k).$$

Proof: Since $f \in L_1(\mathbf{R})$,

$$\sum_{k=-\infty}^{\infty} \int_0^1 |f(x+k)|dx = \int_{-\infty}^{\infty} |f(x)|dx < \infty.$$

Thus the sum

$$\sum_{k=-\infty}^{\infty} f(x+k)$$

converges absolutely almost everywhere and defines an absolutely integrable function $g(x)$ on $[0,1]$. $g(x)$ is normalized and of bounded variation so that it can be expanded in a Fourier series [54]:

$$\sum_{k=-\infty}^{\infty} f(x+k) = \sum_{k=-\infty}^{\infty} \hat{f}(k) e^{-j2\pi kx}.$$

In particular, this can be evaluated at $x = 0$.

□

Poisson's summation formula easily generalizes using Theorem A.4(viii) to give

$$\Delta \sum_{k=-\infty}^{\infty} f(x+k\Delta) = \sum_{k=-\infty}^{\infty} \hat{f}\left(\frac{k}{\Delta}\right) e^{-j2\pi kx/\Delta}.$$

The formula may also be turned around to give the following:

Lemma A.1 *Suppose $f \in L_1(\mathbf{R})$ is of bounded variation and normalized. Then*

$$\sum_{k=-\infty}^{\infty} f(k) e^{-j2\pi kx} = \sum_{k=-\infty}^{\infty} \hat{f}(x-k).$$

Proof: $f(z)e^{-j2\pi zx}$ is in $L_1(\mathbf{R})$, of bounded variation and normalized for any given $x \in \mathbf{R}$.

□

Theorem A.8 *If $f \in L_1(\mathbf{R})$ is of bounded variation and normalized then*

$$\hat{f} \star W_1$$

is a tempered distribution.

Proof: By Definition A.5 we need only show that the sequence of partial sums

$$\hat{f}(t) \star \sum_{k=-n}^{n} \delta(t-k) = \sum_{k=-n}^{n} \hat{f}(t-k)$$

converges as $n \to \infty$ and that the limit is a tempered distribution. The function f satisfies Lemma A.1 and thus

$$\sum_{k=-\infty}^{\infty} f(k) e^{-j2\pi kx} = \sum_{k=-\infty}^{\infty} \hat{f}(x-k).$$

Thus the sequence of partial sums converges for almost everywhere, thereby defining a locally integrable periodic function $g = \hat{f} \star W_1$ which, in turn, defines a tempered distribution.

□

Note that the Lemma provides an alternative means of calculating the convolution.

Finally we can introduce the following novel definition of a product of generalized functions:

Definition A.11 *If $f \in L_1(\mathbf{R})$ is of bounded variation and normalized we define the product*

$$fW_1 = [\hat{f} \star W_1]^{\check{}}.$$

This definition has the trivial generalization

$$fW_\Delta = \frac{1}{\Delta}[\hat{f} \star W_{\frac{1}{\Delta}}]^\vee.$$

Note that f need not be continuous at multiples of Δ.

As an example, consider Eq. (3.2). $f(w') = \Delta\Pi_\Delta(w' - w)$ is absolutely integrable, normalized and of bounded variation so that its product with W_Δ is well-defined in the above sense. Its Fourier transform is

$$\hat{f}(u_{w'}) = \Delta \sum_{k=-\infty}^{\infty} \text{sinc}\,(u_{w'}) e^{-j2\pi w u_{w'}}.$$

Applying Lemma A.1,

$$\begin{aligned}
[\hat{f} \star W_{\frac{1}{\Delta}}](u_{w'}) &= \sum_{k=-\infty}^{\infty} \Delta\Pi_\Delta(k\Delta - w) e^{-j2\pi\Delta k u_{w'}} \\
&= \begin{cases} \frac{1}{2}\left(e^{-j2\pi\Delta n u_{w'}} + e^{-j2\pi\Delta(n+1)u_{w'}}\right), & w = \frac{2n+1}{2}\Delta, n \in \mathbf{Z}, \\ e^{-j2\pi\Delta\left\lfloor \frac{w}{\Delta} + \frac{1}{2}\right\rfloor}, & \text{otherwise}, \end{cases}
\end{aligned}$$

so that

$$[fW_\Delta](w') = \begin{cases} \frac{1}{2}\left[\delta(w' - n\Delta) + \delta(w' - (n+1)\Delta)\right], & w = \frac{2n+1}{2}\Delta, n \in \mathbf{Z}, \\ \delta\left(w' - \Delta\left\lfloor \frac{w}{\Delta} + \frac{1}{2}\right\rfloor\right), & \text{otherwise}, \end{cases}$$

which is the expected output cpdf for a stochastic quantizer.

Eqs. (3.4) and (3.5) can be handled in the same fashion since

$$\hat{f}(u_{w'}) = \text{sinc}\,(u_{w'}) P_{w,\nu,x}(u_{w'} + u_w, u_\nu, u_x)$$

is the Fourier transform of

$$\begin{aligned}
f(w') &= \Pi_\Delta(w') \star p_{w,\nu,x}(w', \nu, x) e^{j2\pi u_w w'} \\
&= \Pi_\Delta(w' - \nu - x) p_\nu(\nu, x) e^{j2\pi u_w (x+\nu)}
\end{aligned}$$

which is an L_1 function of w'. Thus the convolution of Eq. (3.5) and the product of Eq. (3.4) are well-defined by Theorem A.8.

Appendix B

Time Averages and NSD Quantizers

It was shown in Section 4.4 that proper non-subtractive dither can render any desired moments of the total error independent of the system input. Furthermore, it can render errors which are separated in time uncorrelated, so that the spectrum of the total error is white.

It has been correctly observed by Lagadec [41, 42] that moments and joint moments are quantities which cannot be absolutely determined by empirical means. In real time, they must be estimated from a finite series of signal values. It is not immediately obvious that such estimation will proceed similarly for, on the one hand, the total error signal from a dithered quantization operation, and on the other hand, an independent reference random process. It is the aim of this appendix to elucidate the question of practical estimation of statistical moments, and to show that for purposes of such estimation no significant distinction exists between iid

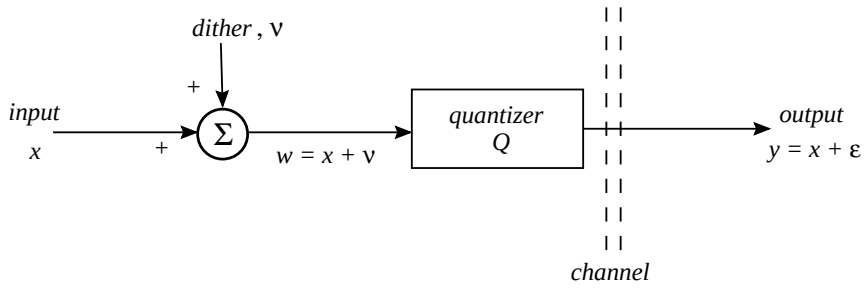


Figure B.1: Schematic of a non-subtractively dithered quantizing system.

noise and the total error produced in a properly dithered quantization operation. In particular, we will allay concerns raised in [41] regarding the convergence of variance estimates in dithered quantizing systems and demonstrate that with regard to moment estimation there is no practical distinction between the total error of a properly dithered quantizing system and an independent iid reference process. Our discussion will be restricted to NSD systems since in SD systems the total error is *precisely* an iid random noise. These investigations have been previously presented by the author in [24].

B.1 Total Error Variance: The Estimation Question

For reference, Fig. B.1 indicates the signals present within an NSD system. Say that, given access to samples of the total error signal, ε, one wishes to calculate its

variance. It is reasonable to hope that a rough estimate of this quantity might be obtained by squaring a set of the sample values (say, N of them) and averaging the results:

$$\text{variance} \approx \frac{1}{N} \sum_{i=0}^{N-1} \varepsilon^2(i). \tag{B.1}$$

One would intuitively expect the accuracy of the result to be better for large N than for small.

