Chapter 19

Dither

Dither inputs are externally applied disturbances that have been used in control systems and in signal processing systems to alleviate the effects of nonlinearity, hysteresis, static friction, gear backlash, quantization, etc. In many cases, dither inputs have been used to "improve" system behavior without there being a clear idea of the nature of the improvement sought and without any method for designing the dither signal other than empiricism. It is the purpose of this chapter to explain the uses of dither signals in systems containing quantizers. We will employ the mathematical methods developed herein for the design of dither signals and for analysis of their benefits and limitations.

19.1 DITHER: ANTI-ALIAS FILTERING OF THE QUANTIZER INPUT CF

When the input signal to a uniform quantizer has statistical properties allowing it to satisfy QT I or QT II, the PQN model can be applied to describe the statistical behavior of the quantizer. This is a linear model, and from the point of view of moments and joint moments, the quantizer acts like a source of additive independent noise. This type of linear behavior would be highly desirable under many circumstances.

When the quantizer input is inadequate for satisfaction of QT II, it is possible to add an independent dither signal to the quantizer input so that the sum of input and dither does satisfy QT II. Then the quantizer exhibits linear behavior and we can say that it is linearized.

Figure 19.1 shows dither d added to the input x. If the CF of the dither alone is bandlimited and satisfies QT II, then the CF of the quantizer input $x + d$ will be bandlimited and will satisfy QT II. The reason is that since x and d are independent, the CF of $x + d$ is the product of the CF of x and the CF of d, and the bandwidth of the product is no wider than the bandwidth of the narrower factor. Therefore, when the bandwidth of the CF of the dither d is narrow enough to satisfy QT II, then the bandwidth of the CF of $x + d$ is also narrow enough to satisfy QT II. The idea is illustrated in Fig. 19.2.

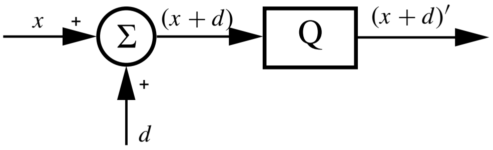


Figure 19.1 Dither added to the quantizer input.

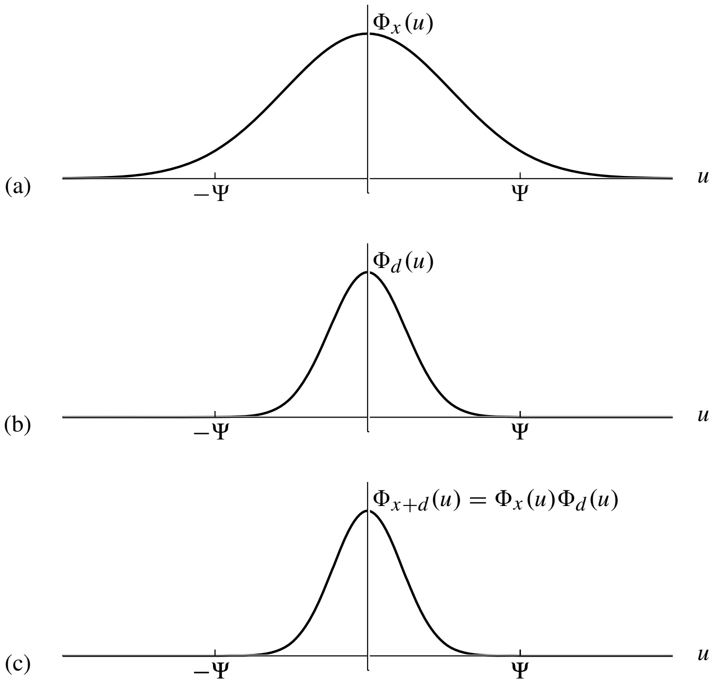


Figure 19.2 Bandlimitation of quantizer input CF resulting from dithering with a bandlimited independent dither signal: (a) CF of x; (b) CF of d; (c) CF of $(x+d)$. $\Psi = 2\pi/q$.

Under these conditions, the PQN model applies and the dither signal has "linearized" the quantizer. On the other hand, if the CF of the input x had a bandwidth sufficiently narrow to satisfy QT II, then the use of the dither would be unnecessary.

Figure 19.3(a) represents quantization of the dithered signal $(x + d)$ as the addition of quantization noise ν. Figure 19.3(b) represents the PQN model that is applicable when $(x + d)$ has a characteristic function sufficiently bandlimited to satisfy QT II.

The addition of a dither signal to the input x serves as anti-alias filtering for its characteristic function. The CF of $x + d$ is bandlimited by the action of the dither.

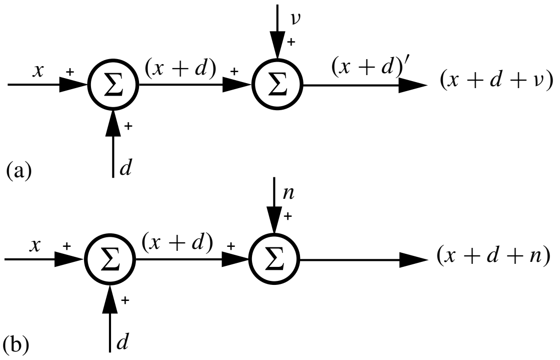


Figure 19.3 Quantization of a dithered signal represented as addition of quantization noise: (a) addition of quantization noise ν; (b) PQN model that is applicable when QT II is satisfied.

This is illustrated in Fig. 19.4. The CF of the dither acts like a lowpass anti-alias filter, and is analogous to a lowpass anti-alias filter commonly used before signal sampling.

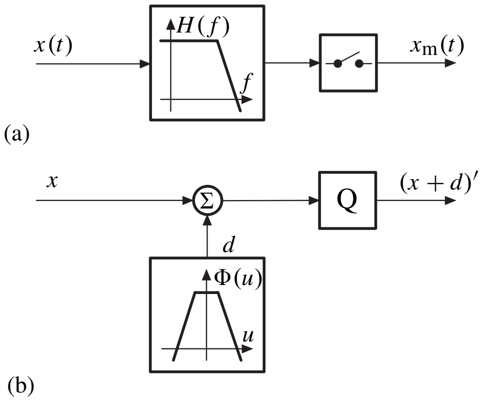


Figure 19.4 Anti-alias filtering in signal sampling, and dithering in quantization: (a) anti-alias; (b) dithering.

All this will be true for dither whose CF satisfies QT I or QT II. These ideas will be extended subsequently for dither signals that do not satisfy QT I or QT II, but do satisfy some of the other quantizing theorems.

19.2 MOMENT RELATIONS WHEN QT II IS SATISFIED

Suppose that the dither d is independent of x, and that the combination $(x + d)$ satisfies QT II. Now refer to Fig. 19.3(a). It is clear that

$$\mathrm{E}\{\nu(x + d)\} = 0\,. \tag{19.1}$$

Therefore,

$$\mathrm{E}\{\nu x\} = -\mathrm{E}\{\nu d\}\,. \tag{19.2}$$

Since x and d are independent, one might expect that

$$\mathrm{E}\{\nu x\} = -\mathrm{E}\{\nu d\} \stackrel{?}{=} 0\,. \tag{19.3}$$

Indeed, if the dither fulfills at least QT IV/A, Eq. (19.3) holds, since QTSD assures both the uncorrelatedness of x and ν (see page 508), and $\mathrm{E}\{\nu\} = 0$. The quantization noise exhibits behavior in a moment sense like independent additive noise, i.e., like n of the PQN model.

To determine the moments and joint moments between $x, d, (x + d), \nu$, and $(x + d)'$, we can make use of the ideas developed in Chapter 7 which are embodied in Eq. (7.82).

In accord with Eq. (7.82), we have

$$\begin{aligned}&\Phi_{(x+d),\nu,(x+d)'}\left(u_{(x+d)}, u_\nu, u_{(x+d)'}\right)\\&\quad= \sum_{l=-\infty}^{\infty} \Phi_{(x+d),n,(x+d+n)}\left(u_{(x+d)}, u_\nu, u_{(x+d)'} + l\Psi\right)\,.\end{aligned} \tag{19.4}$$

Pertaining to the block diagram of quantization in Fig. 19.3(a), the characteristic function

$$\Phi_{(x+d),\nu,(x+d)'}\left(u_{(x+d)}, u_\nu, u_{(x+d)'}\right) \tag{19.5}$$

is the joint characteristic function between $(x+d)$, ν, and $(x+d)'$. Pertaining to the PQN model of Fig. 19.3(b), the characteristic function

$$\Phi_{(x+d),n,(x+d+n)}\left(u_{(x+d)}, u_n, u_{(x+d+n)}\right) \tag{19.6}$$

is the joint characteristic function between $(x + d), n$, and $(x + d + n)$.

Eq. (19.4) relates the joint CF for quantization to the joint CF of the PQN model. The quantization CF is the same as the PQN CF repeated and summed along the $u_{(x+d)'}$-axis. The quantization CF is periodic along the $u_{(x+d)'}$-axis, while the PQN CF is aperiodic. If QT I is satisfied, the periodic sections of the quantization CF do not overlap and all moments and joint moments for quantization are the same as the corresponding ones for PQN. If QT I is not satisfied but QT II is satisfied, then the periodic sections overlap but the moment relations still correspond.

Equation (19.4) can be generalized to encompass all of the variables of dithered quantization represented in Fig. 19.3(a). Accordingly,

$$\begin{aligned}
&\Phi_{x,d,(x+d),\nu,(x+d)'}\left(u_x, u_d, u_{(x+d)}, u_\nu, u_{(x+d)'}\right)\\
&= \sum_{l=-\infty}^{\infty} \Phi_{x,d,(x+d),n,(x+d+n)}\left(u_x, u_d, u_{(x+d)}, u_\nu, u_{(x+d)'} + l\Psi\right)\\
&= \sum_{l=-\infty}^{\infty} \Phi_x\left(u_x + u_{(x+d)} + u_{(x+d)'} + l\Psi\right)\Phi_d\left(u_d + u_{(x+d)} + u_{(x+d)'} + l\Psi\right)\\
&\qquad\qquad \times\Phi_n\left(u_\nu + u_{(x+d)'} + l\Psi\right) . \qquad (19.7)
\end{aligned}$$

If QT I or QT II is satisfied by the quantizer input $(x + d)$, inspection of Eq. (19.7) indicates that when calculating the moments, only the $l = 0$ member of the sum can yield nonzero derivatives, so that all joint moments of x, d, $(x + d)$, ν, and $(x + d)'$ are the same as the corresponding joint moments of x, d, $(x + d)$, n, and $(x + d + n)$. Regarding all moments, therefore, the quantization noise ν behaves like the independent noise n of the PQN model. Furthermore, we now realize that Eq. (19.3) is true.

19.3 CONDITIONS FOR STATISTICAL INDEPENDENCE OF x AND ν, AND d AND ν

The joint probability density of x and ν can be regarded as a marginal density of the joint density of x, d, $(x + d)$, ν, and $(x + d)'$. The corresponding characteristic functions are

$$\Phi_{x,\nu}(u_x, u_\nu) = \Phi_{x,d,(x+d),\nu,(x+d)'}(u_x, 0, 0, u_\nu, 0) . \qquad (19.8)$$

Now making use of Eq. (19.7), we obtain

$$\begin{aligned}
\Phi_{x,\nu}(u_x, u_\nu) &= \sum_{l=-\infty}^{\infty} \Phi_{x,d,(x+d),n,(x+d+n)}(u_x, 0, 0, u_\nu, l\Psi)\\
&= \sum_{l=-\infty}^{\infty} \Phi_x\left(u_x + l\Psi\right)\Phi_d\left(l\Psi\right)\Phi_n\left(u_\nu + l\Psi\right) . \qquad (19.9)
\end{aligned}$$

When QT I is satisfied, $\Phi_{x,\nu}(u_x, u_\nu)$ can be simply expressed as the 0th element of the sum, i.e.,

$$\Phi_{x,\nu}(u_x, u_\nu) = \Phi_{x,d,(x+d),n,(x+d+n)}(u_x, 0, 0, u_\nu, 0)$$

$$\begin{aligned} &= \Phi_{x,n}(u_x, u_\nu) \\ &= \Phi_x\,(u_x)\,\Phi_n\,(u_\nu)\;. \end{aligned} \tag{19.10}$$

Because x and n are statistically independent, it is clear that x and ν are also statistically independent and that ν is uniformly distributed between $\pm q/2$.

When QT I is not satisfied but QT II is satisfied, the elements of the sum in Eq. (19.9) do overlap, but all of the derivatives of

$$\Phi_{x,\nu}(u_x, u_\nu) = \sum_{l=-\infty}^{\infty} \Phi_{x,d,(x+d),n,(x+d+n)}(u_x, 0, 0, u_\nu, l\Psi) \tag{19.11}$$

with respect to u_x and u_ν in the vicinity of $u_x = 0$ and $u_\nu = 0$ will be the same as the corresponding derivatives of $\Phi_{x,n}(u_x, u_\nu) = \Phi_{x,d,(x+d),n,(x+d+n)}(u_x, 0, 0, u_\nu, 0)$. Assuming that the function $\Phi_{x,d,(x+d),n,(x+d+n)}$ is differentiable at its origin,

$$\begin{aligned} &\frac{\partial^{r+t}}{\partial u_x^r \partial u_\nu^t} \sum_{l=-\infty}^{\infty} \Phi_{x,d,(x+d),n,(x+d+n)}\,(u_x, 0, 0, u_\nu, l\Psi)\Big|_{\substack{u_x=0\\u_\nu=0}} \\ &= \frac{\partial^{r+t}}{\partial u_x^r \partial u_\nu^t} \Phi_{x,d,(x+d),n,(x+d+n)}\,(u_x, 0, 0, u_\nu, 0)\Big|_{\substack{u_x=0\\u_\nu=0}}\;. \end{aligned} \tag{19.12}$$

If all the moments exist, combining this with Eq. (19.9) yields

$$\Phi_{x,\nu}(u_x, u_\nu) = \Phi_{x,n}(u_x, u_\nu)\,. \tag{19.13}$$

Once again, since x and n are statistically independent, x and ν are also statistically independent and ν is uniformly distributed between $\pm q/2$.

Analogous arguments can be made regarding the statistical independence of ν and d. It can be seen that

$$\Phi_{d,\nu}(u_d, u_\nu) = \Phi_{d,n}(u_d, u_\nu) \tag{19.14}$$

when QT I is satisfied, so that ν is statistically independent of d under this condition. The quantization noise ν is therefore a statistically independent additive noise, uniformly distributed between $\pm q/2$.

When a dither signal is applied to the input of a quantizer, as in Fig. 19.1, the total noise at the quantizer output will be the difference between the quantizer output $(x+d)'$ and the quantizer input x. Referring to Fig. 19.3(a), the quantizer output $(x+d)'$ is seen to be $(x+d+\nu)$. Accordingly, the total quantizer output noise is

$$\begin{aligned} (x+d)' - x &= (x+d+\nu) - x \\ &= d+\nu\,. \end{aligned} \tag{19.15}$$

This noise can be represented as

$$\xi = d + \nu \, . \tag{19.16}$$

When the quantizer input $(x + d)$ meets the conditions for satisfaction of either QT I or QT II, both d and ν are statistically independent of x. Their sum $(d + \nu)$, the total quantizer output noise, is also statistically independent of the input x.

