# Chapter 12

# *Basics of Floating-Point Quantization*

Representation of physical quantities in terms of floating-point numbers allows one to cover a very wide dynamic range with a relatively small number of digits. Given this type of representation, roundoff errors are roughly proportional to the amplitude of the represented quantity. In contrast, roundoff errors with uniform quantization are bounded between $\pm q/2$ and are not in any way proportional to the represented quantity.

Floating-point is in most cases so advantageous over fixed-point number representation that it is rapidly becoming ubiquitous. The movement toward usage of floating-point numbers is accelerating as the speed of floating-point calculation is increasing and the cost of implementation is going down. For this reason, it is essential to have a method of analysis for floating-point quantization and floating-point arithmetic.

## 12.1 THE FLOATING-POINT QUANTIZER

Binary numbers have become accepted as the basis for all digital computation. We therefore describe floating-point representation in terms of binary numbers. Other number bases are completely possible, such as base 10 or base 16, but modern digital hardware is built on the binary base.

The numbers in the table of Fig. 12.1 are chosen to provide a simple example. We begin by counting with nonnegative binary floating-point numbers as illustrated in Fig. 12.1. The counting starts with the number 0, represented here by 00000. Each number is multiplied by $2^E$, where $E$ is an exponent. Initially, let $E = 0$. Continuing the count, the next number is 1, represented by 00001, and so forth. The counting continues with increments of 1, past 16 represented by 10000, and goes on until the count stops with the number 31, represented by 11111. The numbers are now reset back to 10000, and the exponent is incremented by 1. The next number will be 32, represented by $10000 \times 2^1$. The next number will be 34, given by $10001 \times 2^1$.

| | | Mantissa | | | | |
|---|---|---|---|---|---|---|
| 0 | | 0 | 0 | 0 | 0 | 0 |
| 1 | | 0 | 0 | 0 | 0 | 1 |
| 2 | | 0 | 0 | 0 | 1 | 0 |
| 3 | | 0 | 0 | 0 | 1 | 1 |
| 4 | | 0 | 0 | 1 | 0 | 0 |
| 5 | | 0 | 0 | 1 | 0 | 1 |
| 6 | | 0 | 0 | 1 | 1 | 0 |
| 7 | | 0 | 0 | 1 | 1 | 1 |
| 8 | | 0 | 1 | 0 | 0 | 0 |
| 9 | | 0 | 1 | 0 | 0 | 1 |
| 10 | | 0 | 1 | 0 | 1 | 0 |
| 11 | | 0 | 1 | 0 | 1 | 1 |
| 12 | | 0 | 1 | 1 | 0 | 0 |
| 13 | | 0 | 1 | 1 | 0 | 1 |
| 14 | | 0 | 1 | 1 | 1 | 0 |
| 15 | | 0 | 1 | 1 | 1 | 1 |
| 16 | $\longrightarrow$ | 1 | 0 | 0 | 0 | 0 |
| 17 | | 1 | 0 | 0 | 0 | 1 |
| 18 | | 1 | 0 | 0 | 1 | 0 |
| 19 | | 1 | 0 | 0 | 1 | 1 |
| 20 | | 1 | 0 | 1 | 0 | 0 |
| 21 | | 1 | 0 | 1 | 0 | 1 |
| 22 | | 1 | 0 | 1 | 1 | 0 |
| 23 | | 1 | 0 | 1 | 1 | 1 |
| 24 | | 1 | 1 | 0 | 0 | 0 |
| 25 | | 1 | 1 | 0 | 0 | 1 |
| 26 | | 1 | 1 | 0 | 1 | 0 |
| 27 | | 1 | 1 | 0 | 1 | 1 |
| 28 | | 1 | 1 | 1 | 0 | 0 |
| 29 | | 1 | 1 | 1 | 0 | 1 |
| 30 | | 1 | 1 | 1 | 1 | 0 |
| 31 | $\longleftarrow$ | 1 | 1 | 1 | 1 | 1 |

$\times 2^E$

**Figure 12.1** Counting with binary floating-point numbers with 5-bit mantissa. No sign bit is included here.


**Figure 12.2** A floating-point quantizer.

This will continue with increments of 2 until the number 62 is reached, represented by $11111 \times 2^1$. The numbers are again re-set back to 10000, and the exponent is incremented again. The next number will be 64, given by $10000 \times 2^2$. And so forth. By counting, we have defined the allowed numbers on the number scale. Each number consists of a mantissa (see page 343) multiplied by 2 raised to a power given by the exponent $E$.

The counting process illustrated in Fig. 12.1 is done with binary numbers having 5-bit mantissas. The counting begins with binary 00000 with $E = 0$ and goes up to 11111, then the numbers are recycled back to 10000, and with $E = 1$, the counting resumes up to 11111, then the numbers are recycled back to 10000 again with $E = 2$, and the counting proceeds.

A floating-point quantizer is represented in Fig. 12.2.[1] The input to this quantizer is $x$, a variable that is generally continuous in amplitude. The output of this quantizer is $x'$, a variable that is discrete in amplitude and that can only take on values in accord with a floating-point number scale. The input–output relation for this quantizer is a staircase function that does not have uniform steps.

Figure 12.3 illustrates the input–output relation for a floating-point quantizer with a 3-bit mantissa. The input physical quantity is $x$. Its floating-point representation is $x'$. The smallest step size is $q$. With a 3-bit mantissa, four steps are taken for each cycle, except for eight steps taken for the first cycle starting at the origin. The spacings of the cycles are determined by the choice of a parameter $\Delta$. A general relation between $\Delta$ and $q$, defining $\Delta$, is given by Eq. (12.1):

$$\Delta \stackrel{\triangle}{=} 2^p q \,, \tag{12.1}$$

where $p$ is the number of bits of the mantissa. With a 3-bit mantissa, $\Delta = 8q$. Figure 12.3 is helpful in gaining an understanding of relation (12.1). Note that after the first cycle, the spacing of the cycles and the step sizes vary by a factor of 2 from cycle to cycle.

[1] The basic ideas and figures of the next sections were first published in, and are taken with permission from Widrow, B., Kollár, I. and Liu, M.-C., "Statistical theory of quantization," *IEEE Transactions on Instrumentation and Measurement* 45(6): 35361. 

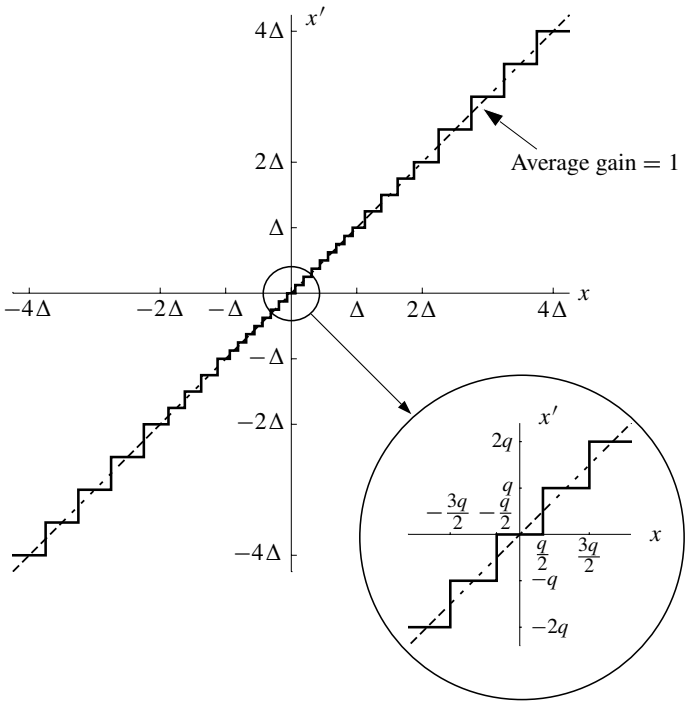


**Figure 12.3** Input–output staircase function for a floating-point quantizer with a 3-bit mantissa.

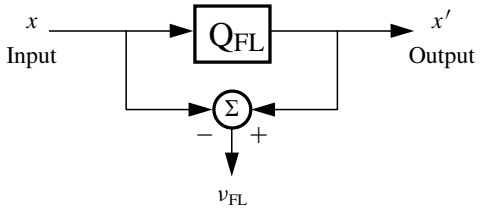


**Figure 12.4** Floating-point quantization noise.

## 12.2 FLOATING-POINT QUANTIZATION NOISE

The roundoff noise of the floating-point quantizer $\nu_{\mathrm{FL}}$ is the difference between the quantizer output and input:

$$\nu_{\mathrm{FL}} = x' - x \,. \tag{12.2}$$

Figure 12.4 illustrates the relationships between $x$, $x'$, and $\nu_{\mathrm{FL}}$.

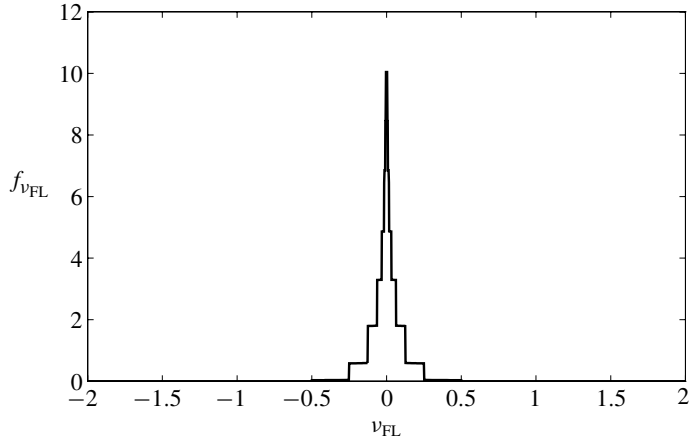


**Figure 12.5** The PDF of floating-point quantization noise with a zero-mean Gaussian input, $\sigma_x = 32\Delta$, and with a 2-bit mantissa.

The PDF of the quantization noise can be obtained by slicing and stacking the PDF of $x$, as was done for the uniform quantizer and illustrated in Fig. 5.2. Because the staircase steps are not uniform, the PDF of the quantization noise is not uniform. It has a pyramidal shape.

With a zero-mean Gaussian input with $\sigma_x = 32\Delta$, and with a mantissa having just two bits, the quantization noise PDF has been calculated. It is plotted in Fig. 12.5. The shape of this PDF is typical. The narrow segments near the top of the pyramid are caused by the occurrence of values of $x$ that are small in magnitude (small quantization step sizes), while the wide segments near the bottom of the pyramid are caused by the occurrence of values of $x$ that are large in magnitude (large quantization step sizes).

The shape of the PDF of floating-point quantization noise resembles the silhouette of a big-city "skyscraper" like the famous Empire State Building of New York City. We have called functions like that of Fig. 12.5 "skyscraper PDFs". Mathematical methods will be developed next to analyze floating-point quantization noise, to find its mean and variance, and the correlation coefficient between this noise and the input $x$.

## 12.3 AN EXACT MODEL OF THE FLOATING-POINT QUANTIZER

A floating-point quantizer of the type shown in Fig. 12.3 can be modeled exactly as a cascade of a nonlinear function (a "compressor") followed by a uniform quantizer (a "hidden" quantizer) followed by an inverse nonlinear function (an "expander"). The

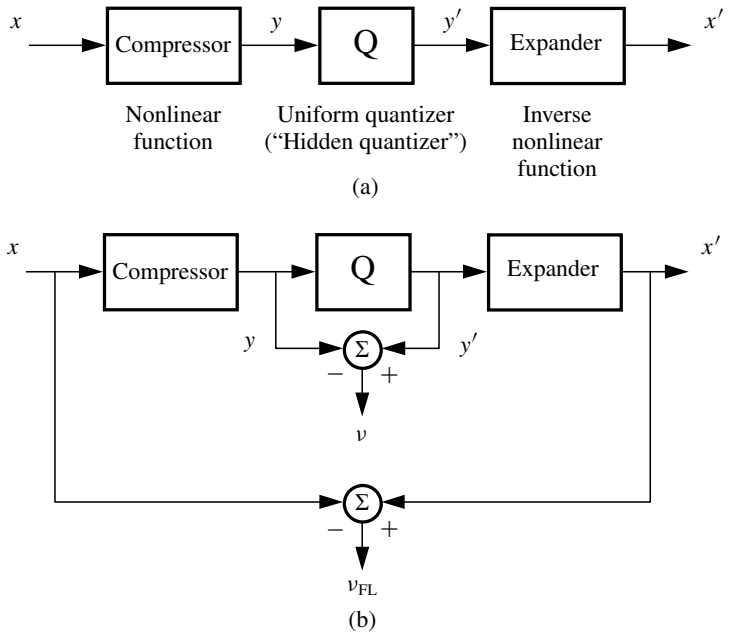


**Figure 12.6** A model of a floating-point quantizer: (a) block diagram; (b) definition of quantization noises.

overall idea is illustrated in Fig. 12.6(a). A similar idea is often used to represent compression and expansion in a data compression system (see e.g. Gersho and Gray (1992), CCITT (1984)).

The input–output characteristic of the compressor ($y$ vs. $x$) is shown in Fig. 12.7, and the input–output characteristic of the expander ($x'$ vs. $y'$) is shown in Fig. 12.8. The hidden quantizer is conventional, having a uniform-staircase input–output characteristic ($y'$ vs. $y$). Its quantization step size is $q$, and its quantization noise is

$$v = y' - y\,. \tag{12.3}$$

Figure 12.6(b) is a diagram showing the sources of $v_{\mathrm{FL}}$ and $v$. An expression for the input–output characteristic of the compressor is

$$
\begin{aligned}
y &= x, && \text{if } -\Delta \le x \le \Delta \\
y &= \tfrac{1}{2}x + \tfrac{1}{2}\Delta, && \text{if } \Delta \le x \le 2\Delta \\
y &= \tfrac{1}{2}x - \tfrac{1}{2}\Delta, && \text{if } -2\Delta \le x \le -\Delta \\
y &= \tfrac{1}{4}x + \Delta, && \text{if } 2\Delta \le x \le 4\Delta \\
y &= \tfrac{1}{4}x - \Delta, && \text{if } -4\Delta \le x \le -2\Delta \\
&\vdots \\
y &= \tfrac{1}{2^k}x + \tfrac{k}{2}\Delta, && \text{if } 2^{k-1}\Delta \le x \le 2^k\Delta \\
y &= \tfrac{1}{2^k}x - \tfrac{k}{2}\Delta, && \text{if } -2^k\Delta \le x \le -2^{k-1}\Delta \\
&\text{or} \\
y &= \tfrac{1}{2^k}x + \operatorname{sign}(x)\cdot\tfrac{k}{2}\Delta, && \text{if } 2^{k-1}\Delta \le |x| \le 2^k\Delta\,,
\end{aligned}
\tag{12.4}
$$

where $k$ is a nonnegative integer. This is a piecewise-linear characteristic.

An expression for the input–output characteristic of the expander is

$$
\begin{aligned}
x' &= y', && \text{if } -\Delta \le y' \le \Delta \\
x' &= 2(y' - 0.5\Delta), && \text{if } \Delta \le y' \le 1.5\Delta \\
x' &= 2(y' - 0.5\Delta), && \text{if } -1.5\Delta \le y' \le -\Delta \\
x' &= 4(y' - \Delta), && \text{if } 1.5\Delta \le y' \le 2\Delta \\
x' &= 4(y' + \Delta), && \text{if } -2\Delta \le y' \le -1.5\Delta \\
&\vdots \\
x' &= 2^k(y' - \tfrac{k}{2}\Delta), && \text{if } \tfrac{k+1}{2}\Delta \le y' \le \tfrac{k+2}{2}\Delta \\
x' &= 2^k(y' + \tfrac{k}{2}\Delta), && \text{if } -\tfrac{k+2}{2}\Delta \le y' \le -\tfrac{k+1}{2}\Delta \\
&\text{or} \\
x' &= 2^k(y' - \operatorname{sign}(y)\cdot\tfrac{k}{2}\Delta), && \text{if } \tfrac{k+1}{2}\Delta \le |y'| \le \tfrac{k+2}{2}\Delta\,,
\end{aligned}
\tag{12.5}
$$

where $k$ is a nonnegative integer. This characteristic is also piecewise linear.

One should realize that the compressor and expander characteristics described above are universal and are applicable for any choice of mantissa size.

The compressor and expander characteristics of Figs. 12.7 and 12.8 are drawn to scale. Fig. 12.9 shows the characteristic ($y'$ vs. $y$) of the hidden quantizer drawn to the same scale, assuming a mantissa of 2 bits. For this case, $\Delta = 4q$.

If only the compressor and expander were cascaded, the result would be a perfect gain of unity since they are inverses of each other. If the compressor is cascaded with the hidden quantizer of Fig. 12.9 and then cascaded with the expander, all in accord with the diagram of Fig. 12.6, the result is a floating-point quantizer. The one illustrated in Fig. 12.10 has a 2-bit mantissa.

