Chapter 17

Roundoff Noise in Digital Feedback Control Systems

Digital control systems are generally feedback systems. Within their feedback loops are parts that are analog and parts that are digital. At certain points within the feedback loop, signals are converted from analog to digital and at other points, signals are converted from digital to analog. Analog-to-digital conversion involves both sampling and quantization. Digital-to-analog conversion involves both quantization and interpolation (see Section 17.2). Analysis of quantization noise in digital control systems must therefore deal with quantization within feedback loops having both digital and analog parts.[1]

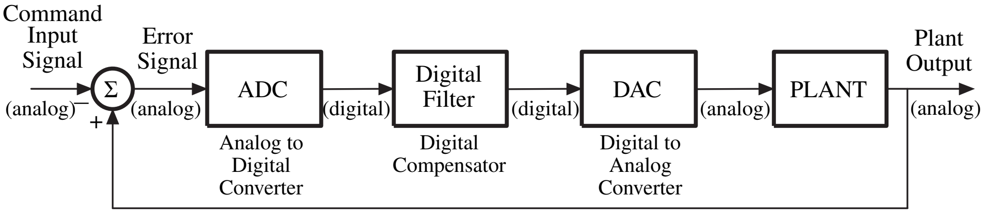

Figure 17.1 A unity-feedback digital control system.

An example of a digital control system is shown in Fig. 17.1. This is a unity-feedback system whose purpose is to have the plant output follow as closely as possible the command input. The plant is the system to be controlled. The command input and the plant output are analog signals. Their difference is the error signal.

[1] This chapter is designed to be self-contained and to present a clear picture of how to analyze the effects of quantization in digital feedback control systems. However, prior study of Chapter 16 on quantization in IIR filters would be very helpful, and prior knowledge of digital control theory would also be helpful. An excellent reference on digital control is the book of Franklin, Powell and Workman (1992).

This error signal is digitized, amplified, and filtered by a digital compensating filter whose output drives a digital-to-analog converter whose output in turn drives the plant input. The goal is to keep the error signal small.

17.1 THE ANALOG-TO-DIGITAL CONVERTER

The analog-to-digital converter (ADC) performs the operations of sampling and quantization. Use of the ADC makes it possible to do the compensator filtering digitally. This often has many advantages over analog filtering. The design of the compensator has important effects on stability of the system and its overall dynamic response.

The analog-to-digital converter is diagrammed in Fig. 17.2. Figure 17.2(a) shows sampling first, then quantization. Figure 17.2(b) shows quantization first, then sampling. From an analytical standpoint, these operators are commutable. Analog-to-digital conversion can be represented either way.

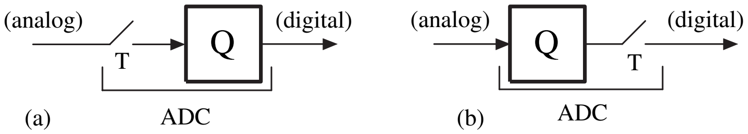

Figure 17.2 Analog-to-digital conversion: (a) sample first, then quantize; (b) quantize first, then sample.

17.2 THE DIGITAL-TO-ANALOG CONVERTER

The digital-to-analog converter (DAC) takes a sampled signal as input, and produces a continuous signal as its output. This output is an interpolation of its input samples. It may have discontinuities in amplitude and/or derivative at the sampling instants, however.

The most common form of DAC is the “zero-order hold.” How this DAC works is illustrated in Fig. 17.3. The figure shows a continuous analog signal being sampled, and the samples being interpolated by a zero-order hold. This form of DAC holds the value of an input sample and produces a corresponding constant output until the next input sample arrives. Then it holds the new input sample and produces a corresponding constant output. The DAC output is a series of step values, a “piecewise constant” signal.

Other forms of DAC do exist, but they are not so common as the zero-order hold. A better DAC would e.g. produce an analog output that would consist of straight line interpolations between the samples. This would generally provide a closer re-creation of the original analog signal. It has the disadvantage of producing

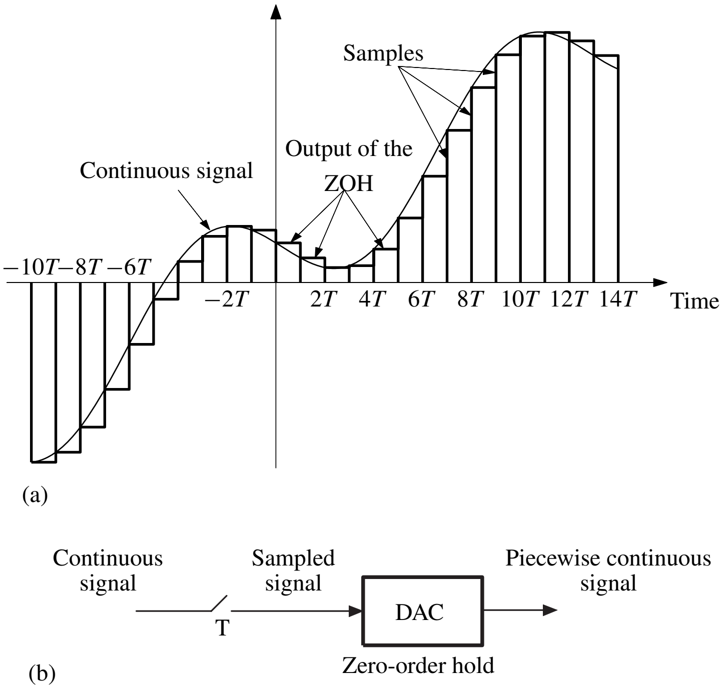

Figure 17.3 Sampling a continuous signal, and interpolation of samples by a zero-order hold: (a) time signals; (b) block diagram.

a delayed response since it is necessary to wait for the next sample in order to do the interpolation.

Analog-to-digital converters generally come on the same circuit board or in the same electronic package with digital-to-analog converters. Similar electronic technology is used in the realization of both converters. The ADC accepts an analog input, generally a voltage signal, and produces an output which is a series of numerical sample values, generally represented in binary form by a finite number of bits, every T seconds. Thus, the ADC output is quantized. The sampling period is T seconds. The zero-order DAC in turn accepts an input every T seconds which is quantized in amplitude and is a series of numerical sample values in binary form. It produces an output that is a piecewise continuous voltage waveform.