We need to note here that this latter independence does not directly follow from the pairwise independences, despite the fact that it seems to be "logical". Only an additional proof can show it (see Section 19.5).[1]

It is possible for d and ν to both be statistically independent of x, and for their sum $d + \nu$ to be dependent on x. To illustrate that the paradoxical case of mutually independent, but in general sense interdependent random variables may easily occur, consider the following example.

Example 19.1 Mutually Independent, but Jointly Interdependent Random Variables

A three-dimensional distribution is illustrated in Fig. 19.5. The random variables take the following joint values with probability 1/4 at the points: $(x = 1, y = 1, z = 1)$, $(x = -1, y = -1, z = 1)$, $(x = 1, y = -1, z = -1)$, and $(x = -1, y = 1, z = -1)$. It is clear that the random variables are not independent, while all the two-dimensional densities, like the one in Fig. 19.5(b), consist of the direct product of the distributions of two independent random variables. In other words, x and y are independent, y and z are independent, and z and x are independent, but x, y, and z are not independent, and $y + z$ and x are also not independent.

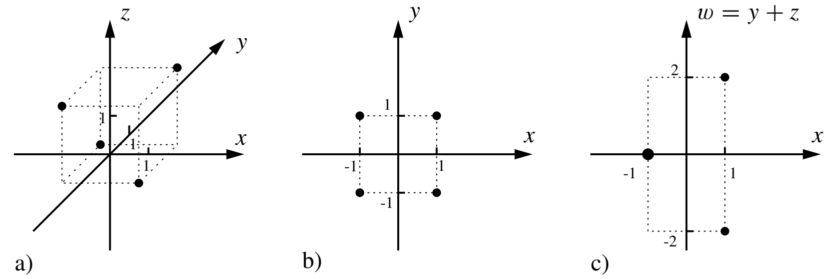


Figure 19.5 Three-dimensional example for pairwise independent random variables which are otherwise not independent: (a) three-dimensional distribution; (b) marginal density $f_{x,y}(x, y)$; (c) joint density $f_{x,y+z}(x, w)$.

[1] It should be noted that under all circumstances, the quantization noise ν is deterministically related to the quantizer input $(x + d)$, although in statistical sense, it may be independent of x, and/or of d, and/or of their sum $(x + d)$.

For the full description of the interrelations of x, d, and ν, we need to investigate the three-dimensional PDF or CF. The latter is given in Addendum S:[2]

$$\begin{aligned}
&\Phi_{x,d,\nu}\left(u_x, u_d, u_\nu\right) \\
&\quad = \sum_{l=-\infty}^{\infty} \Phi_x(u_x + l\Psi)\, \Phi_{\mathrm{d}}(u_d + l\Psi)\, \Phi_n(u_\nu + l\Psi) \\
&\quad = \sum_{l=-\infty}^{\infty} \Phi_x(u_x + l\Psi)\, \Phi_{\mathrm{d}}(u_d + l\Psi)\, \mathrm{sinc}\left(\frac{q(u_\nu + l\Psi)}{2}\right) .
\end{aligned} \tag{19.17}$$

If the conditions of QT I or QT II are satisfied, the joint CF of x, ν, and d becomes the product of the three CFs around zero, respectively, therefore the random variables behave exactly like the ones in PQN.

The derivation is similar for the CF of the triplet (x, d, ξ). The joint CF is (see Addendum S in the website of the book):

$$\begin{aligned}
&\Phi_{x,d,\xi}\left(u_x, u_d, u_\xi\right) \\
&\quad = \sum_{l=-\infty}^{\infty} \Phi_x(u_x + l\Psi)\, \Phi_{\mathrm{d}}(u_d + u_\xi + l\Psi)\, \mathrm{sinc}\left(\frac{q(u_\xi + l\Psi)}{2}\right) .
\end{aligned} \tag{19.18}$$

If the conditions of QT I or QT II are fulfilled, the joint moments are again the same as those of PQN.

19.4 MOMENT RELATIONS AND QUANTIZATION NOISE PDF WHEN QT III OR QT IV IS SATISFIED

The conditions for satisfaction of quantizing theorems QT III or QT IV are summarized in Sections 7.5 and 7.6. The consequences of their satisfaction are also summarized in the same section. These conditions apply to the characteristic function of the quantizer input. They all involve zero values of the characteristic function, and/or zero values of its derivatives. Since the dither d is constructed to be independent of the input x, satisfaction of any or all of the quantizing theorems by the dither alone guarantees the same satisfaction by $(x + d)$: in other words, $(x + d)$, $(x + d)'$ and $\nu = (x + d)' - (x + d)$ behave like described in the previous chapters.[3] The reason is that the CF of $(x + d)$ is the product of the CF of x and the CF of d. Therefore, regardless of the CF of x, the CF of $(x + d)$ will have at least the same zero values

[2] see the web page `http://www.mit.bme.hu/books/quantization/`

[3] The total quantization error, $\xi = d+\nu$, has more complex behavior, as we will see it in Section 19.5.

and zero derivative values as the CF of d alone, at least when QT I, QT II, or QT III are fulfilled. This is not true for QT IV/B, since zero derivative of $\Phi_d(u)$ does not by itself imply zero derivative of the product $\Phi_d(u)\Phi_x(u)$. Therefore, when the dither d fulfills QT IV/B, the sum $(x + d)$ does not necessarily fulfill QT IV/B.

Precise mathematical analysis of the total quantization error $\xi = \nu + d$, along with exact conditions of its independence of d and x, is given in Section 19.5.

19.5 STATISTICAL ANALYSIS OF THE TOTAL QUANTIZATION ERROR $\xi = d + \nu$

In the analysis of $\xi = d + \nu$ ("non-subtractive dither"), thoroughly investigated by Wannamaker (1994), Wannamaker, R. A., Lipshitz, S. P., Vanderkooy, and Wright (2000), and Gray and Stockham (1993), the basic question is not whether ν is independent of $x + d$ or not. The total quantization output noise $\xi = d + \nu$ is the quantity which corrupts the quantized output. We need to know the properties of this variable, and its relationships to x and d.

One might think that since the properties of ν are known from quantization theory discussed until now, $\xi = d + \nu$ is thoroughly known. Unfortunately, this is not the case. Even if both d and ν are independent of x, *their sum is not necessarily independent of it.* This paradox can be best illustrated by an example.

Example 19.2 Dependent Sum of Variables which are Independent of a Third One

Heuristically, one might think that if two variables are each independent from a third one, they have nothing to do with it, so their sum is also independent from it. This example shows that this is not true, especially for quantization with dither.

Let us consider the input variable x as in Fig. 19.6(a), dithered by the dither d, uniform in $[-q/2, q/2]$. It is clear from the figure that x and d are independent. The quantized output $(x+d)'$ is illustrated in Fig. 19.6(b). The quantization error is illustrated in Fig. 19.6(c). As it will be proved on page 508, ν is independent of the input, since the dither is zero-order. However, the total quantization error, $d + \nu$, does depend on x, as Fig. 19.6(d) illustrates. The explanation is given in Fig. 19.6(e): the dither d and the quantization error ν strongly depend on each other, although their correlation coefficient is zero, and their interdependence is different for different values of x, so their sum does depend on it.

The interrelation of ν and d is difficult to analyze, but, fortunately, we do not need detailed knowledge of this. Instead, we consider $\xi = d + \nu$ as a whole. The best idea is to consider the *conditional* CF of ξ:[4]

[4] see the web page: `http://www.mit.bme.hu/books/quantization/`

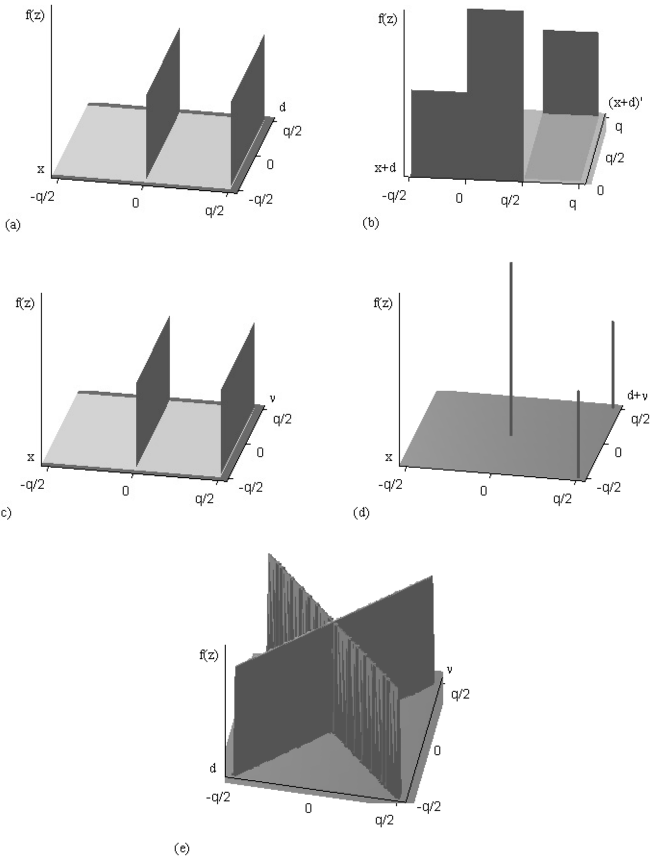


Figure 19.6 The total quantization error, $\xi = d + \nu$, is not obligatorily independent from the input x when the components d and ν of the sum are both independent of the input: (a) joint distribution of the independent x and d; (b) mapping of $x + d$ to $(x + d)'$; (c) joint distribution of x and ν which illustrates that they are independent; (d) joint distribution of x and $d + \nu$ which illustrates that they are not independent; (e) joint distribution of d and ν, illustrating how ν depends on d.

$$\begin{aligned}\Phi_{\xi|x}(u_\xi) &= \sum_{l=-\infty}^{\infty} \Phi_{\mathrm{d}}\left(u_\xi + l\Psi\right) e^{jl\Psi x}\,\mathrm{sinc}\left(\frac{q(u_\xi + l\Psi)}{2}\right)\\ &= \Phi_{\mathrm{d}}\left(u_\xi\right)\mathrm{sinc}\left(\frac{q u_\xi}{2}\right)\\ &\quad + \sum_{\substack{l=-\infty\\ l\neq 0}}^{\infty} \Phi_{\mathrm{d}}\left(u_\xi + l\Psi\right) e^{jl\Psi x}\,\mathrm{sinc}\left(\frac{q(u_\xi + l\Psi)}{2}\right). \qquad (19.19)\end{aligned}$$

Equation (19.19) yields simple and straightforward implications. First, it is clear that the moments of ξ depend on x unless the effect of the last sum disappears at zero from all derivatives. Apart from very special distributions of x (when all values of x are equal to $x_0 + kq$, with x_0 being a deterministic constant), independence is only possible if the dither satisfies QT II.

In Fig. 19.7, probability density functions are given for Example 19.2. The variance of ξ depends on the value of x, despite of uniform dithering.

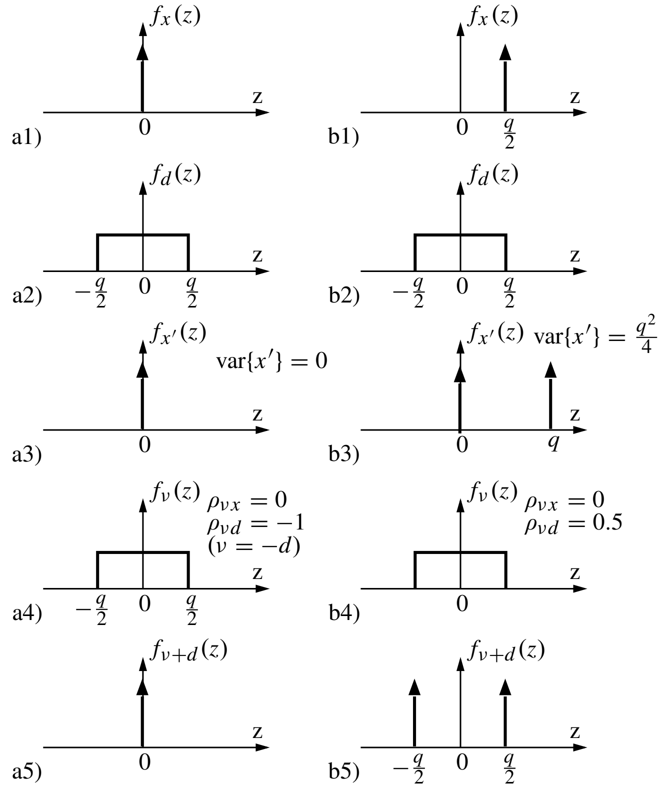


Figure 19.7 Quantization of an "almost constant" signal with uniform dither: (a1–a5) with E$\{x\} = 0$; (b1–b5) with E$\{x\} = q/2$.

The purpose of dithering can be to eliminate or decrease the dependence of the moments of ξ on x. To investigate this, consider that the moments of a variable can be calculated by taking the value of the appropriate derivative of the CF at zero. The last sum in Eq. (19.19) disappears if the CF of the dither is appropriate. Taking into account that $\text{sinc}(l\pi) = 0$ for $l = \pm 1, \pm 2, \ldots$, this immediately leads to the following theorems.

Quantizing Theorem III for Dither (QTD III)

The necessary and sufficient condition of the moments $\mathrm{E}\{\xi^m\}$*, where* $m = 1, 2, \ldots, r+1$*, and r is a nonnegative integer, to be functionally independent*[5] *of the input variable x which does not have a discrete distribution on a grid*[6] $(x_0 + kq)$*, where* $k = 0, \pm 1, \pm 2, \ldots$*, is that*

$$\left.\frac{\mathrm{d}^{m-1}\,\Phi_d(u)}{\mathrm{d}u^{m-1}}\right|_{u=l\Psi} = 0\,, \quad \text{for } l = \pm 1, \pm 2, \ldots \tag{19.20}$$

*for all above values of m (r*th*-order dither). The values of these moments then equal* $\mathrm{E}\{(d+n)^m\}$*.*

An equvalent formulation of condition (19.20) is

$$\left.\frac{\mathrm{d}^m\left(\Phi_d(u)\,\mathrm{sinc}\left(\frac{qu}{2}\right)\right)}{\mathrm{d}u^m}\right|_{u=l\Psi} = 0\,, \quad \text{for } l = \pm 1, \pm 2, \ldots . \tag{19.20b}$$

for values $m = 1, 2, \ldots, r+1$.