Let us proceed with this approach for a system using RPDF dither with a static (dc) system input signal of the form:

$$x(i) = a\Delta,$$

where Δ denotes one LSB of the system (after quantization) and where a is a constant such that $-1/2 \leq a < +1/2$. We will estimate the total error variance and see how the value we obtain changes with N. Fig. B.2 shows results for twenty trials using different, randomly chosen values of a. The curves were produced by evaluating Eq. (B.1) at values of N equal to successive powers of two. For small values of N, the estimates exhibit a broad range of values which sometimes fluctuate wildly as N increases. For $N > 8$ the fluctuations die down and all of the estimates lie roughly in the range from 0 to $\Delta^2/4$, but they show no sign of converging to a single value. (We will see that the reason for this is that the total error variance for an RPDF dithered quantizing system depends upon the value of the static system input value $a\Delta$.)

Now we will try the same experiment with 2RPDF (i.e., TPDF) dither (which is the sort of dither recommended for use in many applications including audio [23, 11, 16, 18]). Fig B.3 shows the results for twenty trials. This time, after initial fluctuations, the variance estimates appear to converge to a value of roughly $\Delta^2/4$.

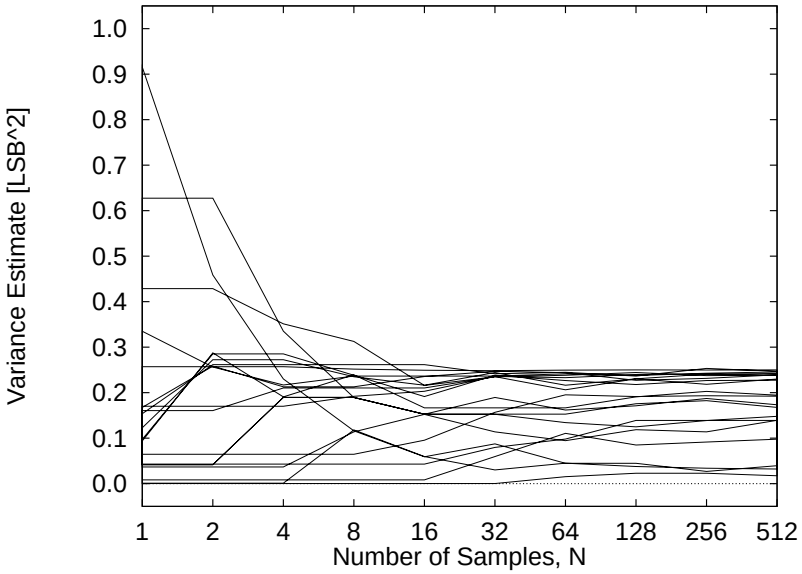


Figure B.2: Total error variance estimates as a function of the number of samples averaged in an RPDF dithered quantizing system. Twenty trials are shown for a system with randomly chosen static input signals of level between -0.5 *and* $+0.5$ *LSB.*

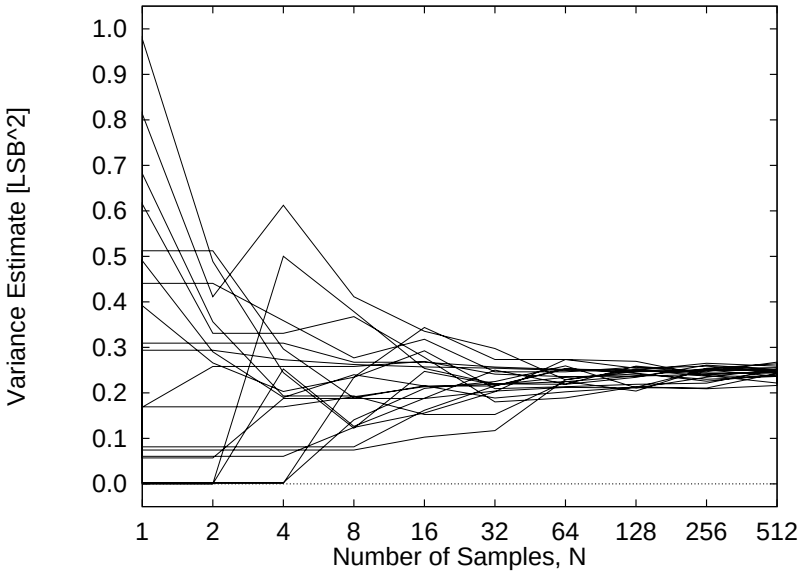


Figure B.3: Total error variance estimates as a function of the number of samples averaged in an 2RPDF dithered quantizing system. Twenty trials are shown for a system with randomly chosen static input signals of level between -0.5 *and* $+0.5$ *LSB.*

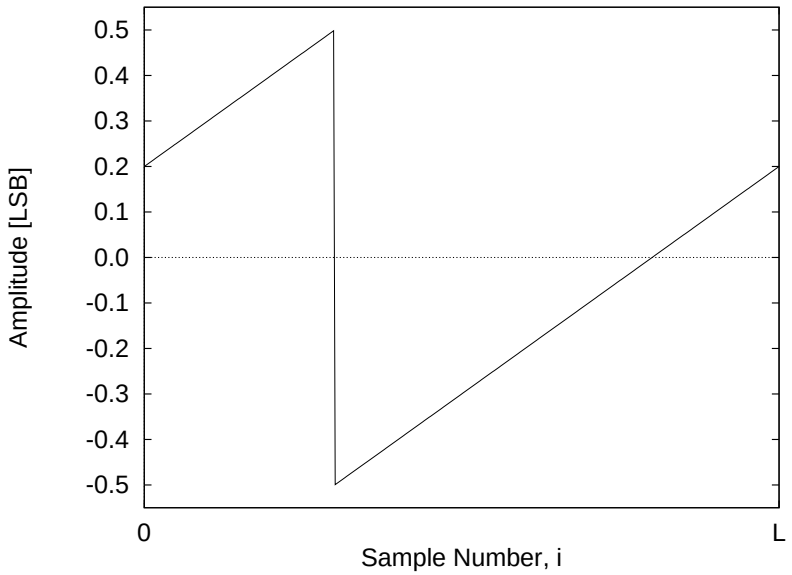


Figure B.4: Periodic bipolar ramp signal ($\alpha = 0.2$).

The observed results can vary with the choice of system input signal. For instance, let us try the above experiments with a system input of the form:

$$x(i) = \Delta \left(\frac{i}{L} + \alpha - \left\lfloor \frac{i}{L} + \alpha + \frac{1}{2} \right\rfloor \right), \tag{B.2}$$

where the "floor" operator $\lfloor\ \rfloor$ returns the greatest integer less than or equal to its argument. The above function is a repeated bipolar ramp of period L samples, amplitude 1 LSB, and starting at a value α, as illustrated in Fig. B.4 (a similar test function was used in [41]). Fig. B.5 shows results for twenty trials using an input ramp signal of period $L = 400$ samples starting at randomly chosen values of α lying between $-1/2$ and $+1/2$. In obtaining this figure RPDF dither was used, while Fig. B.6 shows results of the same experiment using 2RPDF dither. With this choice of input signal, both sets of estimates appear to converge to particular values (of roughly $\Delta^2/6$ and $\Delta^2/4$ respectively), but the 2RPDF curves do so more rapidly with increasing N.