Zero-order digital-to-analog conversion involves both quantization and interpolation as illustrated in Fig. 17.4(a),(b). These operations are commutable from an analytical standpoint. The DAC of the zero-order-hold type has an internal digital memory that holds the previous sample value until the next sample arrives. This memory has a given word length, i.e., it stores numerical values rounded to a fi-

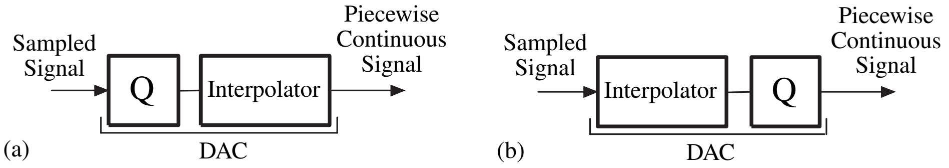

Figure 17.4 Digital-to-analog conversion: (a) quantization and interpolation; (b) interpolation and quantization.

nite number of bits. The binary input to the interpolator in Fig. 17.4(a) is therefore quantized. The interpolator itself is linear.

17.3 A CONTROL SYSTEM EXAMPLE

The control system shown in Fig. 17.5(a) is of the same form as that of Fig. 17.1. It will serve as a specific example for the development of mathematical techniques for analysis of the effects of quantization within control loops that are part analog and part digital. For this example, the plant has a transfer function of $\left(\frac{1}{s(s+1)}\right)$, the digital compensating filter has a transfer function of $0.2\left(\frac{z+1/2}{z-1/2}\right)$, and the sampling period T is 0.1 s. This system was simulated with high precision, so that quantization effects would be negligible, and the step response from command input to plant output is shown in Fig. 17.5(b). The system is stable, with damped oscillations, and has unit gain at zero frequency.

Figure 17.6(a) shows more detail about the analog-to-digital converter, and Fig. 17.6(b) shows how this function can be rearranged in a more convenient form for purposes of analysis. In Fig. 17.6(a), analog-to-digital conversion is represented by sampling and quantization with sampling first, then quantization. Since sampling is linear (the samples of the sum are equal to the sum of the samples), the sampler after the summer can be removed and replaced by two samplers, one in line with the command input and the other in line with the feedback signal. This is done without changing the system response, and is shown in Fig. 17.6(b). The two samplers must sample synchronously, at the same instants.

Generally, the word length at the output of the compensator in Figs. 17.6(a),(b). would be much greater than the length of the binary words fed as inputs to the digital-to-analog converter. The output of the compensator filter might have 16-bit or perhaps 32-bit accuracy, or the compensator could be realized with double-precision floating-point, for example. In any event, the compensator output is assumed to be essentially infinitely fine in amplitude. The DAC might have only 10-bit accuracy. The compensator output samples, although in digital form, would need to be rounded to fit the range of the DAC. The quantizer that does this can be associated with the DAC and it is labeled Q_2. The quantizer associated with the ADC is labeled Q_1.

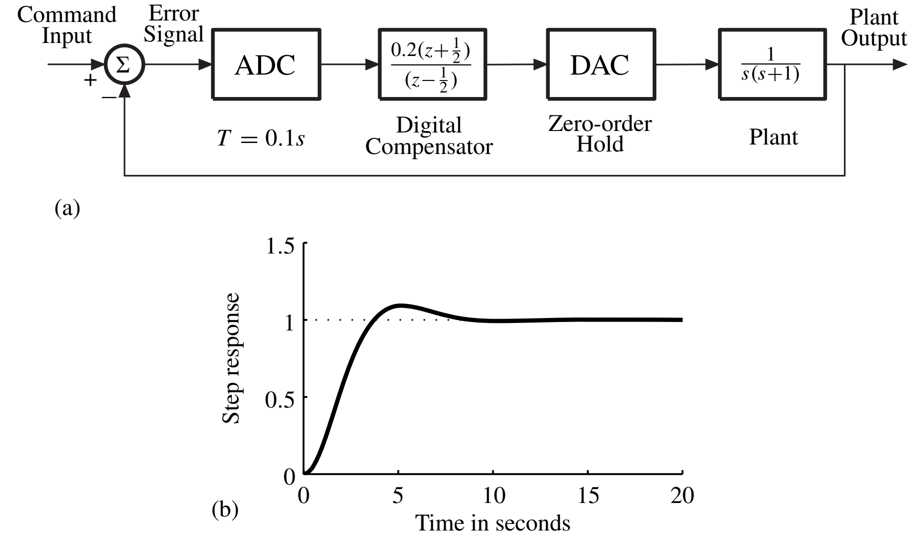

Figure 17.5 A unity-feedback example: (a) block diagram; (b) high-precision step response.

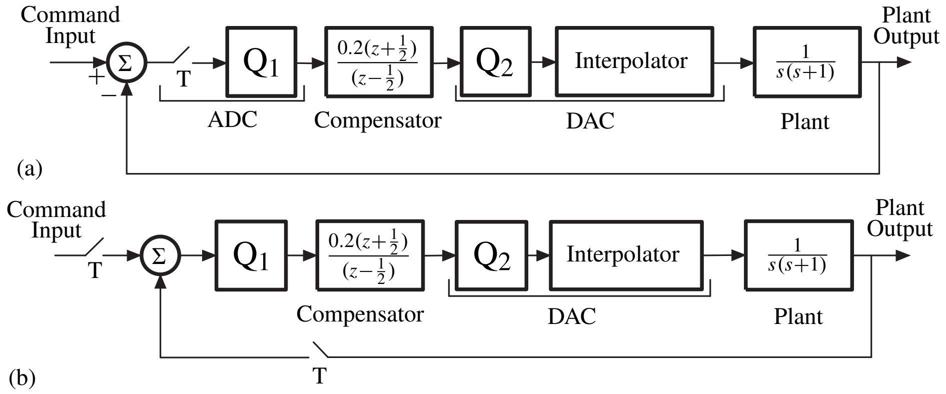

Figure 17.6 Replacement of error signal sampling with synchronous sampling of the command input and the feedback: (a) sampling after the summer; (b) sampling of the command input and the feedback signal.