Functional independence has an important implication: if this holds, then

$$\mathrm{E}\{\xi^m x^t\} = \mathrm{E}\{\xi^m\}\mathrm{E}\{x^t\}\,, \ \text{for } t = 1, 2, \ldots \tag{19.21}$$

For $r = 0$ ($m = 1$; zero-order dither, see Eq. (19.52) on page 508) and $t = 1$, this means uncorrelatedness of ξ and x. However, in this case the second moment of ξ may still depend on x (noise modulation). To avoid also this, first-order dither needs to be used. This will also assure functional independence for the AC power, that is, for the variance (we hear the audible AC power in audio, which is equal to the part of the variance (power), which is within the audible band.

If only one moment of ξ is considered, the theorem can be further refined:

Quantizing Theorem IV for Dither (QTD IV)

The necessary and sufficient condition of the moment $\mathrm{E}\{\xi^m\}$*, where m is a positive integer, to be functionally independent of the input variable x which does not have a discrete distribution on a grid* $(x_0 + kq)$*, where* $k = 0, \pm 1, \pm 2, \ldots$*, is that*

[5] Functional independence means that for any fixed value of x, the moments of the quantization error are the same. This is an important property. Without this, it is possible e.g. that the variance of the quantization error is different for different almost-constant values of x (see Fig. 19.7), causing a *modulated noise*, audible as background noise changing its power with time.

[6] When all values of x are on a grid $(x_0 + kq)$, the expression (19.19) does not depend on x at all. This is also a sufficient condition to ensure functional independence.

$$\frac{\mathrm{d}^m \left(\Phi_d(u)\,\mathrm{sinc}\left(\frac{qu}{2}\right)\right)}{\mathrm{d}u^m}\Bigg|_{u=l\Psi} = 0\,, \quad \text{for } l = \pm 1, \pm 2, \ldots \tag{19.22}$$

The value of this moment then equals $\mathrm{E}\{(d+n)^m\}$.

Again, if this holds, then

$$\mathrm{E}\{\xi^m x^t\} = \mathrm{E}\{\xi^m\}\mathrm{E}\{x^t\}\,, \text{ for } t = 1, 2, \ldots \tag{19.23}$$

For $m = 1$ (zero-order dither, see Eq. (19.52) on page 508) and $t = 1$, this means uncorrelatedness of ξ and x. The second moment of ξ, its power, may still depend on x (noise modulation). To avoid this, condition (19.22) with $m = 2$ needs to be fulfilled. For functional independence of the variance, the best is to assure the condition (19.22) both with $m = 1$ and $m = 2$ (QTD III/B).

The condition of QTD IV is somewhat difficult to check, since the derivatives in (19.22) are not easy to evaluate if condition (19.20) is not fulfilled, but QTD III is a very useful theorem. The immediate consequences are:

- because of functional dependence of the second-order moment on x, a dither uniform between $(\pm q/2)$ is not sufficient to decouple the variance of ξ from x, since the derivatives of its CF are not zero at the required places, but the uncorrelatedness of ξ and x is provided,
- a triangular dither between $(\pm q)$ is appropriate to decouple the variance of ξ from x. for this purpose: its application guarantees e.g. that $\mathrm{var}\{(x+d)'\} = \mathrm{var}\{x\} + \mathrm{var}\{d\} + q^2/12$. Using triangular dither, noise modulation does not occur.

For dithers which are not at least first-order (see page 496), the variance of the total error varies with x. Its lower bound is zero, as it is for almost zero input and small dither; its upper bound is $\mathrm{var}\{\xi\} \le \max(\mathrm{var}\{d+\nu\})\}) = (\sigma_d + \sigma_\nu)^2$, and it is reached for the correlation coefficient equal to 1.

19.6 IMPORTANT DITHER TYPES

In practice, QT I and QT II are not satisfied perfectly, but only approximately so. The other quantizing theorems may, with specially designed dithers, be perfectly satisfied. Of interest are special dithers that have probability density functions which are uniformly distributed between $\pm q/2$, and those that have probability density functions which are triangularly distributed between $\pm q$.

19.6.1 Uniform Dither

If an independent dither has a probability density function that is uniform between $\pm q/2$, its characteristic function is a sinc function that has zero values at $u = l\Psi$,

$l = \pm 1, \pm 2, \ldots$ This dither meets the conditions for perfect satisfaction of QT IV/A. This ensures that the PDF of the quantization noise ν will be uniformly distributed between $\pm q/2$, so that $\mathrm{E}\{\nu\} = 0$, and $\mathrm{E}\{\nu^2\} = q^2/12$. However, in general, the quantization noise ν will be correlated with d. In most applications, this form of correlation would be undesirable. Moreover, noise modulation can be present, that is, slowly changing x can cause changing power of the total quantization error (see page 496), audible and annoying in many applications. In such cases, triangularly distributed dither is preferable.

Example 19.3 Stochastic–Ergodic Conversion

Uniform dither has a special application in the so-called stochastic–ergodic converter (Wehrmann, 1973; Tumfart, 1976). This principle was popular in the nineteen-sixties, as a building block of simple and cheap analog to digital converters, but because of its relatively large variance, it disappeared when sigma–delta converters were introduced.

If zero-mean uniform dither in the form of a periodic triangular wave with

$$|d| \le a/2 \tag{19.24}$$

is added to a signal, having the property

$$|x| \le a/2\,, \tag{19.25}$$

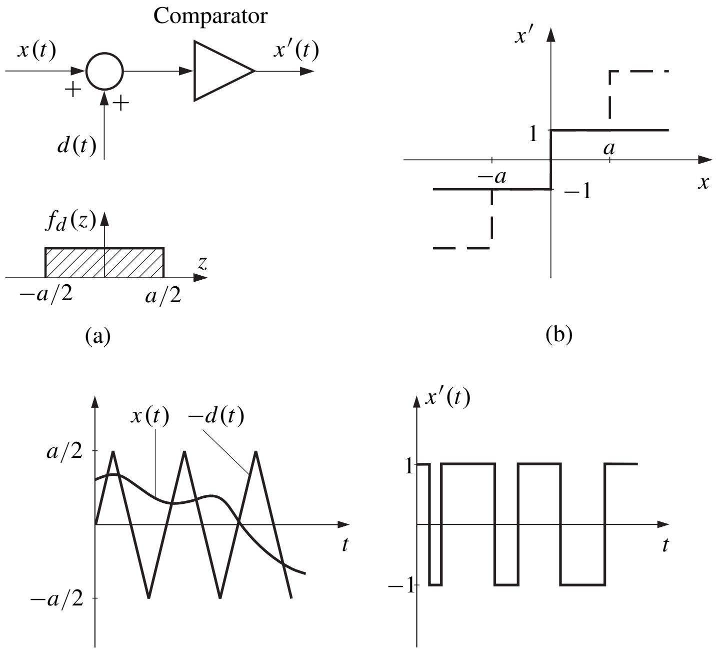
an interesting observation can be made (Fig. 19.8): the sum of signal and dither *never surpasses* the two borders at $\pm a$. This means that from the comparison levels, only the one at zero is used. Consequently, from the point of view of signal treatment, it is irrelevant how the characteristic of the quantizer behaves outside the interval $(-a, a)$: it may even be a limiter (a simple comparator), while from a theoretical point of view, we may consider it as a central portion of a uniform quantizer with $q = a$. In other words: if a uniformly distributed dither is added to the input signal x, and (19.25) is fulfilled, the quantizer may be substituted by a comparator. This is the basic idea of the stochastic–ergodic converter.

The block diagram is shown in Fig. 19.8(a). The output is a binary signal, usable for digital processing. Since the dither fulfils QT III/A, the mean values of x and x' are proportional, so the information on the input mean value is "coded" in the average value of the output signal:

$$\lim_{T\to\infty} \frac{1}{T} \int\limits_{-T/2}^{T/2} x'(t)\,\mathrm{d}t = \frac{\mu_x}{a/2}\,. \tag{19.26}$$

This equation allows simple analysis of the basic properties of the binary signal. Its values are ± 1, and the probabilities are linear functions of the mean value:

$$p = \mathrm{P}(x' = 1) = 0.5(1 + \mathrm{E}\{x'\}) = 0.5 + \frac{\mu_x}{a}$$



(a) (b)

(c)

Figure 19.8 The stochastic–ergodic converter, a comparator C which realizes a fraction of a uniform quantizer, with uniform dither at the input: (a) block diagram, (b) transfer characteristic, (c) signal waveforms.

$$1 - p = \mathrm{P}(x' = -1) = 0.5(1 - \mathrm{E}\{x'\}) = 0.5 - \frac{\mu_x}{a} . \tag{19.27}$$

This determines the variance, since for a binary distribution,

$$\mathrm{var}\{x'\} = 4p(1-p) = 1 - \left(\frac{\mu_x}{a/2}\right)^2 , \quad \mathrm{var}\{ax'\} = a^2 - 4\mu_x^2 . \tag{19.28}$$

Sigma–delta converters have a much smaller variance than this (see Addendum at the Website of this book), therefore SEM-based processing has disappeared from applications today.

As illustrated by (19.28), second and higher moments are generally biased (even after the application of Sheppard's corrections), since the dither fulfills only QT III/A. Therefore, the mean square value is not equal to the sum of second moments of x, d, and n:

$$\mathrm{E}\{x'^2\} = 1 , \quad \mathrm{E}\{(ax')^2\} = a^2 \neq \mu_x^2 + \mathrm{E}\{d^2\} + \frac{a^2}{12} = \mu_x^2 + \frac{a^2}{6} . \tag{19.29}$$

19.6.2 Triangular Dither

If two uniformly distributed dither signals are independent of x and of each other, adding them together creates a new dither signal whose PDF is the convolution of two rectangles. This is a triangular PDF distributed between $\pm q$. The new dither signal has a CF that is a sinc^2 function. It has zero values at $u = l\Psi$, $l = \pm 1, \pm 2, ...$, and has zero derivative values at $u = l\Psi$, $l = \pm 1, \pm 2, ...$. This kind of dither is of very great practical interest. First of all, it meets the conditions for QT IV/A. The quantization noise ν that results when such a dither is used is uniformly distributed between $\pm q/2$, so that $\mathrm{E}\{\nu\} = 0, \mathrm{E}\{\nu^2\} = q^2/12,$. Furthermore, ν is both orthogonal to and uncorrelated with $(x + d)$. The quantization noise ν is orthogonal to and uncorrelated with d, regardless of the nature of x, and furthermore, ν is orthogonal to and uncorrelated with x itself. From the point of view of second-order moments, ν behaves therefore like n of the PQN model. In most applications, this is highly desirable.

The triangular PDF distributed between $\pm q$ also meets the conditions for QT III with $r = 1$ (QT III/B). Therefore,

$$\mathrm{E}\{(x+d)\nu^t\} = \mathrm{E}\{(x+d)n^t\}\,, \quad t = 1, 2, ... \tag{19.30}$$

for all x. Furthermore,

$$\mathrm{E}\{x\nu^t\} = \mathrm{E}\{xn^t\}\,, \quad t = 1, 2, ... \tag{19.31}$$

(see QTSD, Eq. (19.52)), and this is equal to zero if t is odd.

Combining Eq. (19.30) with (19.31), we obtain

$$\mathrm{E}\{d\nu^t\} = \mathrm{E}\{dn^t\}\,, \quad t = 1, 2, ... \tag{19.32}$$

which is equal to zero since the mean value of the dither is zero.

Triangular dither is very important, since it is first-order (its CF and the first derivative of its CF are equal to zero for $u = l\Psi, l = \pm 1, \pm 2, \ldots$, see page 496), has theoretically the smallest variance among all first-order dithers, and it is simple to generate. Since it is first-order, it eliminates noise modulation (see page 496), therefore it is very useful in audio. One of its first applications took place during the recording of the album Tusk (Fleetwood Mac, 1980), but this was not published until 1993 (Gray and Stockham, 1993) because of commercial reasons.

Example 19.4 The Sum of Two Independent, Uniformly Distributed Dithers

One can generate a triangularly distributed dither sequence by adding two independent, uniformly distributed noise sequences. This generates white dither (its spectrum is uniform).

Example 19.5 A Simple High-Pass, Triangularly Distributed Dither

Triangular-PDF dither can be generated by the addition or by the subtraction of two independent uniform samples. Let us assume that the samples of a uniform number generator are independent, and let us generate the triangular samples from a single source of uniform samples by means of the difference equation

$$d_\mathrm{t}(k) = d_\mathrm{u}(k) - d_\mathrm{u}(k-1)\,. \tag{19.33}$$

It is straightforward that this dither is triangular, but it is not white: $R(1) = R(-1) = -\mathrm{var}\{d_\mathrm{u}\} = -q^2/12$. Its spectrum is

$$S(f) = \frac{q^2}{6f_\mathrm{c}}(1-\cos(2\pi f T_\mathrm{c}))\,, \tag{19.34}$$

where $T_\mathrm{c} = 1/f_\mathrm{c}$ is the clock interval (Fig. 19.9).

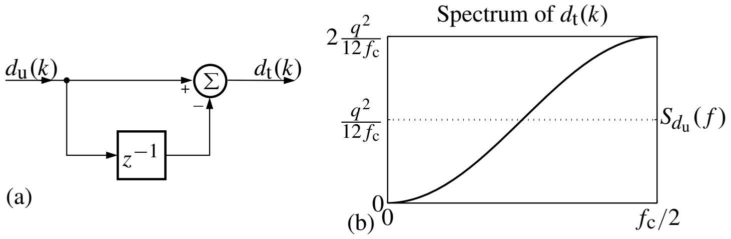


Figure 19.9 High-pass dither: (a) generation; (b) power spectral density.

By this arrangement, most of the power of the dither is removed from low frequencies, and concentrated at high ones. The dither is not white, its high-pass nature can be very useful in many practical situations.

19.6.3 Triangular plus Uniform Dither

The convolution of the triangular PDF with the rectangular PDF gives a dither PDF that is distributed between $\pm 3q/2$. The CF is a sinc^3 function, and it meets the conditions for QT III with $r = 2$. Therefore,

$$\mathrm{E}\{(x+d)^2\nu^t\} = \mathrm{E}\{(x+d)^2 n^t\}\,, \quad t = 1, 2, \ldots \tag{19.35}$$

for all x, and when $x = 0$,

$$\mathrm{E}\{d^2\nu^t\} = \mathrm{E}\{d^2 n^t\}\,, \quad t = 1, 2, \ldots \tag{19.36}$$

Assume that Eq. (19.36) is true for all x. Subtracting both sides of (19.36) from (19.35) one obtains

$$\mathrm{E}\{x^2\nu^t\} + 2\mathrm{E}\{xd\nu^t\} = \mathrm{E}\{x^2 n^t\} + 2\mathrm{E}\{xdn^t\}\,, \quad t = 1, 2, \ldots\,. \tag{19.37}$$

Since x, d, and n are independent of each other, and $\mathrm{E}\{d\} = 0$,

$$\mathrm{E}\{xdn^t\} = 0\,. \tag{19.38}$$

Since x and d are independent of each other but not independent of ν, it is not clear whether

$$\mathrm{E}\{xd\nu^t\} \overset{?}{=} 0\,, \quad t = 1, 2, \ldots \tag{19.39}$$

In light of Eqs. (19.32) and (19.31), it is probable that (19.39) is true or at least approximately true. Combining (19.39) with (19.37), one obtains

$$\mathrm{E}\{x^2\nu^t\} \overset{?}{=} \mathrm{E}\{x^2n^t\}\,, \quad t = 1, 2, \ldots \tag{19.40}$$

This is true, because of QTSD (Eq. (19.52) is fulfilled).