The cascaded model of the floating-point quantizer shown in Fig. 12.6 becomes very useful when the quantization noise of the hidden quantizer has the properties of

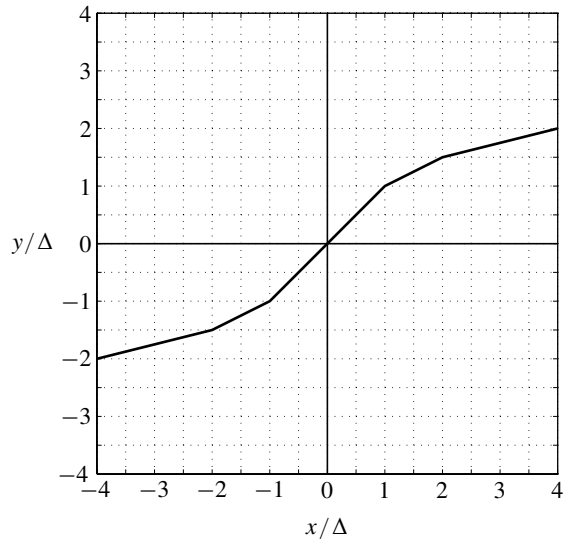


**Figure 12.7** The compressor's input–output characteristic.

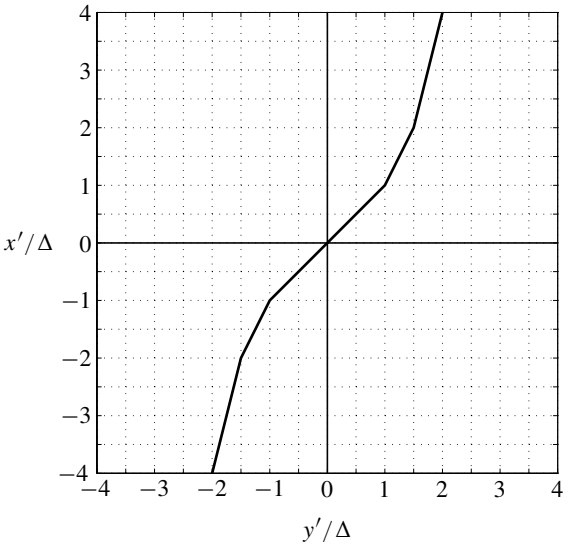


**Figure 12.8** The expander's input–output characteristic.

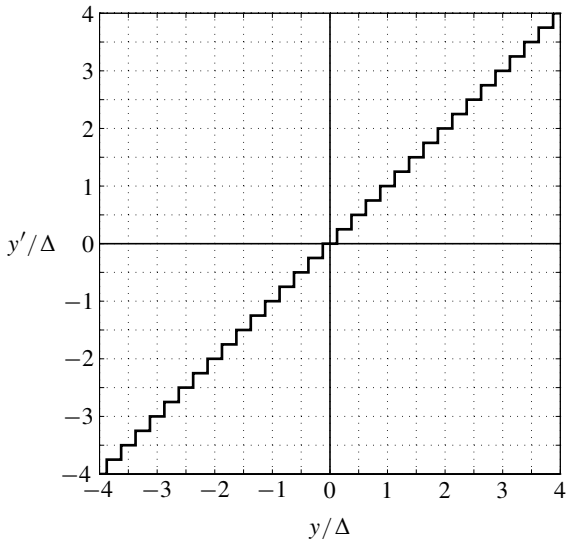


**Figure 12.9** The uniform "hidden quantizer." The mantissa has 2 bits.

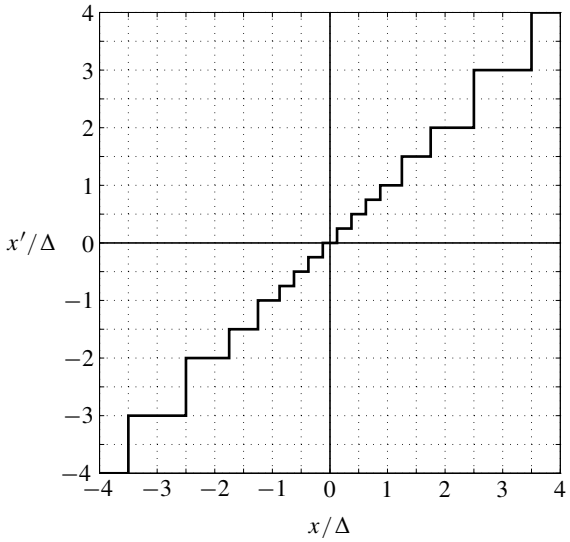


**Figure 12.10** A floating-point quantizer with a 2-bit mantissa.

PQN. This would happen if QT I or QT II were satisfied at the input $y$ of the hidden quantizer. Testing for the satisfaction of these quantizing theorems is complicated by the nonlinearity of the compressor. If $x$ were a Gaussian input to the floating-point quantizer, the input to the hidden quantizer would be $x$ mapped through the compressor. The result would be a non-Gaussian input to the hidden quantizer.

In practice, inputs to the hidden quantizer almost never perfectly meet the conditions for satisfaction of a quantizing theorem. On the other hand, these inputs almost always satisfy these conditions approximately. The quantization noise $\nu$ introduced by the hidden quantizer is generally very close to uniform and very close to being uncorrelated with its input $y$. The PDF of $y$ is usually sliced up so finely that the hidden quantizer produces noise having properties like PQN.

## 12.4 HOW GOOD IS THE PQN MODEL FOR THE HIDDEN QUANTIZER?

In previous chapters where uniform quantization was studied, it was possible to define properties of the input CF that would be necessary and sufficient for the satisfaction of a quantizing theorem, such as QT II. Even if these conditions were not obtained, as for example with a Gaussian input, it was possible to determine the errors that would exist in moment prediction when using the PQN model. These errors would almost always be very small.

For floating-point quantization, the same kind of calculations for the hidden quantizer could in principle be made, but the mathematics would be far more difficult because of the action of the compressor on the input signal $x$. The distortion of the compressor almost never simplifies the CF of the input $x$ nor makes it easy to test for the satisfaction of a quantizing theorem.

To determine the statistical properties of $\nu$, the PDF of the input $x$ can be mapped through the piecewise-linear compressor characteristic to obtain the PDF of $y$. In turn, $y$ is quantized by the hidden quantizer, and the quantization noise $\nu$ can be tested for similarity to PQN. The PDF of $\nu$ can be determined directly from the PDF of $y$, or it could be obtained by Monte Carlo methods by applying random $x$ inputs into the compressor and observing corresponding values of $y$ and $\nu$. The moments of $\nu$ can be determined, and so can the covariance between $\nu$ and $y$. For the PQN model to be usable, the covariance of $\nu$ and $y$ should be close to zero, less than a few percent, the PDF of $\nu$ should be almost uniform between $\pm q/2$, and the mean of $\nu$ should be close to zero while the mean square of $\nu$ should be close to $q^2/12$.

In many cases with $x$ being Gaussian, the PDF of the compressor output $y$ can be very "ugly." Examples are shown in Figs. 12.11(a)–12.14(a). These represent cases where the input $x$ is Gaussian with various mean values and various ratios of $\sigma_x/\Delta$. With mantissas of 4 bits or more, these ugly inputs to the hidden quantizer cause it to produce quantization noise which behaves remarkably like PQN. This is difficult to prove analytically, but confirming results from slicing and stacking

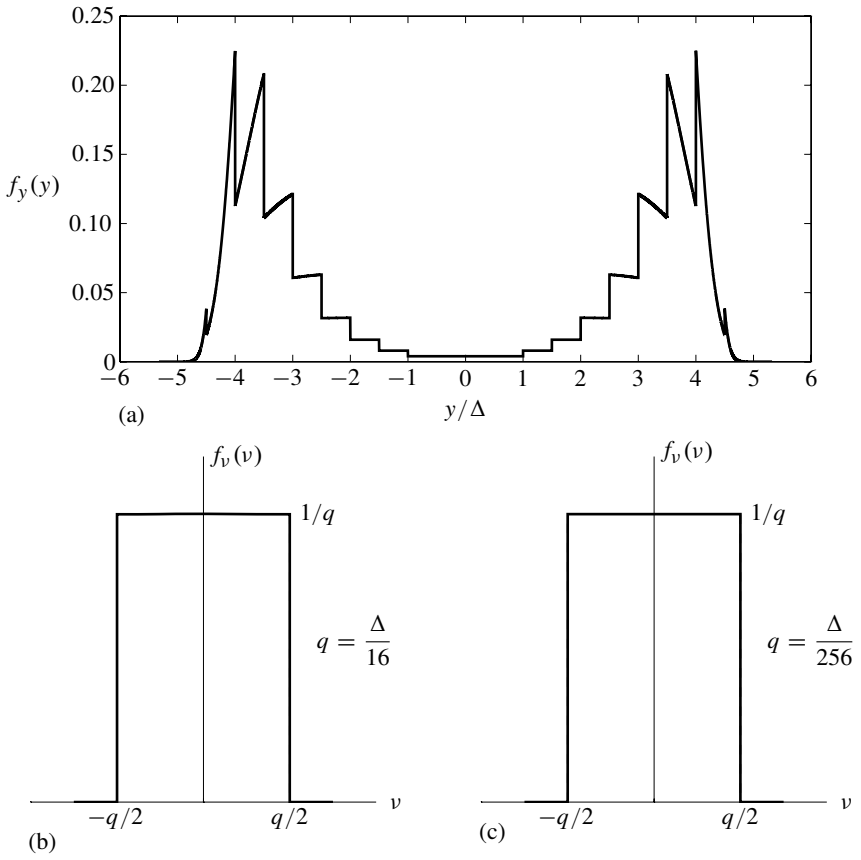


**Figure 12.11** PDF of compressor output and of hidden quantization noise when $x$ is zero-mean Gaussian with $\sigma_x = 50\Delta$: (a) $f_y(y)$; (b) $f_v(v)$ for $p = 4$ ($q = \Delta/16$); (c) $f_v(v)$ for $p = 8$ ($q = \Delta/256$).

these PDFs are very convincing. Similar techniques have been used with these PDFs to calculate correlation coefficients between $v$ and $y$. The noise $v$ turns out to be essentially uniform, and the correlation coefficient turns out to be essentially zero.

Consider the ugly PDF of $y$ shown in Fig. 12.11(a), $f_y(y)$. The horizontal scale goes over a range of $\pm 5\Delta$. If we choose a mantissa of 4 bits, the horizontal scale will cover the range $\pm 80q$. If we choose a mantissa of 8 bits, $q$ will be much smaller and the horizontal scale will cover the range $\pm 1280q$. With the 4-bit mantissa, slicing and stacking this PDF to obtain the PDF of $v$, we obtain the result shown in Fig. 12.11(b). From the PDF of $v$, we obtain a mean value of $1.048 \cdot 10^{-6}q$, and a mean square of $0.9996q^2/12$. The correlation coefficient between $v$ and $y$ is $(1.02 \cdot 10^{-3})$. With an 8-bit mantissa, $f_v(v)$ shown in Fig. 12.11c is even more uniform, giving a mean value for $v$ of $-9.448 \cdot 10^{-9}q$, and a mean square of $1.0001q^2/12$. The correlation

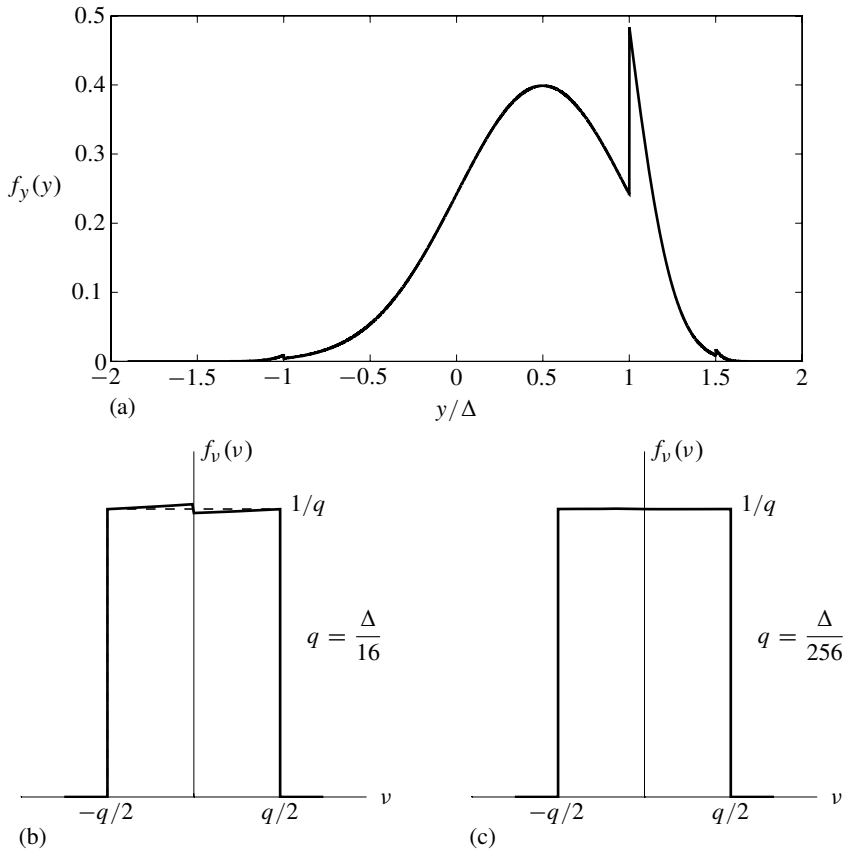


**Figure 12.12** PDF of compressor output and of hidden quantization noise when $x$ is Gaussian with $\sigma_x = \Delta/2$ and $\mu_x = \sigma_x$: (a) $f_y(y)$; (b) $f_\nu(\nu)$ for 4-bit mantissas, $(q = \Delta/16)$; (c) $f_\nu(\nu)$ for 8-bit mantissas, $(q = \Delta/256)$.

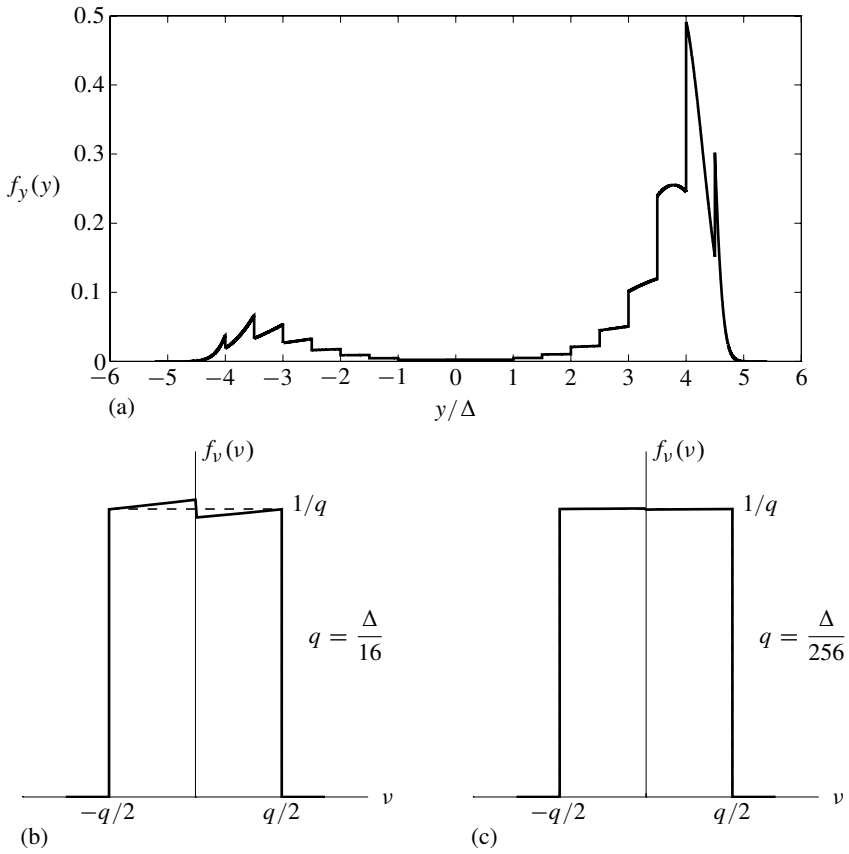


**Figure 12.13** PDF of compressor output and of hidden quantization noise when $x$ is Gaussian with $\sigma_x = 50\Delta$ and $\mu_x = \sigma_x$: (a) $f_y(y)$; (b) $f_v(v)$ for 4-bit mantissas, $(q = \Delta/16)$; (c) $f_v(v)$ for 8-bit mantissas, $(q = \Delta/256)$.

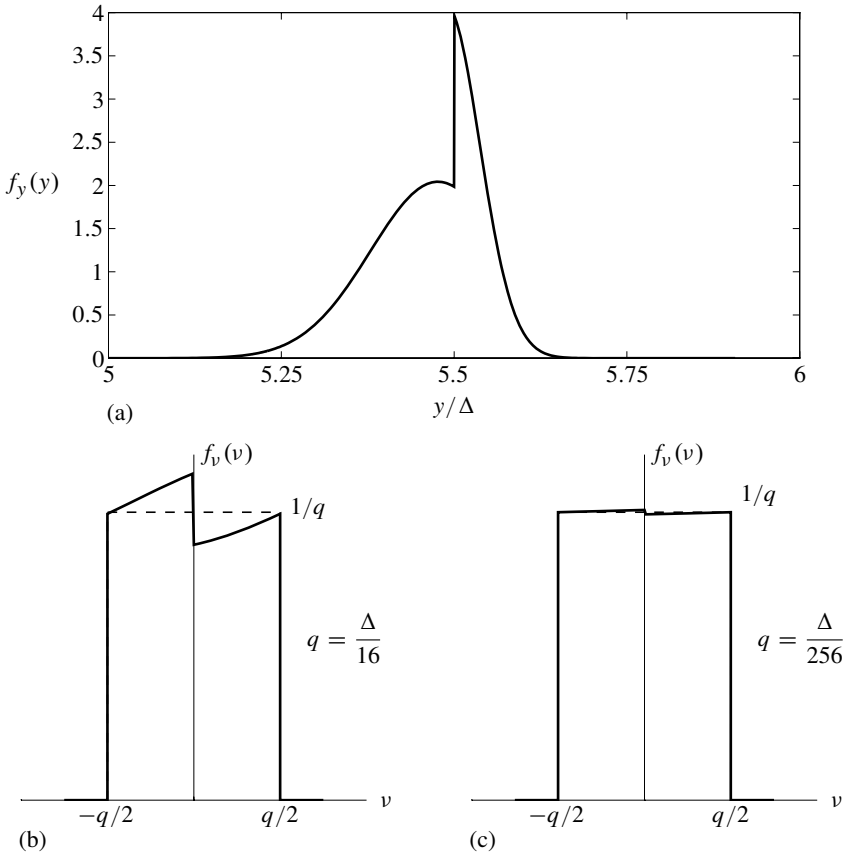


**Figure 12.14** PDF of compressor output and of hidden quantization noise when $x$ is Gaussian with $\sigma_x = 50\Delta$ and $\mu_x = 10\sigma_x = 500\Delta$: (a) $f_y(y)$; (b) $f_\nu(\nu)$ for 4-bit mantissas, $(q = \Delta/16)$; (c) $f_\nu(\nu)$ for 8-bit mantissas, $(q = \Delta/256)$.

