Appendix E

Limits of Applicability of the Theory – Caveat Reader

The theory described in this book is very general: it describes well most quantization and roundoff situations. However, assumptions have been made which are necessary for proper application. In this appendix, some of these will be briefly described.

E.1 LONG-TIME VS. SHORT-TIME PROPERTIES OF QUANTIZATION

Statistical theory of quantization deals with statistics of signals: PDFs, CFs and moments. The basic idea is coined at the beginning of Chapter 4: "Instead of devoting attention to the signal being quantized, let its probability density function be considered."

Having extensively explored the theory, a few basic questions need to be discussed, like:

- when may the PDF be used?
- what are the consequences of the use of the PDF in the application of the theory?
- what has to be done if the PDF may not used, e.g. when the signal is deterministic?

For the introduction of the probability density function, Fig. 3.1(a) shows an ensemble of random time functions. These random time functions are realizations of a *stochastic process*. In practice, usually just one realization is measured, thus averaging is performed as averaging in time – which means that time averages are assumed to be equal to ensemble averages. In this case, the process is called *ergodic* (see page 43).

Quantization theory deals with statistical properties of ergodic processes. Fulfillment of QT I ensures that the PDF of x can be determined from the PDF of x'. Fulfillment of QT I or QT II ensures that the moments of these processes follow Sheppard's corrections. The behavior of statistical descriptors is examined, characterizing time records of infinite length.

In practice, however, finite-length measured records are analyzed. Thus, applicability of our results depends on how much the moments and the PDF of the infinitely long input signal are relevant to our time record. In quantization theory, it is tacitly assumed that the PDF and the moments are usable for the measured data. This is usually true, but not in every case.

Moreover, for deterministic signals, the use of the PDF needs some explanation. A "PDF" can be assigned to the time domain signal, as if this was a random realization of a stochastic process. The ratio of the time while the signal stays in a given amplitude range is taken, relative to the full time length, as "probability":

$$f(x_0) = \lim_{\Delta x \to 0} \frac{\mathrm{P}(x_0 \le x(t) < x_0 + \Delta x)}{\Delta x} \triangleq \lim_{\Delta x \to 0} \frac{\frac{T_{x_0 \le x(t) < x_0 + \Delta x}}{T_{\text{record}}}}{\Delta x} . \tag{E.1}$$

By this, similar statements can be made as for stochastic signals. These will be valid for the whole (infinitely long) signal, since the PDF is also defined for the whole one.

Quantizing theorems refer to global properties: to PDF of the whole signal, or moments as results of averaging over the whole signal. Thus, from the quantization theory point of view, there is no difference e.g. among the signals illustrated in Fig. E.1(a)–(c). They are designed to have identical PDFs. However, it can be observed that while for the signal in Fig. E.1(a) the PDF remains the same if we take the half of the record, for Fig. E.1(b)–(c) this is not true. The statistical properties valid for the full record are not necessarily inherent to parts of it. The statements which are valid for the full record (like the QTs) are not necessarily valid for parts of the record. Therefore, in order to apply QTs and the PQN model with a sound basis, we must measure long enough records for which the PDF is really applicable. If this is not done for some reason, we need to look into short-time properties of quantized data.

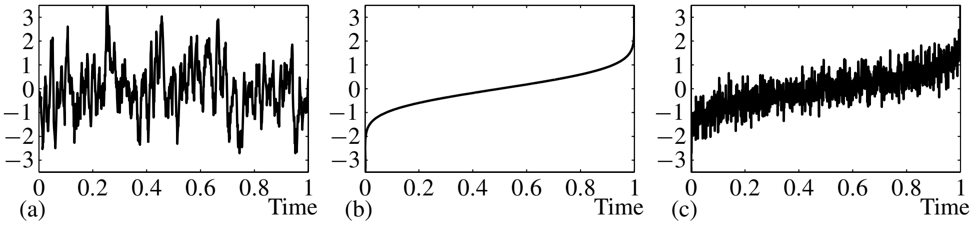

Figure E.1 Different time domain signals having the same PDF: (a) Gaussian signal; (b) inverse of the Gaussian distribution function; (c) sum of the two.