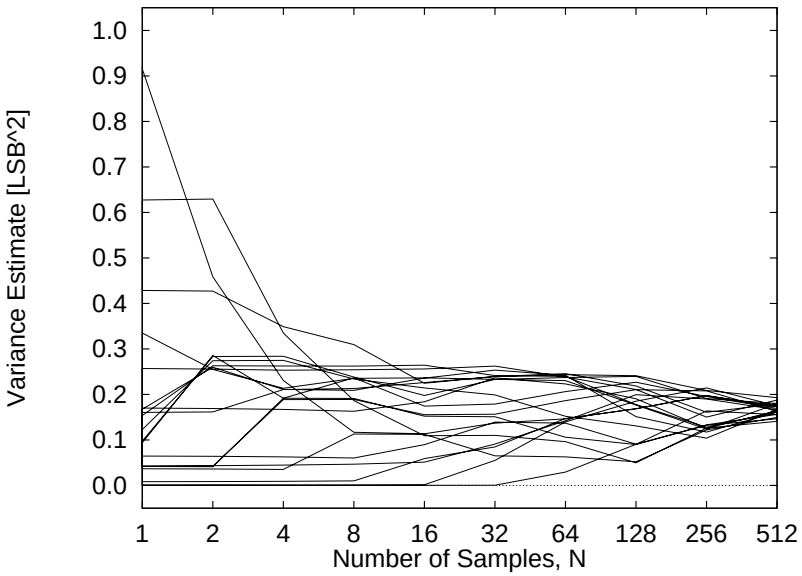


Figure B.5: Total error variance estimates as a function of the number of samples averaged in an RPDF dithered quantizing system. Twenty trials are shown for a system with a repeated ramp input signal with period $L = 400$ samples and randomly chosen starting values α between -0.5 and $+0.5$ LSB.

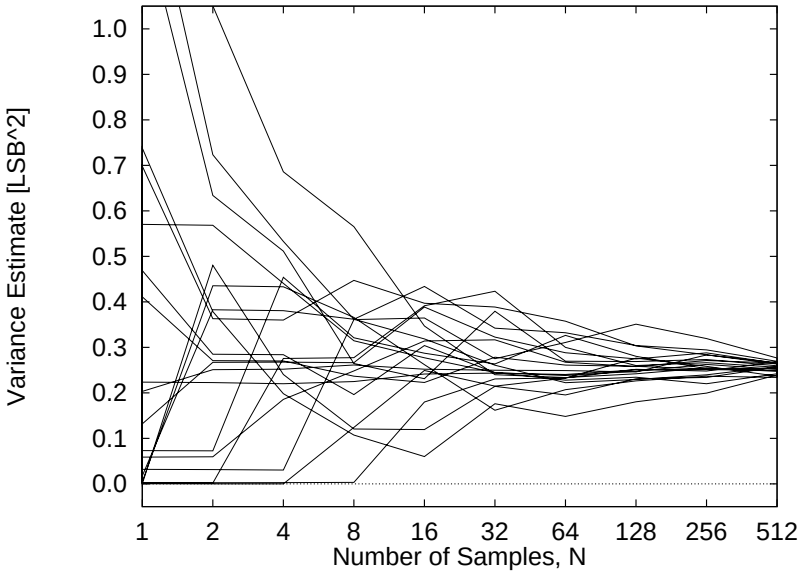


Figure B.6: Total error variance estimates as a function of the number of samples averaged in an 2RPDF dithered quantizing system. Twenty trials are shown for a system with a repeated ramp input signal with period $L = 400$ samples and randomly chosen starting values α.

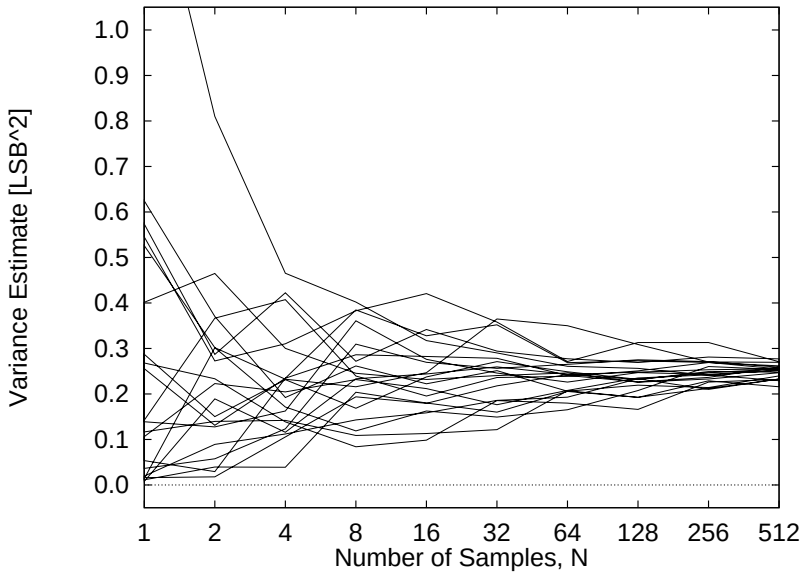


Figure B.7: Total error variance estimates as a function of the number of samples averaged for an iid 3RPDF random noise process. Twenty trials are shown.

It is of interest to compare these curves to similar ones for a random noise process which is not associated with quantization, and whose samples are statistically independent of one another. Fig. B.7 shows the results of twenty trials at estimating the variance of such a process with a piecewise-parabolic pdf (3RPDF or PPDF).[1] The curves appear to converge to a value of roughly $\Delta^2/4$ and the convergence is qualitatively similar to that shown in Fig. B.6.

What should we conclude from these results? Obviously, estimates of the total error variance in dithered systems converge differently given different dither or input signals. In particular, Figs. B.2 and B.5 differ markedly in appearance, although in each case RPDF dither was used. Figs. B.3 and B.6 are more comparable

[1]The reason for this choice of pdf will be clarified in Section B.4.3, although the qualitative appearance of Fig. B.7 would be similar for any independent stationary random process regardless of its distribution.

in their broad features, but how comparable are they to the curves for the independent noise process of Fig. B.7? What are the audible consequences, if any, of the differences? Appropriate dither is supposed to eliminate audible relationships between the system input and the total error. In view of the results obtained above, can we say that the dither is doing its job properly?

The remainder of this appendix attempts to demonstrate that, subject to the choice of an appropriate dither signal, estimates of statistical quantities such as the total error variance converge in a fashion which is not significantly different from the convergence of such estimates for an independent stationary random noise, hence answering the estimation questions raised by Lagadec in [41]. On the other hand, for instance, the use of RPDF dither does not render the total error variance independent of the system input, so that estimates of this quantity are input dependent. This is observed in Fig. B.2 and in [41], which investigated *only* RPDF dithered systems [42]. 2RPDF dither, on the other hand, eliminates all such *noise modulation* (i.e., fluctuations in the error variance), yielding a constant variance of $\Delta^2/4$. Under such conditions, estimates of the total error variance always converge to this value in a well-behaved fashion, as observed, for example, in Fig. B.3.

B.2 Time Averages

In addition to *ensemble averages* represented by expectation values, we can define, for any stochastic process, *time averages* of the form

$$\langle f \rangle \triangleq \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} f(x_i), \tag{B.3}$$

where we have assumed a discrete time variable $t \geq 0$. (Recall that $x_i = x(\zeta, t_i)$, a random variable; see Section 2.1.) Although $\langle f \rangle$ is not time dependent, it is still dependent on ζ and is hence, in general, a random variable.

For many important stochastic processes, however, $\langle f \rangle$ turns out (in the limit) to be independent of ζ so that it is just a number. In particular, processes for which

$$E[f](t) = \langle f \rangle$$

is a numerical constant, independent of ζ and t, for any function f of the random variable, are said to be *ergodic*. The precise conditions for ergodicity are discussed in, for instance, [32]. The essence of one sufficient condition is that

- each sample function $x(\zeta_j, t)$ displays, somewhere in the interval $0 \leq t < \infty$, all the same statistical behaviour as every other sample function (a condition which is assumed to be satisfied in practice), and
- the stochastic process x is *stationary in the strict sense*.

If the relation $E[f](t) = \langle f \rangle$ holds only for some particular f, then the stochastic process is said to be *ergodic in* f. The conditions for this to be true depend on the choice of f and will generally be weaker than the conditions for general ergodicity. Such conditions ensure not only that the mean of finite (N-term) time averages, considered as random variables, tends to the required expectation value as $N \to \infty$, but also that their variance tends to zero.