**TABLE 12.1** Properties of the hidden quantization noise $\nu$ for Gaussian input

| (a) $p = 4$ | $\mathrm{E}\{\nu\}/q$ | $\mathrm{E}\{\nu^2\}/(q^2/12)$ | $\rho_{y,\nu}$ |
|---|---|---|---|
| $\mu_x = 0,\ \sigma_x = 50\Delta$ | 0.0002 | 0.9997 | 0.0140 |
| $\mu_x = \sigma_x,\ \sigma_x = 0.5\Delta$ | −0.0015 | 0.9998 | −0.0069 |
| $\mu_x = \sigma_x,\ \sigma_x = 50\Delta$ | −0.0019 | 0.9994 | −0.0089 |
| $\mu_x = \sigma_x,\ \sigma_x = 500\Delta$ | −0.0023 | 0.9987 | −0.0077 |

| (b) $p = 8$ | $\mathrm{E}\{\nu\}/q$ | $\mathrm{E}\{\nu^2\}/(q^2/12)$ | $\rho_{y,\nu}$ |
|---|---|---|---|
| $\mu_x = 0,\ \sigma_x = 50\Delta$ | $4.25 \cdot 10^{-5}$ | 0.9995 | $8.68 \cdot 10^{-4}$ |
| $\mu_x = \sigma_x,\ \sigma_x = 0.5\Delta$ | $-2.24 \cdot 10^{-4}$ | 1.0005 | $-2.35 \cdot 10^{-4}$ |
| $\mu_x = \sigma_x,\ \sigma_x = 50\Delta$ | $-3.72 \cdot 10^{-4}$ | 0.9997 | $-1.30 \cdot 10^{-3}$ |
| $\mu_x = \sigma_x,\ \sigma_x = 500\Delta$ | $-1.92 \cdot 10^{-4}$ | 0.9985 | $-8.77 \cdot 10^{-4}$ |

coefficient between $\nu$ and $y$ is $(2.8 \cdot 10^{-4})$. With a mantissa of four bits or more, the noise $\nu$ behaves very much like PQN.

This process was repeated for the ugly PDFs of Figs. 12.12(a), 12.13(a), and 12.14(a). The corresponding PDFs of the quantization noise of the hidden quantizer are shown in Figs. 12.12(b), 12.13(b), and 12.14(b). With a 4-bit mantissa, Table 12.1(a) lists the values of mean and mean square of $\nu$, and correlation coefficient between $\nu$ and $y$ for all the cases. With an 8-bit mantissa, the corresponding moments are listed in Table 12.1(b).

Similar tests have been made with uniformly distributed inputs, with triangularly distributed inputs, and with sinusoidal inputs. Input means ranged from zero to one standard deviation. The moments of Tables 12.1(a),(b) are typical for Gaussian inputs, but similar results are obtained with other forms of inputs.

For every single test case, the input $x$ had a PDF that extended over at least several multiples of $\Delta$. With a mantissa of 8 bits or more, the noise of the hidden quantizer had a mean of almost zero, a mean square of almost $q^2/12$, and a correlation coefficient with the quantizer input of almost zero.

The PQN model works so well that we assume that it is true as long as the mantissa has 8 or more bits and the PDF of $x$ covers at least several $\Delta$-quanta of the floating-point quantizer. It should be noted that the single precision IEEE standard (see Section 13.8) calls for a mantissa with 24 bits. When working with this standard, the PQN model will work exceedingly well almost everywhere.

## 12.5 ANALYSIS OF FLOATING-POINT QUANTIZATION NOISE

We will use the model of the floating-point quantizer shown in Fig. 12.6 to determine the statistical properties of $\nu_{\mathrm{FL}}$. We will derive its mean, its mean square, and its correlation coefficient with input $x$.

The hidden quantizer injects noise into $y$, resulting in $y'$, which propagates through the expander to make $x'$. Thus, the noise of the hidden quantizer propagates through the nonlinear expander into $x'$. The noise in $x'$ is $\nu_{\mathrm{FL}}$.

Assume that the noise $\nu$ of the hidden quantizer is small compared to $y$. Then $\nu$ is a small increment to $y$. This causes an increment $\nu_{\mathrm{FL}}$ at the expander output. Accordingly,

$$\nu_{\mathrm{FL}} = \nu \left( \frac{\mathrm{d}x'}{\mathrm{d}y'} \right)_y . \tag{12.6}$$

The derivative is a function of $y$. This is because the increment $\nu$ is added to $y$, so $y$ is the nominal point where the derivative should be taken.

Since the PQN model was found to work so well for so many cases with regard to the behavior of the hidden quantizer, we will make the assumption that the conditions for PQN are indeed satisfied for the hidden quantizer. This greatly simplifies the statistical analysis of $\nu_{\mathrm{FL}}$.

Expression (12.6) can be used in the following way to find the crosscorrelation between $\nu_{\mathrm{FL}}$ and input $x$.

$$\begin{aligned} \mathrm{E}\{\nu_{\mathrm{FL}} x\} &= \mathrm{E}\left\{ \nu \cdot \left( \frac{\mathrm{d}x'}{\mathrm{d}y'} \right)_y \cdot x \right\} \\ &= \mathrm{E}\{\nu\}\, \mathrm{E}\left\{ \left( \frac{\mathrm{d}x'}{\mathrm{d}y'} \right)_y \cdot x \right\} \\ &= 0 . \end{aligned} \tag{12.7}$$

When deriving this result, it was possible to factor the expectation into a product of expectations because, from the point of view of moments, $\nu$ behaves as if it is independent of $y$ and independent of any function of $y$, such as $\left( \frac{\mathrm{d}x'}{\mathrm{d}y'} \right)_y$ and $x$. Since the expected value of $\nu$ is zero, the crosscorrelation turns out to be zero.

Expression (12.6) can also be used to find the mean of $\nu_{\mathrm{FL}}$. Accordingly,

$$\begin{aligned} \mathrm{E}\{\nu_{\mathrm{FL}}\} &= \mathrm{E}\left\{ \nu \cdot \left( \frac{\mathrm{d}x'}{\mathrm{d}y'} \right)_y \right\} \\ &= \mathrm{E}\{\nu\} \mathrm{E}\left\{ \left( \frac{\mathrm{d}x'}{\mathrm{d}y'} \right)_y \right\} \\ &= 0 . \end{aligned} \tag{12.8}$$

So the mean of $\nu_{\text{FL}}$ is zero, and the crosscorrelation between $\nu_{\text{FL}}$ and $x$ is zero. Both results are consequences of the hidden quantizer behaving in accord with the PQN model.

Our next objective is the determination of $\text{E}\{\nu_{\text{FL}}^2\}$. We will need to find an expression for $\left(\frac{\text{d}x'}{\text{d}y'}\right)_y$ in order to obtain quantitative results. By inspection of Figs. 12.7 and 12.8, which show the characteristics of the compressor and the expander, we determine that

$$\begin{aligned}
\frac{\text{d}x'}{\text{d}y'} &= 1 \quad \text{when } 0.5\Delta < x < \Delta \\
\frac{\text{d}x'}{\text{d}y'} &= 2 \quad \text{when } \Delta < x < 2\Delta \\
\frac{\text{d}x'}{\text{d}y'} &= 4 \quad \text{when } 2\Delta < x < 4\Delta \\
&\vdots \\
\frac{\text{d}x'}{\text{d}y'} &= 1 \quad \text{when } -\Delta < x < -0.5\Delta \\
\frac{\text{d}x'}{\text{d}y'} &= 2 \quad \text{when } -2\Delta < x < -\Delta \\
\frac{\text{d}x'}{\text{d}y'} &= 4 \quad \text{when } -4\Delta < x < -2\Delta \\
&\vdots \\
\frac{\text{d}x'}{\text{d}y'} &= 2^k \quad \text{when } 2^{k-1}\Delta < |x| < 2^k\Delta
\end{aligned} \tag{12.9}$$

where $k$ is a nonnegative integer.

These relations define the derivative as a function of $x$, except in the range $-\Delta/2 < x < \Delta/2$. The probability of $x$ values in that range are assumed to be negligible. Having the derivative as a function of $x$ is equivalent to having the derivative as a function of $y$ because $y$ is a monotonic function of $x$ (see Fig. 12.7).

It is useful now to introduce the function $\log_2(x/\Delta) + 0.5$. This is plotted in Fig. 12.15(a). We next introduce a new notation for uniform quantization,

$$x' = Q_q(x) \ . \tag{12.10}$$

The operator $Q_q$ represents uniform quantization with a quantum step size of $q$. Using this notation, we can introduce yet another function of $x$, shown in Fig. 12.15(b), by adding the quantity 0.5 to $\log_2(x/\Delta)$ and uniformly quantizing the result with a unit quantum step size. The function is, accordingly, $Q_1\big(\log_2(x/\Delta) + 0.5\big)$.

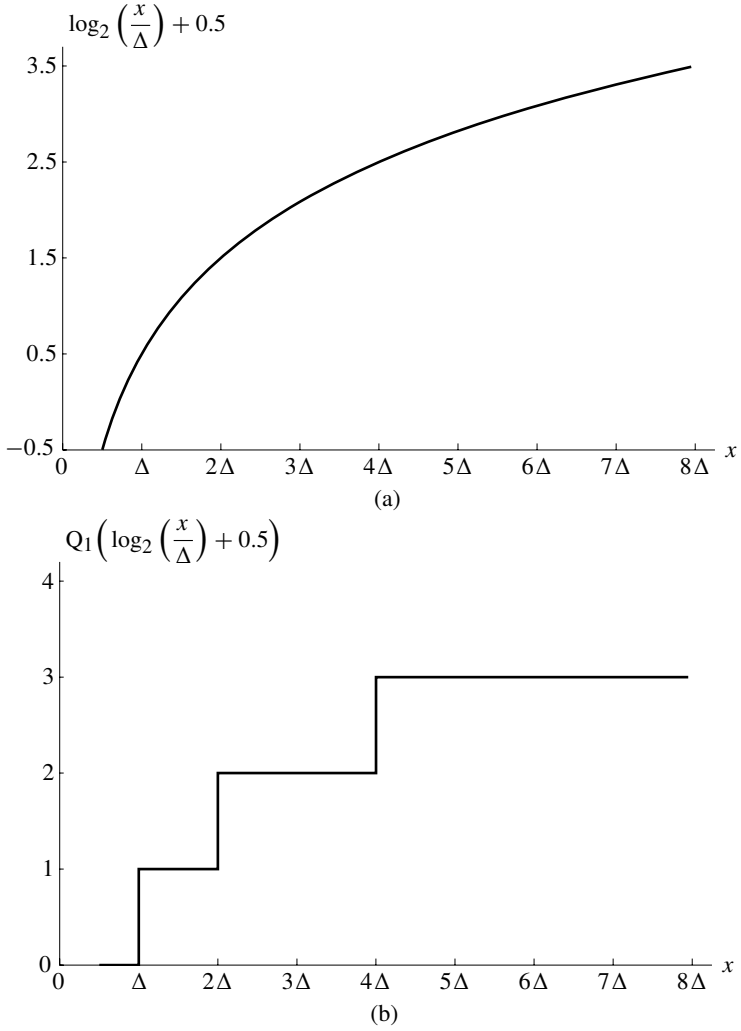


**Figure 12.15** Approximate and exact exponent characteristics: (a) logarithmic approximation; (b) exact expression, $Q_1\Big(\log_2(x/\Delta)+0.5\Big)$ vs. $x$.

Referring back to (12.9), it is apparent that for positive values of $x$, the derivative can be represented in terms of the new function as

$$\left(\frac{\mathrm{d}x'}{\mathrm{d}y'}\right)_x = 2^{Q_1\left(\left(\log_2 \frac{x}{\Delta}\right)+0.5\right)}, \quad x > \Delta/2\,. \tag{12.11}$$

For negative values of $x$, the derivative is

$$\left(\frac{\mathrm{d}x'}{\mathrm{d}y'}\right)_x = 2^{Q_1\left(\left(\log_2 \frac{-x}{\Delta}\right)+0.5\right)}, \quad x < -\Delta/2\,. \tag{12.12}$$

Another way to write this is

$$\left(\frac{\mathrm{d}x'}{\mathrm{d}y'}\right)_x = 2^{Q_1\left(\left(\log_2 \frac{|x|}{\Delta}\right)+0.5\right)}, \quad |x| > \Delta/2\,. \tag{12.13}$$

These are exact expressions for the derivative.

From Eqs. (12.13) and Eq. (12.6), we obtain $\nu_{\mathrm{FL}}$ as

$$\nu_{\mathrm{FL}} = \nu \cdot 2^{Q_1\left(\left(\log_2 \frac{|x|}{\Delta}\right)+0.5\right)}, \quad |x| > \Delta/2\,. \tag{12.14}$$

One should note that the value of $\Delta$ would generally be much smaller than the value of $x$. At the input level of $|x| < \Delta/2$, the floating-point quantizer would be experiencing underflow. Even smaller inputs would be possible, as for example when input $x$ has a zero crossing. But if the probability of $x$ having a magnitude less than $\Delta$ is sufficiently low, Eq. (12.14) could be simply written as

$$\nu_{\mathrm{FL}} = \nu \cdot 2^{Q_1\left(\left(\log_2 \frac{|x|}{\Delta}\right)+0.5\right)}. \tag{12.15}$$

The exponent in Eq. (12.15) is a quantized function of $x$, and it can be expressed as

$$Q_1\left(\left(\log_2 \frac{|x|}{\Delta}\right) + 0.5\right) = \left(\log_2 \frac{|x|}{\Delta} + 0.5\right) + \nu_{\mathrm{EXP}}\,. \tag{12.16}$$

The quantization noise is $\nu_{\mathrm{EXP}}$, the noise in the exponent. It is bounded by $\pm 0.5$.

The floating-point quantization noise can now be expressed as

$$\begin{aligned}\nu_{\mathrm{FL}} &= \nu \cdot 2^{Q_1\left(\left(\log_2 \frac{|x|}{\Delta}\right)+0.5\right)} \\ &= \nu \cdot 2^{\left(\log_2 \frac{|x|}{\Delta} + 0.5\right)+\nu_{\mathrm{EXP}}} \\ &= \nu \cdot \frac{|x|}{\Delta} 2^{0.5} 2^{\nu_{\mathrm{EXP}}}\end{aligned}$$

$$= \sqrt{2}\nu \cdot \frac{|x|}{\Delta} 2^{\nu_{\mathrm{EXP}}} . \tag{12.17}$$

The mean square of $\nu_{\mathrm{FL}}$ can be obtained from this.

$$\begin{aligned} \mathrm{E}\{\nu_{\mathrm{FL}}^2\} &= 2\mathrm{E}\left\{\nu^2 \frac{x^2}{\Delta^2} 2^{2\nu_{\mathrm{EXP}}}\right\} \\ &= 2\mathrm{E}\{\nu^2\}\mathrm{E}\left\{\frac{x^2}{\Delta^2} 2^{2\nu_{\mathrm{EXP}}}\right\} . \end{aligned} \tag{12.18}$$

The factorization is permissible because by assumption $\nu$ behaves as PQN. The noise $\nu_{\mathrm{EXP}}$ is related to $x$, but since $\nu_{\mathrm{EXP}}$ is bounded, it is possible to upper and lower bound $\mathrm{E}\{\nu_{\mathrm{FL}}^2\}$ even without knowledge of the relation between $\nu_{\mathrm{EXP}}$ and $x$. The bounds are

$$\mathrm{E}\{\nu^2\}\mathrm{E}\{\frac{x^2}{\Delta^2}\} \le \mathrm{E}\{\nu_{\mathrm{FL}}^2\} \le 4\mathrm{E}\{\nu^2\}\mathrm{E}\left\{\frac{x^2}{\Delta^2}\right\} . \tag{12.19}$$

These bounds hold without exception as long as the PQN model applies for the hidden quantizer.