From the point of view of these higher-order moments, the quantization noise ν behaves like PQN.

19.6.4 Triangular plus Triangular Dither

The convolution of the triangular PDF with the triangular PDF gives a dither PDF that is distributed between $\pm 2q$. The CF is a sinc^4 function which meets the conditions for QT III with $r = 3$. Following similar reasoning,

$$\mathrm{E}\{(x+d)^3\nu^t\} = \mathrm{E}\{(x+d)^3n^t\}\,, \quad t = 1, 2, \ldots \tag{19.41}$$

$$\mathrm{E}\{d^3\nu^t\} = \mathrm{E}\{d^3n^t\}\,, \quad t = 1, 2, \ldots \tag{19.42}$$

The following equation is true because of QTSD (Eq. (19.52) is fulfilled):

$$\mathrm{E}\{x^3\nu^t\} = \mathrm{E}\{x^3n^t\}\,, \quad t = 1, 2, \ldots \tag{19.43}$$

The quantization noise ν behaves like PQN from the point of view of these higher-order moments. Dither signals having triangular PDFs, or triangular convolved with rectangular, or triangular convolved with triangular are useful in the audio field and in other fields. Dither signals of this type may be generated by adding combinations of independent rectangularly distributed signals.

19.6.5 Gaussian Dither

The Gaussian PDF and CF do not satisfy the conditions for any of the quantizing theorems. However, a Gaussian dither with a standard deviation σ that is equal to or greater than $q/2$, half a quantum step, would come exceedingly close to satisfying QT I or QT II. Regarding all moments and joint moments, the quantization noise

ν would closely approximate the behavior of PQN. Evidence of this is given e. g. in Fig 5.6, in Tables 5.5, 5.2, 6.1, 6.2, 6.3, and 6.4, and throughout Chapter 11. Independent Gaussian noise would serve very well as a dither signal.

An additional benefit of Gaussian dither is that its PDF smoothly touches the horizontal axis at the edges, therefore its properties are not sensitive to the value of the ratio of its standard deviation and the quantum size q. Therefore, e.g. when nonlinearity of ADCs is considered (see Section E.3), Gaussian dither is preferable over uniformly or triangularly distributed dither.

19.6.6 Sinusoidal Dither

A sine wave is easy to generate, and it too could serve as an effective dither. The PDF of the sine wave, shown in Fig. I.6, does not satisfy the conditions for any of the quantizing theorems. However, if the peak-to-peak amplitude of the sine wave covers five to ten quantization boxes or more, the moments and joint moments of the quantization noise ν would have values approximating those of PQN n. Moments of the noise ν are shown in Fig. G.1 and Fig. G.11, for the sinusoidal dither acting alone. Example 20.9 describes a nice application of sinusoidal dither.

The dither is often generated by digital means. Therefore, the properties of *digital dither* need to be studied with special care. This is done in Appendix J.

19.6.7 The Use of Dither in the Arithmetic Processor

The somewhat deterministic pattern of the quantization errors (e.g. in the FFT) can be broken up by applying *uniform dither* before each roundoff operation (Horváth, 1987). The effectiveness of the use of uniform dither may be somewhat surprising, since theoretically, only triangularly distributed dither assures that ν and d will be uncorrelated (uniform dither assures that ξ and x are uncorrelated, but this is not sufficient, since the second moment of ξ depends on x). Simulations show that this is not a significant problem in the noise of FFT calculations. One might speculate that there are several roundoff noise sources, and their *average* behaves similarly to the average of independent noises. Dithering with uniform dither could be a true improvement in many kinds of arithmetic calculations. Its implementation is however not straightforward, since the dither needs to be added to the results *after* the operation, but *before* roundoff.[7]

Moreover, there is another important problem. If new dither is generated for each FFT, the results are not precisely reproducible for the same sequence, therefore testing becomes difficult. On the other hand, if the samples of the dither are fixed for an FFT to assure reproducibility, the average properties of the dither cannot be

[7] An alternative is to add the dither before the operations (scaling, addition or multiplication), but then each operation needs to be executed with inputs of increased precision. Furthermore, in multiplication, this increases the average variance.

exploited. A possibility is to “freeze” the dither sequence for testing only, but this is an additional complication usually not accepted.

An alternative improvement is to apply convergent rounding (see page 396).

19.7 THE USE OF DITHER FOR QUANTIZATION OF TWO OR MORE VARIABLES

A stochastic dither d is generally constructed to be independent of the input signal x. When two or more variables $x_1, x_2, ..., x_N$ are quantized, the dithers $d_1, d_2, ..., d_N$ can be constructed so that each of these dither signals are independent of each other and of all of the input variables $x_1, x_2, ..., x_N$. If all of the input variables $x_1, x_2, ..., x_N$ have zero values, the dither signals will be quantized, and the resulting quantization noises $\nu_1, \nu_2, ..., \nu_N$ will be independent of each other. If all of the input signals $x_1, x_2, ..., x_N$ being independent of $d_1, d_2, ..., d_N$ have other values, not all zero, whether stochastic or deterministic, the quantization noises $\nu_1, \nu_2, ..., \nu_N$ will not necessarily be independent of each other.

Suppose however that the dither signals $d_1, d_2, ..., d_N$ meet the conditions for multidimensional QT I (or, that they are independent, and each of them meets the conditions of QT I). The quantization noises $\nu_1, \nu_2, ..., \nu_N$ can be shown to be independent of each other, regardless of the nature of the inputs $x_1, x_2, ..., x_N$ as long as the dither signals are independent of these inputs. The quantization noises will have joint moments with the dither signals and the input signals that would be the same as if the quantization noises were replaced with independent PQN noises. The same results can be shown to be obtained when QT I is not satisfied but QT II is satisfied by the dither signals.

When the dither signals $d_1, d_2, ..., d_N$ do not meet the conditions for either QT I or QT II, but do meet conditions for other of the quantizing theorems, it is not clear whether the quantization noises $\nu_1, \nu_2, ..., \nu_N$ will be independent of each other or uncorrelated with each other, even when the dither signals $d_1, d_2, ..., d_N$ are all independent of each other. The input signals $x_1, x_2, ..., x_N$ are generally correlated with each other, and their characteristics may dominate and determine the statistical relations between $\nu_1, \nu_2, ..., \nu_N$.

Of greatest interest are conditions under which the quantization noises $\nu_1, \nu_2, ..., \nu_N$ are uncorrelated with each other. This happens for instance when all of the dither signals satisfy multidimensional QT IV/A. Then, although ν_1 is deterministically related to $(d_1 + x_1)$, it is true that

$$\mathrm{E}\{\nu_1(d_1 + x_1)\} = 0 \ , \quad \mathrm{E}\{\nu_1 d_1\} = 0 \ , \quad \text{and} \quad \mathrm{E}\{\nu_1 x_1\} = 0 \, . \tag{19.44}$$

Also, it is true that

$$\mathrm{E}\{\nu_2(d_2 + x_2)\} = 0 \ , \quad \mathrm{E}\{\nu_2 d_2\} = 0 \ , \quad \text{and} \quad \mathrm{E}\{\nu_2 x_2\} = 0 \, . \tag{19.45}$$

Although x_1 and x_2 are correlated with each other, ν_1 and ν_2 are uncorrelated with x_1 and x_2 respectively.

Therefore ν_1 and ν_2 are uncorrelated with each other and, by similar reasoning, we may conclude that $\nu_1, \nu_2, ..., \nu_N$ are mutually uncorrelated.

If $x_1, x_2, \ldots\ x_N$ are samples of a time function and $d_1, d_2, ..., d_N$ are independent additive dither samples that are independent or at least uncorrelated over time (white), and the dither satisfies any or all of the quantizing theorems QT I, QT II, QT III/B, then the quantization noise will be uniformly distributed between $\pm q/2$, uncorrelated with the input x and the dither d, and will be uncorrelated over time (white).

The precise condition of whiteness can also be given. Let us look for the condition that for two different time samples, $\mathrm{E}\{\xi_1\xi_2\} = \mathrm{E}\{d_1 d_2\}$. Wannamaker (1992) proved the following statement, using the conditional 2-D CF of ν:

Autocorrelation Function of ξ at Nonzero Lag Values
For independent dither, the necessary and sufficient condition of

$$\mathrm{E}\{\xi_1\xi_2\} = \mathrm{E}\{d_1 d_2\}\,, \tag{19.46}$$

is that each of the following equations is true:

$$\begin{aligned}
\Phi_{d_1,d_2}(l_1\Psi_1, l_2\Psi_2) &= 0 \quad \textit{for} \quad (l_1, l_2) \neq (0,0)\\
\left.\frac{\partial\, \Phi_{d_1,d_2}(l_1\Psi_1, u_{d_2})}{\partial u_{d_2}}\right|_{u_{d_2}=0} &= 0 \quad \textit{for} \quad l_1 \neq 0\\
\left.\frac{\partial\, \Phi_{d_1,d_2}(u_{d_1}, l_2\Psi_2)}{\partial u_{d_1}}\right|_{u_{d_1}=0} &= 0 \quad \textit{for} \quad l_2 \neq 0\,.
\end{aligned} \tag{19.47}$$

Example 19.6 Fulfillment of Conditions (19.47)
A simple application of this theorem is the use of independent, uniformly distributed dithers to both inputs. In this case

$$\Phi_{d_1,d_2}(u_{d_1}, u_{d_2}) = \Phi_{d_1}(u_{d_1})\,\Phi_{d_2}(u_{d_2}) = \mathrm{sinc}\left(\frac{q_1 u_1}{2}\right) \mathrm{sinc}\left(\frac{q_2 u_2}{2}\right)\,,$$

and it can be seen that the conditions (19.47) fulfill.
This means e.g. that if both inputs are converted via two independent stochastic–ergodic converters (see Section 19.3), the product is an unbiased estimate of the correlation:

$$\mathrm{E}\left\{A_1 x_{q1} \cdot a_2 x_{q2}\right\} = \mathrm{E}\left\{x_1 x_2\right\}\,, \tag{19.48}$$

since $\mathrm{E}\{d_i\} = 0$.

Example 19.7 Defocusing the Optics
Photos can be used to determine the two-dimensional position of an object. Paradoxically, resolution can be improved by defocusing the optics. This is approximately equivalent to replacing the picture of each point by a uniformly illuminated round spot. The resulting picture is a two-dimensional convolution of the true picture and a function uniformly distributed above a circle. Therefore, this defocused spot acts like two-dimensional dither, since it is a two-dimensional lowpass filter. See Exercise 8.6 for some details.

19.8 SUBTRACTIVE DITHER

When the input to a quantizer cannot be relied upon to meet the criteria for a suitable quantizing theorem, a dither signal can be added to the quantizer input to guarantee satisfaction. Although the dither is beneficial in that it linearizes the quantizer and ensures known properties for the quantization noise, it does make the quantizer output more noisy. A clever idea was proposed by Roberts (1962) to overcome this drawback. He added a dither to the quantizer input, and subtracted it from the quantizer output. The dither thus acted like a catalyst in a chemical process, making the process work better but not appearing in the process output. Roberts' subtractive dither idea should be used whenever it is possible.

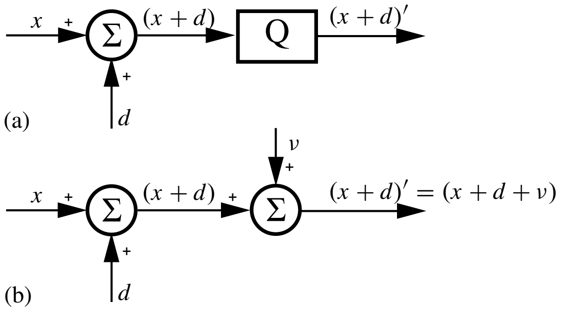


Figure 19.10 Dither without subtraction: (a) a quantizer with dither; (b) quantizer replaced by additive quantization noise.

Figure 19.10(a) shows the quantizer with a dithered input, and Fig. 19.10(b) shows how the dither contributes to the quantizer output noise. Without subtractive dither, the quantizer output is $x+d+\nu$. Accordingly, when d and ν are uncorrelated,

$$\begin{aligned}\begin{pmatrix}\text{quantizer}\\ \text{output}\\ \text{noise power}\end{pmatrix} &= \mathrm{E}\{\nu^2\}+\mathrm{E}\{d^2\}\\ &= q^2/12+\mathrm{E}\{d^2\}\,.\end{aligned} \tag{19.49}$$

The power of the dither adds to the quantization noise power.

Figure 19.11(a) illustrates the idea of subtractive dither. Fig. 19.11(b) shows subtractive dither with the quantizer represented by additive quantization noise. The result of dither subtraction on the quantizer output is shown in Fig. 19.11(c). With subtractive dither, the quantizer output is $x+\nu$. The power of the dither does not add to the quantization noise power, but if the dither satisfies QT IV/A, the PQN model is perfectly usable, independently of the properties of x. The benefit of subtractive dither is clear.

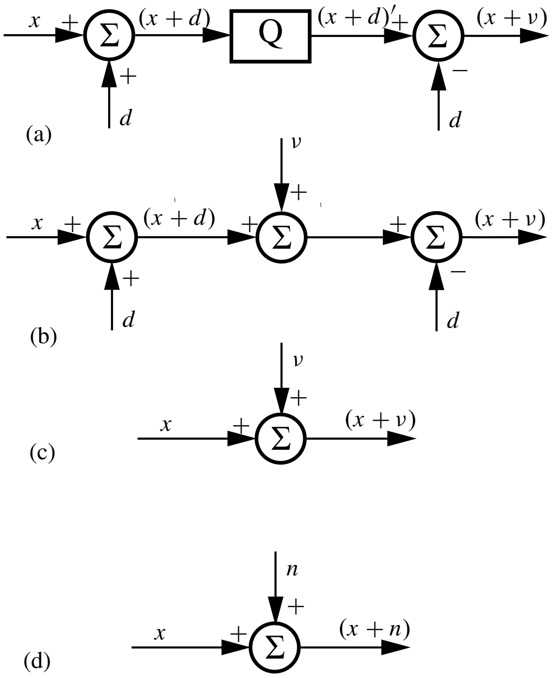


Figure 19.11 Dither with subtraction: (a) a quantizer with subtractive dither; (b) quantizer replaced by additive quantization noise, with subtractive dither; (c) equivalent diagram of quantizer with subtractive dither; (d) PQN model.