Both quantizers are fixed-point since present day ADC devices quantize uniformly, and DAC devices produce uniform steps of output voltage. Both quantizers are of course nonlinear devices. On the other hand, the sampler and the zero-order hold interpolator perform linear operations on signals.

Figure 17.6(b) can be redrawn as shown in Fig. 17.7(a), to provide both the continuous-time and the sampled plant output. One can easily see the equivalence of the systems of Fig. 17.6(b) and Fig. 17.7(a).

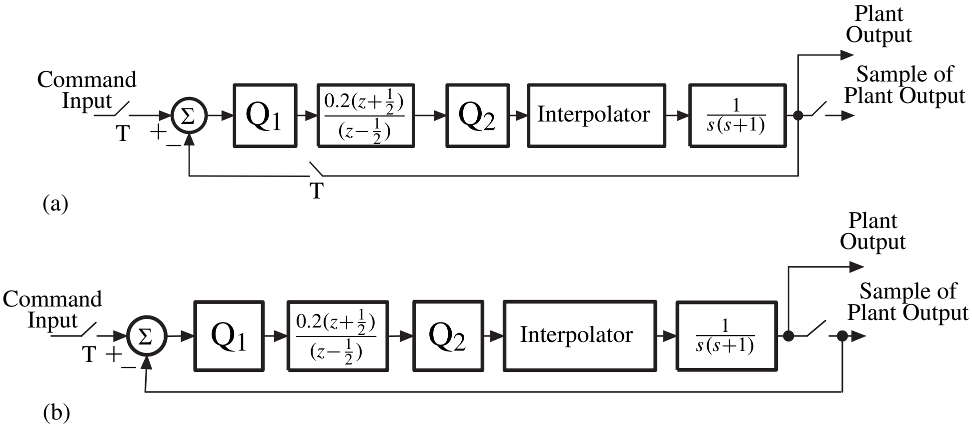

Figure 17.7 A reconfiguration of the system of Fig. 17.6 which obtains the plant output and samples of the plant output: (a) obtaining samples of the plant output; (b) feedback of the samples to the summer.

Figure 17.7(b) is equivalent to Fig. 17.7(a). It is this drawing that will be most useful for the analysis of quantization noise in the plant output signal. This noise will be present in the samples of the plant output, and it will be possible to calculate the variance of the noise in the output samples.

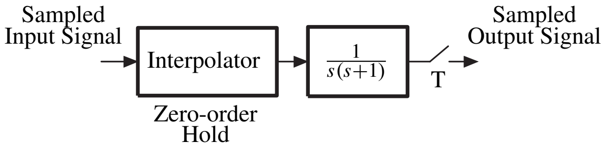

Figure 17.8 A subsystem of Fig. 17.7.

Before proceeding further, it will be useful to take an aside and study the behavior of the subsystem shown in Fig. 17.8. This subsystem appears within the feedforward part of the loop shown in Fig. 17.7(b). This subsystem is linear and it does not involve quantization. Its input is sampled, and its output consists of samples. Because it is linear and its input and output are sampled, the entire subsystem can be replaced by a single linear digital filter having a z-transform transfer function. This digital filter will have an impulse response that can be obtained by applying a single unit impulse at time zero to the input of the subsystem, and observing the string of impulses that then will emerge at the subsystem output.

First we apply a unit impulse to the input of the zero-order interpolator. Its output will be a single rectangular pulse of unit height and of duration T, as illustrated in Fig. 17.9. This rectangular pulse will be applied as input to $\frac{1}{s(s+1)}$.

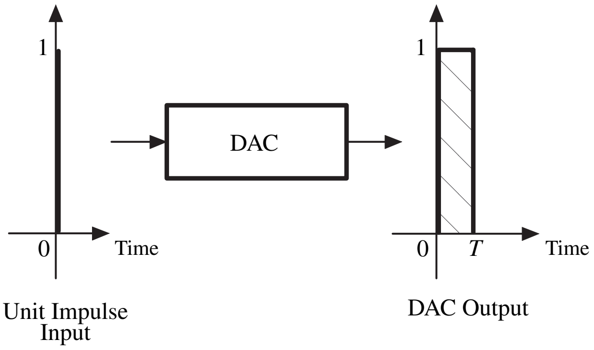

Figure 17.9 Impulse response of the zero-order hold interpolator.

The next step is to find the response of $\frac{1}{s(s+1)}$ to the rectangular pulse. The rectangular pulse can be regarded as a sum of a positive unit step function and a negative unit step function delayed by one sample period T. This is illustrated in Fig. 17.10. The response of $\frac{1}{s(s+1)}$ to the rectangular pulse will be found by calculating its step response, and subtracting from it this step response delayed by T.

The step response of $\frac{1}{s(s+1)}$ has a Laplace transform $\frac{1}{s^2(s+1)}$, which can be expressed by the following partial fraction expression:

$$\frac{1}{s^2(s+1)} = \frac{1}{s^2} - \frac{1}{s} + \frac{1}{s+1}\,. \tag{17.1}$$

In the time domain, this step response is

$$\text{step response} = \begin{cases} t - 1 + e^{-t}\,, & t \geq 0\,, \\ 0\,, & t < 0\,. \end{cases} \tag{17.2}$$

The step response is sampled every T seconds, with $T = 0.1$ s. The z-transform of the samples of the step response is

$$\begin{aligned}\begin{pmatrix} z\text{-transform} \\ \text{of step response} \end{pmatrix} &= Tz^{-1}\left(\frac{1}{1-z^{-1}}\right)^2 - \left(\frac{1}{1-z^{-1}}\right) + \left(\frac{1}{1-e^{-T}z^{-1}}\right) \\ &= \frac{(T+e^{-T}-1)z^{-1} + (1-e^{-T}-Te^{-T})z^{-2}}{(1-z^{-1})^2(1-e^{-T}z^{-1})}\end{aligned}$$