If, in addition to this, the PQN model applies to the quantized exponent in Eq. (12.15), then a precise value for $\mathrm{E}\{\nu_{\mathrm{FL}}^2\}$ can be obtained. With PQN, $\nu_{\mathrm{EXP}}$ will have zero mean, a mean square of 1/12, a mean fourth of 1/80, and will be uncorrelated with $\log_2\left(\frac{|x|}{\Delta}\right)$. From the point of view of moments, $\nu_{\mathrm{EXP}}$ will behave as if it is independent of $x$ and any function of $x$. From Eq. (12.17),

$$\nu_{\mathrm{FL}} = \sqrt{2} \cdot \nu \cdot \frac{|x|}{\Delta} 2^{\nu_{\mathrm{EXP}}} = \sqrt{2} \cdot \nu \cdot \frac{|x|}{\Delta} e^{\nu_{\mathrm{EXP}} \ln 2} . \tag{12.20}$$

From this, we can obtain $\mathrm{E}\{\nu_{\mathrm{FL}}^2\}$:

$$\begin{aligned} \mathrm{E}\{\nu_{\mathrm{FL}}^2\} &= 2\mathrm{E}\left\{\nu^2 \frac{x^2}{\Delta^2} e^{2\nu_{\mathrm{EXP}} \ln 2}\right\} \\ &= 2\mathrm{E}\left\{\nu^2 \frac{x^2}{\Delta^2}\Big(1 + 2\nu_{\mathrm{EXP}} \ln 2 + \frac{1}{2!}(2\nu_{\mathrm{EXP}} \ln 2)^2 + \frac{1}{3!}(2\nu_{\mathrm{EXP}} \ln 2)^3 \right. \\ &\quad \left. + \frac{1}{4!}(2\nu_{\mathrm{EXP}} \ln 2)^4 + \cdots\Big)\right\} \\ &= \mathrm{E}\{\nu^2\}\mathrm{E}\left\{\frac{x^2}{\Delta^2}\right\} \mathrm{E}\left\{2 + 4\nu_{\mathrm{EXP}} \ln 2 + 4\nu_{\mathrm{EXP}}^2 (\ln 2)^2 + \frac{8}{3}\nu_{\mathrm{EXP}}^3 (\ln 2)^3 \right. \\ &\quad \left. + \frac{4}{3}\nu_{\mathrm{EXP}}^4 (\ln 2)^4 + \cdots\right\} . \end{aligned} \tag{12.21}$$

Since the odd moments of $\nu_{\mathrm{EXP}}$ are all zero,

$$\mathrm{E}\{\nu_{\mathrm{FL}}^2\} = \mathrm{E}\{\nu^2\}\mathrm{E}\left\{\frac{x^2}{\Delta^2}\right\} \left(2 + 4(\frac{1}{12})(\ln 2)^2 + (\frac{4}{3})(\frac{1}{80})(\ln 2)^4 + \cdots\right)$$

$$= 2.16 \cdot \mathrm{E}\{\nu^2\}\mathrm{E}\left\{\frac{x^2}{\Delta^2}\right\} . \tag{12.22}$$

This falls about half way between the lower and upper bounds.

A more useful form of this expression can be obtained by substituting the following,

$$\mathrm{E}\{\nu^2\} = \frac{q^2}{12}, \quad \text{and} \quad \frac{q}{\Delta} = 2^{-p} . \tag{12.23}$$

The result is a very important one:

$$\mathrm{E}\{\nu_{\mathrm{FL}}^2\} = 0.180 \cdot 2^{-2p} \cdot \mathrm{E}\{x^2\} . \tag{12.24}$$

**Example 12.1 Roundoff noise when multiplying two floating-point numbers**

Let us assume that two numbers: $1.34 \cdot 10^{-3}$ and 4.2 are multiplied using IEEE double precision arithmetic. The exact value of the roundoff error could be determined by using the result of more precise calculation (with $p > 53$) as a reference. However, this is usually not available.

The mean square of the roundoff noise can be easily determined using Eq. (12.24), without the need of reference calculations. The product equals $5.63 \cdot 10^{-3}$. When two numbers of magnitudes similar as above are multiplied, the roundoff noise will have a mean square of $\mathrm{E}\{\nu_{\mathrm{FL}}^2\} \approx 0.180 \cdot 2^{-2p} \cdot 3.2 \cdot 10^{-5}$. Although we cannot determine the precise roundoff error value for the given case, we can give its expected magnitude.

**Example 12.2 Roundoff noise in 2nd-order IIR filtering with floating point**

Let us assume one evaluates the recursive formula

$$y(n) = -a(1)y(n-1) - a(2)y(n-2) + b(1)x(n-1) .$$

When the operations (multiplications and additions) are executed in natural order, all with precision $p$, the following sources of arithmetic roundoff can be enumerated:

1. roundoff after the multiplication $a(1)y(n-1)$:
   $\mathrm{var}\{\nu_{\mathrm{FL}1}\} \approx 0.180 \cdot 2^{-2p} \cdot a(1)^2 \cdot \mathrm{var}\{y\}$,
2. roundoff after the storage of $-a(1)y(n-1)$:
   $\mathrm{var}\{\nu_{\mathrm{FL}2}\} = 0$, since if the product in item 1 is rounded, the quantity to be stored is already quantized,
3. roundoff after the multiplication $a(2)y(n-2)$:
   $\mathrm{var}\{\nu_{\mathrm{FL}3}\} \approx 0.180 \cdot 2^{-2p} \cdot a(2)^2 \cdot \mathrm{var}\{y\}$,
4. roundoff after the addition $-a(1)y(n-1) - a(2)y(n-2)$:
   $\mathrm{var}\{\nu_{\mathrm{FL}4}\} \approx 0.180 \cdot 2^{-2p} \big(a(1)^2 \cdot \mathrm{var}\{y\} + a(2)^2 \cdot \mathrm{var}\{y\} + 2a(1)a(2) \cdot C_{yy}(1)\big)$,
5. roundoff after the storage of $-a(1)y(n-1) - a(2)y(n-2)$:
   $\mathrm{var}\{\nu_{\mathrm{FL}5}\} = 0$, since if the product in item 4 is rounded, the quantity to be stored is already quantized,

6. roundoff after the multiplication $b(1)x(n-1)$:
   $\text{var}\{\nu_{\text{FL}6}\} \approx 0.180 \cdot 2^{-2p} \cdot b(1)^2 \cdot \text{var}\{x\}$,
7. roundoff after the last addition:
   $\text{var}\{\nu_{\text{FL}7}\} \approx 0.180 \cdot 2^{-2p} \cdot \text{var}\{y\}$,
8. roundoff after the storage of the result to $y(n)$:
   $\text{var}\{\nu_{\text{FL}8}\} = 0$, since if the sum in item 7 is rounded, the quantity to be stored is already quantized.

These variances need to be added to obtain the total variance of the arithmetic roundoff noise injected to $y$ in each step.

The total noise variance in $y$ can be determined then by adding the effect of each injected noise. This can be done using the response to an impulse injected to $y(0)$: if this is given by $h_{yy}(0), h_{yy}(1), \ldots$, including the unit impulse itself at time 0, the variance of the total roundoff noise can be calculated by multiplying the above variance of the injected noise by $\sum_{n=0}^{\infty} h_{yy}^2(n)$.

For all these calculations, one also needs to determine the variance of $y$, and the one-step covariance $C_{yy}(1)$.

If $x$ is sinusoidal with amplitude $A$, $y$ is also sinusoidal with an amplitude determined by the transfer function: $\text{var}\{y\} = |H(f_1)|^2 \,\text{var}\{x\} = |H(f_1)|^2 A^2/2$. $C_{yy}(1) \approx \text{var}\{y\}$ for low-frequency sinusoidal input.

If $x$ is zero-mean white noise, $\text{var}\{y\} = \text{var}\{x\} \sum_{n=0}^{\infty} h^2(n)$, with $h(n)$ being the impulse response from $x$ to $y$. The covariance is about $\text{var}\{x\} \sum_{n=0}^{\infty} h(n)h(n+1)$.

Numerically, let us use single precision ($p = 24$), and let $a(1) = -0.9$, $a(2) = 0.5$, and $b(1) = 0.2$, and let the input signal be a sine wave with unit power ($A^2/2 = 1$), and frequency $f_1 = 0.015 f_{\text{s}}$. With these, $H(f_1) = 0.33$, and $\text{var}\{y\} = 0.11$, and the covariance $C_{yy}(1)$ approximately equals $\text{var}\{y\}$. Evaluating the variance of the floating-point noise,

$$\text{var}\{\nu_{\text{FL}}\} = \sum_{n=1}^{8} \text{var}\{\nu_{\text{FL}n}\} \approx 1.83 \cdot 10^{-16}\,. \tag{12.25}$$

*The use of extended-precision accumulator and of multiply-and-add operation*

In many modern processors, an accumulator is used with extended precision $p_{\text{acc}}$, and multiply-and-add operation (see page 374) can be applied. In this case, multiplication is usually executed without roundoff, and additions are not followed by extra storage, therefore roundoff happens only for certain items (but, in addition to the above, $\text{var}\{\nu_{\text{FL}2}\} \neq 0$, since $\nu_{\text{FL}1} = 0$). The total variance of $y$, caused by arithmetic roundoff, is

$$\begin{aligned}\text{var}_{\text{inj}} &= \text{var}\{\nu_{\text{FL}2}\} + \text{var}\{\nu_{\text{FL}4}\} + \text{var}\{\nu_{\text{FL}7}\} + \text{var}\{\nu_{\text{FL}8}\} \\ &= 0.180 \cdot 2^{-2p_{\text{acc}}} \cdot a(1)^2 \,\text{var}\{y\}\end{aligned}$$

$$+ 0.180 \cdot 2^{-2p_{\text{acc}}} \left( a(1)^2 \operatorname{var}\{y\} + a(2)^2 \operatorname{var}\{y\} + 2a(1)a(2)C_{yy}(1) \right)$$
$$+ 0.180 \cdot 2^{-2p_{\text{acc}}} \cdot \operatorname{var}\{y\}$$
$$+ 0.180 \cdot 2^{-2p} \cdot \operatorname{var}\{y\} . \tag{12.26}$$

In this expression, usually the last term dominates. With the numerical data, $\operatorname{var}\{\nu_{\text{FL}}\text{-ext}\} \approx 7.1 \cdot 10^{-17}$.

Making the same substitution in Eq. (12.19), the bounds are

$$\frac{1}{12} \cdot 2^{-2p} \cdot \mathrm{E}\{x^2\} \le \mathrm{E}\{\nu_{\text{FL}}^2\} \le \frac{1}{3} \cdot 2^{-2p} \cdot \mathrm{E}\{x^2\} . \tag{12.27}$$

The signal-to-noise ratio for the floating-point quantizer is defined as

$$\text{SNR} \stackrel{\triangle}{=} \frac{\mathrm{E}\{x^2\}}{\mathrm{E}\{\nu_{\text{FL}}^2\}} . \tag{12.28}$$

If both $\nu$ and $\nu_{\text{EXP}}$ obey PQN models, then Eq. (12.24) can be used to obtain the SNR. The result is:

$$\text{SNR} = 5.55 \cdot 2^{2p} . \tag{12.29}$$

This is the ratio of signal power to quantization noise power. Expressed in dB, the SNR is

$$\begin{aligned} \text{SNR, dB} &\approx 10 \log \left( (5.55) 2^{2p} \right) \\ &= 7.44 + 6.02p . \end{aligned} \tag{12.30}$$

If only $\nu$ obeys a PQN model, then Eq. (12.27) can be used to bound the SNR:

$$12 \cdot 2^{2p} \ge \text{SNR} \ge 3 \cdot 2^{2p} . \tag{12.31}$$

This can be expressed in dB as:

$$4.77 + 6.02p \le \text{SNR, dB} \le 10.79 + 6.02p . \tag{12.32}$$

From relations Eq. (12.29) and Eq. (12.31), it is clear that the SNR improves as the number of bits in the mantissa is increased. It is also clear that for the floating-point quantizer, the SNR does not depend on $\mathrm{E}\{x^2\}$ (as long as $\nu$ and $\nu_{\text{EXP}}$ act like PQN). The quantization noise power is proportional to $\mathrm{E}\{x^2\}$. This is a very different situation from that of the uniform quantizer, where SNR increases in proportion to $\mathrm{E}\{x^2\}$ since the quantization noise power is constant at $q^2/12$.

When $\nu$ satisfies a PQN model, $\nu_{\text{FL}}$ is uncorrelated with $x$. The floating-point quantizer can be replaced for purposes of least-squares analysis by an additive independent noise having a mean of zero and a mean square bounded by (12.27). If in ad-

dition $\nu_{\text{EXP}}$ satisfies a PQN model, the mean square of $\nu$ will be given by Eq. (12.24). Section 12.6 discusses the question about $\nu_{\text{EXP}}$ satisfying a PQN model. This does happen in a surprising number of cases. But when PQN fails for $\nu_{\text{EXP}}$, we cannot obtain $\text{E}\{\nu_{\text{FL}}^2\}$ from Eq. (12.24), but we can get bounds on it from (12.27).

## 12.6 HOW GOOD IS THE PQN MODEL FOR THE EXPONENT QUANTIZER?

The PQN model for the hidden quantizer turns out to be a very good one when the mantissa has 8 bits or more, and: (a) the dynamic range of $x$ extends over at least several times $\Delta$, (b) for the Gaussian case, $\sigma_x$ is at least as big as $\Delta/2$. This model works over a very wide range of conditions encountered in practice. Unfortunately, the range of conditions where a PQN model would work for the exponent quantizer is much more restricted. In this section, we will explore the issue.

### 12.6.1 Gaussian Input

The simplest and most important case is that of the Gaussian input. Let the input $x$ be Gaussian with variance $\sigma_x^2$, and with a mean $\mu_x$ that could vary between zero and some large multiple of $\sigma_x$, say $20\sigma_x$. Fig. 12.16 shows calculated plots of the PDF of $\nu_{\text{EXP}}$ for various values of $\mu_x$. This PDF is almost perfectly uniform between $\pm\frac{1}{2}$ for $\mu_x = 0, 2\sigma_x$, and $3\sigma_x$. For higher values of $\mu_x$, deviations from uniformity are seen.

The covariance of $\nu_{\text{EXP}}$ and $x$, shown in Fig. 12.17, is essentially zero for $\mu_x = 0, 2\sigma_x$, and $3\sigma_x$. But for higher values of $\mu_x$, significant correlation develops between $\nu_{\text{EXP}}$ and $x$. Thus, the PQN model for the exponent quantizer appears to be intact for a range of input means from zero to $3\sigma_x$. Beyond that, this PQN model appears to break down. The reason for this is that the input to the exponent quantizer is the variable $|x|/\Delta$ going through the logarithm function. The greater the mean of $x$, the more the logarithm is saturated and the more the dynamic range of the quantizer input is compressed.

The point of breakdown of the PQN model for the exponent quantizer is not dependent on the size of the mantissa, but is dependent only on the ratios of $\sigma_x$ to $\Delta$ and $\sigma_x$ to $\mu_x$.

When the PQN model for the exponent quantizer is applicable, the PQN model for the hidden quantizer will always be applicable. The reason for this is that the input to both quantizers is of the form $\log_2(x)$, but for the exponent quantizer $x$ is divided by $\Delta$, making its effective quantization step size $\Delta$, and for the hidden quantizer, the input is not divided by anything and so its quantization step size is $q$. The quantization step of the exponent quantizer is therefore coarser than that of the hidden quantizer by the factor

$$\frac{\Delta}{q} = 2^p \,. \tag{12.33}$$

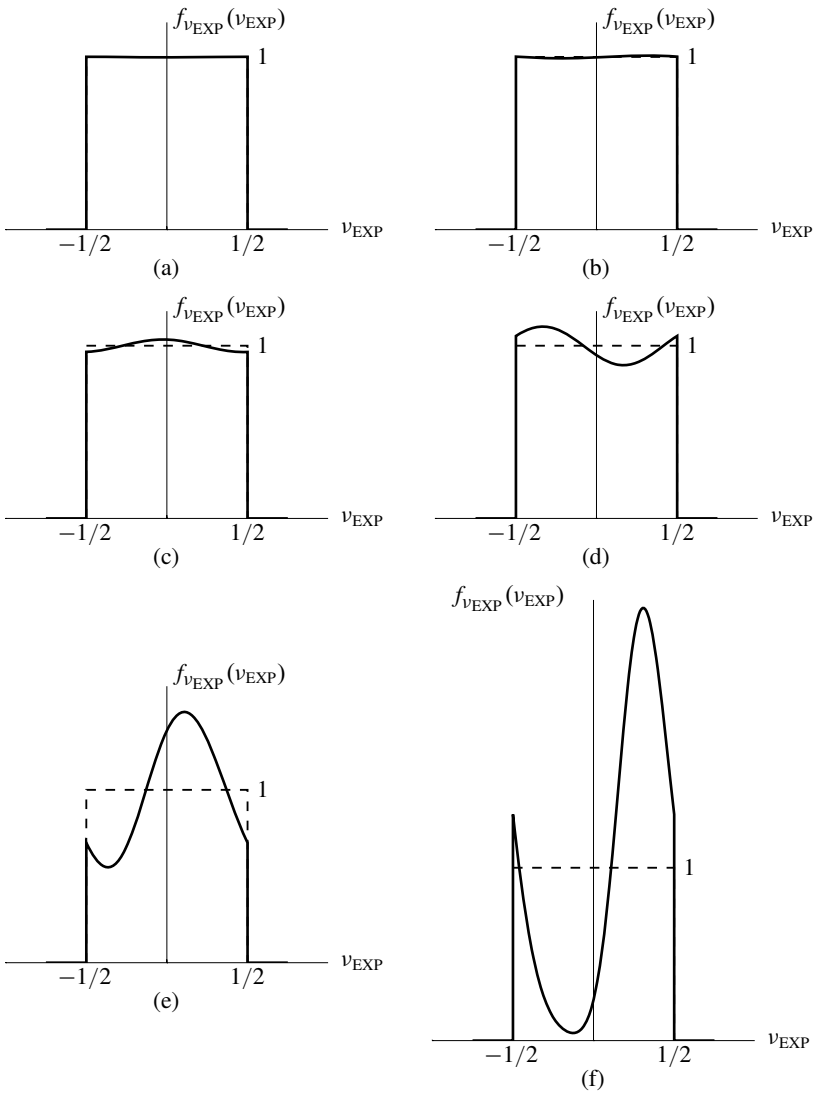


**Figure 12.16** PDFs of noise of exponent quantizer, for Gaussian input $x$, with $\sigma_x = 512\Delta$: (a) $\mu_x = 0$; (b) $\mu_x = \sigma_x$; (c) $\mu_x = 2\sigma_x$; (d) $\mu_x = 3\sigma_x$; (e) $\mu_x = 5\sigma_x$; (f) $\mu_x = 9\sigma_x$.