A good example of the contradiction between long-term and short-term properties becomes evident in digitally recorded music. PQN is usually perfectly applicable to the full recording. However, based on quantization theory, this is true *in average* only. Nothing is said about short-time behavior. What we hear as noise, however, is short-time (< 0.1 s) noise power. This should be small, and what is audible from it should stay at a constant level, independent of the input signal, otherwise it becomes annoying.

An example is shown in Fig. E.2. In Fig. E.2(a), a short-time record of a Gaussian distributed signal is shown. This has been quantized with $q = \sigma_x$, that is, the quantization theorem is well fulfilled. The quantized signal is shown in Fig. E.2(b). Between 1.4 s and 2 s, there is no change at the output, while between 2 s and 2.5 s there is noise, clearly audible in the audio frequency range.

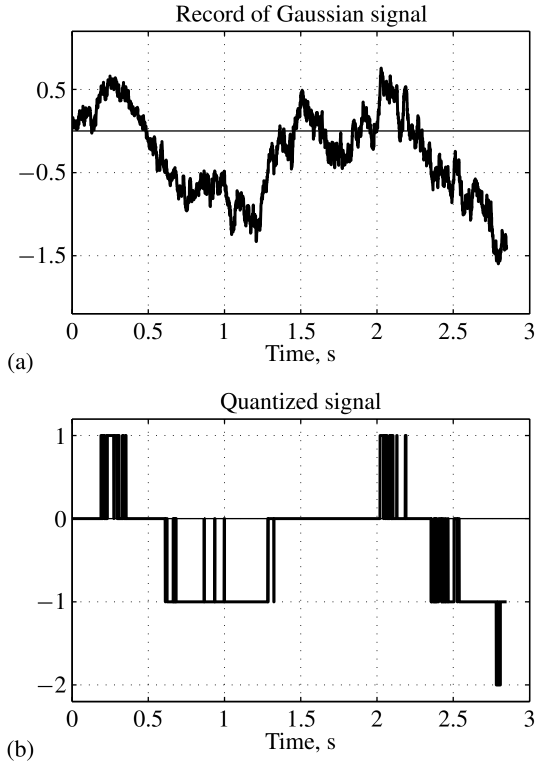

Figure E.2 Noise modulation: (a) time record of a Gaussian signal; (b) result of quantization with $q = \sigma_x$.

Short-time dependence of the moments of ν on x is possible even if their long-time correlation is zero, so it deserves analysis on its own.

E.1.1 Mathematical Analysis

The tool of this investigation is the conditional PDF of the quantization noise, with the condition being the instantaneous value of the input signal:

$$f_{\nu|x}(z) = \operatorname{rect}\left(\frac{z}{q}\right) \sum_{m=-\infty}^{\infty} \delta(x+z+mq)\,. \tag{E.2}$$

The conditional CF is the Fourier transform of this:

$$\begin{aligned}\Phi_{\nu|x}(u) &= \int_{-\infty}^{\infty} f_{\nu|x}(z)\, e^{jzu}\,\mathrm{d}z = \operatorname{sinc}\left(\frac{qu}{2}\right) \star \left(\sum_{l=-\infty}^{\infty} e^{-jxu}\delta(u+l\Psi)\right) \\ &= \sum_{l=-\infty}^{\infty} \operatorname{sinc}\left(\frac{q(u+l\Psi)}{2}\right) e^{jxl\Psi}\,. \end{aligned} \tag{E.3}$$