In order to prove the independence of ν of input x for a dither fulfilling QT IV/A, the joint CF of the input signal and the noise can be investigated:

$$\begin{aligned}\Phi_{x,\nu}\,(u_x, u_\nu) &= \sum_{l=-\infty}^{\infty} \Phi_x(u_x + l\Psi)\,\Phi_{\mathrm{d}}(l\Psi)\,\Phi_n(u_\nu + l\Psi)\\ &= \sum_{l=-\infty}^{\infty} \Phi_x(u_x + l\Psi)\,\Phi_{\mathrm{d}}(l\Psi)\,\mathrm{sinc}\left(\frac{q(u_\nu + l\Psi)}{2}\right). \qquad (19.50)\end{aligned}$$

Studying this, an important consequence can be drawn. When the dither satisfies QT IV/A, then all $l \neq 0$ members disappear in the sum:

$$\Phi_{x,\nu}(u_x, u_\nu) = \Phi_x(u_x) \cdot \Phi_n(u_\nu)\,, \tag{19.51}$$

which is the CF of the sum of two independent random variables. Therefore, the following theorem holds:

Quantizing Theorem for Subtractive Dither (QTSD)
If in dithered quantization

$$\Phi_{\mathrm{d}}(l\Psi) = 0 \quad \textit{for} \quad l = \pm 1, \pm 2, \ldots, \tag{19.52}$$

then the quantization noise $\nu = (x + d)' - (x + d)$ *will be independent of the input signal* x*, and it will be uniform between* $(\pm q/2)$.

If (19.52) is fulfilled, we call d a *zero-order dither*.

We mention here that independence of x and ν does not imply the independence of d and ν, but for subtractive dither this is not required, since d is eliminated from the output variable. It is however possible that *both x and d* fulfill the conditions of QT IV/A. In this case, x, ν and d are all pairwise independent. As we have seen in Example 19.1 (page 491), this is not enough by itself to assure that the triplet x, d, and ν be independent. The key to this is studying the derivatives of the joint CF. This, and a general theorem on joint moments is given in the Addendum, readable in this book's website.

19.8.1 Analog-to-Digital Conversion with Subtractive Dither

Subtractive dither can only be utilized when the exact dither signal is available for subtraction after quantization. A dither signal from a pseudo-random generator would generally be used. Since the generating algorithm would be known, the dither sequence could be recreated for subsequent subtraction. This could be used with digital recording of audio or video signals. The dither would be added to the analog signal before quantization, and then be subtracted after playback and digital-to-analog conversion.

The subtractive dither signal would need to be generated by a computer or by some form of digital apparatus so that it could be recorded and repeated. Injecting the dither into the input of an analog-to-digital converter (ADC) would require the dither to be in analog form. If generated by computer, the digital dither signal could be converted to analog by a zero-order-hold digital-to-analog converter (DAC). Care would be needed in doing this to ensure that the sampling instants of the ADC would not correspond to the transition instants of the DAC.

Suppose that the ADC is a 16-bit converter. Then the dither must exist on a much finer scale with much finer quantization steps so that it approximates an analog dither. For example, let the dither exist on a 24-bit scale that covers the same dynamic range as the 16-bit scale of the ADC. So the steps of the dither will be $2^8 = 256$ times finer than the quantization steps of the ADC. The dither signal in analog form is added to the input x, and the sum is quantized. The quantizer output, in digital form, becomes the input to a computer. The dither, now in digital form, must be subtracted in the computer from the computer input. Refer to Fig. 19.12.

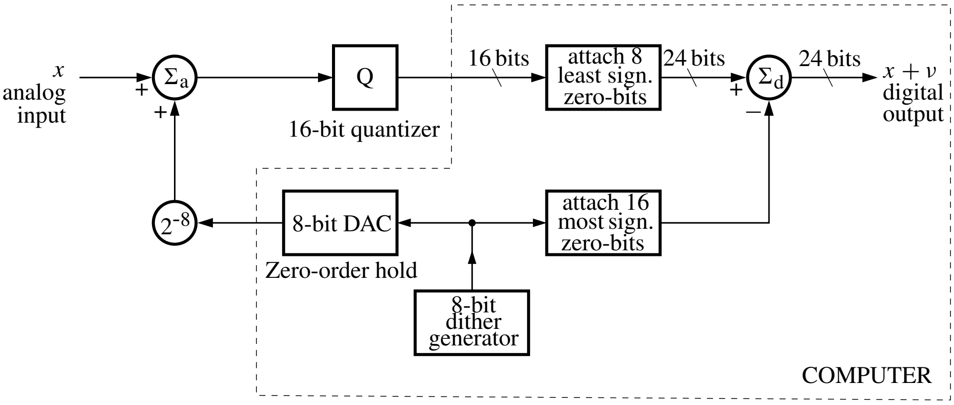


Figure 19.12 A fixed-point analog-to-digital converter with subtractive dither. Σ_a denotes analog summation, and Σ_d denotes digital summation. The abbreviation “sign.” stands for “significant.”

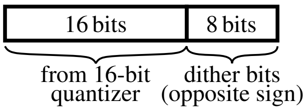


Figure 19.13 Formation of the 24-bit output of the analog-to-digital converter of Fig. 19.12.

In Fig. 19.12, an 8-bit pseudo-random generator generates digital dither that might be, for example, uniformly distributed, or triangular distributed, or Gaussian, and so forth. The dither signal goes through an attenuator with an attenuation of 2^{-8}, and is added to the analog input x. The dithered input is quantized to 16-bits. The quantizer output is then given eight additional bits in the least significant bit positions, all set to zero. The quantizer output now is 24-bit. The 8-bit dither signal with sixteen zero bits attached to the most significant positions is then subtracted from the quantizer output signal. Thus, a 24-bit dither is subtracted from the 24-bit quantizer signal to produce the 24-bit quantized output. Figure 19.13 shows the formation of the bits of the quantized output.

In actual practice, some care will be needed to choose the gain of the analog attenuator. The gain shown in Fig. 19.12, 2^{-8}, is based on assuming integer numbers and a gain of unity for the DAC. This is not realistic, since a commercial 8-bit DAC whose maximum number magnitude would be $2^7 - 1$ might produce a maximum output of 5 volts, but illustrates the need of proper scaling of the dither. The gain needs to be chosen to assure proper dither level at the input and at the output. It is also important to add dither d at the quantizer input while the *same* d is subtracted from the quantizer output.

The converter of Fig. 19.12 has an analog input and a 24-bit digital output. It should be understood that this system is a 16-bit converter with 16-bit resolution. The quantization noise power, $q^2/12$, corresponds to that of the 16-bit quantizer. With a 24-bit output, the quantizer of Fig. 19.12 has nicer quantization noise properties when used in many applications such as in digital video and digital audio. The disadvantage is that the digitized samples have 24 bits rather than 16 bits. This increases the cost of storage and processing.

This could be avoided by storing the 16-bit ADC output, and regenerate and subtract the dither only when the samples are used. But with integrated circuits, bits and memory have become very cheap. Analog-to-digital conversion is still expensive.

Experiments were done with a quantization system similar to that of Fig. 19.12. The input signal x was sinusoidal, and scaled so that its peak-to-peak range covered five quantization steps.

Figure 19.14(a) shows the original sinusoid. Figure 19.14(b) shows its power spectrum. Figure 19.14(c) shows the quantized sinusoid, quantized without dither. Figure 19.14(d) shows the spectrum of the sinusoid, quantized without dither. Figure 19.14(e) shows the quantized sinusoid, quantized with subtractive dither. Figure 19.14(f) shows the spectrum of the sinusoid, quantized with subtractive dither. The frequency of the sinusoid was chosen to be a subharmonic of the sampling frequency, and the dither was chosen to be uniformly distributed over $\pm q/2$ and uncorrelated over time. The spectrum of the original sinusoid shows a single line. The spectrum of the quantized sinusoid, without dither, shows the distortion. There are harmonics of the original sinusoid, and the average amplitude of the fundamental is somewhat smaller in this case than the corresponding amplitude of the original sinusoid. (In other cases, the amplitude of the fundamental could be larger.) The spectrum of the quantized sinusoid with subtractive dither shows a fundamental and no harmonics. The average fundamental amplitude is the same as that of the original sinusoid. But now, the effects of random quantization noise are seen in the baseline of the spectrum of Fig. 19.14(f). This is white $q^2/12$ noise, corresponding to ν. Without dither, the distortion of the sinusoid due to quantization as manifested in the harmonics shown in Fig. 19.14(d) and in the change in amplitude of the fundamental together add to a total power level of approximately $q^2/12$. This is approximate

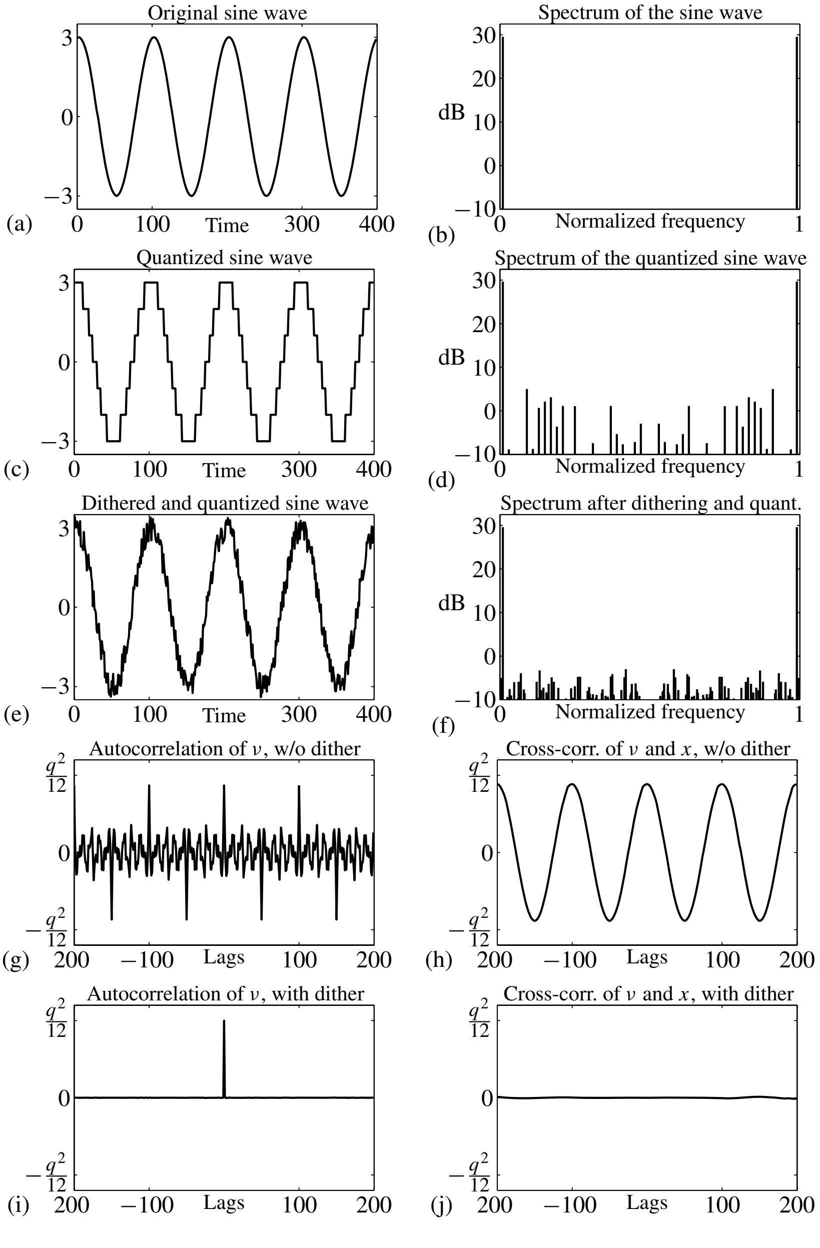


Figure 19.14 Quantization of a sinusoidal signal, without dither and with subtractive dither: (a) the original sine wave; (b) the spectrum of original sine wave; (c) the quantized sine wave; (d) spectrum of the quantized sine wave; (e) sine wave quantized with subtractive dither; (f) spectrum of sine wave quantized with subtractive dither; (g) autocorrelation function between ν and x without dither; (h) crosscorrelation function between ν and x without dither; (i) autocorrelation function of quantization noise with subtractive dither; (j) crosscorrelation function between ν and x with subtractive dither.

because the sinusoid by itself does not satisfy any quantizing theorem at the quantizer input.

Figures 19.14(g) and 19.14(i) show autocorrelation functions of the quantization noises without dither and with subtractive dither. Without dither, the periodic components of the quantization noise are evident. With subtractive dither, the autocorrelation shows no periodic components and shows the quantization noise to be simply white noise having power of $q^2/12$. With no dither, the autocorrelation for zero lag shows the total quantization noise power to be just slightly less than $q^2/12$ (in other cases, it could be higher than $q^2/12$).

The use of dither does not generally reduce the power of quantization noise, and indeed, without subtraction, adds its power to the quantization noise power making things potentially worse. Dither is used to condition the quantization noise, to make it white and to cause its joint moments with the input x to be like that of PQN.

Figure 19.14(h) shows the crosscorrelation function between ν and x for quantization without dither. The correlation between periodic quantization noise and the sinusoidal input is clearly evident. This result is very different from that of Fig. 19.14(j), showing the crosscorrelation function between ν and x for quantization with subtractive dither. In this case, the quantization noise is white noise and is not periodic. It does not correlate with the input x. From the standpoint of moments, it behaves like PQN.

19.9 DITHER WITH FLOATING-POINT

19.9.1 Dither with Floating-Point Analog-to-Digital Conversion

At this time, almost all analog-to-digital converters convert on a fixed-point scale with uniform quantization. However, some analog-to-digital converters are in development that convert an analog input directly to floating-point form. We anticipate that in the future such converters will be available and therefore we write this section to analyze their behavior with and without dither.

A floating-point quantizer is diagrammed in Fig. 19.15(a). Dither d is applied to its input in Fig. 19.15(b). Generation of the dither is shown in Fig. 19.15(c). A zero mean random noise signal d_y having a fixed PDF is multiplied by 2^E to obtain the dither d.

The floating-point conversion and the dither process are done in two steps. First, without dither, the input x is converted to floating-point digital form. The result is a mantissa and an exponent E. With the second step, the dither d is added to x and the conversion is repeated giving a new mantissa and exponent. The desired conversion result is the new mantissa and exponent. The first step is done just to get the exponent E, which is used in the formation of the dither for the second step.