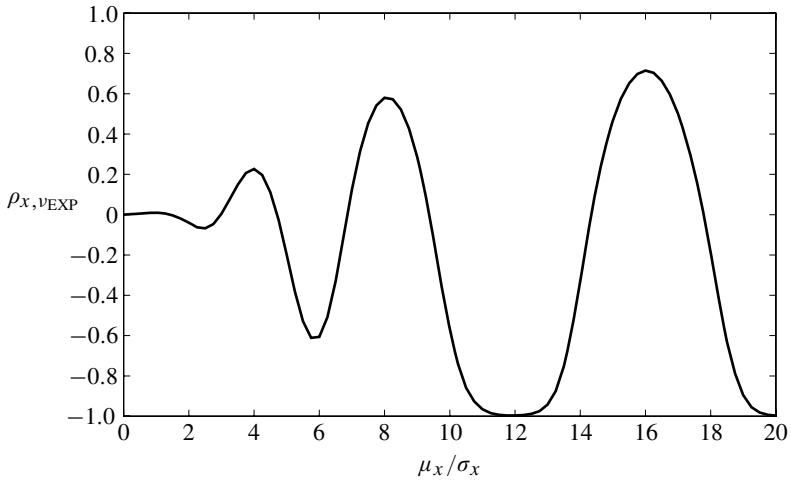


**Figure 12.17** Correlation coefficient between $v_{\mathrm{EXP}}$ and $x$ versus $\mu_x/\sigma_x$, for Gaussian input $x$ with $\sigma_x = 512\Delta$.

With an 8-bit mantissa, this ratio would be 256. So if the conditions for PQN are met for the exponent quantizer, they are easily met for the hidden quantizer.

For the Gaussian input case, the PQN model for the exponent quantizer works to within a close approximation for $0 \leq \mu_x < 3\sigma_x$. The correlation coefficient between $v_{\mathrm{EXP}}$ and $x$, plotted versus $\mu_x/\sigma_x$ in Fig. 12.17, indicates very small correlation with $\mu_x$ in this range. One would therefore expect formula (12.24) for $\mathrm{E}\{v_{\mathrm{FL}}^2\}$ to be very accurate in these circumstances, and it is.

Eq. (12.24) may be rewritten in the following way:

$$0.180 = 2^{2p}\left(\frac{\mathrm{E}\{v_{\mathrm{FL}}^2\}}{\mathrm{E}\{x^2\}}\right). \tag{12.34}$$

This form of (12.24) suggests a new definition:

$$\begin{pmatrix}\text{normalized}\\ \text{floating-point}\\ \text{quantization}\\ \text{noise power}\end{pmatrix} \triangleq 2^{2p}\left(\frac{\mathrm{E}\{v_{\mathrm{FL}}^2\}}{\mathrm{E}\{x^2\}}\right). \tag{12.35}$$

The normalized floating-point quantization noise power (NFPQNP) will be equal to the "magic number" 0.180 when a PQN model applies to the exponent quantizer and, of course, another PQN model applies to the hidden quantizer. Otherwise the NFPQNP will not be exactly equal to 0.180, but it will be bounded. From Eq. (12.27),

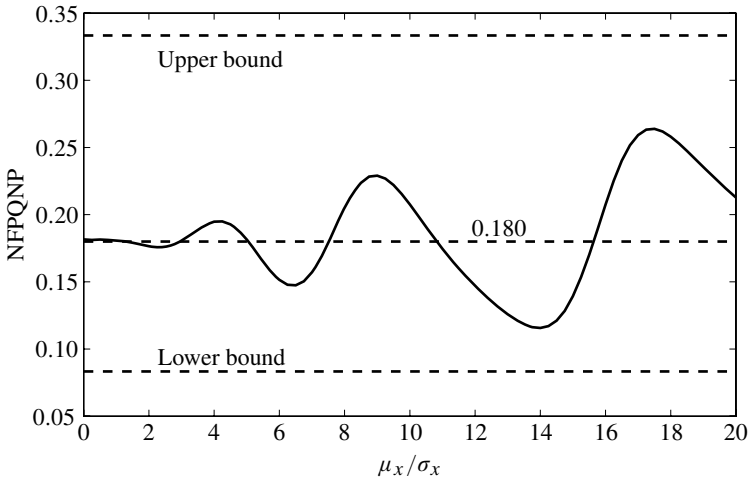


**Figure 12.18** Normalized floating-point quantization noise power (NFPQNP) versus $\mu_x/\sigma_x$, for Gaussian input $x$ with $\sigma_x = 512\Delta$ and 8-bit mantissas.

$$\frac{1}{12} \leq \left( \begin{array}{c} \text{normalized} \\ \text{floating-point} \\ \text{quantization} \\ \text{noise power} \end{array} \right) \leq \frac{1}{3}\,. \tag{12.36}$$

These bounds rely on the applicability of a PQN model for the hidden quantizer.

For the Gaussian input case, the NFPQNP is plotted versus the ratio of input mean to input standard deviation in Fig. 12.18. The value of NFPQNP is very close to 0.180 for $0 \leq (\mu_x/\sigma_x) < 3$. When $\mu_x$ is made larger, the NFPQNP departs from its nominal value but, in any event, stays within the bounds.

Now that we have a good idea about the floating-point quantization noise power, we need to determine the mean of the floating-point quantization noise and its correlation with the input $x$. Fig. 12.19 is a plot of $\mathrm{E}\{\nu_{\mathrm{FL}}\}$, normalized with respect to $\mathrm{E}\{x\}$, versus $\mu_x/\sigma_x$, and Fig. 12.20 is a plot of the correlation coefficient $\rho_{\nu_{\mathrm{FL}},x}$ versus $\mu_x/\sigma_x$. When the mean of $x$ is not zero, the mean of $\nu_{\mathrm{FL}}$ remains very close to zero, never more than a tiny fraction of $\mu_x$, less than one part in $10^5$. The correlation between $\nu_{\mathrm{FL}}$ and $x$ also remains very close to zero, generally less than 0.5% when $\mu_x$ varies between zero and $20\sigma_x$.

That the mean of $\nu_{\mathrm{FL}}$ is very close to zero and that the correlation between $\nu_{\mathrm{FL}}$ and $x$ is very close to zero results from the hidden quantizer behaving in accord with the PQN model. This is illustrated in Figs. 12.21 and 12.22.

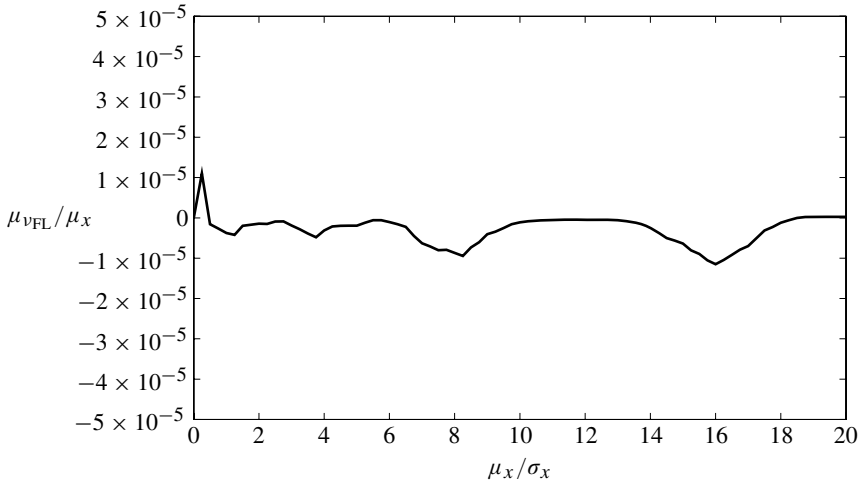


**Figure 12.19** Relative mean of the floating-point noise: $\mu_{v_{\mathrm{FL}}}/\mu_x$ versus $\mu_x/\sigma_x$, for Gaussian input $x$, with $\sigma_x = 512\Delta$, and 8-bit mantissas.

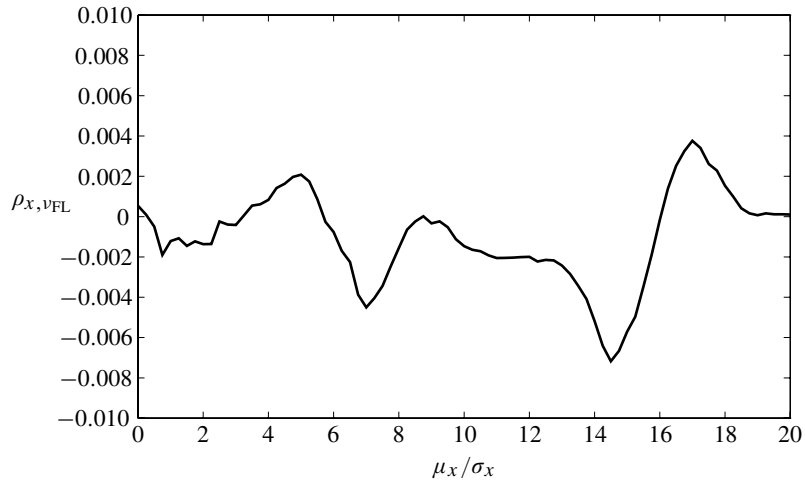


**Figure 12.20** Correlation coefficient of floating-point quantization noise and input $x$: $\rho_{v_{\mathrm{FL}},x}$ versus $\mu_x/\sigma_x$, for Gaussian input $x$, with $\sigma_x = 512\Delta$, and 8-bit mantissas.

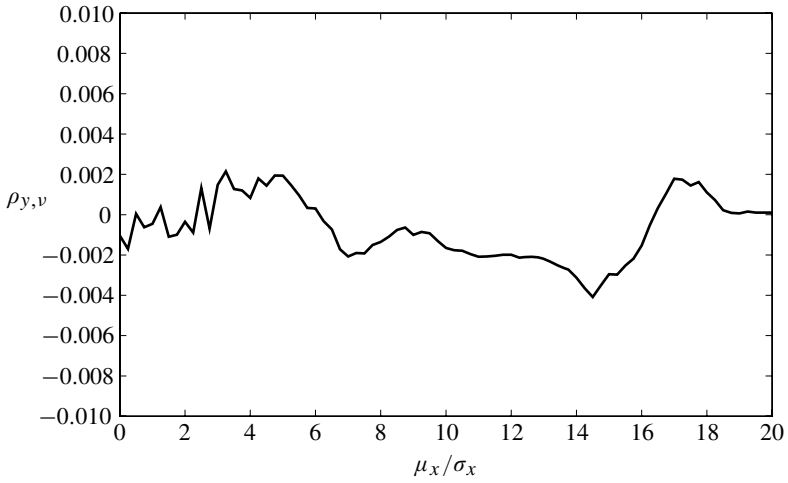


**Figure 12.21** Correlation coefficient of the hidden quantization noise and input $y$: $\rho_{v,y}$ versus $\mu_x/\sigma_x$, for Gaussian input $x$, with $\sigma_x = 512\Delta$, and 8-bit mantissas.

For purposes of analysis, it is clear that the floating-point quantizer can be replaced by a source of additive independent noise with zero mean and mean square given by Eq. (12.24) or bounded by (12.27).

### 12.6.2 Input with Triangular Distribution

A similar set of calculations and plots have been made for an input $x$ having a triangular PDF with variable mean. The results are shown in Figs. 12.23–12.28. Fig. 12.23 shows PDFs of $\nu_{\text{EXP}}$ for various mean values of input $x$. The amplitude range of $x$ is $\pm A + \mu_x$. The correlation coefficient between $\nu_{\text{EXP}}$ and $x$ is moderately small for $0 \leq \mu_x < A$, the PDFs of $\nu_{\text{EXP}}$ are close to uniform in that range, and they become non uniform outside this range.

Thus, the PQN model works well for the exponent quantizer in the range $0 \leq \mu_x < A$. Outside this range, the PQN model begins to break down. This is confirmed by inspection of the normalized floating-point quantization noise power which has been calculated and plotted in Fig. 12.25.

The mean of $\nu_{\text{FL}}$ is very close to zero. When normalized with respect to the mean of $x$, the relative mean remains less than a few parts per $10^5$ over the range of input means from $0 < \mu_x < 20A$. The correlation coefficient between $\nu_{\text{FL}}$ and $x$ is of the order of 1% over the same range of input means.

With a triangular input PDF, the floating-point quantizer may be replaced for purposes of moment analysis by an independent source of additive noise having zero mean and a mean square given by Eq. (12.24) or otherwise bounded by (12.27).

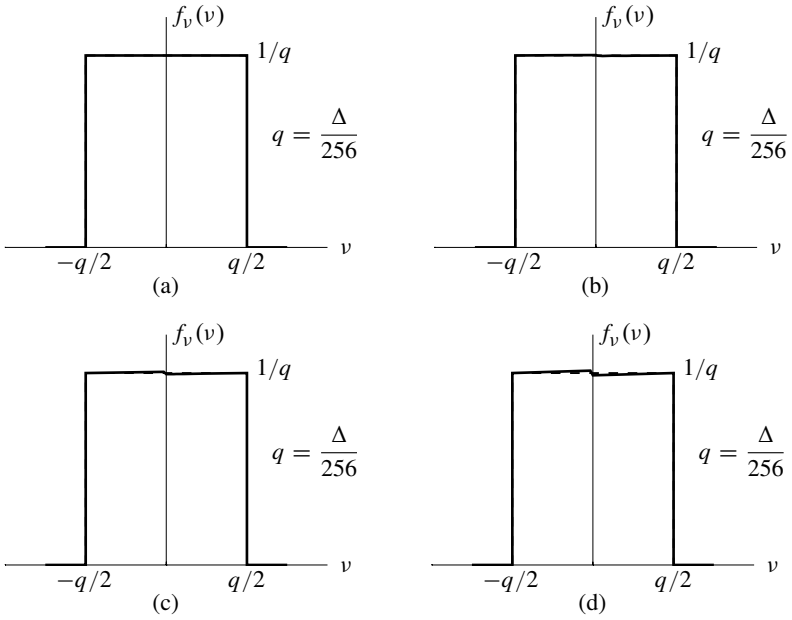


**Figure 12.22** Noise PDFs for hidden quantizer, Gaussian input PDF with $\sigma_x = 512\Delta$, and 8-bit mantissas, (a) $\mu_x = 0$; (b) $\mu_x = 5\sigma_x$; (c) $\mu_x = 14.5\sigma_x$; (d) $\mu_x = 17\sigma_x$.

### 12.6.3 Input with Uniform Distribution

A more difficult case is that of input $x$ having a uniform PDF. The abrupt cutoffs at the edges of this PDF cause the CF of $x$ to have wide "bandwidth." A study of the floating-point quantizer having an input of variable mean and a uniform PDF has been done, and results are shown in Figs. 12.29–12.35. The width of the uniform PDF is $2A$.

Fig. 12.29 gives PDFs of the noise of the exponent quantizer for various mean values of $x$. Fig. 12.30 shows the correlation coefficient of $\nu_{\text{EXP}}$ and input $x$ for $\mu_x$ in the range $0 < \mu_x < 20A$. Both the PDFs and the correlation coefficient show that the PQN model for the exponent quantizer works for a zero-mean input. However, even with a mean of $\mu_x = A$, the exponent quantizer's PQN model begins to break down. This is confirmed by inspection of Fig. 12.31. The normalized floating-point quantization noise power is reasonably close to the "magic value" of 0.180 for $\mu_x = 0$, and this is reasonably so even for $\mu_x = A$. Beyond that range, the values of NFPQNP can differ significantly from 0.180, although they are still bounded.

