Appendix J

Digital Dither

Quantization theory deals primarily with continuous-amplitude signals and continuous-amplitude dither. However, within a digital signal processor or a digital computer, both the signal and the dither are represented with finite word length. Examples are digital FIR and IIR filtering, digital control, and numerical calculations. In these cases, intermediate results (e.g. products of numbers) whose amplitude is discrete, have excess bit length, so they must be re-quantized to be stored with the bit number of the memory. Before re-quantization, digital dither may be added to the signal, or sometimes this is even necessary to avoid limit cycles and hysteresis (see Fig. J.1, and Exercises 17.10–17.12, page 462).

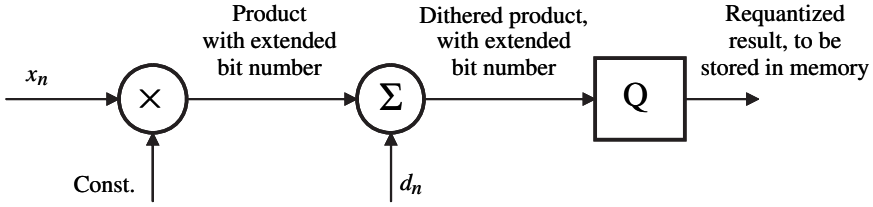

Figure J.1 Application of digital dither within a computer, after multiplication.

Another scenario, when the dither is digital, is when the dither is generated within the computer for the quantization of analog signals. This usually means that each dither sample is produced by a pseudo-random number generator, and a D/A converter is used to convert the number to an analog level to be added to the input of the quantizer before quantization.

In both cases, it is good to know the properties of the most common digital dithers. Therefore, in this appendix we will investigate the properties of digital dither which is desired to be added to a digital signal before requantization.[1]

[1] A part of this appendix was first published in, and is reprinted with permission, from Kollár, I., "Digital non-subtractive dither: Necessary and sufficient condition for unbiasedness, with implementation issues," *Proceedings of the 23rd IEEE Instrumentation and Measurement Technology Conference,* Sorrento, Italy, 24-27 April 2006, pp. 140–145.

J.1 QUANTIZATION OF REPRESENTABLE SAMPLES

An interesting approach was presented by Wannamaker, R. A., Lipshitz, S. P., Vanderkooy, and Wright (2000). They have recognized that in general, no digital dither can completely remove quantization bias. Therefore, they looked into the possibility of removing the bias for all the input numbers *representable*[2] with the given bit number at the input of the quantizer. Inspired by their work, we state a theorem here, which has a condition which can be fulfilled by practical dithers:

Quantizing Theorem for Digital Dither (QTDD)
For a digital system in which re-quantization is used to remove the L least significant bits of binary data, $\mathrm{E}\{\xi^m|x\}$ *has the same value for all representable values of x for* $m = 1, 2, \ldots, r$*, if a non-subtractive digital dither (with the same precision as the input data) is applied for which*

$$\left.\frac{\mathrm{d}^t\,\Phi_\mathrm{d}(u)}{\mathrm{d}u^t}\right|_{u=l\Psi} = 0\,, \tag{J.1}$$

for $t = 0, 1, \ldots, r-1$*, at* $l = 1, 2, \ldots, 2^L - 1$ *((*$r-1$*)th-order digital dither).*

The proof follows from examination of the conditional CF of ξ, given in the Addendum. The required moments are not influenced by the value of x in the infinite sum of the last part, because

- For the first moment, we need to examine the sum for the values $x = kq_\mathrm{d} = k2^{-L}q$, where q_d is the quantum size of the digital dither. We will look at the terms for which $l \neq 2^L\lambda$ (λ is an integer), and at the terms for which $l = 2^L\lambda$, separately.