We will sometimes find it useful to denote the k-th moment of a stochastic process which is ergodic in x^k as m_k. Hence, for processes whose k-th moment is constant with respect to time, we will write that

$$m_k = E[x^k] = \langle x^k \rangle.$$

B.3 Estimators

In monitoring the statistical properties of a stochastic process, no real-time system (e.g., the human ear) can rely on either ensemble or infinite time averages. The pdf of the process at a given time is not usually known *a priori*, so expectation values cannot be computed, and an infinitude of samples is not available for time averaging (neither would one want to wait forever to get the result). In practice, statistical quantities such as moments must be approximated using some practical time-limited algorithm. For instance, we might reasonably hope to arrive at an approximate value, $\hat{m}_k$, of the k-th moment of the stochastic process x by using Eq. (B.3) truncated at the N-th term to give the following formula:

$$\hat{m}_k \triangleq \frac{1}{N} \sum_{i=0}^{N-1} x_i^k. \tag{B.4}$$

We say that the rule assigning a value to $\hat{m}_k$ is an *estimator* for m_k, and that $\hat{m}_k$ is an *estimate* thereof.

Eq. (B.4) represents Eq. (B.1) generalized to estimate an arbitrary (k-th) moment and recast in the parlance of random variables. That is, the terms in Eq. (B.1) were simple numbers, whereas those in Eq. (B.4) are random variables whose statistical properties are described by associated pdf's. It thus captures the properties of not just a single trial estimation, but of such estimations in general.

Observe that Eq. (B.4) *assumes* that the moment to be estimated, $E[x^k](t)$, is at least roughly constant for $0 \leq i \leq N-1$, otherwise the estimate will not represent a meaningful quantity. Also note that, due to the finite number of terms in the summation, the estimate $\hat{m}_k$ is itself a random variable even if the stochastic process in question is ergodic in m_k (i.e., in practice the estimate depends on the sample function obtained). The statistical behaviour of this random variable is obviously

of considerable practical interest, and depends on the choice of estimator.

One desirable property in an estimator is that it yield, on average, the correct result. In particular, it would be nice if

$$E[\hat{m}_k] = m_k.$$

Such an estimator is said to be *unbiased*. We can easily see whether or not the estimator Eq. (B.4) has this property by using the linearity of the expectation value operator:

$$\begin{aligned} E[\hat{m}_k] &= E\left[\frac{1}{N}\sum_{i=0}^{N-1} x_i^k\right] \\ &= \frac{1}{N}\sum_{i=0}^{N-1} E\left[x_i^k\right]. \end{aligned} \tag{B.5}$$

If $E[x_i^k]$ is not a constant for $0 \leq i \leq N-1$ then Eq. (B.5) cannot be further simplified. On the other hand, if the process is ergodic in m_k (or if, at least, the m_k is constant over the time interval of the estimation), then

$$\begin{aligned} E[\hat{m}_k] &= E\left[x_i^k\right] \\ &= m_k \end{aligned} \tag{B.6}$$

so that $\hat{m}_k$ is unbiased.

An estimator may be unbiased, but yield wildly fluctuating results with successive trials. A common measure of the consistency of an estimator is its mean-square error (MSE):

$$\mathrm{MSE}[\hat{m}_k] \triangleq E\Big[(\hat{m}_k - m_k)^2\Big].$$

Note that this is only a meaningful quantity if the process is ergodic in m_k.

What is the MSE of $\hat{m}_k$ as defined by Eq. (B.4)? Assuming that $m_k = E[x_i^k]$ is a constant for $0 \leq i \leq N-1$, then

$$\begin{aligned} \mathrm{MSE}[\hat{m}_k] &= E\left[\left(\frac{1}{N}\sum_{i=0}^{N-1} x_i^k - m_k\right)^2\right] \\ &= \frac{1}{N^2}\sum_{i,j=0}^{N-1} E[x_i^k x_j^k] - \frac{2m_k}{N}\sum_{i=0}^{N-1} E[x_i^k] + m_k^2 \\ &= \frac{1}{N^2}\sum_{i,j=0}^{N-1} E[x_i^k x_j^k] - m_k^2. \end{aligned} \tag{B.7}$$

Now x_i^k and x_j^k $(i \neq j)$ are said to be *uncorrelated* if

$$E[x_i^k x_j^k] = E[x_i^k]E[x_j^k].$$

If this is the case for i and j between 0 and $N-1$ where $i \neq j$, then Eq. (B.7) reduces to

$$\mathrm{MSE}[\hat{m}_k] = \frac{1}{N}\left[\frac{1}{N}\sum_{i=0}^{N-1} E[x_i^{2k}] - m_k^2\right]. \tag{B.8}$$

Finally, if

$$E[x_i^{2k}] = m_{2k}$$

is a numerical constant independent of time for $0 \leq i \leq N-1$, then

$$\mathrm{MSE}[\hat{m}_k] = \frac{1}{N}\left[m_{2k} - m_k^2\right]. \tag{B.9}$$

Eqs. (B.7), (B.8), and (B.9) are of crucial importance for the treatment of moment estimation in dithered systems which is to follow. A noteworthy feature of each is the nature of its dependence upon N, which affects the relationship between the accuracy of an estimate and the number of data points used to produce it. We will refer to the function $\mathrm{MSE}[\hat{m}_k](N)$ as the *convergence curve* for $\hat{m}_k$. In particular, Eq. (B.9) implies that, for any process which is strict-sense stationary,

the convergence curve decreases as $1/N$ with increasing N. It is to this convergence behaviour that we must compare the convergence of moment estimates in dithered quantizing systems.

Let us then proceed to apply an estimator in the form of Eq. (B.4) to the total error process of a dithered quantizing system. We will seek to determine whether or not the resultant estimate converges and, if so, to what and how rapidly, for systems using different types of dither. Any conclusions will be compared to an independent stationary stochastic process.

B.4 Moment Estimation In Dithered Systems

Each signal present in a quantizing system can be considered as a stochastic process, but we will limit our discussion primarily to the statistical properties of ν and ε. (We will henceforth drop from ε_i the subscript i, associating it with time t_i, unless it is specifically required.)

It has been shown (see Eq. (4.48)) that the conditional pdf of ε is

$$p_{\varepsilon|x}(\varepsilon, x) = [\Delta\Pi_\Delta \star p_\nu](\varepsilon) W_\Delta(\varepsilon + x), \tag{B.10}$$

where p_ν is the pdf of the dither. Note that this function is periodic with period Δ. Eq. (B.10) shows that the conditional pdf of ε is functionally dependent on x regardless of the choice of p_ν, so that as x varies with time so do the statistical properties of ε. This is a reflection of the fact that, in an NSD system, ε can never be made a stationary random process independent of the system input.

We will find opportunity to use the following *input-averaged total error pdf*:

$$\begin{aligned}\bar{p}_\varepsilon(\varepsilon) &= \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} p_{\varepsilon|x}(\varepsilon, x)dx \\ &= [\Delta\Pi_\Delta \star p_\nu](\varepsilon).\end{aligned}$$

In the most general case, all of the moments of ε will be time varying, so that estimates of them will be at best approximate and at worst meaningless. Theorem 4.7 indicates however, that using iid nRPDF non-subtractive dither renders the first n moments of the total error independent of the system input, and results, for $n \geq 2$, in a total error power of $(n+1)\Delta^2/12$. The moments of the total error are then given, for $1 \leq k \leq n$, by Eq. (4.31). Of particular usefulness are the expressions for $k = 1$ and $k = 2$:

$$E[\varepsilon] = E[\nu] \tag{B.11}$$

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

Furthermore, Eq. (4.35) shows that such dither will render

$$E[\varepsilon_i^k \varepsilon_j^\ell] = E[\varepsilon_i^k]E[\varepsilon_j^\ell]$$

(i.e., it will render ε_i^k and ε_j^ℓ uncorrelated) for positive integers $k, \ell \leq n$ and $i \neq j$.

These properties will prove sufficient to make several important statements concerning the estimation of statistical quantities in systems using practical dither signals. We will thus proceed to consider systems using three common types of dither: null dither (i.e., undithered systems with $p_\nu(\nu) = \delta(\nu)$), RPDF dither, and 2RPDF (i.e., TPDF) dither. We see from Theorem 4.7 that null dither will not render any moments of the total error independent of the system input (since $P_\nu(u) = 1$). RPDF dither, however, will render (only) the first moment independent, and 2RPDF dither will render (only) the first and second moments independent.