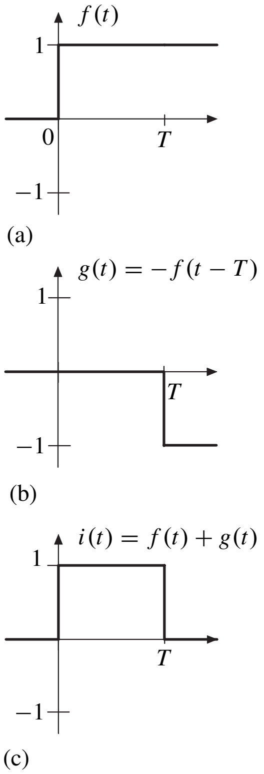

Figure 17.10 The rectangular pulse as a sum of a positive step and a negative delayed step.

$$= \frac{0.00484z^{-1} + 0.00468z^{-2}}{(1-z^{-1})^2(1-0.905z^{-1})} \,. \tag{17.3}$$

This result can now be used to find the z-transform of the samples of the rectangular-pulse response. Since the rectangular pulse is the sum of a positive step and a negative step delayed by one sample period, the desired z-transform is the z-transform of Eq. (17.3) multiplied by $(1 - z^{-1})$. Therefore, the z-transform of the samples of the output of $\frac{1}{s(s+1)}$ when its input is the rectangular pulse of Fig. 17.10 is

$$\begin{pmatrix} z\text{-transform} \\ \text{of response} \\ \text{to rect pulse} \end{pmatrix} = \frac{0.00484z^{-1} + 0.00468z^{-2}}{(1-z^{-1})(1-0.905z^{-1})} \,. \tag{17.4}$$

Figure 17.11(a) shows the subsystem described above. A unit impulse applied to this subsystem causes a string of output samples whose z-transform is given by Eq. (17.4). Figure 17.11(b) shows an equivalent linear discrete filter whose impulse response is identical to that of the subsystem. The equivalent linear filter is a digital filter whose transfer function may be designated by $H_{\mathrm{EQ}}(z)$. This transfer function is, from Eq. (17.4), given by

$$H_{\mathrm{EQ}}(z) = \frac{0.00484z^{-1} + 0.00468z^{-2}}{(1 - z^{-1})(1 - 0.905z^{-1})} . \tag{17.5}$$

A good reference on z-transforms and on the zero-order hold is the book of Franklin, Powell and Workman (1992).

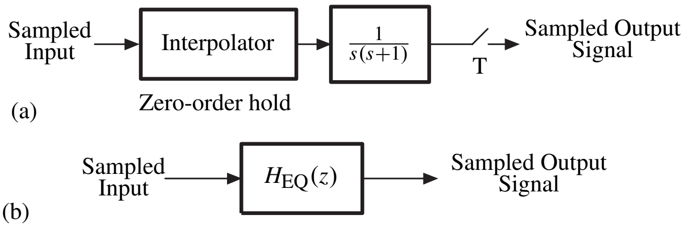

Figure 17.11 Equivalence between a subsystem of Fig. 17.7 and a simple digital filter: (a) the subsystem of Fig. 17.7; (b) an equivalent linear digital filter.

The two systems shown in Fig. 17.11 have the same impulse response and transfer function. Since they are linear systems, any train of input pulses applied to both will cause identical trains of output pulses. Therefore, the subsystem of Fig. 17.11(a) can be replaced by the digital filter of Fig. 17.11(b) having the transfer function given by Eq. (17.5). This substitution was made to the system of Fig. 17.7(b), and the resulting system is shown in Fig. 17.12. The feedback loop of the system of Fig. 17.12 is now all-discrete and is in a form that allows easy analysis of the effects of quantization.

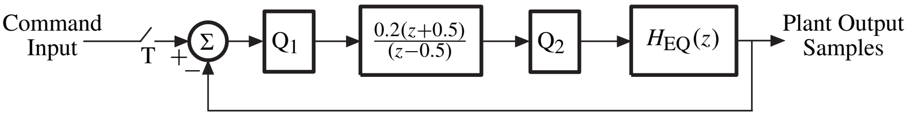

Figure 17.12 System devised for the calculation of samples of the plant output.

In most circumstances, QT II will be satisfied to very close approximation at the inputs of both of the quantizers shown in Fig. 17.12. The two quantizers will be injecting into the control system stationary white noises whose statistical properties will be pre-determined and known. For purposes of calculation of quantization noise variance at the plant output, the two quantizers may be replaced by two sources of additive PQN. This change has been made with the system of Fig. 17.12, and the result is shown in Fig. 17.13. The system of Fig. 17.13 is linear, and may be used to calculate quantization noise power at the plant output. It may also be used to determine stability of the control system and to obtain the step response from the command input to the samples of the plant output.

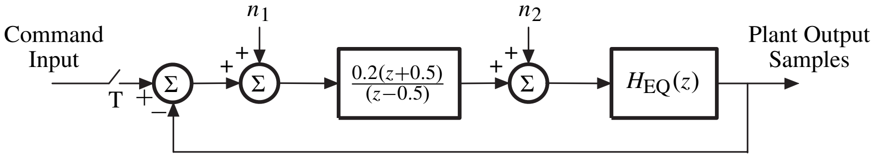

Figure 17.13 System devised for the calculation of step response, output noise power, and stability of the control system.

Referring to Fig. 17.13, the discrete transfer function from input samples to output samples is

$$
\begin{aligned}
H_{\text{io}}(z) &= \frac{0.2\dfrac{z+0.5}{z-0.5}\,H_{\text{EQ}}(z)}{1+0.2\,\dfrac{z+0.5}{z-0.5}\,H_{\text{EQ}}(z)} \\
&= \frac{0.2\dfrac{z+0.5}{z-0.5}\,\dfrac{0.00484(z+0.967)}{(z-1)(z-0.905)}}{1+0.2\dfrac{z+0.5}{z-0.5}\,\dfrac{0.00484(z+0.967)}{(z-1)(z-0.905)}} \\
&= \frac{0.000968(z+0.5)(z+0.967)}{(z-0.4933)(z-(0.955-0.0603j))(z-(0.955+0.0603j))}\,.
\end{aligned}
\tag{17.6}
$$

The discrete impulse response from input to output, corresponding to the transfer function $H_{\text{io}}(z)$, is plotted in Fig. 17.14(a). It is apparent from this plot that the system is stable and slightly underdamped. The discrete step response of this transfer function is plotted in Fig. 17.14(b). The static error for a step input is zero. This results from integration that exists within the plant itself.