$$\begin{aligned}\mathrm{E}\{\xi|x\} &= \frac{1}{j}\frac{\mathrm{d}}{\mathrm{d}u_\xi}\left(\sum_{l=-\infty}^{\infty}\Phi_\mathrm{d}\left(u_\xi + l\Psi\right)e^{jl\Psi x}\,\mathrm{sinc}\left(\frac{q(u_\xi + l\Psi)}{2}\right)\right)\Bigg|_{u_\xi=0}\\ &= \mu_\mathrm{d}\\ &\quad + \frac{1}{j}\sum_{\substack{l=-\infty\\ l\neq 2^L\lambda}}^{\infty}\Phi_\mathrm{d}\left(l\Psi\right)e^{jl\frac{2\pi}{q}kq_\mathrm{d}}\frac{\mathrm{d}}{\mathrm{d}u}\,\mathrm{sinc}\left(\frac{q(u_\xi + l\Psi)}{2}\right)\Bigg|_{u_\xi=0}\\ &\quad + \frac{1}{j}\sum_{\substack{\lambda=-\infty\\ \lambda\neq 0}}^{\infty}\Phi_\mathrm{d}\left(2^L\lambda\Psi\right)e^{j2^L\lambda\frac{2\pi}{q}kq_\mathrm{d}}\frac{\mathrm{d}}{\mathrm{d}u}\,\mathrm{sinc}\left(\frac{q(u_\xi + 2^L\lambda\Psi)}{2}\right)\Bigg|_{u_\xi=0}\end{aligned}$$

[2]Representable are the numbers which are possible for the input signal of the quantizer, that is, all the numbers that may be given by a combination of the input bits.

$$
\begin{aligned}
&= \mu_{\mathrm{d}} \\
&\quad + \frac{1}{j} \sum_{\substack{l=-\infty \\ l \neq 2^L \lambda}}^{\infty} \Phi_{\mathrm{d}} \left(l \frac{2\pi}{q} \right) e^{jlk \frac{2\pi}{2^L}} \frac{q}{\pi} \frac{1}{l} \\
&\quad + \frac{1}{j} \sum_{\substack{\lambda=-\infty \\ \lambda \neq 0}}^{\infty} \Phi_{\mathrm{d}} \left(\lambda \frac{2\pi}{q_{\mathrm{d}}} \right) e^{j\lambda k 2\pi} \frac{q_{\mathrm{d}}}{\pi} \frac{1}{\lambda} \,. \qquad \text{(J.2)}
\end{aligned}
$$

The first sum equals zero because of the theorem's condition (J.1) for $m = 1$, and the second sum does not depend on k (and thus, does not depend on x).[3] This proves the theorem for $m = 1$.

- Similarly, for $m > 1$, independence of x (or of k) will be provided, since the terms for $l \neq 2^L \lambda$ disappear because of (J.1), and x disappears from each term corresponding to $l = 2^L \lambda$:

$$
e^{jl\Psi x} = e^{j2^L \lambda \frac{2\pi}{q} k d_{\mathrm{d}}} = e^{j\lambda k 2\pi} \equiv 1 \,. \qquad \text{(J.3)}
$$

While functional independence of x is provided, the sum usually does not equal zero. For $m = 1$, an extra condition was given to assure zero value (footnote 2 on page 687). For $m = 2$, the nonzero value is more common. If (J.1) is fulfilled for $r = 2$, it is enough to examine the possibly nonzero elements:

$$
\begin{aligned}
\mathrm{E}\{\xi^2 | x\} &= \frac{1}{j^2} \frac{\mathrm{d}^2}{\mathrm{d}u_\xi^2} \left(\sum_{l=-\infty}^{\infty} \Phi_{\mathrm{d}} \left(u_\xi + l\Psi \right) e^{jl\Psi x} \operatorname{sinc} \left(\frac{q(u_\xi + l\Psi)}{2} \right) \right) \Big|_{u_\xi = 0} \\
&= \mathrm{E}\{d^2\} + \mathrm{E}\{n^2\} \\
&\quad - \sum_{\substack{\lambda=-\infty \\ \lambda \neq 0}}^{\infty} \dot{\Phi}_{\mathrm{d}} \left(2^L \lambda \Psi \right) e^{j2^L \lambda \frac{2\pi}{q} k d_{\mathrm{d}}} \frac{\mathrm{d}}{\mathrm{d}u_\xi} \operatorname{sinc} \left(\frac{q(u_\xi + 2^L \lambda \Psi)}{2} \right) \Big|_{u_\xi = 0} \\
&\quad - \sum_{\substack{\lambda=-\infty \\ \lambda \neq 0}}^{\infty} \Phi_{\mathrm{d}} \left(2^L \lambda \Psi \right) e^{j2^L \lambda \frac{2\pi}{q} k d_{\mathrm{d}}} \frac{\mathrm{d}^2}{\mathrm{d}u_\xi^2} \operatorname{sinc} \left(\frac{q(u_\xi + 2^L \lambda \Psi)}{2} \right) \Big|_{u_\xi = 0} \\
&= \mathrm{E}\{d^2\} + \mathrm{E}\{n^2\}
\end{aligned}
$$