B.4.1 Undithered Systems

We wish to compare moment estimation in undithered systems to moment estimation for some stationary random process. The question naturally arises as to what pdf is appropriate for this *reference process*. We argue that the appropriate reference process is *uniformly distributed;* that is, it has a pdf p_{ref_0} of the form

$$p_{ref_0}(w,t) \stackrel{\triangle}{=} \Pi_\Delta(w).$$

Indeed, the Classical Model of Quantization assumes that the total error in an undithered system has precisely this pdf. Furthermore, if the conditional pdf of the total error in such a system is averaged over all possible input levels, a rectangular function Π_Δ is the result.

The moments of the above reference process are:

$$\begin{aligned} E[\varepsilon] &= 0 \\ E[\varepsilon^2] &= \frac{\Delta^2}{12} \\ E[\varepsilon^k] &= \begin{cases} \dfrac{1}{k+1}\left(\dfrac{\Delta}{2}\right)^k, & k \text{ even}, \\ 0, & k \text{ odd}. \end{cases} \end{aligned}$$

These moments are all time invariant so that for such a reference process we can use Eq. (B.9) to write that

$$\mathrm{MSE}[\hat{m}_k] = \frac{1}{N}\left[m_{2k} - m_k^2\right].$$

How do these results compare with those for an undithered quantizing system? In such a system, the total error is a deterministic function of the input. Hence, for an arbitrary time varying input all moments of the error are time dependent and the MSE of estimates thereof will be ill defined.

On the other hand, for static system inputs, the error is a constant numerical value, ε, so that

$$E[\varepsilon_i^k] = \varepsilon^k.$$

Thus, all estimates of the error will converge *immediately* (i.e., $\mathrm{MSE}[\hat{m}_k](N) = 0$) for any static input. This is little consolation for the fact that the mean error is generally non-zero. The reader should by now be well aware that that undithered quantizing systems produce distortion of signals passing through them. Obviously, the total error in an undithered system behaves very little like an independent stationary random process with respect to moment estimation, but this is not surprising.

B.4.2 Rectangular-pdf Dithered Systems

We argue that the appropriate reference process to which an RPDF dithered system should be compared has a triangular pdf of 2 LSB peak-to-peak amplitude (i.e., 2RPDF). Such a process corresponds to one which would be produced by summing the notional statistically independent uniformly distributed processes associated with the dither and the idealized quantization error of the CMQ. The relevant pdf is

$$\begin{aligned} p_{ref_1}(w,t) &= [\Pi_\Delta \star \Pi_\Delta](w) \\ &= \begin{cases} \dfrac{1}{\Delta}\left(1 - \dfrac{|w|}{\Delta}\right), & 0 \le |w| < \Delta, \\ \\ 0, & \text{otherwise}, \end{cases} \end{aligned}$$

with associated moments

$$E[\varepsilon] = 0 \tag{B.13}$$

$$E[\varepsilon^2] = \frac{\Delta^2}{6} \tag{B.14}$$

$$E[\varepsilon^k] = \begin{cases} \dfrac{\Delta^k}{(k+1)(k+2)}, & k \text{ even}, \\ 0, & k \text{ odd}. \end{cases} \tag{B.15}$$

We must treat the mean error in an RPDF dithered system differently from the higher moments, since it is a constant,

$$m_1 = 0,$$

according to Theorem 4.7 and Eq. (B.11). Also, in such a system we have

$$E[\varepsilon_i \varepsilon_j] = E[\varepsilon_i]E[\varepsilon_j]$$

for $i \neq j$, so that, according to Eq. (B.8), we can write

$$\mathrm{MSE}[\hat{m}_1] = \frac{1}{N}\left[\frac{1}{N}\sum_{i=0}^{N-1} E[\varepsilon_i^2]\right]. \tag{B.16}$$

Unfortunately, $E[\varepsilon_i^2]$ is not constant for time-varying inputs. While this means that the $\mathrm{MSE}[\hat{m}_1]$ does not in general decrease as $1/N$, we can at least compute bounds for it by using Eq. (B.10) to find the variance of ε as a function of x:

$$\begin{aligned} E[\varepsilon^2|x] &= \int_{-\infty}^{\infty} \varepsilon^2 p_{\varepsilon|x}(\varepsilon, x) d\varepsilon \\ &= \begin{cases} x(-x+\Delta), & 0 \leq x < \Delta, \\ E[\varepsilon^2|x-\ell\Delta], & \ell\Delta \leq x < (\ell+1)\Delta. \end{cases} \end{aligned} \tag{B.17}$$

That is, for $0 \leq x < \Delta$, $E[\varepsilon^2|x]$ is a section of a parabola, which is periodically repeated outside this interval as shown in Fig. B.8. The maximum and mini-

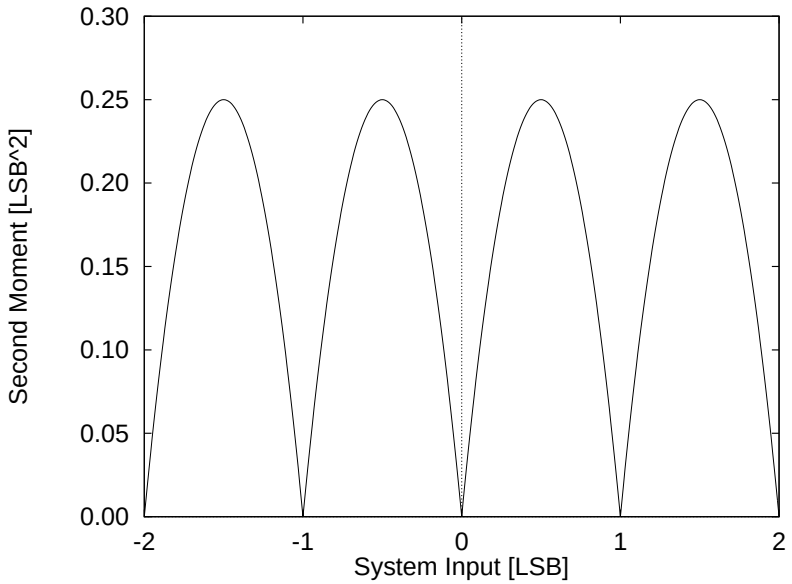


Figure B.8: $E[\varepsilon^2|x]$ as a function of x for an RPDF dithered quantizing system.

mum values of this function are $\Delta^2/4$ and 0, respectively. We may conclude from Eq. (B.16) that $\mathrm{MSE}[\hat{m}_1]$ always lies between the curves

$$f_{min_1}(N) = 0 \qquad \text{and} \qquad f_{max_1}(N) = \frac{\Delta^2}{4N}.$$

The convergence curve $\mathrm{MSE}[\hat{m}_1](N)$ for the reference process is given by Eqs. (B.13), (B.14) and (B.9) as

$$f_{ref_1} = \frac{\Delta^2}{6N}.$$

It is straightforward to calculate from Eq. (B.17) that the average value of $E[\varepsilon^2|x]$ in an RPDF dithered system is $\Delta^2/6$. Substituting this value into Eq. (B.16) for $E[\varepsilon_i^2]$ yields the *average* convergence curve, which is identical to $f_{ref_1}(N)$.