The relative mean of $\nu_{\text{FL}}$ is plotted in Fig. 12.32 versus $\mu_x/A$. The values are larger than for the Gaussian and triangular cases, but are still very small, less than 6 parts per $10^5$. The correlation coefficient between $\nu_{\text{FL}}$ and $x$, shown in Fig. 12.33, is of the order of 2% or less until $\mu_x/A \approx 6$, but then it becomes quite a bit larger. To

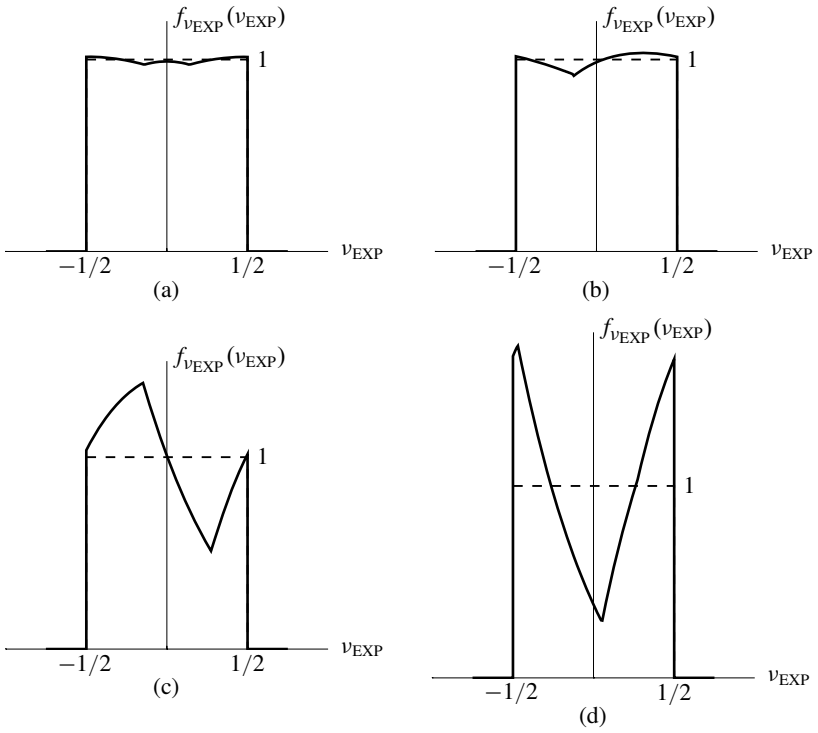


**Figure 12.23** PDFs of noise of exponent quantizer, for triangular input PDF, with $A = 400\Delta$: (a) $\mu_x = 0$; (b) $\mu_x = A$; (c) $\mu_x = 2A$; (d) $\mu_x = 2.5A$.

explore this, we plotted the correlation coefficient between the hidden quantization noise $v$ and the hidden quantizer input $y$. This is shown in Fig. 12.34. The correlation coefficient is less than 2% until $\mu_x/A \approx 6$. It appears that PQN for the hidden quantizer is beginning to breakdown when $\mu_x/A$ becomes greater than 6.

PDFs of $v$ have been plotted to study the break down of PQN, and the results are shown in Fig. 12.35. The PDFs are perfectly uniform or close to uniform, but irregularities appear for $\mu_x/A$ higher than 6. The correlation coefficients $\rho_{v,y}$ and $\rho_{v_{FL},x}$ also indicate PQN break down for $\mu_x/A$ greater than 6. By increasing the number of mantissa bits by a few, these breakdown indicators, which are slight, disappear altogether.

For purposes of moment analysis, the floating-point quantizer can be replaced by a source of additive zero-mean independent noise whose mean square is given by Eq. (12.24) or is bounded by (12.27). One must be sure, however, that the mantissa has enough bits to allow the PQN model to apply to the hidden quantizer. Then $v_{FL}$ is uncorrelated with $x$, and the quantizer replacement with additive noise is justified.

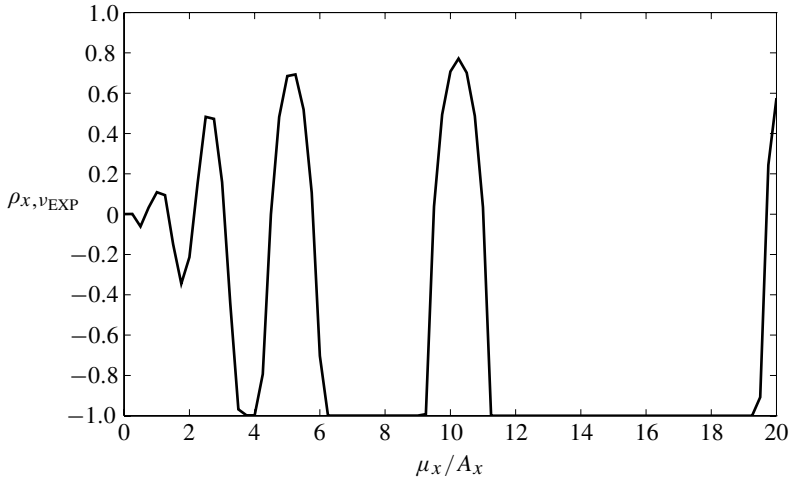


**Figure 12.24** Correlation coefficient between $v_{\mathrm{EXP}}$ and $x$ versus $\mu_x/A$, for triangular input PDF with $A = 400\Delta$.

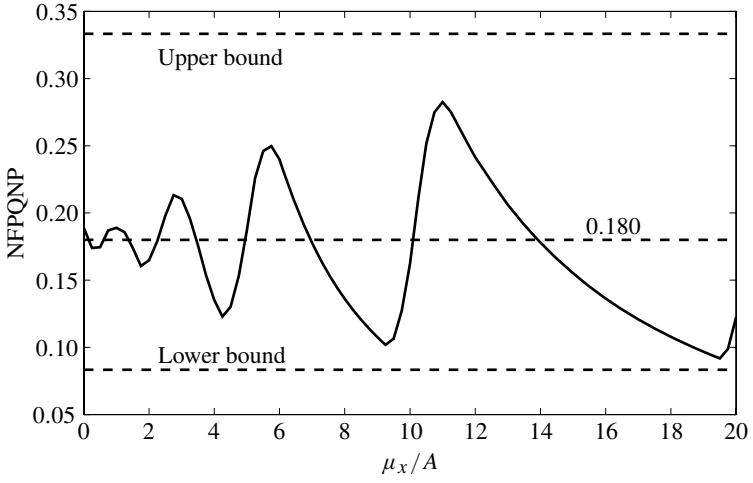


**Figure 12.25** Normalized floating-point quantization noise power (NFPQNP) versus $\mu_x/A$, for triangular input PDF with $A = 400\Delta$, and 8-bit mantissas.

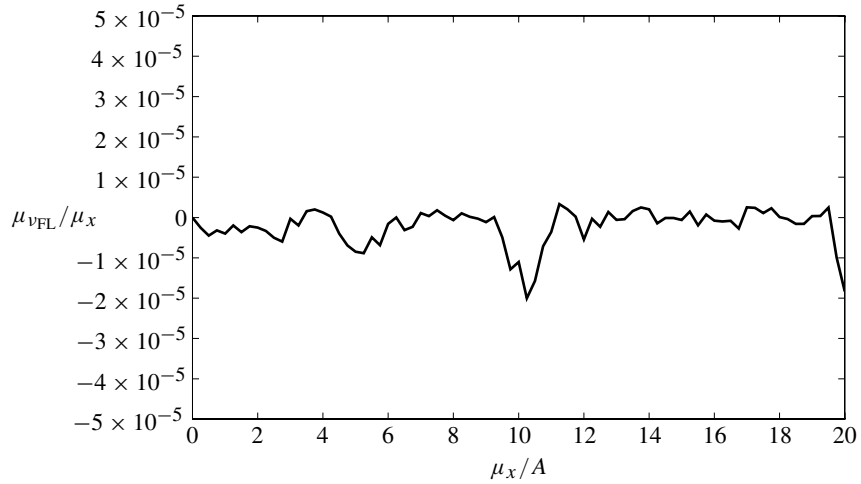


**Figure 12.26** Relative mean of the floating-point noise: $\mu_{v_{\mathrm{FL}}}/\mu_x$ versus $\mu_x/A$, for triangular input PDF with $A = 400\Delta$, and 8-bit mantissas.

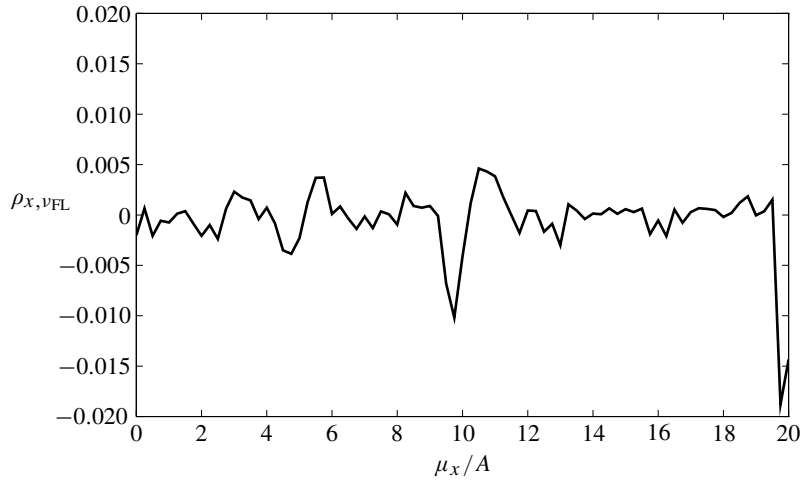


**Figure 12.27** Correlation coefficient of floating-point quantization noise and input $x$: $\rho_{v_{\mathrm{FL}},x}$ versus $\mu_x/A$, for triangular input PDF with $A = 400\Delta$, and 8-bit mantissas.

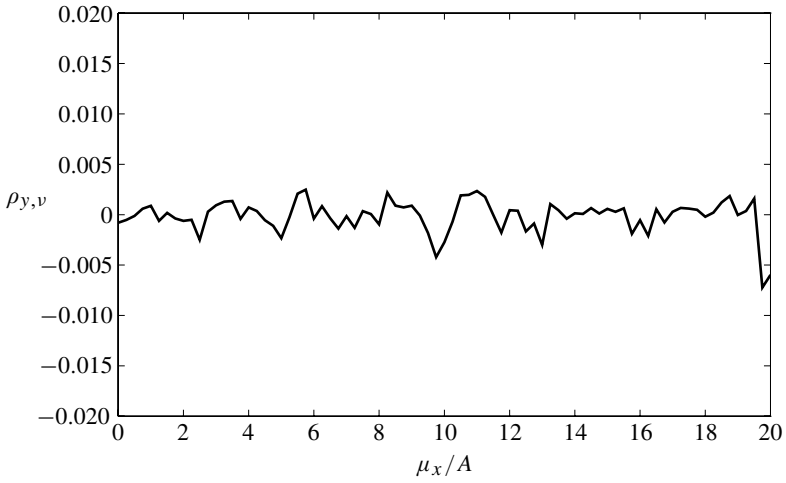


**Figure 12.28** Correlation coefficient of the hidden quantization noise and input $y$: $\rho_{v,y}$ versus $\mu_x/A$, for triangular input PDF with $A = 500\Delta$, and 8-bit mantissas.

### 12.6.4 Sinusoidal Input

The PDF of the sinusoidal input, like the uniform PDF, has abrupt cutoff at its edges. Its shape is even worse than that of the uniform PDF from the point of view of satisfying conditions for PQN. Results of a study of a floating-point quantizer with a sinusoidal input and a variable mean are shown in Figs. 12.36–12.42. The zero-to-peak amplitude of the sinusoidal input is $A$.

Fig. 12.36 gives PDFs of the noise of the exponent quantizer for various mean values of $x$. None of these PDFs suggest PQN conditions for the exponent quantizer, even when $\mu_x = 0$. This is testimony to the difficulty of satisfying PQN condition for the case of the sinusoidal input. Fig. 12.37 shows the correlation coefficient between $v_{\text{EXP}}$ and input $x$ for various values of the input mean $\mu_x$. The correlation coefficient is most often very far from zero, further indicating that PQN conditions are not met for the exponent quantizer. Accordingly, it will not be possible to get an accurate value of $\text{E}\{v_{\text{FL}}^2\}$ from Eq. (12.34), but it will be possible to bound $\text{E}\{v_{\text{FL}}^2\}$ by using (12.36).

Fig. 12.38(a) shows the normalized floating-point quantization noise power for a variety of input mean values. The mantissa has $p = 8$ bits. As long as PQN is satisfied for the hidden quantizer, the normalized quantization noise power stays within the bounds. This seems to be the case. With the mantissa length doubled to $p = 16$ bits, the normalized floating-point noise power measurements, plotted in Fig. 12.38(b), came out the same. PQN is surely satisfied for the hidden quantizer when given such a long mantissa. More evidence will follow.

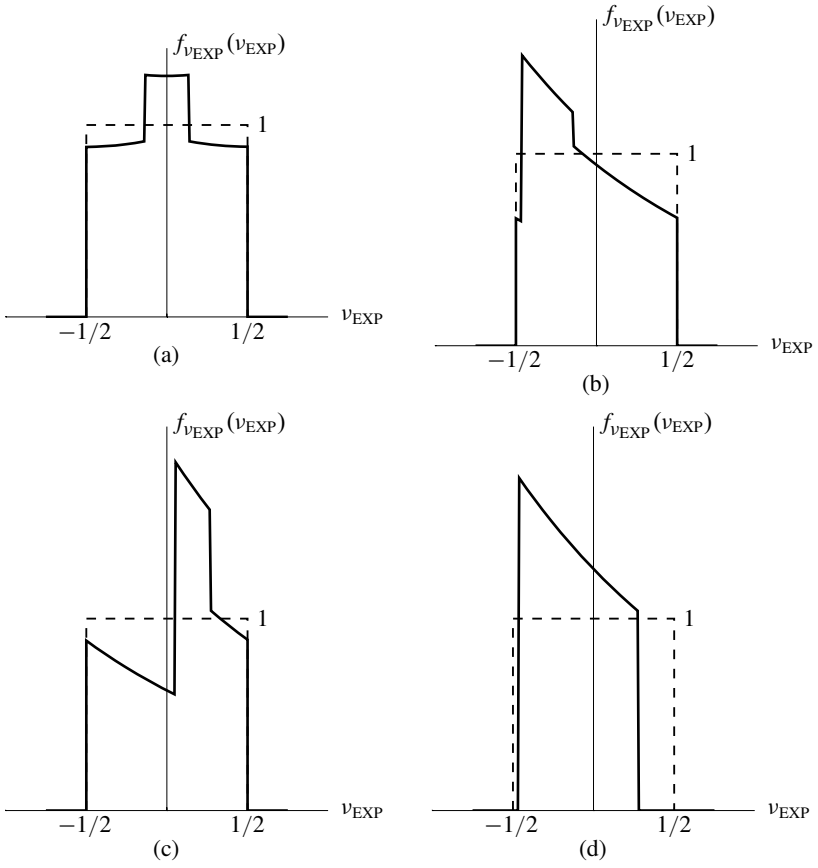


**Figure 12.29** PDFs of noise of exponent quantizer, for uniform input PDF with $A = 400\Delta$: (a) $\mu_x = 0$; (b) $\mu_x = 1.5A$; (c) $\mu_x = 2.5A$; (d) $\mu_x = 4A$.

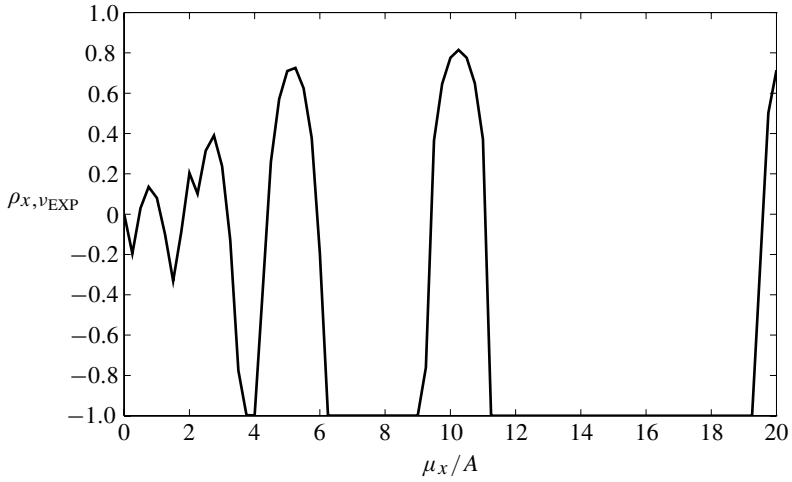


**Figure 12.30** Correlation coefficient between $v_{\mathrm{EXP}}$ and $x$ versus $\mu_x/A$, for uniform input PDF with $A = 400\Delta$.

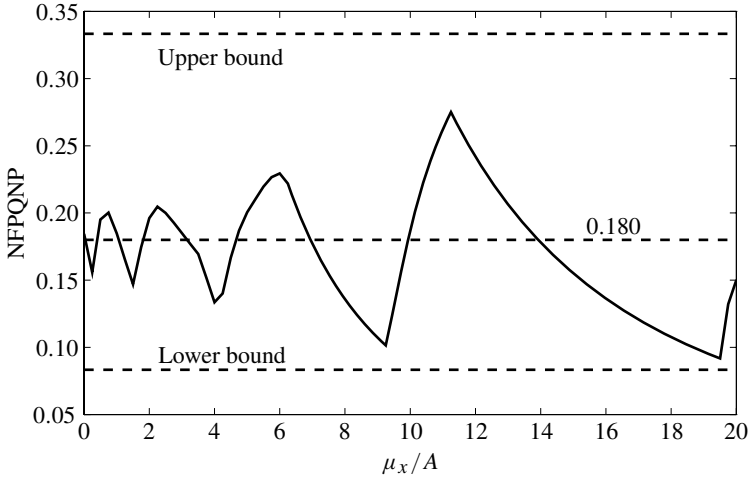


**Figure 12.31** Normalized floating-point quantization noise power (NFPQNP) versus $\mu_x/A$, for uniform input PDF with $A = 400\Delta$ and 8-bit mantissas.