[3] In addition, if the dither has a distribution symmetric to zero, or a distribution symmetric to any of $nq_{\mathrm{d}}/2$, $n = \pm 1, \pm 2, \ldots$, the second sum also equals zero since

$$
e^{-jn \frac{q_{\mathrm{d}}}{2} \frac{2\pi}{q_{\mathrm{d}}}} \Phi_{\mathrm{d}} \left(\lambda \frac{2\pi}{q_{\mathrm{d}}} \right) = e^{jn \frac{q_{\mathrm{d}}}{2} \frac{2\pi}{q_{\mathrm{d}}}} \Phi_{\mathrm{d}} \left(-\lambda \frac{2\pi}{q_{\mathrm{d}}} \right), \quad \text{that is,} \quad \Phi_{\mathrm{d}} \left(\lambda \frac{2\pi}{q_{\mathrm{d}}} \right) = \Phi_{\mathrm{d}} \left(-\lambda \frac{2\pi}{q_{\mathrm{d}}} \right),
$$

therefore the terms in the sum are cancel out pairwise for $\pm\lambda$.

$$- \sum_{\substack{\lambda=-\infty \\ \lambda\neq 0}}^{\infty} \dot{\Phi}_{\mathrm{d}}\left(\lambda\frac{2\pi}{q_{\mathrm{d}}}\right)\frac{q_{\mathrm{d}}}{\pi}\frac{1}{\lambda} + \sum_{\substack{\lambda=-\infty \\ \lambda\neq 0}}^{\infty} \Phi_{\mathrm{d}}\left(\lambda\frac{2\pi}{q_{\mathrm{d}}}\right)\frac{q_{\mathrm{d}}^2}{\pi^2}\frac{1}{\lambda^2}\,. \tag{J.4}$$

This is often somewhat larger than $\mathrm{E}\{d^2\}+\mathrm{E}\{n^2\}$. However, the deviation is relatively small, since the first sum is zero for a dither which has a distribution symmetric to zero and has values only on the grid kq_{d} (or at least only on the grid $kq_{\mathrm{d}}/2$, $k = 0, \pm1, \pm2, \ldots$, see Exercise 19.29), while the second sum can be upper bounded since $|\,\Phi(u)| \leq 1$:

$$\mathrm{E}\{\xi^2|x\} \leq \frac{q_{\mathrm{d}}^2}{3}\,, \tag{J.5}$$

which is usually negligible when compared to $\mathrm{E}\{d^2\} + \mathrm{E}\{n^2\} = \mathrm{E}\{d^2\} + q^2/12$. Since this deviation does not depend on x, it can be corrected for.

From QTDD, the important consequence is that the resolution (LSB) of the digital dither should be the same as of the data to be quantized, $q_{\mathrm{d}} = 2^{-L}q$. The theorem provides that this is sufficient. Finer resolution of the dither would be superfluous, coarser resolution would not be sufficient. This theorem is useful for dithering in digital computers and digital signal processors.

J.1.1 Dirac Delta Functions at $q/2 + kq$

While the above proof is correct, the theorem has an important application limitation. Quantization theory is considered as area sampling of a smooth PDF. When at the edge of such an area there is a Dirac delta function, it is tacitly assumed in the derivation that half of the integral of the Dirac delta belongs to this area, and half of it to the next area. This is a property of Fourier transform pairs, and the proofs are based upon Fourier transform. This corresponds to random-direction quantization of input values equal to $(\text{integer} + 0.5)q$: half of the values at the comparison levels are rounded downwards, half of them are rounded upwards.