Quantization noise at the plant output comes from both quantizers, Q_1 and Q_2. Referring to Fig. 17.13, the corresponding noise sources n_1 and n_2 are injected into the feedback system and propagate to the system output. In accord with our assumption that QT II is satisfied at the inputs of both quantizers, noise n_2 will be uncorrelated with the input to quantizer Q_2 and will therefore be uncorrelated with noise n_1. Since n_1 and n_2 are uncorrelated with each other, their respective responses at the system output will also be uncorrelated with each other. Since these noises have zero means and are mutually uncorrelated, the mean square of the output quantization noise will be the sum of the mean squares of the noises due to n_1 and n_2. The two noise components can be calculated separately and then combined. The

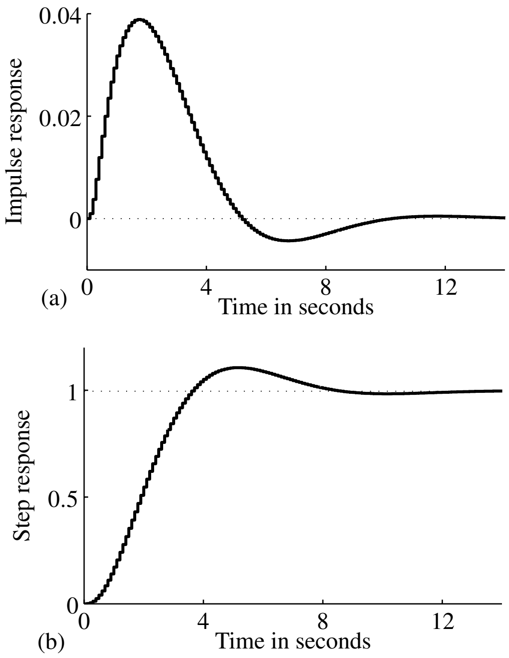

Figure 17.14 Discrete impulse and step responses of the system of Fig. 17.13: (a) impulse response; (b) step response.

methods used here are very similar to those used in Chapter 16 to determine output quantization noise with IIR digital filters.

The impulse response from the point of injection of noise n_1 to the system output is the same as the discrete impulse response from the system input to output. The transfer function is the same as Eq. (17.6), and is therefore

$$H_{n_1 o}(z) = \frac{0.000968(z+0.5)(z+0.967)}{(z-0.4933)(z-(0.955-0.0603j))(z-(0.955+0.0603j))}\,. \tag{17.7}$$

The impulse response was calculated by inverse z-transform, and the sum of squares of the impulses of this impulse response was calculated to be

$$\begin{pmatrix} \text{sum of squares} \\ \text{of impulse response} \\ \text{from } n_1 \text{ to output} \end{pmatrix} = 0.0333\,. \tag{17.8}$$

The noise injected by Q_1 has a mean square value of $q_1^2/12$, where q_1 is the step size of Q_1. Therefore the mean square of the output quantization noise due to Q_1 is

$$\begin{pmatrix} \text{mean square of} \\ \text{output noise due} \\ \text{to } Q_1 \end{pmatrix} = 0.0333 \frac{q_1^2}{12} \,. \tag{17.9}$$

The transfer function from the point of injection of noise n_2 to the system output is the same as $H_{n_1 o}(z)$ multiplied by the reciprocal of $\frac{0.2(z+0.5)}{(z-0.5)}$. This is

$$H_{n_2 o} = \frac{0.00484(z-0.5)(z+0.967)}{(z-0.493)(z-(0.955-0.0603j))(z-(0.955+0.0603j))} \,. \tag{17.10}$$

The sum of squares of the impulses of the impulse response from n_2 to the system output has been calculated as

$$\begin{pmatrix} \text{sum of squares} \\ \text{of impulse response} \\ \text{from } n_2 \text{ to output} \end{pmatrix} = 0.0935 \,. \tag{17.11}$$

Accordingly,

$$\begin{pmatrix} \text{mean square of} \\ \text{output noise due} \\ \text{to } Q_2 \end{pmatrix} = 0.0935 \, \frac{q_2^2}{12} \,. \tag{17.12}$$

The mean square of the total quantization noise in the plant output can now be calculated. This is

$$\begin{pmatrix} \text{mean square of} \\ \text{output noise due} \\ \text{to } \mathrm{Q}_1 \text{ and } \mathrm{Q}_2 \end{pmatrix} = \frac{0.033 q_1^2 + 0.0935 \, q_2^2}{12} \,. \tag{17.13}$$

The plant output is continuous. The quantization noise in the plant output is also continuous. The mean square of the continuous quantization noise in the plant output is given approximately by Eq. (17.13), as this is the mean square of the output quantization noise observed at the sampling instants.

17.4 SIGNAL SCALING WITHIN THE FEEDBACK LOOP

An electronic analog-to-digital converter has a finite dynamic range. The maximum input voltage that would result in an output having the largest binary number is specified by the manufacturer. The negative input voltage that causes the smallest binary output number is similarly specified. The number of bits in the output number is also specified. One bit is the sign bit. The remaining bits correspond to the magnitude of the output number.

A 10-bit ADC would have 2^{10} quantum steps covering its input range. If the input range were ±5 V for example, each quantum step would have a voltage value of

$$q = \frac{10\,\mathrm{V}}{2^{10}} = \frac{10\,\mathrm{V}}{1024} = 0.009766\,\mathrm{V}\,. \tag{17.14}$$

Suppose that a 10-bit ADC with an input range of ±5 V were to be used with the feedback control system shown in Fig. 17.1. Suppose further that the details of the system are as specified in the diagram of Fig. 17.6(a). Now, let the command input be a sinusoid of amplitude 5 $\mathrm{V_{RMS}}$. Assume that the frequency of this sinusoid could range from near zero up to half the sampling frequency.