Evaluation of Eqs. (E.2) and (E.3) seems to be awkward, since because of the deterministic nature of quantization, these describe simple one-to-one relationships. However, for slightly random x, these functions will be smeared, and the functions of ν are obtained:

$$\begin{aligned} f_\nu(z) &= \int_{-\infty}^{\infty} f_{\nu|x}(z) f(x)\,\mathrm{d}x\,, \\ \Phi_\nu(z) &= \int_{-\infty}^{\infty} \Phi_{\nu|x}(z) f(x)\,\mathrm{d}x\,. \end{aligned} \tag{E.4}$$

The conditional moments of ν can be calculated by differentiating Eq. (E.3), and by setting u to zero:

$$\begin{aligned} \mathrm{E}\{\nu|x\} &= \frac{1}{j}\frac{\mathrm{d}\,\Phi_{\nu|x}(u)}{\mathrm{d}u}\bigg|_{u=0} \\ &= \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \frac{q}{2\pi j}\frac{(-1)^l}{l}\, e^{jxl\Psi} \\ &= \sum_{l=1}^{\infty} \frac{q}{\pi}\frac{(-1)^l}{l}\sin(xl\Psi)\,. \end{aligned} \tag{E.5}$$

$$
\begin{aligned}
\mathrm{E}\{\nu^2|x\} &= \frac{1}{j^2}\frac{\mathrm{d}^2\,\Phi_{\nu|x}(u)}{\mathrm{d}u^2}\bigg|_{u=0} \\
&= \frac{q^2}{12} + \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \frac{q^2}{2\pi^2}\frac{(-1)^l}{l^2}\, e^{jxl\Psi} \\
&= \frac{q^2}{12} + \sum_{l=1}^{\infty} \frac{q^2}{\pi^2}\frac{(-1)^l}{l^2}\cos(xl\Psi)\,.
\end{aligned} \tag{E.6}
$$

$$
\begin{aligned}
\mathrm{E}\{\nu\} &= \int_{-\infty}^{\infty} \mathrm{E}\{\nu|x\} f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \sum_{l=1}^{\infty} \frac{q}{\pi}\frac{(-1)^l}{l}\sin(xl\Psi) f(x)\,\mathrm{d}x\,, \\
\mathrm{E}\{\nu^2\} &= \int_{-\infty}^{\infty} \mathrm{E}\{\nu^2|x\} f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{q^2}{12} + \sum_{l=1}^{\infty} \frac{q^2}{\pi^2}\frac{(-1)^l}{l^2}\cos(xl\Psi) f(x)\,\mathrm{d}x\,, \\
&\;\;\vdots \\
M_{i,\nu} &= \int_{-\infty}^{\infty} M_{i,\nu|x} f(x)\,\mathrm{d}x, \quad i = 1, 2, \ldots
\end{aligned} \tag{E.7}
$$

The above expressions clearly show that the moments of the quantization noise may depend on x. This is not in contradiction with PQN theory described earlier. When x sufficiently varies, these dependencies will be smoothed out to zero. QT I or QT II are sufficient to eliminate moment dependency for records long enough to have reasonable PQN.

However, if x is almost constant in a finite-length interval, the variance will behave as illustrated in Fig. E.3(c). This is not very disturbing, since the peaks in the variance have zero width and x practically never takes this value. However, if x has a small variation, like in Fig. E.3(d)–(f), the variance shows noticeable changes. This phenomenon, the so-called noise modulation, is very annoying in audio applications when a slowly changing input value with a small additive noise causes a changing noise power. If this happens, an appropriate (e.g. triangular) dither has to be used (see Section 19.5) to assure short-time fulfillment of the QT conditions, especially for soft, low-level recordings.