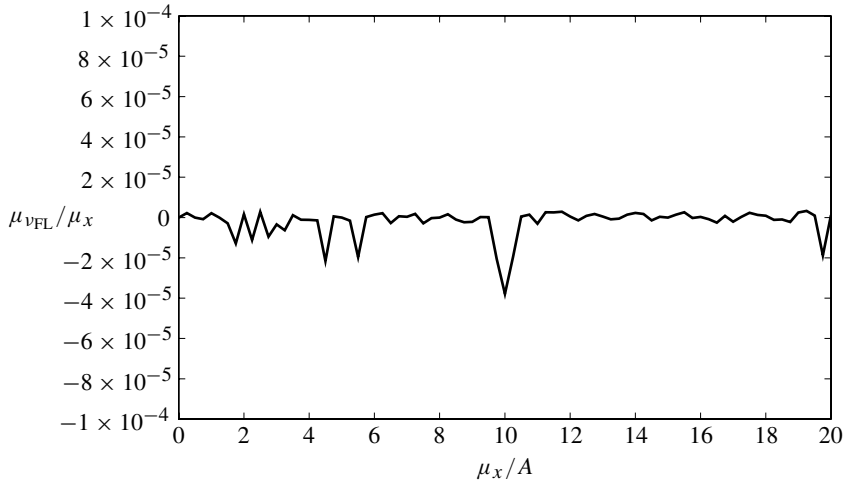


**Figure 12.32** Relative mean of the floating-point noise: $\mu_{v_{\mathrm{FL}}}/\mu_x$ versus $\mu_x/A$, for uniform input PDF with $A = 400\Delta$ and 8-bit mantissas.

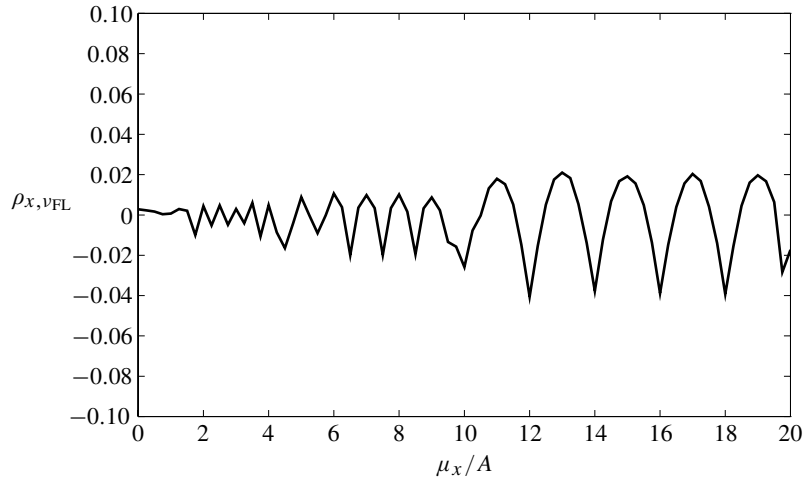


**Figure 12.33** Correlation coefficient of floating-point quantization noise and input $x$: $\rho_{v_{\mathrm{FL}},x}$ versus $\mu_x/A$, for uniform input PDF with $A = 400\Delta$ and 8-bit mantissas.

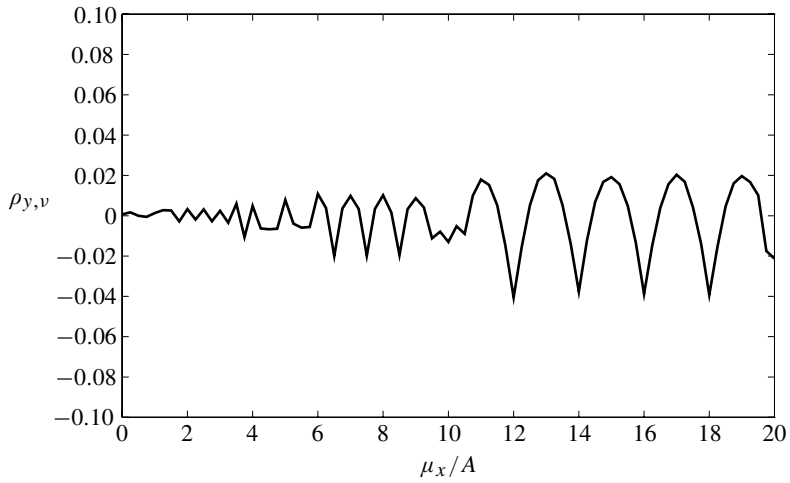


**Figure 12.34** Correlation coefficient of hidden quantization noise $v$ and hidden quantizer input $y$: $\rho_{v,y}$ versus $\mu_x/A$, for uniform input PDF with $A = 400\Delta$ and 8-bit mantissas.

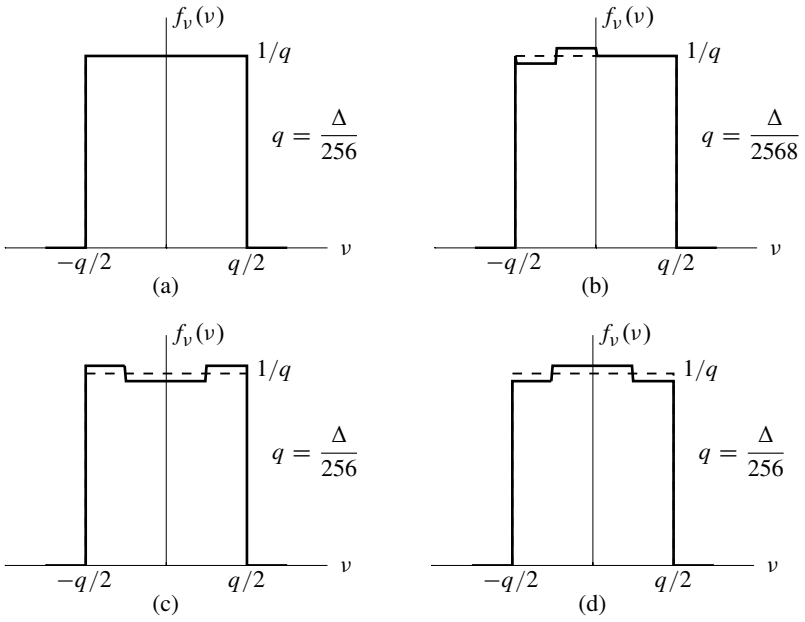


**Figure 12.35** Noise PDFs for hidden quantizer, uniform input PDF with $A = 200\Delta$, 8-bit mantissas, (a) $\mu_x = 0$; (b) $\mu_x = 10A$; (c) $\mu_x = 15A$; (d) $\mu_x = 16A$.

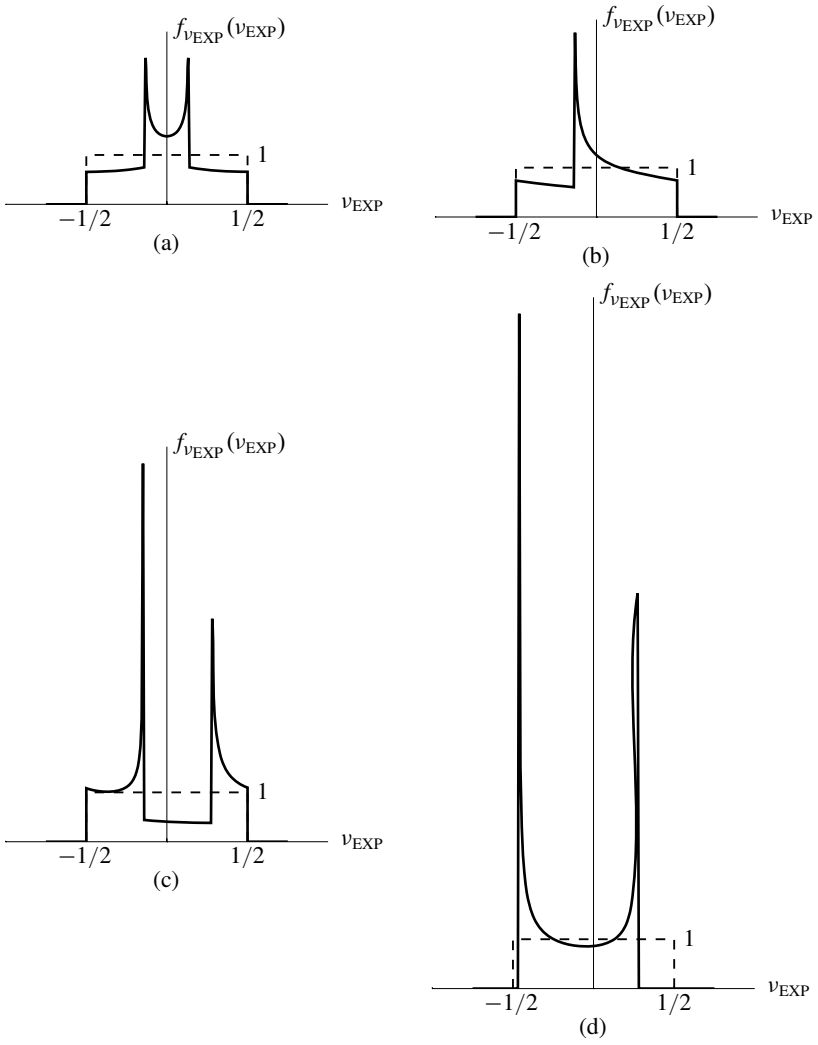


**Figure 12.36** PDFs of noise of the exponent quantizer, for sinusoidal input with $A = 400\Delta$: (a) $\mu_x = 0$; (b) $\mu_x = A$; (c) $\mu_x = 2A$; (d) $\mu_x = 4A$.

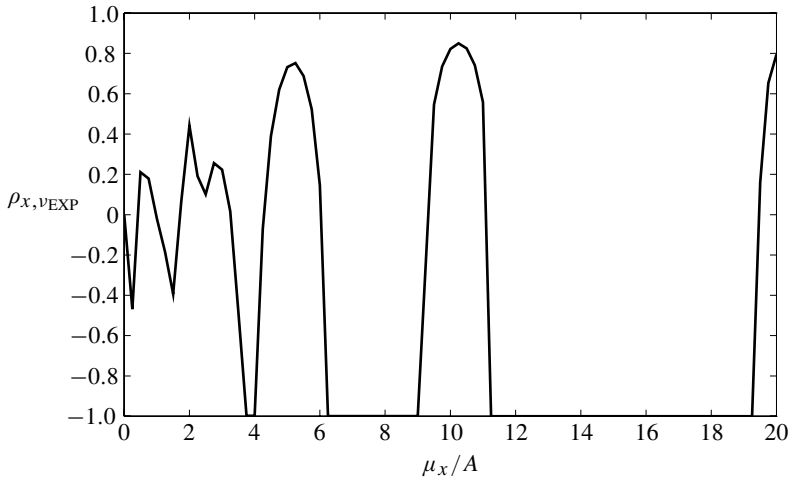


**Figure 12.37** Correlation coefficient between $\nu_{\mathrm{EXP}}$ and $x$ versus $\mu_x/A$, for sinusoidal input, $A = 400\Delta$.

The relative mean of $\nu_{\mathrm{FL}}$ with an 8-bit mantissa is plotted in Fig. 12.39(a) for a variety of input mean values. The relative mean is very small, a few parts in 10 000 at worst. The mean of $\nu_{\mathrm{FL}}$ being close to zero is consistent with PQN being satisfied for the hidden quantizer. Doubling the length of the mantissa to 16 bits, the relative mean, plotted in Fig. 12.39(b), becomes very very small, a few parts in 10 000 000, for a range of input mean values.

Fig. 12.40 shows the correlation coefficient of $\nu_{\mathrm{FL}}$ and $x$ for various mean values of $x$. Fig. 12.40(a) corresponds to $p = 8$ bits. The correlation coefficient is highly variable as $\mu_x$ is varied, and at worst could be as high as 25%. As such, one could not replace the floating-point quantizer with a source of additive independent noise. PQN for the hidden quantizer fails. So the correlation coefficient was re-evaluated with a 16-bit mantissa, and the result, shown in Fig. 12.40(b), indicates a corresponding worst-case correlation coefficient of the order of 1%. With this mantissa, it seems reasonable that one could replace the floating-point quantizer with a source of additive independent noise. The resulting analysis would be a very good approximation to the truth. PQN for the hidden quantizer is a good approximation.

Fig. 12.41 shows correlation coefficients between the hidden quantization noise and the hidden quantizer input for various values of $\mu_x$. These figures are very similar to Figs. 12.40(a)-(b), respectively. Relationships between $\rho_{\nu_{\mathrm{FL}},x}$ and $\rho_{\nu,y}$ will be discussed below.

One final check will be made on the PQN hypothesis for the hidden quantizer. Fig. 12.42 shows the PDF of $\nu$ for the hidden quantizer for various values of input mean. These PDFs are not quite uniform, and in some cases they are far from uni-

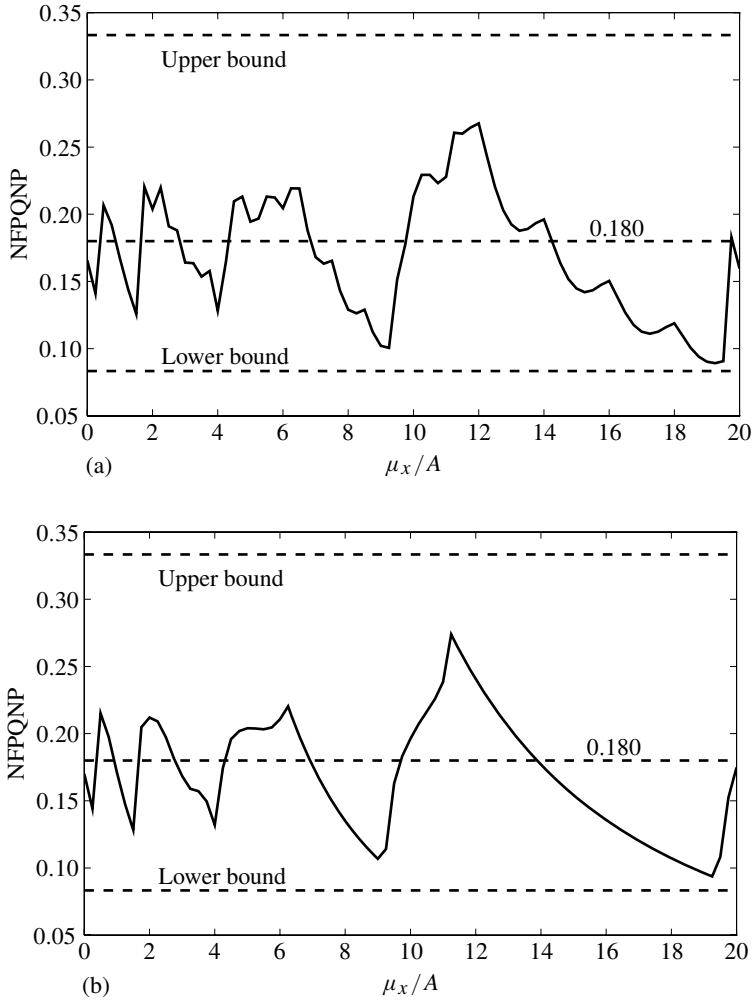


**Figure 12.38** Normalized floating-point quantization noise power (NFPQNP) vs. $\mu_x/A$, for a sinusoidal input with $A = 400\Delta$: (a) 8-bit mantissa; (b) 16-bit mantissa.

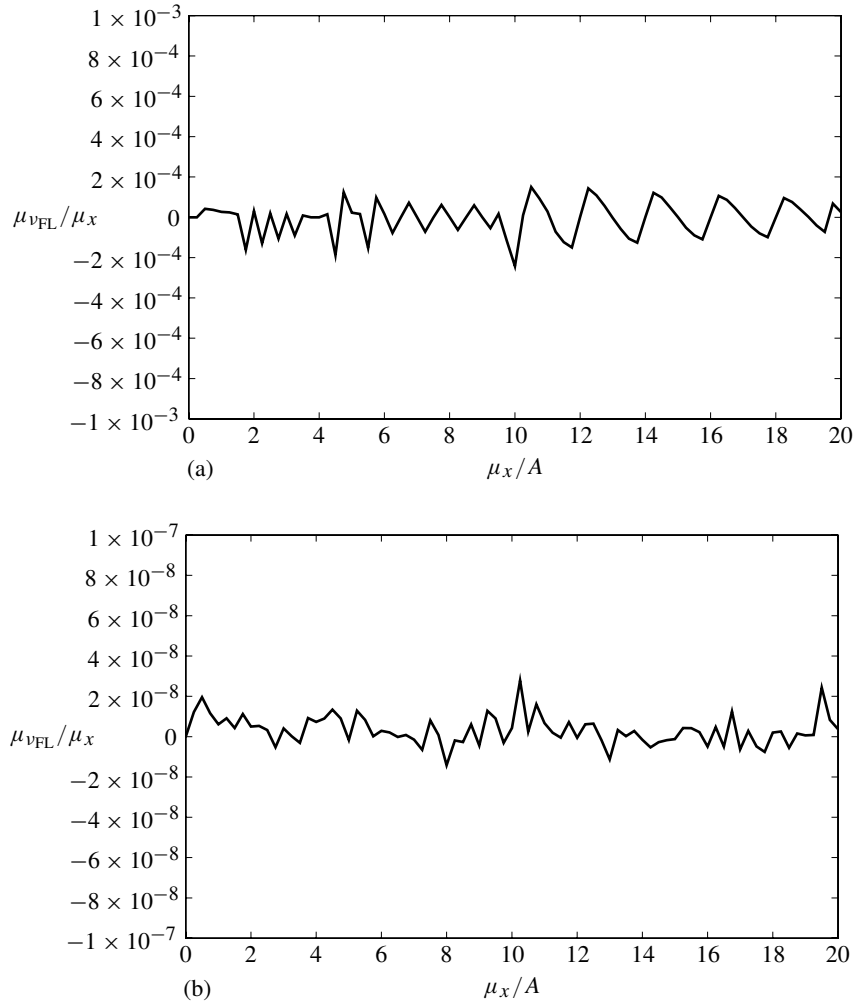


**Figure 12.39** Relative mean of the floating-point noise, $\nu_{\mathrm{FL}}/\mu_x$ versus $\mu_x/A$, for sinusoidal input with $A = 400\Delta$: (a) 8-bit mantissa; (b) 16-bit mantissa.