The existence of such Dirac delta functions is a common case in re-quantization. In practical processors, however, as well as in simulations in MATLAB, a deterministic algorithm is implemented (see Exercise 1.3): such values are either rounded always upwards, or always downwards, or towards zero, or towards $\pm\infty$ (like in MATLAB's `round(·)` function), or convergent rounding is implemented (rounding towards the closest *even* number when the input is exactly at 0.5 LSB distance from two representable numbers). Quantization theory *does not deal with these cases*. Therefore, we have to content ourselves by

- either accepting that convergent rounding averages out the bias for the given input signal,
- or assuming that for the given input, the probability of the values just at 0.5 LSB from two representable values is very small,

- or saying that if $q_\mathrm{d} \ll q$, deviation between theory and practice is negligible,
- or implementing in the processor (in the simulation program) a modification of rounding to correspond to theory: when having a number which equals (integer + 0.5) LSB, either additional dither should determine if rounding is done in the upwards or downwards direction, or the program takes care to do upwards/downwards rounding alternately for the same level.

J.2 DIGITAL DITHER WITH APPROXIMATELY NORMAL DISTRIBUTION

In a computer, it is easy to generate normally distributed numbers. Either a pseudo-random number generator can be used, or several independent, identically distributed random numbers can be added. These normally distributed numbers are then quantized to q_d to make a digital dither.

The characteristic function of the approximately normally distributed digital dither is

$$\Phi_\mathrm{d}(u) = \sum_{\lambda=-\infty}^{\infty} e^{j(u+\lambda\Psi_\mathrm{d})\mu} e^{-\frac{\sigma^2(u+\lambda\Psi_\mathrm{d})^2}{2}} \operatorname{sinc}\left(\frac{q_\mathrm{d}(u+\lambda\Psi_\mathrm{d})}{2}\right), \qquad \text{(J.6)}$$

with $\Psi_\mathrm{d} = 2\pi/q_\mathrm{d}$.

If the common rule $\sigma > q$ is followed for the normally distributed dither, and its mean value is equal to zero, the moments of the dither can be well reconstructed from samples of the digital dither, using Sheppard's corrections. Some similar corrections can be used between moments of $(x+d)'$ and of $(x+d)$, or between moments of $(x+d)' - d$, and of x.

For Gaussian dither with $\sigma > q$, the condition (J.1) of QTDD is fulfilled with good approximation, thus the moments are independent of x, if x is representable on the grid kq_d. This does not mean however that for *any* value of x, the moments would be unbiased. It is heuristically clear that the digital dither has "roughness" q_d, therefore, if x is arbitrary, the error in Sheppard's first correction may reach $\pm q_\mathrm{d}/2$, and in Sheppard's second correction it may be in the order of magnitude of $q_\mathrm{d}^2/6$ (Exercise 19.34). We cannot go into these details, the errors of this kind can be studied in detail by investigation of the corresponding CFs, by making use of the dither CF, given in Eq. (J.6).

J.3 GENERATION OF DIGITAL DITHER

Let us turn now to the generation of digital dither. Random number generators can be realized based on different principles (Godfrey, 1993). One of the most popular methods is based on feedback shift registers. These generate $2^N - 1$ pseudo-random

bits, where N is the register length, and these can be used to generate pseudo-random numbers. If 2^N is not very large, we notice that the generated dither has a periodic nature, and that using a full period, it goes through every individual step. This latter fact may be used for increased efficiency in averaging: in this special case the sequence to be averaged contains all possible values just once. Therefore, the result of averaging is exact, with no randomness due to dithering.

The first thing we have to decide is the distribution of the dither. We can approximate any distribution by digital means; however, uniform and triangular dithers are by far the most popular ones. We will deal here with these. Gaussian and sinusoidal dithers are also usual and reasonable choices. Their properties can be determined with similar analysis.

As for the number of bits, according to QTDD (see page 686), it is not reasonable to use a dither which has finer resolution than the variable to be quantized. Thus, the difference of the bit numbers of the accumulator and of the memory (the storage bit number) determine the reasonable bit number of the dither.

When the number of bits is known, the digital representation of the distribution is to be selected. We can consider digital dither as a *finely quantized* version of the continuous one. Therefore, using at least a few bits, we can apply the approximation that the variance is $\mathrm{var}\{d\} \approx \mathrm{var}\{d_\mathrm{c}\} + q_\mathrm{d}^2/12 \approx \mathrm{var}\{d_\mathrm{c}\}$, with d_c being the continuous-time dither, and q_d denoting the dither LSB.