In many practical implementations of floating-point dithering, the signal is first compressed, and the already compressed y is then uniformly quantized. Therefore,

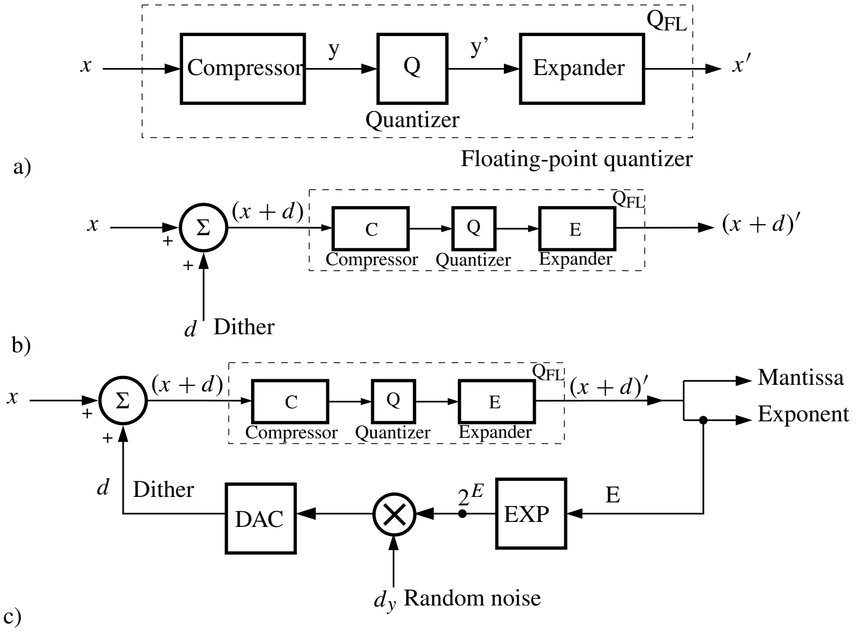


Figure 19.15 Two-step floating-point quantizer with dither: (a) the floating-point quantizer; (b) quantizer with dither; (c) dithered quantizer showing formation of the dither.

we can add the dither directly to y, and avoid the complications caused by dithering the input signal x, see Fig. 19.16.

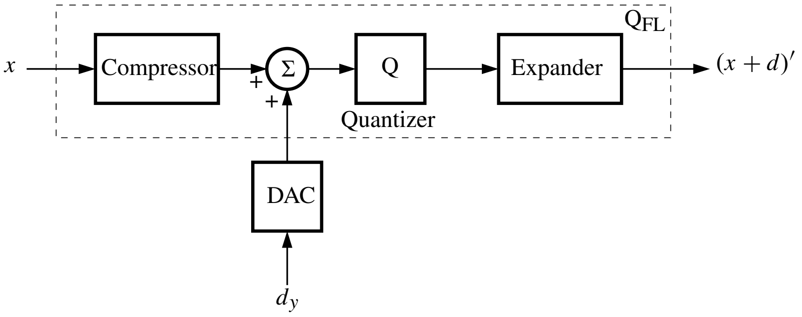


Figure 19.16 Floating-point quantizer with dither applied directly to the compressed signal.

An example of a floating-point quantizer's input–output staircase function is shown in Fig. 12.3. To provide dither for such a quantizer, the range of the input x would need to be known so that the standard deviation of the dither could be adjusted for the step size. This is accomplished by multiplying the random noise d_y by 2^E in order to obtain the dither d.

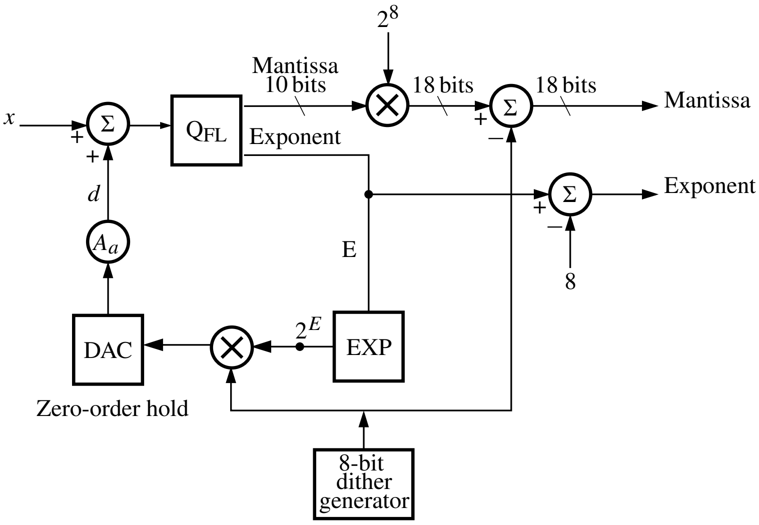


Figure 19.17 Floating-point ADC with subtractive dither. A_a denotes an analog attenuator.

Another example of a floating-point quantizer's input–output staircase function is shown in Fig. 12.10. Its compressor characteristic is shown in Fig. 12.7, its expander characteristic is shown in Fig. 12.8, and its hidden quantizer characteristic is shown in Fig. 12.9. If the random noise d_y is uniformly distributed over the range $\pm q/2$ (the smallest step of the floating-point quantizer), multiplying it by 2^E will cause the additive dither at the hidden quantizer input to be uniformly distributed over plus and minus half a step for all amplitudes of the input x, with rare exceptions at exponent transitions. Then QT IV/A would be satisfied at the input of the hidden quantizer, with the consequence that the quantization noise ν_{FL} would be uncorrelated with the input x. If on the other hand the random noise d_y is triangular distributed over the range $\pm q$, then multiplying it by 2^E will cause the additive dither at the hidden quantizer input to be triangular distributed over the range $\pm q$. QT IV would be satisfied and the quantization noise ν_{FL} would once again be uncorrelated with the input x.

Dither d_y with a triangular PDF will work better than dither d_y with a rectangular PDF. Dither with a triangular PDF generally works better than dither with a uniform PDF for uniform quantization. For floating-point quantization, there is an additional reason for the superiority of dither with a triangular PDF. The additional reason is that the tapered triangular PDF handles edge effects more gracefully, when x is at a level where the mantissa is ready to overflow and cause the exponent to go up by one. Likewise, a Gaussian dither also works well. The uniformly distributed dither however has the advantage (except for edge effects) of being bounded by the narrowest limits.

19.9.2 Floating-Point Quantization with Subtractive Dither

A floating-point analog-to-digital converter with subtractive dither would need to be constructed by combining the system functions illustrated in Figs. 19.12 and 19.15. Figure 19.17 shows a system that does this. The input x is analog, and the output is digital, consisting of an exponent and a mantissa. The floating point quantizer $\mathrm{Q_{FL}}$ is assumed, for purposes of illustration, to produce an output with a 10-bit mantissa. For this illustration, the dither is assumed to be 8-bit, including one sign bit. At the converter output, the mantissa has 18 bits. But it should be realized that the basic resolution of this converter is that of a 10-bit mantissa, corresponding to the resolution of $\mathrm{Q_{FL}}$. The purpose of the additional mantissa bits is to allow the quantization noise to be well behaved, as is the case with subtractive dither.

In a floating-point ADC, dither can be added to the already compressed signal, thereby simplifying the structure, see Fig. 19.18.

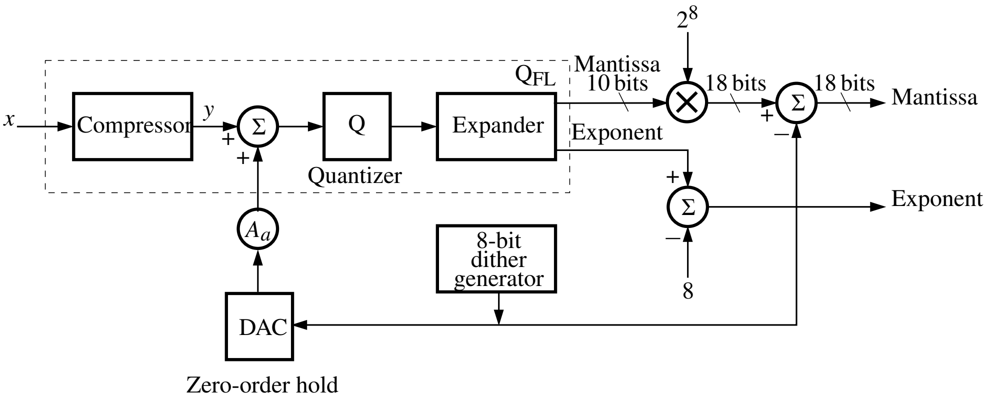


Figure 19.18 Floating-point ADC with subtractive dither, applied to the mantissa. A_a denotes an analog attenuator.

The eight bits of the pseudo-random dither generator may be assumed for this illustration to be independently randomly chosen with 50% probability of 0's and 50% probability of 1's. Equally likely negative and positive numbers will be generated whose values will be distributed over the range ± 127. This approximates a uniformly distributed dither. Other distributions are possible, but this one is a good and convenient choice.

Subtracting the dither at the output of $\mathrm{Q_{FL}}$ can be done in the following way. The 10-bit mantissa of $\mathrm{Q_{FL}}$ is made into an 18-bit mantissa by appending eight zeros beyond its least significant bit. This is indicated by multiplication of the mantissa by

2^8. (This scaling is compensated for by subtracting 8 from the exponent.) The 8-bit dither is then subtracted from the enlarged mantissa. Taking into account the sign bit of the dither, its seven bits are subtracted (if the dither is positive) or added (if the dither is negative) to the seven least significant bits of the enlarged mantissa. The range of the subtracted dither is plus or minus one half of the least significant bit of the mantissa of $\mathrm{Q_{FL}}$.

Adding the dither at the input to $\mathrm{Q_{FL}}$ can be done as follows. The 8-bit dither is multiplied by 2^E (its bits are shifted left by the amount E) and fed to a DAC. The DAC output is scaled by a gain factor to form the input dither d. The exact scale factor cannot be predicted because the gain of the DAC (this is the ratio of the largest output voltage to the largest input binary number) would need to be known. The scale factor must be chosen so that the dither subtracted at the output of $\mathrm{Q_{FL}}$ is equal to the dither added at its input.

Simulation experiments were done with a floating-point quantizer without dither and with subtractive dither. The input signal was a sine wave, shown in Fig. 19.19(a). Its spectrum is shown in Fig. 19.19(b). With a mantissa length of $p = 2$, the quantized sine wave was visibly distorted, as shown in Fig. 19.19(c). The spectrum of the quantized sine wave in Fig. 19.19(d) shows a fundamental and its harmonics. Figure 19.19(e) shows the result of floating-point quantization of the sine wave with subtractive dither. Although the distortion in this case looks worse than without dither, the spectrum of the quantized sine wave in Fig. 19.19(f) shows a fundamental and no harmonics. The quantization noise ν_{FL} is essentially white and not periodic. The autocorrelation function of ν_{FL} without dither, shown in Fig. 19.19(g), shows periodicity in the quantization noise, while the autocorrelation function of ν_{FL} with a subtractive dither shown in Fig. 19.19(i) is the autocorrelation function of white noise. The crosscorrelation function between ν_{FL} and x for floating-point quantization without dither is periodic, as can be seen in Fig. 19.19(h). The same crosscorrelation function for floating-point quantization with subtractive dither is shown in Fig. 19.19(j). Since in this case ν_{FL} is essentially white noise, it does not correlate hardly at all with the periodic input x, and the crosscorrelation function is close to zero for all lags.

With floating-point quantization like with fixed-point uniform quantization, subtractive dither of proper design makes the quantization noise behave like PQN from the point of view of moments and joint moments.

19.9.3 Dithered Roundoff with Floating-Point Computation

When pairs of floating-point numbers are added or multiplied in a computer, it is generally necessary to quantize the sum or product in order to preserve the length of the mantissa.

When floating-point numbers are multiplied, the mantissa of the product is the product of the mantissas, and the exponent of the product is the sum of the exponents.

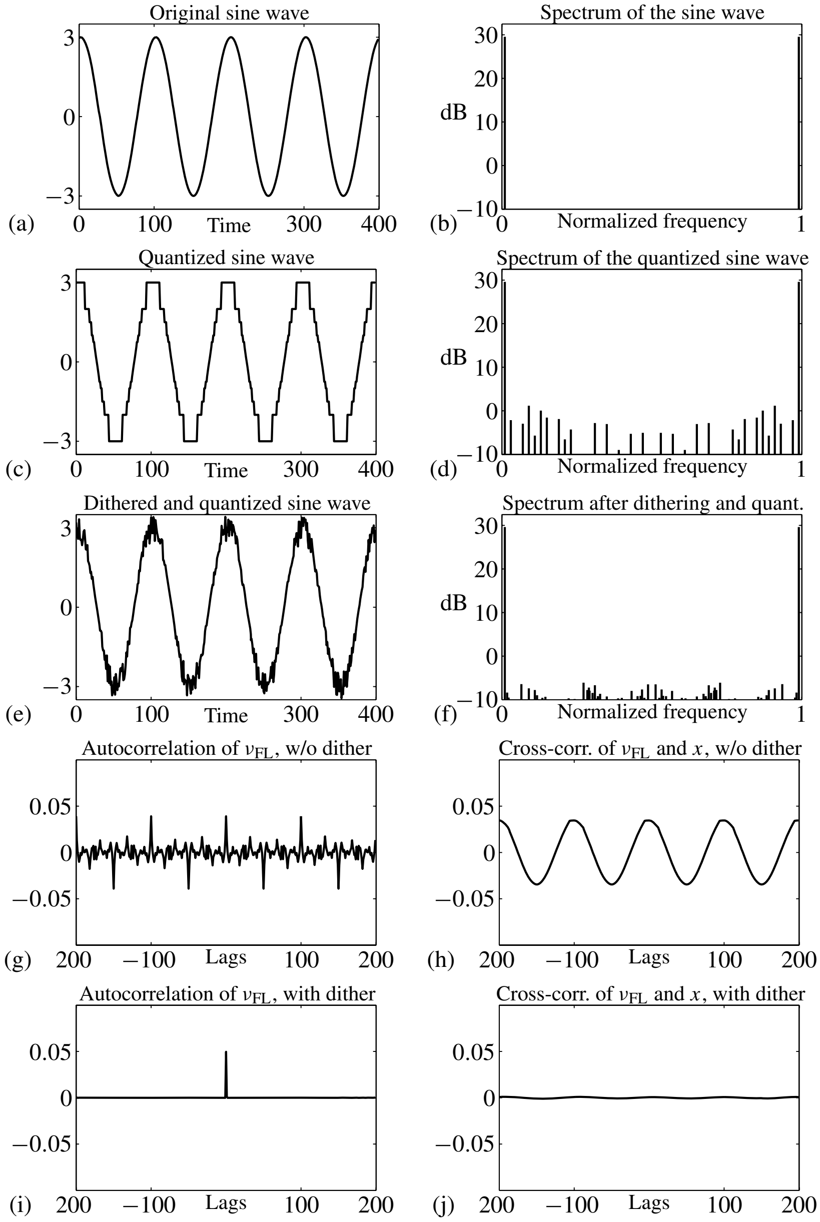


Figure 19.19 Floating-point quantization of a sinusoidal signal with $p = 2$ (bits of possible normalized mantissa values: -1.1, -1.0, $+1.0$, $+1.1$), exponent $-3, -2, \ldots, 3$, without dither and with subtractive dither: (a) the original sine wave; (b) spectrum of the original sine wave; (c) the quantized sine wave; (d) spectrum of the quantized sine wave; (e) sine wave quantized with subtractive dither; (f) spectrum of a sine wave quantized with subtractive dither; (g) autocorrelation function of quantization noise without dither; (h) crosscorrelation function between ν_{FL} and x for quantization without dither; (i) autocorrelation function of quantization noise with subtractive dither; (j) crosscorrelation function between ν_{FL} and x for quantization with subtractive dither (note the small correlation values: their deviations from zero are caused by the finite sample number used in simulation).