Our first objective will be to design the system so that quantizer Q_1 will not be overloaded, but that the maximum range of its input voltage will be just within its design range ±5 V. This would minimize the effects of quantization noise without overloading the quantizer. The range of the command input is $\pm 5\,\mathrm{V} \cdot \sqrt{2}\mathrm{V} = \pm 7.07\,\mathrm{V}$. We will first need to find the gain or the value of the transfer function magnitude from input to error (input of Q_1) at low frequency, and at the highest frequency, i.e., half the sampling frequency.

The transfer function from samples of the command input to the input of Q_1 (the error signal) can be obtained from the block diagrams of Fig. 17.12 and 17.13. Using the feedback formula, this is

$$\begin{aligned} &H_{\mathrm{ie}}(z) \\ &= \frac{1}{1 + \frac{0.2(z+0.5)}{(z-0.5)} H_{\mathrm{EQ}}(z)} \\ &= \frac{1}{1 + \frac{0.2(z+0.5)(0.00484z^{-1} + 0.00468z^{-2})}{(z-0.5)(1-z^{-1})(1-0.905z^{-1})}} \\ &= \frac{(1-z^{-1})(z-0.5)(1-0.905z^{-1})}{(1-z^{-1})(z-0.5)(1-0.905z^{-1}) + 0.2(z+0.5)(0.00484z^{-1} + 0.00468z^{-2})} \end{aligned} \tag{17.15}$$

This transfer function is a function of frequency, since $z = e^{-sT}$, and with sinusoidal inputs, $z = e^{-j\omega T}$ where ω is the input sinusoidal frequency in radians per second. At low frequencies, $\omega \approx 0$ and therefore $z \approx 1$. Accordingly,

$$H_{\mathrm{ie}}(z)\Big|_{\text{low frequency}} \approx 0\,. \tag{17.16}$$

This corresponds to the error being zero at zero frequency, which is a characteristic of feedback systems with integration within the loop (the plant has a pole at $s = 0$).

The transfer function is shown in Fig. 17.15. At low frequency, the 10 V sinusoidal input will cause an error of essentially zero volts. This will not overload Q_1. It may in fact cause an underload problem, which we will address below. But in any event, Q_1 will not be overloaded.

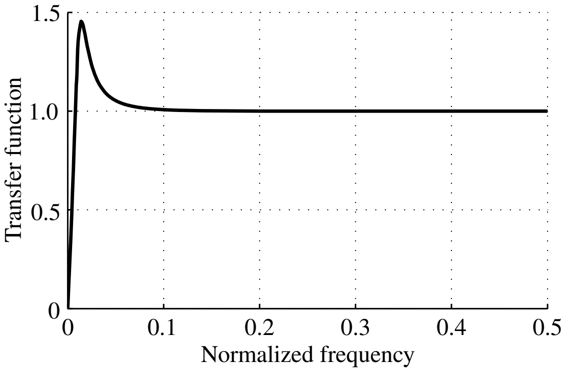

Figure 17.15 Transfer function from the command input to the input of Q_1.

At the highest frequency, corresponding to half the sampling frequency, $\omega = \pi/T$. This gives a value of $z = e^{-j\omega T} = e^{-j\pi} = -1$. At this frequency, the transfer function is

$$H_{\text{ie}}(z)\Big|_{\substack{\text{half the sampling}\\ \text{frequency}}} = 1.000\,0028 \approx 1\,. \tag{17.17}$$

The fact that the input-to-error transfer function has a unit value means that the plant output is zero and that the system does not respond at such a high frequency.

The maximum of the transfer function is about 1.45, at $f =\approx 0.015 f_{\text{s}}$.

To keep the quantizer Q_1 from overloading, its input signal level will need to be reduced. As is, this input level could be 1.45 times larger than that of the command input, i.e., within the range of ± 10.25 V. Accordingly, it will be necessary to insert an analog attenuator having the gain of $5/10.25 = 0.49$. Let this gain be designated by $a = 0.49$. To restore the proper gain level in the system, an additional gain of $1/a = 2.04$ will need to be incorporated into the digital compensator. Figure 17.6(a) will now appear as shown in Fig. 17.16.

This design assures that the ADC is not overloaded for any sine wave at the command input with amplitude below 7.07 V. However, overload is still possible with other signal forms for which $|x(t)| < 0.707$ V holds. In Fig. 17.12, the maximum sample at the input of Q_1 occurs when all samples of the command input are equal to either -7.07 V, or $+7.07$ V, so that the sample equals to $\sum_k |h_{\text{ADC}}(k)| \cdot 7.07$ V, where $h_{\text{ADC}}(k)$ denotes the impulse response from the command input to the input of

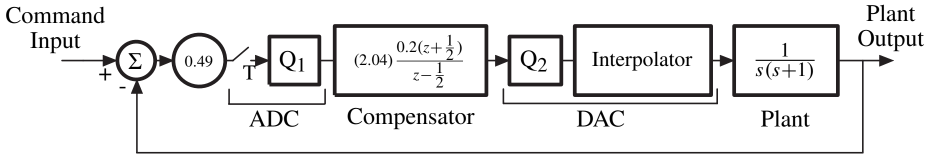

Figure 17.16 The control system of Fig. 17.6(a) with signal scaling to prevent overload of Q_1 with a 5 V_{RMS} sinusoidal input.

Q_1. Numerical calculation shows that this is equal to 15.0 V. Therefore, theoretically a gain of 1/3 (and a compensation gain of 3) should be introduced instead of 0.49 and 2.04, respectively, to have absolutely no overload.

The quantizer Q_2 is not a physical quantizer. It is an arithmetic quantizer. It is implemented digitally by the same computer that implements the digital compensator. This compensator produces a binary output that will usually have many more bits than the binary input to the DAC. As such, the most significant bits of the compensator output will be mapped directly as input bits to the DAC. The compensator output bit corresponding to the least significant DAC input bit will be rounded up or down by the computer, based on the remaining compensator output bits of lower significance.