J.3.1 Uniformly Distributed Digital Dither

For uniform dither, we have a few, almost equivalent, solutions (Fig. J.2).

In Fig. J.2(a), the digital dither clearly has a bias of $\mathrm{E}\{d\} = -q_\mathrm{d}/2$. The representation is simple and straightforward. The number of different values is 2^L, with $L = \log_2(q/q_\mathrm{d})$. The variance is $\mathrm{var}\{d\} = \mathrm{var}\{d_\mathrm{c}\} - q_\mathrm{d}^2/12 = q^2/12 - q_\mathrm{d}^2/12$. In $1/2^L$ part of the cases $x + d$ will be equal to (integer + 0.5) LSB (see the remark above section J.2).

In Fig. J.2(b), we have removed the bias. The dither can be represented with L bits, keeping in mind that each dither sample has an additional 1 at the bit position $0.5\,\mathrm{LSB}_\mathrm{dither}$.[4]

In Fig. J.2(c), the digital dither needs $L + 1$ bits for representation, since it can have $2^L + 1$ different values. The 0.5 LSB problem arises also here similarly to the case of Fig. J.2(a).

All three cases behave similarly.

In Fig J.3 we have illustrated the behavior of the most important characteristics of the quantization noise of the dither of Fig. J.2(b). We can observe that even for a few-bit dither, some of the characteristics of the noise are good enough, but the variance still can have large variations: it changes between $[0, q^2/4]$. The cause of

[4] In this case, for proper quantization we need to round values $x + d = (k + 0.5)q$ upwards. This is even simpler to implement than up or down rounding with probability 0.5-0.5.

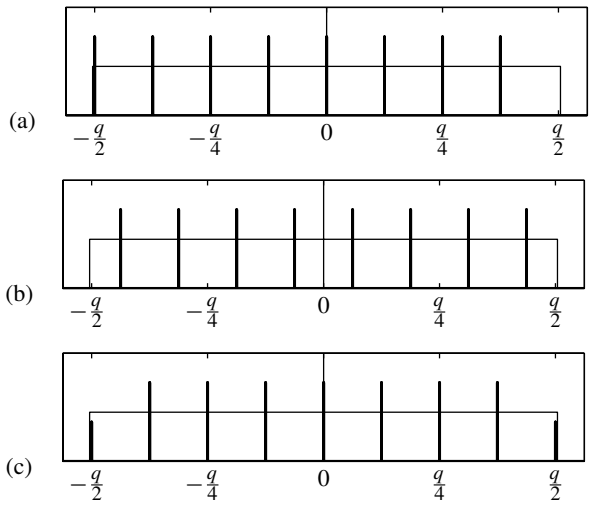

Figure J.2 Discrete uniform dithers with $L = 3$ ($q_\mathrm{d} = q/2^3$): (a) simple (two's complement) binary representation which has mean value $-q_\mathrm{d}/2$; (b) unbiased (shifted) binary representation; (c) unbiased binary representation with half-probability boundary samples (needs $L + 1$ bits for coding all the possible values).

the anomaly is that the dither is only zero-order. The CF of the dither in Fig. J.3(b) is:

$$\begin{aligned}\Phi_\mathrm{d}(u) &= \int_{-\infty}^{\infty} \sum_{n=-2^{L-1}+1}^{2^{L-1}} \frac{q_\mathrm{d}}{q}\,\delta\Big(x-(i-0.5)q_\mathrm{d}\Big)e^{jux}\,\mathrm{d}x \\ &= \frac{q_\mathrm{d}}{q}\,\frac{e^{j\frac{uq}{2}}-e^{-j\frac{uq}{2}}}{e^{j\frac{uq_\mathrm{d}}{2}}-e^{-j\frac{uq_\mathrm{d}}{2}}} \\ &= \frac{q_\mathrm{d}}{q}\,\frac{\sin\left(\dfrac{qu}{2}\right)}{\sin\left(\dfrac{q_\mathrm{d}u}{2}\right)} \\ &= \frac{\sin\left(\dfrac{qu}{2}\right)}{2^L \sin\left(2^{-L}\dfrac{qu}{2}\right)}\,. \end{aligned} \tag{J.7}$$

The values of the characteristic function of the dither are zero at $l \cdot 2\pi/q$, $l = \pm 1, \pm 2, \ldots$ except when $l = k \cdot 2^L$, $k = \pm 1, \pm 2, \ldots$ Therefore, this dither is only *digitally zero-order* dither. The exceptional peaks (see Fig. J.3f) have no influence on the first moment, see Eq. (J.2). However, the derivatives are not zero, allowing for significant correlation values between d and ν.