The product of the mantissas has a length of the order of the sum of the lengths of the two mantissas, and this necessitates roundoff. When floating-point numbers are summed, the mantissas need to be aligned relative to each other taking into account the exponent values. A large difference in the exponents could cause the sum to have a mantissa length much greater than that of either of the mantissas. Generally all numbers in the computer have mantissas of the same standard length. As such, the mantissa of the sum of two numbers has a length of the order of the machine's standard mantissa length plus the difference in the exponents.

In general, the mantissas of the sum, difference, product, or quotient of two floating-point numbers will be longer than the standard machine mantissa. To reduce the length of the mantissa in order to conform to the standard, one could truncate the mantissa by throwing away the excess least significant bits. However, a much better approach would be to round the least significant bit of the standard word up or down depending on the bits to be eliminated. (Do they represent a quantity that is greater or less than one half the value of the least significant bit of the standard?)

Rounding is what is done most commonly. Before rounding, a dither could be added. The purpose is to make sure that the quantization noise ν_{FL} is uncorrelated with the input x. How this might be done is illustrated in Fig. 19.20.

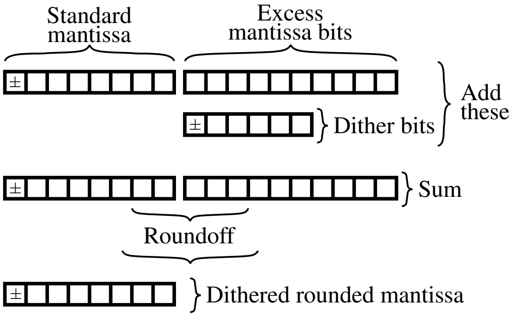


Figure 19.20 Dithered quantization of a floating-point number.

In Fig. 19.20, the most significant part of the mantissa to be rounded consists of 8 bits, and the least significant part consists of 10 bits. The 18-bit number is to be rounded so that it will contain just 8 bits. Before rounding, a dither is added. For this example, the dither is digital and contains 5 bits plus a sign bit. All 6 bits may be chosen randomly, independently, and with equal probability of 1s and 0s. By adding or subtracting the 5 dither bits with the bit alignment illustrated in Fig. 19.20, a dither that is approximately uniformly distributed between plus and minus one half the value of the least significant bit is added to the 18-bit mantissa. The resulting number is then rounded to 8 bits.

In digital signal processing, we can generate and use digital dither before arithmetic quantization. This sounds very reasonable, but *implementation* is not trivial at all. In most computers, quantization is an inherent part of arithmetic operations, since the results are calculated with the available memory precision only. In such cases, we obtain the already quantized samples, therefore we cannot add the dither *before* quantization. Therefore an indirect solution needs to be chosen. For example, by modification of the calculation algorithm, we could generate lower bits of the result separately, in order to prepare for dithering. However, this is cumbersome and often dubious.

The situation is much better when the results are generated in an accumulator with higher precision than that of the memory. If these excess bits are available, they can be used, combined with dither, to decrease the quantization distortion. If the excess bits are not directly available, usually there is a roundabout way to indirectly obtain them, but these possibilities are typically not worth doing.

An additional observation that can be made is that the dither amplitude is usually at the LSB level. For digital dither, its resolution is certainly finer than the LSB of storage. Therefore, the dither cannot be subtracted after quantization, since the result would have again bits which cannot be stored. Consequently, in calculations usually only *non-subtractive dither* can be used.

The use of this kind of dither causes the quantization noise of the hidden quantizer ν to be uncorrelated with its input y. The independent dither is also uncorrelated with y. Based on the theory of floating-point quantization, the floating-point quantization noise ν_{FL} and the dither at the floating-point quantizer output are uncorrelated with each other and are both uncorrelated with the input x. Since the quantization noise and the dither are both uniformly distributed between plus and minus one half of a quantization step of the hidden quantizer, they contribute equally in the mean square sense to the net distortion of the quantization process.

By dithering before rounding in floating-point arithmetic, we ensure that the quantization noise is uncorrelated with the input x. The net distortion, the sum of the quantization noise and the dither, will also be uncorrelated with the input x, and will have a mean square value equal to twice that of the quantization noise alone. The doubling of the distortion power is the price paid for uncorrelatedness between the distortion and input x. Sometimes this is worthwhile.

The question is, could the doubling of the distortion power be avoided by subtractive dithering? The answer is yes, but the standard mantissa length would need to be increased. Contemplating this, one would probably be better off using a longer mantissa and rounding, with or without dither.

Finally, we need to mention an important disadvantage of such dithered calculations. If the dither is random or pseudo-random, the result of each calculation becomes slightly random, depending on the dither values. This means that when repeating the same calculation, the result may slightly differ. This makes the proper testing of otherwise deterministic algorithms very difficult. Therefore, dither in arithmetic calculations is not used in practice.

19.10 THE USE OF DITHER IN NONLINEAR CONTROL SYSTEMS

An important area of application for dither signals is to nonlinear feedback control systems. Here, nonlinearity can appear as a result of quantization in digital control systems, and it could result from sensors such as proximity detectors or shaft encoders. Nonlinearity could be in the form of a stairstep of uniform quantization, or it could have the characteristics of on–off two-level (signum) error detectors used in contactor control, or three or more level detectors. Hysteresis might be present. Some of these characteristics are shown in Figs. 1.4–1.7.

Purposes of introducing dither into feedback control systems with nonlinearity are, among others, linearization, stabilization, and elimination of “limit cycles” and static offsets. Engineers have used dither for years, on an empirical basis, to “improve” the behavior of nonlinear control systems. Now, with the theory to explain the effects of dither and how to design it, dither can be used more effectively and with greater confidence. The subject is discussed in Chapter 17.

19.11 SUMMARY

If the quantization noise ν has PQN-properties, the quantizer, a nonlinear device, behaves in a statistical sense like a linear device. The quantizer is then a source of additive noise whose statistical properties are known and fixed: mean $= 0$, variance $= q^2/12$, uncorrelated with the quantizer input x.

If the quantizer input x has a PDF that does not satisfy any of the quantizing theorems, the quantization noise ν will not have properties like PQN. These properties can be obtained, if desired, by the addition of a suitably designed independent dither signal d to the quantizer input.

When dither is added, the input to the quantizer is $(x+d)$. The output is $(x+d)'$. The quantizer output is the sum of quantizer input x, the quantizer noise ν, and the dither d, i.e., $(x + d)' = x + \nu + d$. If the PDF of d satisfies QT I for example, the sum $(x + d)$ will automatically satisfy QT I. The CF of d will be bandlimited, and the CF of $(x + d)$, the product of the CFs, will be bandlimited at least to the same extent, regardless of the properties of the quantizer input x. If the PDF of d does not satisfy QT I but satisfies QT II, then the PDF of $(x + d)$ will satisfy QT II, by similar reasoning. In fact, since all of the quantizing theorems involve regions of zero-value and zero derivatives for the CF of x, satisfaction of any of the quantizing theorems QT I, QT II or QT III by the dither d will ensure satisfaction of the same quantizing theorem by $(x + d)$.

If dither d has a CF with a narrow enough bandwidth to satisfy QT I, then its addition to x before quantization ensures that the quantizer input $(x + d)$ will have a CF with a narrow enough bandwidth to satisfy QT I. The addition of dither to the quantizer input acts in a way that is analogous to lowpass filtering commonly used before signal sampling to prevent aliasing. The dither is an anti-alias filter for quantization.

Dither can be designed to satisfy QT I or QT II only approximately, not perfectly. A Gaussian dither with $\sigma > q/2$ comes very close to perfection however. With such a dither, the PQN model works almost perfectly. The quantization noise ν has essentially zero mean, a variance of $q^2/12$, and is uncorrelated with input x. The total output noise $(\nu + d)$ is also essentially uncorrelated with input x. In fact, if QT II were perfectly satisfied, the quantization noise ν would be independent of x, and the total output noise $(\nu + d)$ would also be independent of x. The quantizer would be "linearized" by the dither, and the PQN model would prevail.

If the quantizer input is a time series, a series of samples over time, a white Gaussian dither with $\sigma > q/2$ would cause the quantization noise to have essentially zero mean, a variance of $q^2/12$, to be uncorrelated with input x, and to be an additive independent white noise.

A sinusoidal dither whose zero-to-peak amplitude covers several quantum boxes, although it would not have a PDF that satisfies any of the quantizing theorems, would cause the quantization noise to have close to PQN properties.

A dither with a triangular PDF whose amplitude range covers $\pm q$ would perfectly satisfy QT III/B. Using this dither would cause ν to have mean of zero, a variance of $q^2/12$, and it would be uncorrelated with $(x + d)$, d, and x. The total noise $(\nu + d)$ at the quantizer output would be uncorrelated with x. From the point of view of second-order moments, use of this dither would ensure that the quantization noise ν would behave exactly like n of the PQN model.

When dither is added to the quantizer input, the total quantizer output noise is equal to $(\nu + d)$. The disadvantage to using dither is that the total output noise includes d. When using a Gaussian dither whose standard deviation is $q/2$ for example, the total output noise power is $q^2/12+q^2/4 = q^2/3$, which is four times greater than that of the quantization noise alone. When using a triangular dither whose amplitude range is $\pm q$ for example, the total output noise power is $q^2/12 + q^2/6 = q^2/4$, which is three times greater than that of the quantization noise alone. The added noise power at the quantizer output represents a significant price to pay for both of these dithers.

In some cases, this additional output noise can be eliminated by subtracting the same dither signal at the quantizer output that had been added in at the quantizer input. The method is called "subtractive dither," and was first proposed by Roberts (1962). It should always be practiced when it is possible to do so. The method removes the dither from the quantizer output. The dither then works like the catalyst of a chemical process. It makes the process work better, but does not appear at the process output.

The dither methodology developed for uniform quantization can be modified for use with floating-point quantization. Dither can also be used in numerical floating-point computation. Frequently, dither is used in nonlinear feedback control systems to achieve linearization or approximate linearization. The theoretical ideas about dither that have been developed in this chapter could be very helpful in designing dither signals for these and other applications. One could use the quantizing theorems QT I, QT II, QT III, and QT IV to design dithers so that the quantization noise would have desired physical and statistical properties. The most popular dithers are Gaussian, sinusoidal, rectangular, and triangular.

19.12 EXERCISES

19.1 Assume that when recording for 16-bit storage on a CD, a dither triangularly distributed in $[-q, q]$ is used. What is the maximum SNR in this case? Theoretically, can an SNR of 96 dB be reached for 16-bit fixed-point representation, using dither?

19.2 The input range of an eight-bit A/D converter (7 bits + sign) is ±1 V. A DC voltage between ±0.5 V is to be measured, the desired accuracy is ±1 mV. Let us suppose that the A/D converter can be well modeled by an ideal uniform quantizer. Use a Gaussian dither.

(**a**) Determine the required minimum standard deviation of the dither, if the mean of the quantization error is to be kept smaller than 0.5 mV.

(**b**) Using the dither determined in part (a), how many independent samples of the quantizer output need to be averaged so that the maximum error within the 95% confidence interval is not larger than 0.5 mV?

19.3 An unknown signal is quantized using a a dither which is a sum of three independent, uniformly distributed variables whose range is $\pm q$ each.

(**a**) Which moments of $\xi = (d + \nu)$ are equal to the corresponding moments of $(n + d)$?

(**b**) Which joint moments of $\xi = (d + \nu)$ with x and d correspond to joint moments of $(d + n)$ with x or d?
Hint: use the joint CFs of x and ϖ, and d and ϖ, respectively, available from the web address
`http://www.mit.bme.hu/books/quantization/.`

19.4 The input signal x satisfies the conditions of QT IV/A. What can we say about ν and about $\xi = d + \nu$? What can we say if d satisfies QT IV/A, but x not?

19.5 Let a dither be generated with the PDF given in Fig. E3.12.1 of Exercise 3.12 (page 55). For which values of q is the dither zero-order, for which values is it first-order?

19.6 Let a dither be generated with the PDF given in Fig. E3.13.1 of Exercise 3.13 (page 55). For which values of q is the dither zero-order, for which values is it first-order?

19.7 Consider the dither PDFs given in Fig. E19.7.1(a)–(d). What is the order of the dither generated with these PDFs? For which values of q is the dither zero-order, for which values is it first-order? Give the expressions of the dither CFs.

19.8 Determine the characteristic function of the quantized output variable $(x + d)'$,

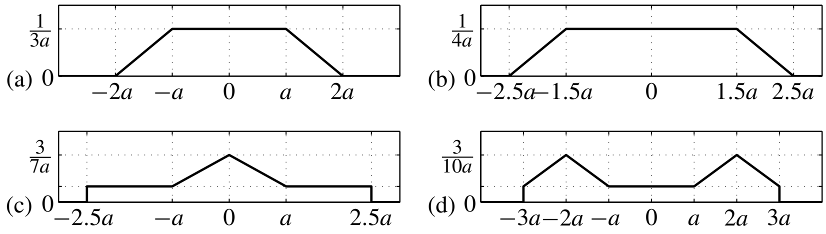


Figure E19.7.1 Probability density functions of dithers to analyze.

(a) starting from the CF of the quantizer output without dither, given by Eq. (4.11),

(b) starting from the joint CF given by Eq. (19.4),

(c) starting from the joint CF given by Eq. (19.17).

19.9 Prove that for first-order dither, d and ν are uncorrelated.

19.10 Prove that for independent first-order dither, $\text{var}\{(x+d)'\} = \text{var}\{x\} + \text{var}\{d\} + q^2/12$.

19.11 If x is uniformly distributed in $(-q/2, q/2)$, and d is also uniformly distributed in $(-q/2, q/2)$, the sum $x + d$ fulfills QT III/B, although d is only zero-order (it is not first-order).

(a) Is the quantization noise $\nu = (x + d)' - (x + d)$ independent of x? Are they correlated?

(b) Is ν independent of d? Are they correlated?

(c) Is ν independent of $x + d$? Are they correlated?

(d) Which moments of $\xi = d + \nu$ are functionally independent of x?

19.12 Assume that the PDF of x can be written as the weighted sum of two shifted PDFs:

$$f_x(z) = 0.5g(z) + 0.5g(z + q/2)\,, \tag{E19.12.1}$$

where $g(z)$ is a PDF, and the dither is uniform in $(-q/4, q/4)$.

(a) Determine which of the QTs for dither (QTD III, or QTD IV, or QTSD, or GQTSD) are fulfilled.

(b) Which QT is fulfilled by $x + d$?

(c) Determine the PDF of the quantization noise $\nu = (x + d)' - (x + d)$.

(d) Is the quantization noise ν independent of x? Are they correlated?

(e) Is ν independent of d? Are they correlated?

(f) Is ν independent of $x + d$? Are they correlated?