It is of interest to analyze how the series in Eqs. (E.5) and (E.6) yield the functions in Fig. E.3. It is clear that the Fourier series in Eq. (E.5) is of a sawtooth. The form of (E.6) is more complex. It is straightforward to see that the two extreme values are 0 and $q^2/4$. These are the squares of the two bounds of $|\nu|$: for $x = 0$, $\nu = 0$, and its square is also 0; for $x = q/2$, the other extreme value is taken. As for the

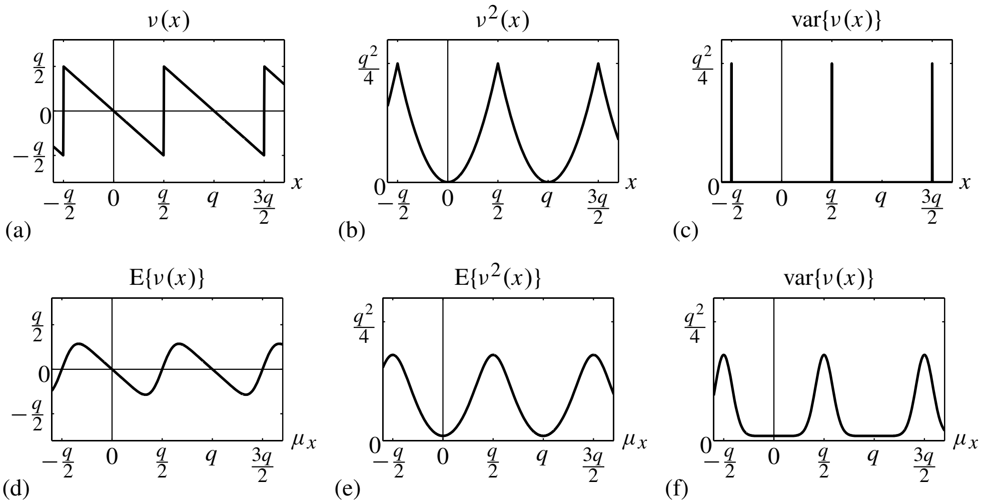

Figure E.3 Dependence of functions of the quantization error on the input variable x:
(I) x =const: (a) $\nu(x)$; (b) $\nu^2(x)$; (c) var$\{\nu(x)\} = \nu^2(x) - (\mathrm{E}\{\nu(x)\})^2$;
(II) noisy input, x normally distributed with $\sigma_x = q/10$, dependence on μ_x: (d) E$\{\nu(x)\}$;
(e) E$\{\nu^2(x)\}$; (f) var$\{\nu(x)\}$.

sum, with $x = 0$ the cosine is one, and the sum gives $-q^2/12$; with $x = q/2$, the cosine is equal to $(-1)^l$, and it is easy to check that

$$\sum_{l=1}^{\infty} \frac{q^2}{\pi^2} \frac{1}{l^2} = \frac{q^2}{6}\,, \tag{E.8}$$

and $q^2/12 + q^2/6 = q^2/4$.

When QT II is fulfilled, the moments take the PQN values, $\mathrm{E}\{\nu\} = 0$, and $\mathrm{E}\{\nu^2\} = q^2/12$. This is true because the moments can be calculated according to Eqs. E.7, and this smoothing of the conditional moments yields the above values.

As a summary of the above discussion, it can be stated that small variability of x, like in short-time records, will not efficiently smear out x-dependent (and time-dependent) behavior of moments of ν, but a variability fulfilling QT I or QT II will smear them out. Short-time average values of the moments of ν may depend on beginning of the time record.

E.2 SATURATION EFFECTS

In the statistical theory of quantization, it is assumed that the quantization characteristic has the same staircase form of infinite length in both directions.

In practice, the finite amplitude range of the input signal restricts the range of interest, therefore it is enough if the quantization characteristic is uniform where it is really used. It may be *thought* that the characteristic continues similarly at places where it is not used.