The characteristic functions of the other two dithers are similar.

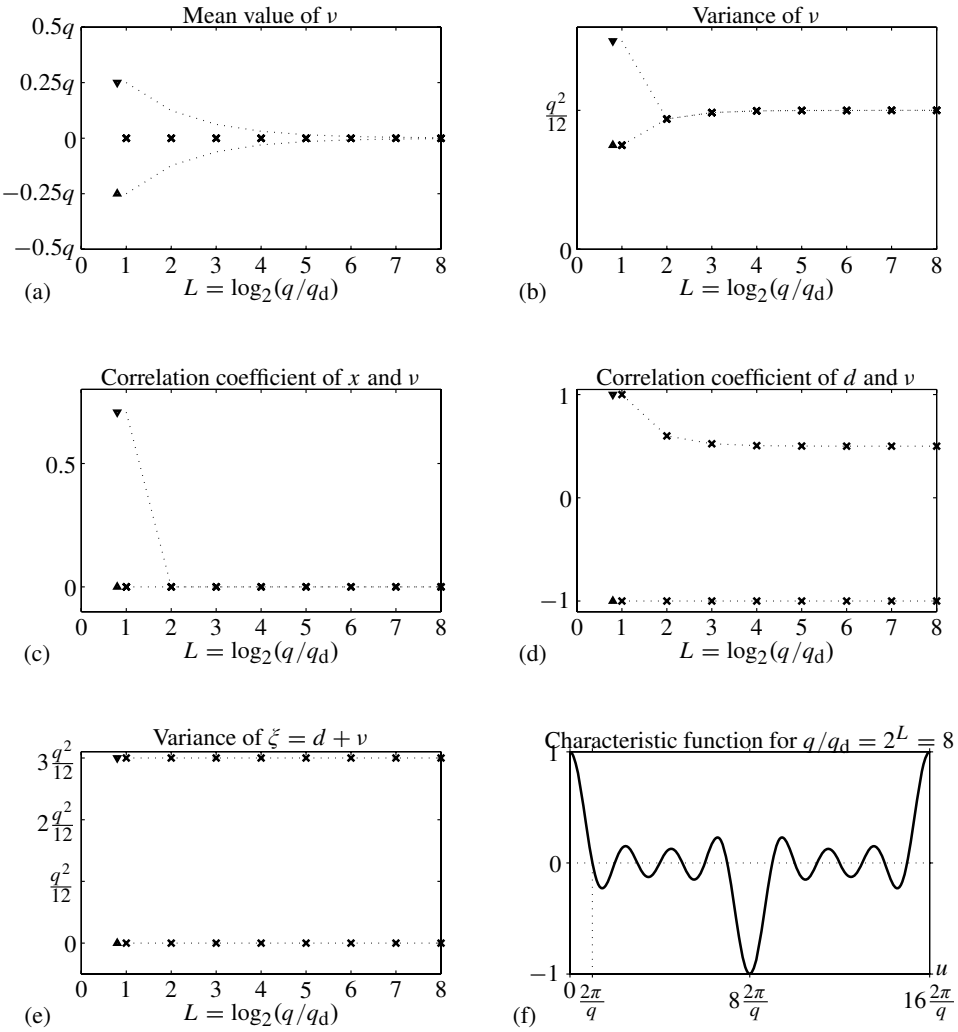

Figure J.3 Quantization noise characteristics for digital uniform dither as in Fig. J.2(b). The **x** marks were calculated for the cases when x is discrete: $k2^{-L}q$, $k = 0, 1, \ldots$; the dotted upper/lower bounds (marked by the triangles) were determined from all (continuous-amplitude) values of x. (a) mean value; (b) variance; (c) correlation coefficient with input x; (d) correlation coefficient with input d; (e) variance of $\xi = d + \nu$; (f) the CF of the dither for $q/q_\mathrm{d} = 2^L = 8$.