Fig. B.9 shows a family of curves generated in a computer experiment which tried to estimate the mean total error of an RPDF dithered quantizing system with a static system input of 0.5 LSB. Each curve corresponds to a separate trial

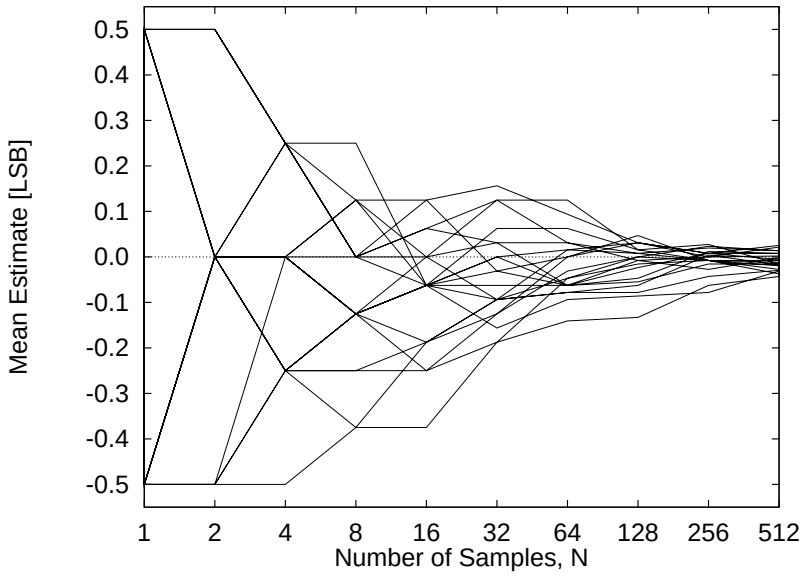


Figure B.9: Estimate of $E[\varepsilon]$ for an RPDF dithered quantizing system with a 0.5 LSB system input, shown as a function of the number of samples used in the estimate.

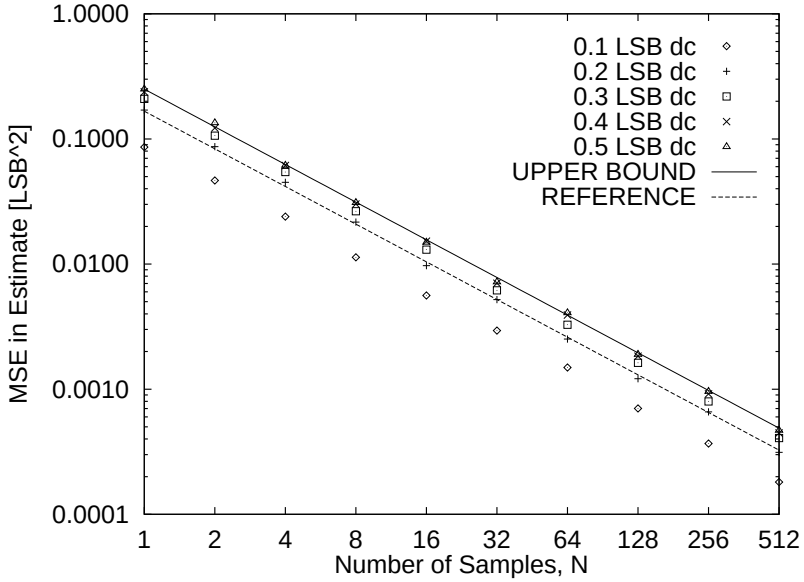


Figure B.10: $MSE[\hat{m}_1](N)$ for an RPDF dithered quantizing system with static system inputs, compared with the theoretical upper bound and reference convergence curves, f_{max_1} and f_{ref_1}. Data averaged over 1000 trials.

using the estimator discussed above, and represents the estimate $\hat{m}_1$ as a function of the number N of samples use to compute it. At any given N, each curve can assume a different value, since the estimate is a random variable, but this would be true even if we were trying to estimate the mean of the reference process. A more meaningful curve to examine is the *mean square* of a large number of such curves, which tends to the convergence curve $\mathrm{MSE}[\hat{m}_1](N)$ (provided that the latter is well-defined). This curve is plotted in Fig. B.10 for a variety of static system inputs. One thousand trials were averaged to yield the data. Note that since the data for a static 0 LSB system input resides at zero it does not appear in the figure. Also shown for comparison are f_{max_1} and f_{ref_1}. Note that for all of the given system inputs, the empirical convergence curves lie on or below the theoretical maximum.

For non-static system input signals, $\mathrm{MSE}[\hat{m}_1](N)$ will not decrease like $1/N$, but will always be bounded by f_{max_1} and f_{min_1}. This is demonstrated in Fig. B.11 for a repeated ramp input (see Eq. (B.2)) with $L = 100$ and $\alpha = 0.0$. We conclude that while we cannot predict the precise functional form of $\mathrm{MSE}[\hat{m}_1]$ for this kind of system, it is bounded from above by a curve which approaches zero at a rate of $1/N$.

Unfortunately, we cannot make similar statements about $\mathrm{MSE}[\hat{m}_k]$ for $k > 1$. For an RPDF dithered system, m_k is not independent of the system input for $k > 1$, so that given a time-varying input signal these moments will also vary with time. Hence, any estimates of such moments will be meaningless. For fluctuating inputs which, in the long run, distribute themselves uniformly over an integral number of quantizing steps, estimates of m_k will tend to converge to the *mean* value of $E[\varepsilon^k|x]$. This is precisely what was observed in Fig. B.5 (similar behaviour was observed in [41]), where the variance estimates slowly converged to a value of $\Delta^2/6$.

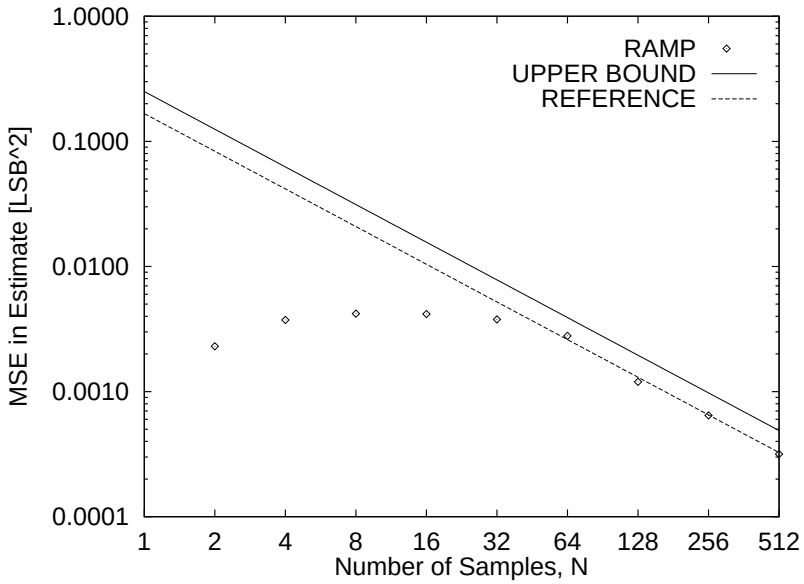


Figure B.11: $MSE[\hat{m}_1](N)$ for an RPDF dithered quantizing system with a repeated ramp system input signal ($L = 200$ and $\alpha = 0.0$). Data averaged over 1000 trials.

Hence this behaviour is a consequence of noise modulation, and is to be expected. For static system inputs, the variance is constant but dependent on the input level, so that estimates thereof will converge but to a different value for different inputs. This is observed in Fig. B.2 where, as we noted, the estimates do not converge to any unique value.

B.4.3 Triangular-pdf Dithered Systems

The appropriate reference process against which to compare the total error of a 2RPDF dithered quantizing system corresponds to the sum of three statistically independent, uniformly distributed random processes, so that it has a piecewise-

parabolic pdf (3RPDF or PPDF) of the form

$$\begin{aligned} p_{ref_2}(w,t) &= [\Pi_\Delta \star \Pi_\Delta \star \Pi_\Delta](w) \\ &= \frac{1}{2\Delta^3} \times \begin{cases} \dfrac{3\Delta^2}{2} - 2w^2, & 0 \le |w| < \dfrac{\Delta}{2}, \\ \left(|w| - \dfrac{3\Delta}{2}\right)^2, & \dfrac{\Delta}{2} \le |w| < \dfrac{3\Delta}{2}, \\ 0, & \text{otherwise}, \end{cases} \end{aligned}$$

with associated moments

$$\begin{aligned} E[\varepsilon] &= 0 \\ E[\varepsilon^2] &= \frac{\Delta^2}{4} \\ E[\varepsilon^4] &= \frac{91\Delta^4}{560} \\ E[\varepsilon^k] &= \begin{cases} \dfrac{3}{4}\dfrac{3^{k+2}-1}{(k+1)(k+2)(k+3)}\left(\dfrac{\Delta}{2}\right)^k, & k \text{ even}, \\ 0, & k \text{ odd}. \end{cases} \end{aligned}$$

In a 2RPDF dithered system, the first two moments of the total error are input independent and given by Eqs. (B.11) and (B.12) as

$$\begin{aligned} m_1 &= 0 \\ m_2 &= \frac{\Delta^2}{4}, \end{aligned}$$

which are equal to the first two moments of the reference process. Hence, Eq. (B.9) allows us to write that

$$\mathrm{MSE}[\hat{m}_1] = \frac{\Delta^2}{4N},$$

which is precisely equal to the convergence function of the first moment of the reference process and independent of the system input signal.