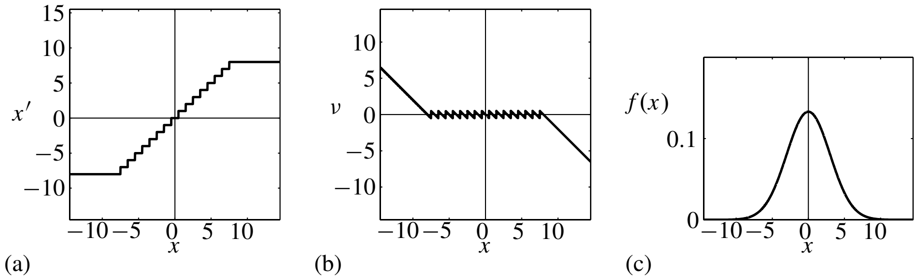

Figure E.4 Saturation in a real quantizer, with $N = 15$: (a) quantizer characteristics; (b) quantization error characteristics; (c) PDF of an input signal.

In the previous investigations, it was simply assumed that although the range of the quantization characteristic is always limited in practice, the input signal is scaled properly to avoid saturation. In this section, this scaling is briefly discussed.

In scaling for quantization, two requirements need to be satisfied:

- the input signal must be small enough to avoid saturation,
- the input signal must be large enough to fulfill an appropriate quantizing theorem, and also to assure good signal-to-noise ratio with respect to the quantization noise.

Therefore, it seems to be best to have the largest possible input signal which still avoids saturation. However, most PDFs, like the Gaussian or the Laplacian, have infinite range, so in theory, saturation will always happen.

This is not as bad as it sounds. Distributions with infinite range are only models of reality with limited precision. For finite sample size, the central limit theorem assures good approximation of the bell curve at the main lobe, but approximation becomes poor at the tails. Therefore, behavior of the tails is not a good model anyhow.

This is true, but it does not give guidance where is the limit above which it may be assumed that the probability is zero: over $2\sigma_x$, $3\sigma_x$, or more? For the Gaussian distribution, this is not a very difficult dilemma, since a few times σ_x is generally a good choice (e.g. $\mathrm{P}\{|x - \mu_x| > 3\sigma_x\} \approx 0.997$). For other distributions, it may be more difficult to determine a good rule.

A systematic approach can be to maximize the SNR, by selecting the ratio of the uniform quantization range, $N \cdot q$ (where N is the number of quantizing intervals), and of the standard deviation of the input signal, for quantization characteristics like in Fig. E.4(a). For fixed N and σ_x, this means proper selection of q in

$$\max_{Nq/\sigma_x} \text{SNR} = \max_{Nq/\sigma_x} 10\log_{10}\left(\frac{\text{E}\{x^2\}}{\text{E}\{\nu^2\}}\right) , \tag{E.9}$$

Thus, there is a small probability that saturation occurs, but its effect is relatively small. Figure E.5 illustrates the optimal ratio as a function of the number of quantization levels N, for two common distributions. It can be observed that in both cases, the optimal saturation level increases with increasing number of quantization intervals. In other words, the range to be quantized is larger when more quantization levels are used. It is surprising to note that the required uniform quantization range seems to increase with N without bound. This is indeed the case for many distributions, as proved by Hui and Neuhoff (2001).

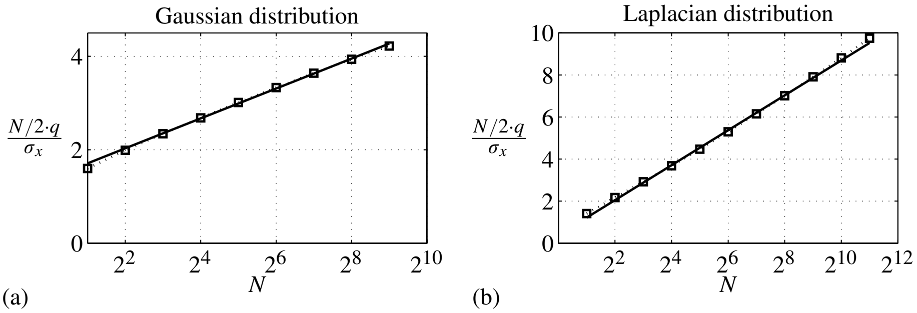

Figure E.5 Optimal setting of uniform quantization (a) for normal distribution, $\frac{N/2\cdot q}{\sigma_x} \approx 1.39 + 0.32\log_2 N$; (b) for Laplacian distribution, $\frac{N/2\cdot q}{\sigma_x} \approx 0.39 + 0.83\log_2 N$.