J.3.2 Triangularly Distributed Digital Dither

For triangular dither, we have again three almost equivalent forms.

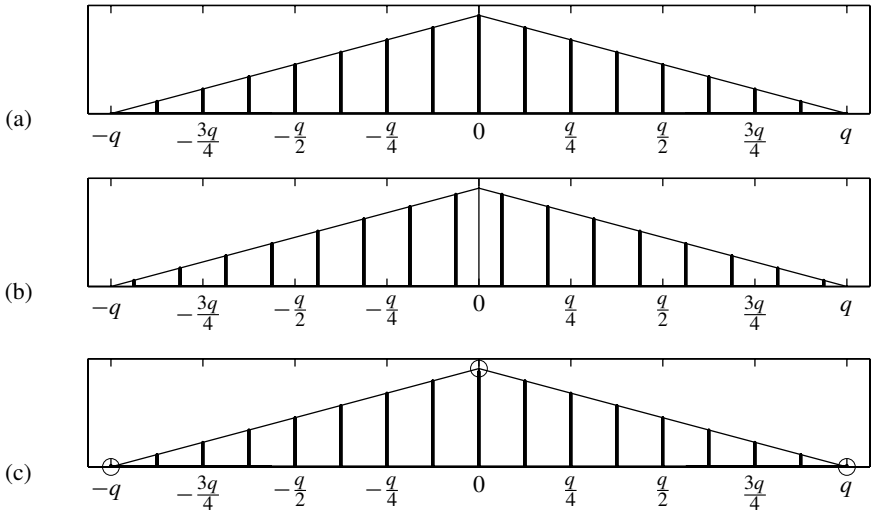

Figure J.4 Discrete triangular dithers with $L = 3$ ($q_\mathrm{d} = q/2^3$): (a) combination of two dithers of Fig. J.2(b), or mean-corrected combination of two dithers of Fig. J.2(a); (b) continuous-amplitude triangular dither, quantized with a mid-riser quantizer to resolution q_d; (c) continuous-amplitude triangular dither, quantized with a mid-tread quantizer to resolution q_d.

In Fig. J.4(a), the digital triangular dither can be obtained by simply adding two dithers of Fig. J.2(b), or by adding two dithers of Fig. J.2(a) and subtracting the bias $-q_\mathrm{d}$. The representation is simple and straightforward. The number of values is $2 \cdot 2^L - 1$, so the necessary number of bits is $L + 1$. The variance is $\mathrm{var}\{d\} = \mathrm{var}\{d_\mathrm{c}\} - 2q_\mathrm{d}^2/12 = 2q^2/12 - 2q_\mathrm{d}^2/12$ (double of variance of the first digital dither).

This digital dither cannot be obtained by direct quantization of the continuous-time triangular one. A possibility to have this is illustrated in Fig. J.4(b). This dither can still be represented with $L + 1$ bits ($2 \cdot 2^L$ different values), keeping in mind that each dither sample has an additional 1 at the bit position 0.5 LSB, like in Fig. J.2(a).

The dither form of Fig. J.4c is the result of mid-tread uniform quantization of the continuous-amplitude dither. Mathematically, the distribution can be obtained by correcting the dither shown in Fig. J.4(a) by subtracting a probability $P_1 = (q_\mathrm{d}/2)/q \cdot 1/q \cdot q_\mathrm{d}/2 = q_\mathrm{d}^2/4q^2$ at the center, and executing similar corrections at the edges. The number of amplitude levels is $2 \cdot 2^L + 1$.

Condition (J.1) is fulfilled for $r = 1$, therefore these dithers are zero-order digital dithers.