We have been assuming that the compensator is implemented with fixed-point arithmetic. If arithmetic is done instead in floating-point, then the computer will make a conversion from floating-point to fixed-point with appropriate roundoff to drive the DAC input. The function of the quantizer Q_2 will be unchanged since the DAC has a uniform quantization scale.

For best operation of this system, we will keep the quantizer Q_2 from just overloading for the various inputs of interest. Working further with the current example, we will assume that the digital-to-analog converter is an 8-bit DAC and that its output voltage range is ±5 V. The problem is to scale the input of Q_2 so that there is no overload for the 5 V_{RMS} command input sinusoids at low frequency and at half the sampling frequency.

The transfer function from samples of the command input to the input of Q_2 can be obtained from the block diagrams of Figs. 17.12 and 17.13. (For this calculation, Q_2 has a gain of unity.) Using the feedback formula, this transfer function is

$$H_{iQ_2}(z) = \frac{0.2(1-z^{-1})(1-0.905z^{-1})(z+0.5)}{(1-z^{-1})(z-0.5)(1-0.905z^{-1})+0.2(z+0.5)(0.00484z^{-1}+0.00468z^{-2})}\,. \tag{17.18}$$

At low frequencies, $z \approx 1$ and

$$H_{iQ_2}(z)\Big|_{\text{low frequency}} = 0\,. \tag{17.19}$$

At half the sampling frequency, $z = -1$ and

$$H_{\mathrm{iQ}_2}(z)\Big|_{\substack{\text{half the sampling}\\ \text{frequency}}} = 0.0667\,. \tag{17.20}$$

The complete transfer function is shown in Fig. 17.17. The maximum value is about 0.87 .

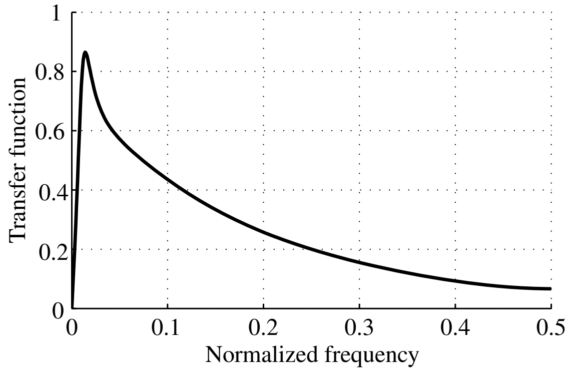

Figure 17.17 Transfer function from the command input to the input of Q_2.

At this frequency, a ±7.07 V peak-to-peak input would therefore cause an output for Q_2 of $(\pm 7.07\,\mathrm{V}) \cdot (0.87) = \pm 6.15\,\mathrm{V}$. This is close to the limit of ±5 V for the DAC output, but some downscaling is necessary to really keep it within the limits. Let us multiply the gain of the digital controller by $5/6.15 = 0.813$, and follow the DAC by a compensating analog gain of $6.15/5 = 1.23$. This is illustrated in Fig. 17.18. For this system, both the ADC and DAC will run without overload with a $5\,\mathrm{V_{RMS}}$ command input, and they will not overload at half the sampling frequency.

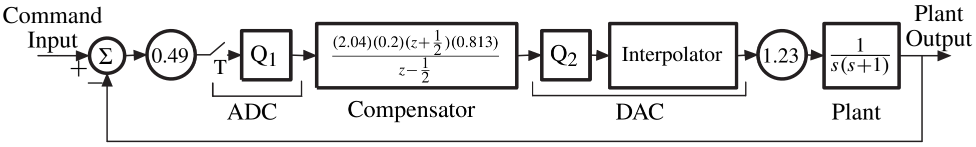

Figure 17.18 The control system of Fig. 17.6(a) with signal scaling to prevent overload of both Q_1 and Q_2 with a $5\,\mathrm{V_{RMS}}$ sinusoidal input.

This design assures that the DAC is not overloaded for any sine wave at the command input with amplitude below 7.07 V. However, overload is still possible with other signal forms for which $|x(t)| < 0.707\,\mathrm{V}$. In Fig. 17.12, the maximum sample at the input of Q_2 occurs when all samples of the command input are equal to either $-7.07\,\mathrm{V}$, or $+7.07\,\mathrm{V}$, so that the sample equals to $\sum_k |h_{\mathrm{DAC}}(k)| \cdot 7.07\,\mathrm{V}$, where $h_{\mathrm{DAC}}(k)$ denotes the impulse response from the command input to the input of Q_2. Numerical calculation shows that this is equal to 9.12 V. Therefore, theoretically a gain of 0.55 (and a compensation gain of 1.82) should be introduced instead of 0.813 and 1.23, respectively, to have absolutely no overload.

At very low frequency, both the ADC and DAC will have extremely small inputs and will be underloaded. This means that their inputs will correspond to less than a single quantum step or less than a very small number of steps. Underloaded feedback systems could develop highly nonlinear behavior that could result in static errors in the feedback loop or could result in spontaneous finite small amplitude oscillations, called limit cycles. However, as soon as the amplitudes of the inputs to both ADC and DAC have amplitudes covering several quantum boxes, linear behavior is restored and limit cycles disappear.

When a step function is applied to the input of the system of Fig. 17.18, the response is essentially linear until the basic transient is over and steady-state is reached. Then, a static error will exist between the input and the plant output. The size of the error depends upon the amplitude of the input step and system initial conditions, and will be bounded (see Section 17.7 below). This particular system does not develop limit cycles. In general, both static error and limit cycles can be prevented by injection of external dither signals. This is illustrated with homework Exercises 17.10–17.12.

The dither is a zero-mean external signal or noise added to the command input. Its purpose is to prevent static error, hysteresis, or limit cycles by linearizing the quantizers, causing satisfaction of QT II at the quantizer inputs. The subject of linearization by addition of dither is treated in detail in Chapter 19.

17.5 MEAN SQUARE OF THE TOTAL QUANTIZATION NOISE AT THE PLANT OUTPUT

To calculate the output noise for the system of Fig. 17.18, we will assume that QT II is satisfied at the inputs of both Q_1 and Q_2. The quantizers can now be replaced by additive noise sources as shown in Fig. 17.19.