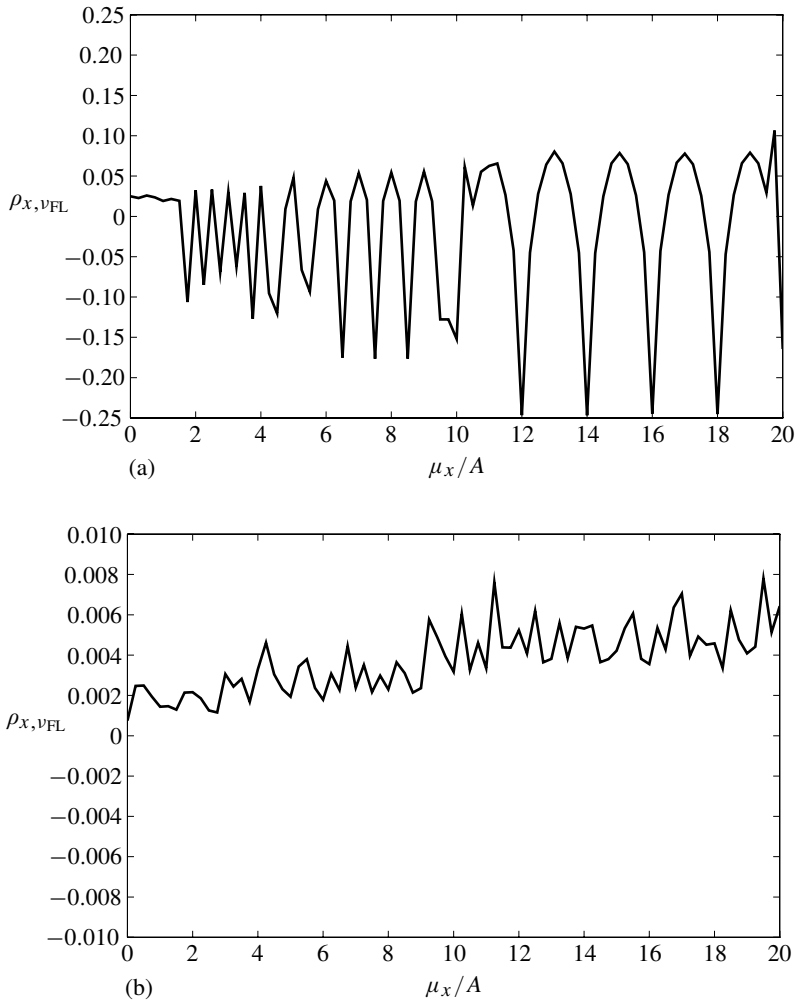


**Figure 12.40** Correlation coefficient of floating-point quantization noise and input $x$, $\rho_{v_{\mathrm{FL}},x}$ versus $\mu_x/A$ for sinusoidal input, with $A = 400\Delta$: (a) 8-bit mantissa; (b) 16-bit mantissa.

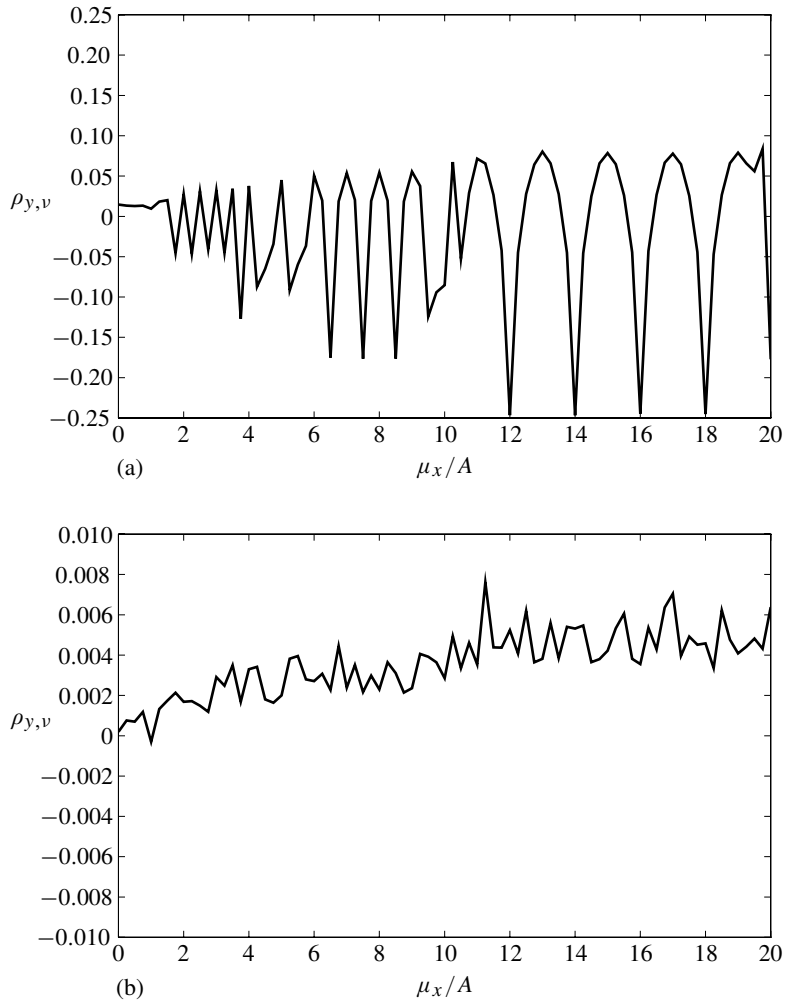


**Figure 12.41** Correlation coefficient of hidden quantization noise $v$ and hidden quantizer input $y$, $\rho_{v,y}$ versus $\mu_x/A$ for sinusoidal input with $A = 400\Delta$: (a) 8-bit mantissa; (b) 16-bit mantissa.

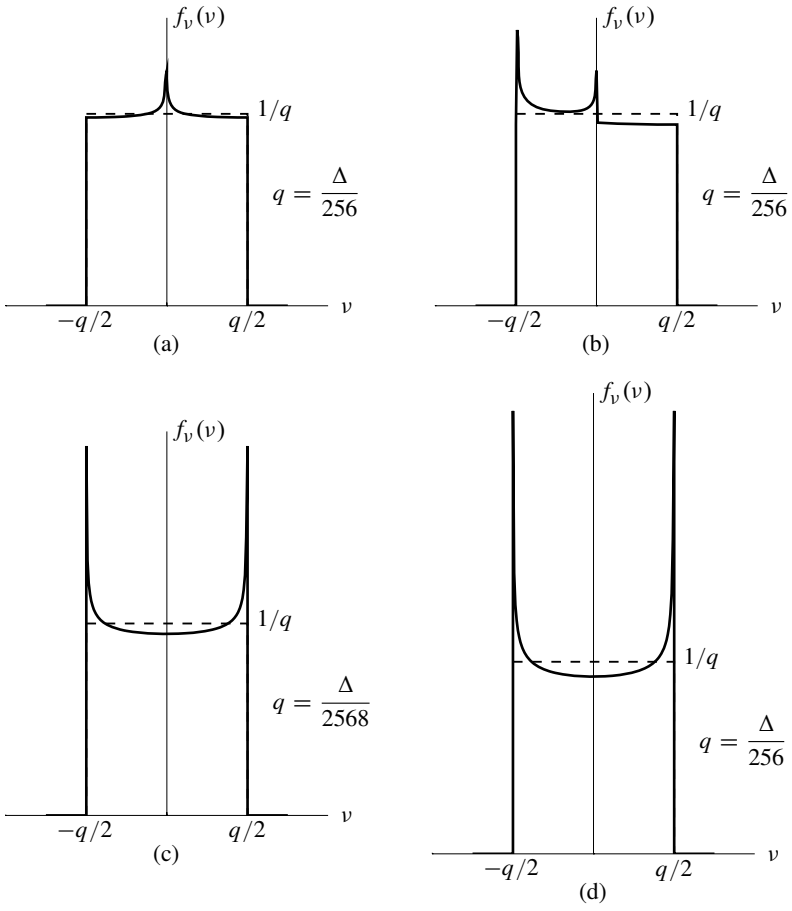


**Figure 12.42** Noise PDFs for the hidden quantizer, sinusoidal input, $A = 400\Delta$, 8-bit mantissas: (a) $\mu_x = 0$; (b) $\mu_x = 1.75A$; (c) $\mu_x = 6.5A$; (d) $\mu_x = 12A$.

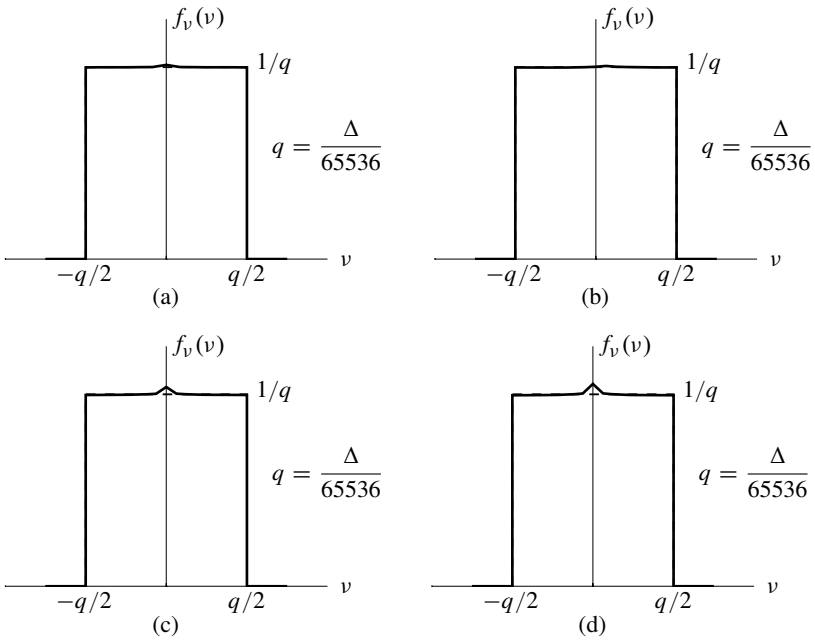


**Figure 12.43** Noise PDFs for the hidden quantizer, sinusoidal input, $A = 400\Delta$, 16-bit mantissas: (a) $\mu_x = 0$; (b) $\mu_x = 1.75A$; (c) $\mu_x = 9.5A$; (d) $\mu_x = 12A$.

form. PQN is not really satisfied. Fig. 12.43 shows PDFs of $\nu$ with a 16-bit mantissa. These PDFs are almost perfectly uniform, and it is safe to say that the PQN hypothesis for the hidden quantizer is very close to being perfectly true. Using it will give excellent analytical results.

## 12.7 A FLOATING-POINT PQN MODEL

When the hidden quantizer behaves in accord with a PQN model, the floating-point quantization noise $\nu_{\text{FL}}$ has zero mean, zero crosscorrelation between $\nu_{\text{FL}}$ and $x$, and a mean square value bounded by Eq. (12.27). For purposes of analysis, the floating-point quantizer can be replaced by a source of additive noise, i.e. floating-point PQN. This comprises a PQN model for the floating-point quantizer.

When the input $x$ is zero-mean Gaussian, the mean square of $\nu_{\text{FL}}$ is given by Eq. (12.24). The normalized floating-point quantization noise power (NFPQNP), defined by Eq. (12.35), is equal to the "magic number" 0.180. When the input $x$ is distributed according to either zero-mean triangular, zero-mean uniform, or zero-

mean sinusoidal, the NFPQNP is close to 0.180. For other cases, as long as the floating-point PQN model applies, the NFPQNP is bounded by

$$1/12 \leq \text{NFPQNP} \leq 1/3\,. \tag{12.37}$$

Comparing the PQN model for floating-point quantization with that for uniform quantization, two major differences are apparent: (a) With floating-point quantization, $\nu_{\text{FL}}$ has a skyscraper PDF, while with uniform quantization, $\nu$ has a uniform PDF. Both PDFs have zero mean. (b) With floating-point quantization, $\text{E}\{\nu_{\text{FL}}^2\}$ is proportional to $\text{E}\{x^2\}$, while with uniform quantization, $\text{E}\{\nu^2\} = q^2/12$, which is fixed. Both $\nu_{\text{FL}}$ and $\nu$ are deterministically related to their respective quantizer inputs, but they are both uncorrelated with these inputs. The fact that the quantization noises are uncorrelated with the quantizer inputs allows the quantizers to be replaced with additive PQN for purposes of calculating moments such as means, mean squares, correlation coefficients, and correlation functions.

## 12.8 SUMMARY

By modeling the floating-point quantizer as a compressor followed by a uniform quantizer followed by an expander, we have been able to analyze the quantization noise of the floating-point quantizer. Quantizing theorems, when satisfied, allow the use of a PQN model. When conditions for the PQN model are met, floating-point quantization noise has zero mean and it is uncorrelated with the input to the floating-point quantizer. Its mean square value is bounded by

$$\frac{1}{12} \cdot 2^{-2p} \cdot \text{E}\{x^2\} \leq \text{E}\{\nu_{\text{FL}}^2\} \leq \frac{1}{3} \cdot 2^{-2p} \cdot \text{E}\{x^2\}\,. \tag{12.27}$$

When, in addition, PQN conditions are met for the "exponent quantization"

$$Q_1\big(\log_2(|x|/\Delta) + 0.5\big),$$

the actual mean square of the floating-point quantization noise is given by

$$\text{E}\{\nu_{\text{FL}}^2\} = 0.180 \cdot 2^{-2p} \cdot \text{E}\{x^2\}\,. \tag{12.24}$$

When the input signal to a floating-point quantizer occupies a range of values so that the quantizer is neither underloaded or overloaded, PQN conditions are almost always met very closely. The bounds (12.27) will therefore apply. When the mantissa is sufficiently long, for example 16 bits or more, the exponent quantization condition will be almost always met very closely, and equation (12.24) will closely approximate $\text{E}\{\nu_{\text{FL}}^2\}$.

The signal-to-noise ratio of the floating-point quantizer is defined by

$$\text{SNR} \triangleq \frac{\text{E}\{x^2\}}{\text{E}\{\nu_{\text{FL}}^2\}}\,. \tag{12.28}$$

When the bounds (12.27) apply, the SNR is bounded by

$$12 \cdot 2^{2p} \geq \text{SNR} \geq 3 \cdot 2^{2p} . \qquad (12.31)$$

When the more exact equation (12.24) applies, the SNR is given by

$$\text{SNR} = 5.55 \cdot 2^{2p} . \qquad (12.29)$$

## 12.9 EXERCISES

**12.1** Two floating-point numbers are added. The first one is approximately 2, the other one is approximately 1. What is the probability that before roundoff for storage the fractional part of the sum exactly equals 0.5 times the least significant bit (LSB)? What is the probability of this when both numbers are approximately equal to 1.5?

**12.2** For zero-mean Gaussian noise with $\sigma = 12.5\Delta$ applied to the input of the floating-point quantizer with an 8-bit mantissa (see page 343), determine numerically

- **(a)** the PDF of $y$,
- **(b)** the PDF of $\nu$,
- **(c)** the variance of $\nu$,
- **(d)** the correlation coefficient of $y$ and $\nu$,
- **(e)** from the results of (b)–(d), does the PQN model apply to the hidden quantizer?

**12.3** Verify the results of Exercise 12.2 by Monte Carlo.

**12.4** For zero-mean Gaussian noise with $\sigma = 12.5\Delta$ applied to the input of the floating-point quantizer with an 8-bit mantissa, determine numerically

- **(a)** the PDF of $\nu_{\text{FL}}$,
- **(b)** the PDF of $\nu_{\text{EXP}}$.

Do these correspond to PQN?

**12.5** Verify the results of Exercise 12.4 by Monte Carlo.

**12.6** Repeat Exercise 12.2 with $\mu = 15\Delta$.

**12.7** Verify the results of Exercise 12.6 by Monte Carlo.

**12.8** Repeat Exercise 12.4 with $\mu = 150\Delta$.

**12.9** Verify the results of Exercise 12.8 by Monte Carlo.

**12.10** Is it possible that the distribution of $\nu_{\text{FL}}$ is uniform, if $x$ extends to more than one linear section of the compressor (that is, its range includes at least one corner point of the piecewise linear compressor)? If yes, give an example.

Is it possible that both $x$ and $\nu_{\text{FL}}$ have uniform distributions at the same time ?

**12.11** How large is the power of $\nu_{\text{FL}}$ when an input with normal distribution having parameters $\mu = 2^{p+1}$ and $\sigma = 2^{p+3}$ is applied to a floating-point quantizer with 6-bit mantissa ($p = 6$)? Determine this theoretically. Verify the result by Monte Carlo.

**12.12** The independent random variables $x_1$ and $x_2$ both are uniformly distributed in the interval (1,2). They are multiplied together and quantized. Determine a continuous approximation of the distribution of the mantissa in the floating-point representation of the product $x_1 x_2$. Note: this non-uniform distribution illustrates why the distribution of mantissas of floating-point calculation results usually cannot be accurately modeled by uniform distribution.