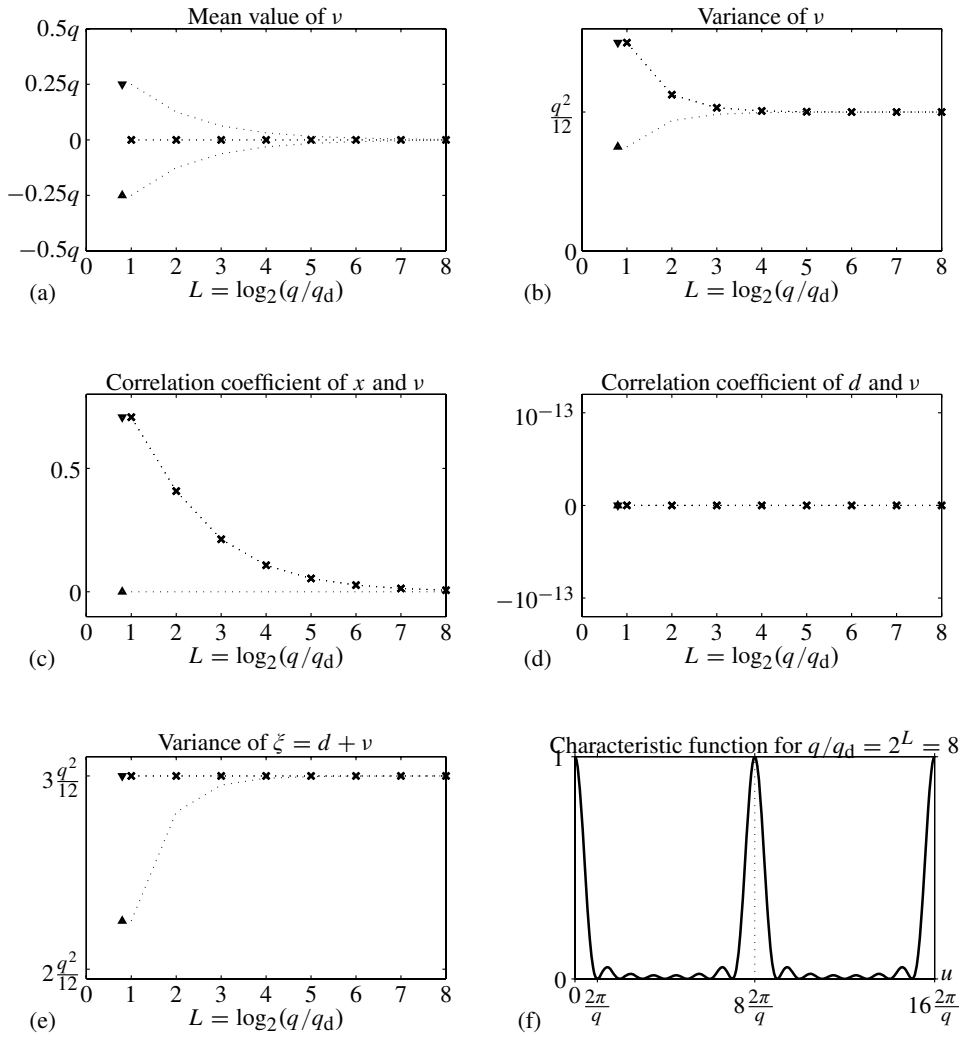

Figure J.5 Quantization noise characteristics for digital triangular dither given in Fig. J.4(a). The **x** marks were calculated for the cases when x is discrete: $k2^{-L}q$, $k = 0, 1, \ldots$; the upper/lower bounds (marked by the triangles) were determined from all (continuous-amplitude) values of x. (a) mean value; (b) variance; (c) correlation coefficient with input x; (d) correlation coefficient with input d; (e) variance of $\xi = d + \nu$; (f) the CF of the dither for $2^L = q/q_d = 8$.

In Fig J.5 we have illustrated the behavior of the most important characteristics of the noise of Fig. J.4(a). The CF of this dither is:

$$\Phi_d(u) = \left(\frac{\sin\left(\frac{qu}{2}\right)}{2^L \sin\left(2^{-L}\frac{qu}{2}\right)} \right)^2 . \tag{J.8}$$

The characteristic functions of the other two dithers are similar but slightly different.

The value and the first derivative of the characteristic function of the dither are equal to zero at the places required in Eq. (J.1), so this dither is a *digital first-order* dither. The peaks shown in the plots have no x-dependent effect on the second moments when the input is digital with LSB $= q_{\mathrm{d}}$, as provided by theorem J.1, therefore they can be corrected for, using knowledge of the dither. This is the dither which can be recommended for digital systems.