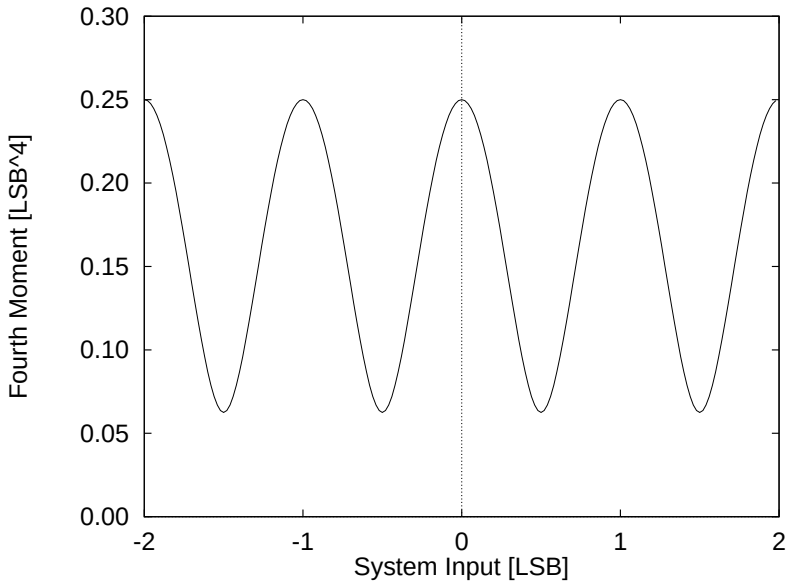


Figure B.12: $E[\varepsilon^4|x]$ as a function of x for a 2RPDF dithered quantizing system.

MSE$[\hat{m}_2]$ depends on the fourth moment of the total error, which is input dependent in this kind of quantizing system, so that the best we can do is set bounds upon it as we did for MSE$[\hat{m}_1]$ in Section B.4.2. Using Eq. (B.10), as before, we find that

$$\begin{aligned} E[\varepsilon^4|x] &= \int_{-\infty}^{\infty} \varepsilon^4 p_{\varepsilon|x}(\varepsilon, x) d\varepsilon \\ &= \begin{cases} 3x^4 - \dfrac{3\Delta^2}{2}x^2 + \dfrac{\Delta^2}{4}, & 0 \le x < \Delta, \\ \quad E[\varepsilon^4|x - \ell\Delta], & \ell\Delta \le x < (\ell+1)\Delta. \end{cases} \end{aligned}$$

This function is shown in Fig. B.12. Its maximum and minimum values are $\Delta^4/4$ and $\Delta^4/16$, respectively. We conclude using Eq. (B.8) that MSE$[\hat{m}_2]$ always lies between the curves

$$f_{min_2}(N) = 0 \qquad \text{and} \qquad f_{max_2}(N) = \frac{3\Delta^4}{16N}.$$

The average value of $E[\varepsilon^4|x]$ is $13\Delta^4/80$, which yields an average convergence curve identical to f_{ref_2}:

$$f_{ref_2} = \frac{\Delta^4}{10N}.$$

Fig. B.13 shows a family of curves generated in a computer experiment which tried to estimate the second moment of the total error in a 2RPDF dithered quantizing system with a static null system input. Fig. B.14 shows 1000-fold averages of the mean-square error in such curves for various static system input values. (The data for a static 0.5 LSB system input resides at zero and hence does not appear in the figure.) Shown for comparison are f_{max_2} and f_{ref_2}.

Again, although we cannot predict the precise form of MSE$[\hat{m}_2]$ for this system, we conclude that its upper bound is a curve which approaches zero as $1/N$. It is now clear why the estimates of Figs. B.3, B.6 and B.13 all converge quickly, and in a similar fashion, to a value of $\Delta^2/4$, in spite of the different system input signal associated with each figure.

Such claims cannot be made about MSE$[\hat{m}_k]$ for $k > 2$, in which case, as for MSE$[\hat{m}_2]$ in an RPDF dithered system, the quantity being estimated is not constant for non-static system input signals. For fluctuating inputs which, in the long run, distribute themselves uniformly over an integral number of quantizing steps, such estimates of m_k will tend to converge to the mean value of $E[\varepsilon^k|x]$.

B.5 Conclusions

Let us try to relate our findings to the questions posed in Section B.1 and the experimental results shown there.

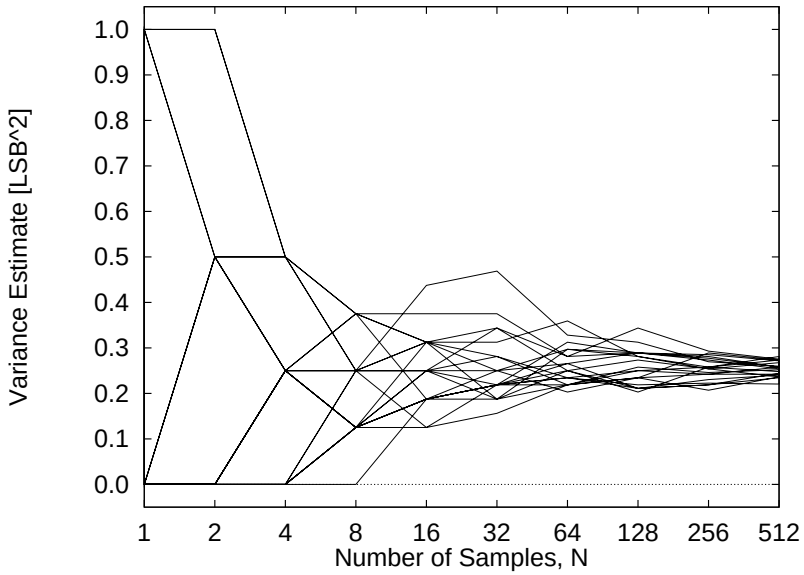


Figure B.13: Estimates of $E[\varepsilon^2]$ for a 2RPDF dithered quantizing system with a 0 LSB system input, shown as a function of the number of samples used in the estimate.

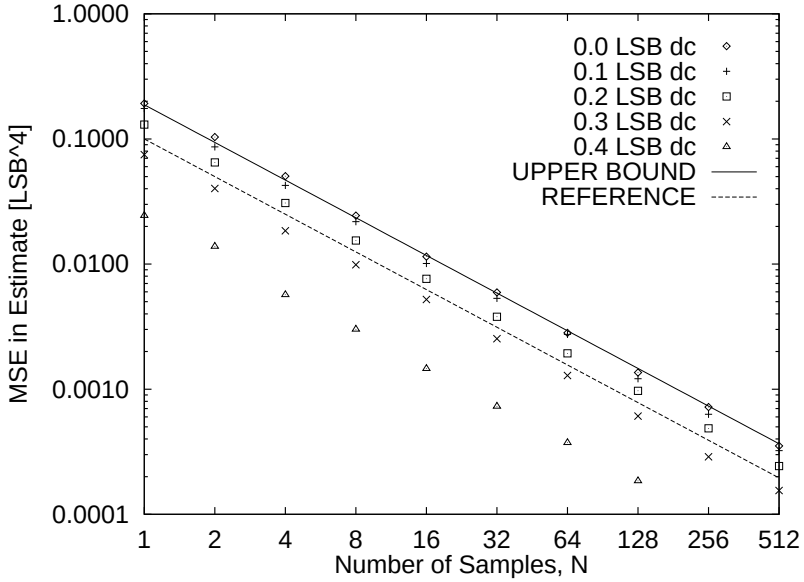


Figure B.14: $MSE[\hat{m}_2](N)$ for a 2RPDF dithered quantizing system with static system inputs, compared with the theoretical upper bound and reference convergence curves, f_{max_2} and f_{ref_2}. Data averaged over 1000 trials.