(g) Which moments of $\xi = d + \nu$ are functionally independent of x?

(h) Determine the PDF of $\xi = d + \nu$ for x binary with equal probabilities at $(0, q/2)$, furthermore at $(q/4, 3q/4)$.

(i) Determine the CF of $\xi = d + \nu$ for these cases.

19.13 For input x like in Eq. (E19.12.1), and a dither with trapezoidal PDF (Fig. E19.13.1):

$$f_d(z) = \begin{cases} 2\left(\frac{z}{q}+\frac{3}{4}\right)\frac{1}{q} & \text{for } -\frac{3}{4}q \le z < -\frac{1}{4}q \\ \frac{1}{q} & \text{for } -\frac{1}{4}q \le z < \frac{1}{4}q \\ 2\left(\frac{3}{4}-\frac{z}{q}\right)\frac{1}{q} & \text{for } \frac{1}{4}q \le z < \frac{3}{4}q \\ 0 & \text{otherwise} \end{cases} \qquad \text{(E19.13.1)}$$

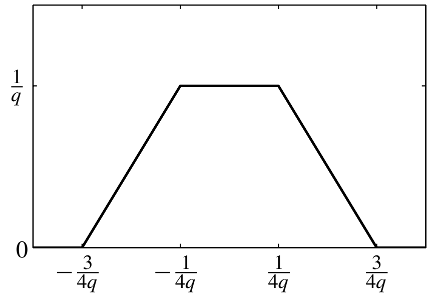


Figure E19.13.1 Trapezoidal PDF.

repeat the questions (c)-(h) of Exercise 19.12.

19.14 Assume that the PDF of x can be written as the sum of three shifted PDFs:

$$f_x(z) = 0.25g(z-q/2) + 0.5g(z) + 0.25g(z+q/2)\,, \qquad \text{(E19.14.1)}$$

where $g(z)$ is a PDF, and the dither is triangular in $(-q/2, q/2)$. Repeat the questions of Exercise 19.12.

19.15 A dither signal satisfies QT IV/B but not QT IV/A. What is assured by the theory?

19.16 Roberts (1962) has introduced a global measure of the distortion of the quantizer with subtractive dither, the so-called D-factor. This is defined as

$$D = \mathrm{E}\{(x - \mathrm{E}\{(x'-d)|x\})^2\}\,. \qquad \text{(E19.16.1)}$$

This is a nonnegative measure of the quantizer distortion with subtractive dither, and it becomes zero exactly when the resulting quantization noise ν is unbiased. Schuchman (1964) has shown that this statement is equivalent to the conditions for QTSD (page 508), therefore the measurement of the quantity D can be used to check for fulfillment of QTSD. Prove this statement.

19.17 At the website of the book, you can find the image `einstein.tif`. Do all of the following:

(a) Read the image into MATLAB and quantize it uniformly to 8 representative levels. Evaluate the mean squared error between the original image and the quantized image.

(b) Generate random noise uniformly distributed over $[-d/2, d/2]$, where d is the step size for the quantization in part (a), and add this noise to the original image. What is the mean squared error between the original image and the dithered image? Quantize this dithered image uniformly to 8 representative levels. Also, evaluate the mean squared error between the original and the dithered and quantized image.

(**c**) In which case, part (a) or part (b), do you get a higher mean squared error? In which case, part (a) or part (b), is the subjective quality of the quantized image better? Explain your answers briefly.

(**d**) The matrix

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \tag{E19.17.1}$$

is a 2-D lowpass filter kernel. Filter the quantized image from (a) using this kernel. Also filter the dithered and quantized image from (b) using this kernel. Evaluate the mean squared error with respect to the original image for both. What effect does the lowpass filtering have, in this case, on subjective quality?

Hint: Use the following code for filtering an image:

```
kernel = [1 2 1; 2 4 2; 1 2 1];
kernel = kernel/sum(sum(kernel));
img_filt = conv2(img, kernel);
img_filt = img_filt(2:(size(img_filt,1)-1),...
   2:(size(img_filt,2)-1));
```

19.18 Derive the equivalent of Sheppard's first 4 corrections (express the moments of x) with independent zero-mean non-subtractive dither applied which satisfies the PQN model.

19.19 Prove that with independent dither triangularly distributed between $\pm q$, the quantization noise ν is orthogonal to and uncorrelated with d, regardless of the nature of x, and furthermore, ν is orthogonal to and uncorrelated with x.

19.20 Verify, by Monte Carlo simulation, Eqs. (19.39), (19.40), and (19.43).

19.21 Using Gaussian and sinusoidal dithers, with various input signals x, check by Monte Carlo simulation how closely the PQN model fits the experimental data regarding the joint moments of x, d, $\xi = d + \nu$, and ν. Specifically, $q = 1$,

(**a**) Gaussian dither, $\mu = 0$, $\sigma = q$,

(**b**) sinusoidal dither, $\mu = 0$, $A_{\text{pp}} = 15q$,

for signal

(**i**) constant x, with value $x = 0.3q$,

(**ii**) sinusoidal x, $\mu = 0.1q$, $A_{\text{pp}} = 0.3q$,

(**iii**) uniform x, $\mu = 0.2q$, $A_{\text{pp}} = q$,

check the moments $\text{E}\{d\xi\}$, $\text{E}\{\nu\}$, $\text{E}\{x\nu\}$.

19.22 Suppose that, as in Section 19.7 (page 504), each of the individual dither signals $d_1, d_2, ..., d_N$ meets the conditions for QT II. Show that the quantization noises $\nu_1, \nu_2, ..., \nu_N$ are independent of each other, regardless of the nature of the inputs $x_1, x_2, ..., x_N$ as long as the dither signals are independent of these inputs.

19.23 Prove that if the dithers d_i, and d_j satisfy the multidimensional QT IV/A, ν_i and ν_j are uncorrelated with x_i and x_j respectively, and that ν_i and ν_j are uncorrelated with each other.

19.24 Verify by computer simulation that with triangular independent dithers distributed between $\pm q$, ν_1 and ν_2 are uncorrelated with x_1, x_2, d_1, and d_2, for x_1 and x_2 being input time series signals as in Exercise 19.21, even if $x_1 = x_2$. Examine $\mathrm{cov}\{x_1, \nu_2\}$, $\mathrm{cov}\{d_1, \nu_1\}$, and $\mathrm{cov}\{d_2, \nu_1\}$.

19.25 Explain why the statement following Eq. (19.13): "since x and n are statistically independent, x and ν are also statistically independent and ν is uniformly distributed between $\pm q/2$" is true, and discuss the form of $\Phi_{x,\nu}(u_x, u_\nu)$.

19.26 Assume that for numbers equal to (integer+0.5)q, an independent dither uniformly distributed in $(-q/2, q/2)$ is added, and afterwards, convergent rounding (see page 396) is implemented. Describe the behavior of ν, $\mathrm{E}\{\nu|x\}$, $\mathrm{E}\{\xi|x\}$, $\mathrm{E}\{\nu^2|x\}$, and $\mathrm{E}\{\xi^2|x\}$, for x uniformly distributed on the discrete values $\{(k + 0.5)q\}$, $k = 1, 2, ..., N$.

19.27 Prove the statement (Autocorrelation Function of ξ at Nonzero Lag Values), given on page 505.

19.28 The distance of the Moon from the Earth can be determined by radar echo measurements (John H. DeWitt, Jr, Jan. 1946, Zoltán Bay, Feb. 1946). However, these echoes are very weak, thus the noise buries them.

In order to improve the SNR, a pseudo-code can be emitted, and the two-way travel time can be determined by crosscorrelation of the pseudo-code to the noisy echo received from the moon.

Simulate the correlation-based experiment, using the following data.

The average Earth–Moon distance is about $D = 385\,000$ km. Therefore, the estimated delay is about $2D/c \approx 2.56$ s.

The radar signal of 100 MHz is switched on and off according to a pseudo-random binary signal (Godfrey, 1993) of length $N = 2^{10} - 1 = 1023$. The clock frequency of the binary signal is $f_\mathrm{c} = 10$ Hz.

Received power as a function of time is calculated, anti-alias filtered, then sampled. At the output of the bandpass receiver ($\Delta f/f_0 \approx 0.01$), the noise is white Gaussian. The estimated SNR (calculated with respect to signal power with "signal on") is about -20 dB. This means that the received signal is very noisy, but, on the other hand, this noise acts like dither, allowing detection of the signal even with very rough quantization.

The sampling frequency is $f_\mathrm{s} = 1$ kHz. Cross-correlation is determined with the input binary signal.

Determine by simulation whether with 8-bit ADC (input range: $\pm 3\sigma$ around the approximate value of the power to be sampled), crosscorrelation is usable to detect the distance of the Moon. If there any difference if quantization of the received noisy signal is made rough, to 4-level, or to 3-level, or even to 2-level? Why do these also work?

19.29 Prove that if a discrete dither has a distribution symmetric to zero, and has values only on the grid kq_d (or at least on the grid $kq_\mathrm{d}/2$), with $k = 0, \pm 1, \pm 2, \ldots$, it is true for the derivative of its characteristic function that $\dot{\Phi}_{\mathrm{d}}\left(\lambda\frac{2\pi}{q_{\mathrm{d}}}\right) = 0$, for $\lambda = \pm 1, \pm 2, \ldots$. Note: these derivative values appear in Eq. (J.4). Examples: distributions shown in Fig. J.4(a)–(c) (Appendix J, page 693).

19.30 (**a**) Determine the variances of the dithers defined in Fig. J.2(a)–(c) (page 691) in the amplitude domain.

(**b**) Determine the characteristic functions of the dithers defined in Fig. J.2(a),(c). (Notice that the distribution in Fig. J.2(c) is a convolution of the distribution shown in Fig. J.2(b) and of the binary distribution[8] at $\pm q_{\mathrm{d}}/2$.)

(**c**) Do the CFs exactly fulfill condition (J.1) for $r = 1$?

(**d**) Use the CFs for an alternative determination of the variances.

19.31 (**a**) Determine the variances of the dithers defined in Fig. J.4(a)–(c) (page 693) in the amplitude domain.

(**b**) Determine the characteristic functions of the dithers defined in Fig. J.4(b),(c). (Notice that the distribution in Fig. J.4(b) is an convolution of the distributions illustrated in Fig. J.2(b) and Fig. J.2(c), furthermore that the distribution in Fig. J.4(c) is an autoconvolution of the distribution of Fig. J.2(c).)

(**c**) Do the CFs exactly fulfill condition (J.1) for $r = 2$?

(**d**) Define in what sense are the CFs for Fig. J.4(b),(c) similar to the characteristic function (J.8).

(**e**) Use the CF for an alternative determination of the variances. Compare the results with the ones for part (a).

19.32 QTDD (see page 686) assures that the moments of ξ can be made functionally independent of x, but does not guarantee their values.

(**a**) Assume that a triangular digital dither of Fig. J.4(a) is applied. Determine the value of the moment $\mathrm{E}\{\xi^2\}$ by

i. theoretical calculations, and/or
ii. by numerical summation of the probabilities, for $L = 3$, and/or
iii. by Monte Carlo, for $L = 3$ (e.g. use roundrand, page 713).

and give its difference from the theoretical value $3q^2/12$.

For each case, determine the maximum possible deviation from the above value, caused by deterministic roundoff of the quantizer input values $(k+0.5)q$, exactly in the midpoint between two representable numbers.

(**b**) Repeat the questions in **(a)** for the dither of Fig. J.4(b).

(**c**) Repeat the questions in **(a)** for the dither of Fig. J.4(c).

19.33 Devise simple means to generate dithers illustrated in

(**a**) Fig. J.2(a)–(c) (page 691),

(**b**) Fig. J.4(a)–(c) (page 693).

[8] Binary distribution is binomial distribution with $n = 1$, see Section 3.7, page 47.

19.34 Determine simple approximations of the errors of the corrected first and second moments when approximately normally distributed digital dither is applied (normally distributed dither with $\mu = 0$, and $\sigma = q$ is quantized with a uniform midtread quantizer with quantum step q_d), with $2^L = q/q_\mathrm{d} = 16$:

(a) when x has a continuous distribution, like exponential, with $\lambda = 1/q$;
(b) when x is representable on the grid kq_d.

19.35 The dither in Fig. J.2(a) (page 691) is asymmetric to the vertical axis (the mean value is $-q_\mathrm{d}$). We try to make it symmetric by deleting the value at $-q/2$, and increasing the probabilities of the other values accordingly.

(a) What do you think, why is this dither not recommended in the book?
(b) Determine the characteristic function of the dither. Does this CF fulfill condition (J.1) for $r = 1$?
(c) Determine the dependence of the error of the first moment on the discrete values $x = kq_\mathrm{d}$, and on x when it has continuous value.
(d) Can this "chopped" dither be used as dither?

19.36 Assume that for numbers equal to (integer + 0.5)q, an independent digital dither as shown in Fig. J.2(b) is added with $q_\mathrm{d} = q/4$, and afterwards, either rounding towards zero (see page 12), or convergent rounding (see page 396) is implemented. Describe the behavior of ν, $\mathrm{E}\{\nu|x\}$, $\mathrm{E}\{\xi|x\}$, $\mathrm{E}\{\nu^2|x\}$, and $\mathrm{E}\{\xi^2|x\}$, for x uniformly distributed on the discrete values $\{(k + 0.5)q\}$, $k = 1, 2, ..., N$.

19.37 Determine the conditional CF $\Phi_{\nu|x}(u)$, and use this for the proof of GQTSD (see the Addendum on this book's website):

General Quantizing Theorem for Subtractive Dither (GQTSD)
If in dithered quantization

$$\frac{\mathrm{d}^m\,\Phi_\mathrm{d}(u_d)}{\mathrm{d}u_d^m}\Bigg|_{u_d=l\Psi} = 0 \tag{E19.37.1}$$

for m being a nonnegative integer, and $l = \pm1, \pm2, \ldots$, then

$$\mathrm{E}\{x^{t_x} d^m \nu^{t_\nu}\} = \mathrm{E}\{x^{t_x} d^m n^{t_\nu}\}$$

for $t_x = 0, 1, \ldots,$ and $t_\nu = 0, 1, \ldots$

19.38 Starting from the joint characteristic function:

$$\begin{aligned}\Phi_{x,d,\nu}\left(u_x, u_d, u_\nu\right) \\ = \sum_{l=-\infty}^{\infty} \Phi_x(u_x + l\Psi)\,\Phi_\mathrm{d}(u_d + l\Psi)\,\mathrm{sinc}\left(\frac{q(u_\nu + l\Psi)}{2}\right)\end{aligned}, \tag{E19.38.1}$$

prove the following equation:

$$\Phi_{x,d,\xi}\left(u_x, u_d, u_\xi\right) = \sum_{l=-\infty}^{\infty} \Phi_x(u_x + l\Psi)\,\Phi_\mathrm{d}(u_d + u_\xi + l\Psi)\,\mathrm{sinc}\left(\frac{q(u_\xi + l\Psi)}{2}\right). \tag{E19.38.2}$$