For the Gaussian distribution, the number of levels $N = 2^8$ requires coverage of the interval $\pm 4\sigma_x$ ($q = 8\sigma_x/2^8$ in this case). This range is larger than it was anticipated above, and increases further with increasing N. This makes us reconsider our common belief of the quantization of the Gaussian distribution.

Proper selection of the uniformly quantized range is important because while the selection of a too fine grain size is attractive, it may easily cause the bias caused by overload to dominate over quantization error.

E.3 ANALOG-TO-DIGITAL CONVERSION: NON-IDEAL REALIZATION OF UNIFORM QUANTIZATION

Quantization most commonly occurs in analog-to-digital converters. These usually have uniform step size, thus it could be thought that the theory in this book applies to them without any further considerations.

Unfortunately, this is not the case. Real ADCs approximately perform uniform quantization, but they are prone to production errors. The characteristics approximate uniform quantization, but in signal processing, the approximation errors must be considered.

In the specification of ADCs, the deviation of their characteristics is given with certain quality figures. Here differential and integral nonlinearities will be briefly discussed. They are both given in LSB, as:

- integral nonlinearity (INL): a measure in LSB of the worst-case deviation from the ideal A/D transfer characteristic. The INL is the worst-case value found by subtracting the actual code transition locations from an end-point or a best-straight-line fit of the characteristics. Each of these approaches can yield different numbers for the same data converter, and do not account for offset and drift.
- differential nonlinearity (DNL): a measure in LSB of the worst-case deviation of quantization interval widths from the ideal size (q). An ideal converter would have each interval exactly the same size, and a DNL of 0 (zero).

Let us discuss a few examples.

Example E.1 QT III in A/D Converters
When QT III/A is fulfilled, theory provides that the mean value of the quantization noise is zero (that is, $\mathrm{E}\{x'\} = \mathrm{E}\{x\}$). However, this is not true when the quantizer is not perfectly uniform. Let us assume that the differential nonlinearity is 0.5 LSB, and the signal is uniformly distributed in an interval of width q. QT III/A is fulfilled for the nominal q, however, according to the specification, there can be a quantization interval which is $2q$ wide (0.5 LSB error on both sides). Consequently, the A/D converter is insensitive to the position of the signal within the interval. The error of the mean value may change between $\pm 0.5q$.

Example E.2 Error of the Mean Value in ADCs
Let us assume that the specified INL is 0.5 LSB. This means that even if the differential nonlinearity is very small (e.g. 0.05 LSB), the measured mean value may differ from the true one by the maximum specified INL. That is, the measured mean value of a signal may be off by $0.5q$, even with a dither normally distributed over several quantization intervals (that is, if $\sigma_x > q$, which is theoretically enough to eliminate the quantization bias).

Nonlinearities in the step sizes are randomly distributed along the staircase. However, it is not assured that random distribution of these nonlinearity errors can be exploited. When the signal (or the signal+dither) extends several quantum boxes, it would be expected that these errors cancel each other, or at least more or less average out. With 0.5 LSB DNL, and randomly positioned errors, one might expect that the mean value of a signal uniformly distributed over 30 quantum boxes, the mean value of the measured signal has an error of $0.5q/\sqrt{30} < 0.1q$. This is unfortunately not true. There is more than one reason for this:

- Nothing assures that the errors work against each other. It is possible that in the range of interest, *all* quantum levels are off by 0.5 LSB, thus the mean value is also off by this quantity.
- Averaging will not help to decrease the error. Even if the error sources are randomly appearing, the characteristic of an ADC is deterministic, thus repeated experiments are prone repeatedly to the same nonlinearity error.