We have found that there exist definite prerequisites for successful moment estimation in dithered quantizing systems. In particular, if a meaningful estimate of m_k is desired then it is necessary that this quantity be rendered independent of the system input and thus constant with respect to time by using an appropriate dither signal. It is now obvious how to interpret Fig. B.2: the variance estimates should not be expected to converge to a unique value since in an RPDF dithered system the total error variance depends upon the value of the static system input applied. By the same token, the variance of the total error in such a system given a ramped input signal is not constant, so the curves of Fig. B.5 ultimately converge to a value representing the average variance of the total error during the time interval of estimation, namely $\Delta^2/6$. In both cases, the behaviour of the estimate is profoundly affected by the presence of noise modulation.

On the other hand, the curves in Figs. B.3 and B.6 all converge to a unique value because, with 2RPDF dither, the variance of the total error is constant at $\Delta^2/4$ for all inputs. How does the convergence compare with that for a stationary random process whose samples are statistically independent of one another? It has been shown that the MSE of variance estimates for such a noise decreases like $1/N$, while the corresponding MSE in a 2RPDF dithered system is bounded from above by a curve which decreases like $1/N$. Hence, although we cannot in general predict the functional form of MSE$[\hat{m}_2]$ for such systems, we can say that it goes to zero *at least as fast* as the MSE of some independent random noise process. Furthermore, we have found that estimates of the total error variance converge on average as rapidly as variance estimates for a piecewise-parabolically distributed noise of variance $\Delta^2/4$, and also that estimates of the mean total error converge *precisely* as rapidly as estimates of the mean of such a noise. We deduce that the input dependence of the estimation process noted in the 1RPDF dithered system

(above and in [41]) would not have been observed had 2RPDF dither been used.

Signal moments higher than the second have not been observed to have perceptual significance in most applications. Indeed, variations in these moments have proven inaudible in a wide variety of listening tests [21]. Hence, the recommended dither for audio applications is 2RPDF [23, 11, 16, 18], since this dither is unique in minimizing the second moment of the total error subject to the restriction that it render both the first and second moments constant with respect to time regardless of the system input (see Theorem 4.11). We have seen that for practical moment estimation purposes the total error in a system using proper 2RPDF dither displays convergence properties which are as good as, or better than, an independent noise signal.

All of these desirable results are contingent upon the choice of proper dither. If 1RPDF dither (or 2RPDF dither of incorrect amplitude) is used, the desired moments will not be constant and estimates thereof will generally be meaningless. We conclude that dither does its job properly, but only if its attributes are properly chosen.

Appendix C

Derivatives of the $\mathrm{sinc}\,(x)$ Function

In this appendix we prove two technical lemmas required in Section 4.4.3.

Lemma C.1 *If*

$$f(x) = \frac{\sin(x)}{x}$$

then for $n \in \mathbf{Z}, n \geq 0$,

$$f^{(n)}(x) = \sum_{i=0}^{n} \frac{n!}{(n-i)!} \frac{\sin\left(x + (n+i)\frac{\pi}{2}\right)}{x^{i+1}}.$$

Proof: We will use induction. We observe that the formula holds for $n = 0$ and suppose that it holds for $n = m$ with the object of proving that it then holds for $n = m + 1$. We also observe that

$$\frac{d}{d\theta}\sin(\theta) = \sin\left(\theta + \frac{\pi}{2}\right).$$

Then, differentiating the expression for $f^{(m)}(x)$, we have

$$\begin{aligned}
& f^{(m+1)}(x) \\
&= \sum_{i=0}^{m} \frac{m!}{(m-i)!} \frac{\sin\left(x+(m+i+1)\frac{\pi}{2}\right)}{x^{i+1}} - \sum_{i=0}^{m} \frac{m!}{(m-i)!} \frac{(i+1)\sin\left(x+(m+i)\frac{\pi}{2}\right)}{x^{i+2}} \\
&= \frac{\sin\left(x+(m+1)\frac{\pi}{2}\right)}{x} + \sum_{i=1}^{m} \frac{1}{x^{i+1}} \left\{ \frac{m!}{(m-i)!} \sin\left(x+(m+i+1)\frac{\pi}{2}\right) \right. \\
&\qquad \left. -\frac{m!}{(m-i+1)!} i \sin\left(x+(m+i-1)\frac{\pi}{2}\right) \right\} - m!(m+1)\frac{\sin\left(x+2m\frac{\pi}{2}\right)}{x^{m+2}} \\
&= \frac{\sin\left(x+(m+1)\frac{\pi}{2}\right)}{x} + \sum_{i=1}^{m} \left[\frac{m!(m-i+1)+m!i}{(m-i+1)!} \right] \frac{\sin\left(x+(m+i+1)\frac{\pi}{2}\right)}{x^{i+1}} \\
&\qquad -(m+1)!\frac{\sin\left(x+2m\frac{\pi}{2}\right)}{x^{m+2}} \\
&= \sum_{i=0}^{m+1} \frac{(m+1)!}{(m+1-i)!} \frac{\sin\left(x+(m+1+i)\frac{\pi}{2}\right)}{x^{i+1}}.
\end{aligned}$$

This proves the lemma.

□

Of course, this implies that for

$$\text{sinc}\,(x) \triangleq \frac{\sin(\pi\Delta x)}{\pi\Delta x}$$

we have

$$\text{sinc}^{(n)}\,(x) = (\pi\Delta)^n \sum_{i=0}^{n} \frac{n!}{(n-i)!} \frac{\sin\left(\pi\Delta x+(n+i)\frac{\pi}{2}\right)}{(\pi\Delta x)^{i+1}}.$$

Lemma C.2 *Suppose $m, n \in \mathbf{Z}$, $m \geq 1$ and $n \geq m$. Then*

$$\frac{d^n}{dx^n}\left[\frac{\sin(x)}{x}\right]^m$$

is non-zero for $x = k\pi$, $k \in \mathbf{Z}_0$.

Proof: Let

$$f(x) = \frac{\sin(x)}{x}.$$

From Lemma C.1 we have

$$f^{(n)}(k\pi) = \sum_{i=0}^{n} \frac{n!}{(n-i)!} \frac{\sin\left(k\pi + (n+i)\frac{\pi}{2}\right)}{(k\pi)^{i+1}}.$$

Suppose, for purposes of contradiction, that $k \in \mathbf{Z}_0$ and that the above expression vanishes. This implies that

$$\sum_{i=0}^{n} \frac{n!}{(n-i)!} \frac{\sin\left((n+i)\frac{\pi}{2}\right)}{(k\pi)^{i+1}} = 0.$$

The left-hand side of this expression is a polynomial in $\frac{1}{k\pi}$ so that $z_0 = \frac{1}{k\pi}$ must be a non-zero root of the equation

$$\sum_{i=0}^{n} \frac{n!}{(n-i)!} \sin\left((n+i)\frac{\pi}{2}\right) z^i = 0.$$

Then, since z_0 is an algebraic number, k must be a transcendental number, contradicting the assumption that k is an integer. Thus no derivatives of $f(x)$ vanish for $x = k\pi$, $k \in \mathbf{Z}_0$.

The extension to powers of $f(x)$ is straightforward. The non-vanishing terms in $\frac{d^n}{dx^n}[f(x)]^m$ will consist entirely of finite products of derivatives of $f(x)$, again resulting in a polynomial in $\frac{1}{k\pi}$ which cannot vanish for $k \in \mathbf{Z}_0$.

□

This, of course, implies that the n-th derivative of $[\mathrm{sinc}\,(x)]^m$ is non-vanishing for $x = \frac{k}{\Delta}$, $k \in \mathbf{Z}_0$, $n \geq m$.