Nonlinear errors can even cause harmonic products above the quantization noise level. In order to gain a feeling of this, let us calculate the following example.

Example E.3 Harmonic Distortion Due to Nonlinearity
Let us assume that the integral nonlinearity of an ADC is 0.3 LSB. A sinusoidal signal $x(t) = A\cos(2\pi f_1 t + \phi)$, $A = 5q$, is coherently sampled ($f_1 = kf_s/N$, that is, no leakage happens in the DFT). Let us consider the spectral error for $k = 7$, $N = 512$, if the comparison levels are

$$\{y_{-4.5}, y_{-3.5}, y_{-2.5}, y_{-1.5}, y_{-0.5}, y_{0.5}, y_{1.5}, y_{2.5}, y_{3.5}, y_{4.5}\} = \{-4.8q, -3.5q, -2.2q, -1.3q, -0.4q, 0.6q, 1.7q, 2.7q, 3.8q, 4.3q\}.$$

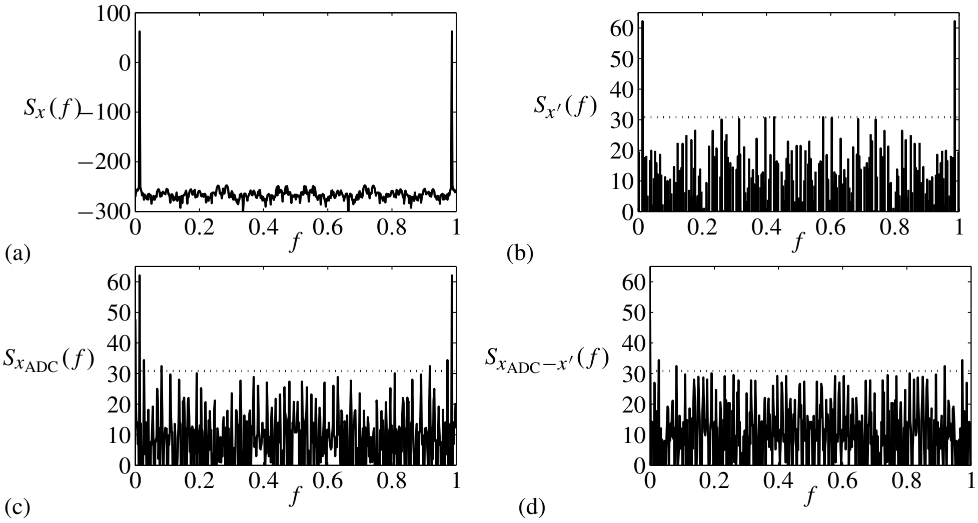

Figure E.6 Spectra of sine wave and quantized sine wave: (a) spectrum of the non-quantized sine; (b) spectrum of the ideally quantized sine; (c) spectrum of the output of the ADC; (d) difference of the spectrum of the output of the ADC and of the ideally quantized sine.

From Fig. E.6 it is observable that the nonlinearity of the ADC causes a peak of 3.5 dB over the largest peak of the quantization error of the ideal quantizer, moreover, at a place where almost no power was present in the quantized sine wave.

Example E.4 Error in a Successive Approximation ADC
An attractive structure for making an ADC is successive approximation. The basic algorithm is as follows. The input sample is compared to half of the allowed input amplitude range value (this is called reference value). If the input sample is larger, then this value is subtracted from the input sample, and bit one is stored, otherwise bit 0 is stored. In the next step, comparison with half of the above reference value follows. By repetition, bits of the converted value are determined in succession.

Let us assume that in a 10-bit successive approximation ADC, each subtraction has an error of 0.2 LSB. Then, the worst case error after 9 subtractions is about 1.8 LSB. The error of successive approximation ADCs grows rapidly with the number of stages.

High-resolution ADCs are implemented using principles which circumvent this error accumulation, like sigma-delta ADCs (see the Addendum at this book's website).