The system of Fig. 17.19 is very similar to that of Fig. 17.13. The injected noises n_1 and n_2 are not the same, respectively, for both of these systems, but they have correspondingly the same moments so the same noise designations have been used for both systems.

The transfer function from the point of injection of n_1 to the plant output is given by Eq. (17.7) for the system of Fig. 17.13. The corresponding transfer function for the system of Fig. 17.19 is the same, except that it must be scaled by the factor 2.04 . Accordingly,

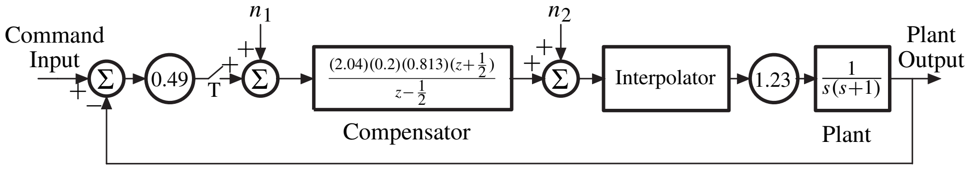

Figure 17.19 The system of Fig. 17.18 with quantizers replaced by sources of additive noise.

$$H_{n_1\text{o}}(z) = \frac{(2.04)0.000968(z+0.5)(z+0.967)}{(z-0.4933)(z-(0.955-0.0603j))(z-(0.955+0.0603j))} \,. \tag{17.21}$$

The sum of squares of the impulses of the corresponding impulse response is the same as Eq. (17.8), except that it must be scaled by $2.04^2 \approx 4.16$. Accordingly,

$$\begin{pmatrix}\text{sum of squares}\\ \text{of impulse response}\\ \text{from } n_1 \text{ to output}\end{pmatrix} = 0.0333 \cdot (2.04^2) = 0.1386 \,. \tag{17.22}$$

The mean square of the output quantization noise due to Q_1 is therefore

$$\begin{pmatrix}\text{mean square}\\ \text{of output noise}\\ \text{due to } \mathrm{Q}_1\end{pmatrix} = 0.1386\frac{q_1^2}{12} \,. \tag{17.23}$$

The mean square of the output quantization noise due to Q_2 may be similarly calculated by multiplying Eq. (17.12) by 1.23^2. Accordingly,

$$\begin{pmatrix}\text{mean square}\\ \text{of output noise}\\ \text{due to } \mathrm{Q}_2\end{pmatrix} = 0.1415\, q_2^2/12 \,. \tag{17.24}$$

For the scaled system of Fig. 17.19, the mean square of the total quantization noise in the plant output is

$$\begin{pmatrix}\text{mean square}\\ \text{of output noise}\\ \text{due to } \mathrm{Q}_1 \text{ and } \mathrm{Q}_2\end{pmatrix} = \frac{0.1386\, q_1^2 + 0.1415\, q_2^2}{12} \,. \tag{17.25}$$

The analog-to-digital converter is a 10-bit ADC whose output range is 10 V (±5 V). This converter has $2^{10} = 1024$ steps. Each step gives an increment to the converter output of $10/1024 = 0.00977\,\mathrm{V}$. Therefore, $q_1 = 0.00977\,\mathrm{V}$. The digital-to-analog converter is an 8-bit DAC whose output range is also $10\,\mathrm{V}$ ($\pm 5\,\mathrm{V}$). Each step gives an increment to its output of $10\,\mathrm{V}/2^8 = 10\,\mathrm{V}/256 = 0.0391\,\mathrm{V}$. Therefore, $q_2 = 0.0391$ V. Using these values of q_1 and q_2, Eq. (17.25) becomes

$$\begin{aligned}\begin{pmatrix}\text{mean square}\\ \text{of output noise}\\ \text{due to } \mathrm{Q}_1 \text{ and } \mathrm{Q}_2\end{pmatrix} &= \frac{(0.1386)\cdot 0.00977^2 + (0.1415)\cdot 0.0391^2}{12}\,\mathrm{V}^2 \\ &= 1.91\cdot 10^{-5}\ \mathrm{V}^2\,.\end{aligned} \tag{17.26}$$

The output quantization noise power is given by Eq. (17.26), and this will remain true for a very wide range of input amplitudes and frequencies, as long as QT II is satisfied to a good approximation at the inputs of Q_1 and Q_2.

The output signal-to-noise ratio can now be calculated. Suppose that the command input is a $5\,\mathrm{V_{RMS}}$ low-frequency sine wave. The output will also be a $5\,\mathrm{V_{RMS}}$ sine wave of the same low frequency, plus quantization noise. The output signal-to-noise ratio will be

$$\begin{pmatrix}\text{output}\\ \text{SNR}\end{pmatrix} = \frac{5^2}{1.91\cdot 10^{-5}} = 1.31\cdot 10^6\,. \tag{17.27}$$

Since Eq. (17.27) is a ratio of output signal power to output quantization noise power, the output SNR may be expressed in decibels as follows:

$$\begin{pmatrix}\text{output}\\ \text{SNR}\end{pmatrix} = 10\log_{10}\left(1.31\cdot 10^6\right) = 61.2\,\mathrm{dB}\,. \tag{17.28}$$

The output quantization noise power will be the same if the amplitude of the sinusoidal input is reduced, as long as it is not made so small that QT II is not satisfied at the inputs of Q_1 and Q_2. With a fixed level of quantization noise power, the output SNR will be reduced if the amplitude of the sinusoidal input is reduced. The SNR given by Eq. (17.28) is the best that is achievable by the system of Fig. 17.19. At higher input frequencies, the plant output drops off, but the quantization noise persists, and the output SNR drops off accordingly.

17.6 SATISFACTION OF QT II AT THE QUANTIZER INPUTS

Conditions for satisfaction of QT II or QT III will generally not be met perfectly at the quantizer inputs. The special PDFs that permit exact satisfaction of QT III, such as uniform, triangular, etc., will not exist at the quantizer inputs within feedback loops of signal processing and control systems. These conditions could be met, however, if external independent additive dither signals capable of satisfying QT II or QT III were applied at the quantizer inputs (see Chapter 19). Without external dither, QT II or QT III will not be perfectly applicable to the two quantizers.

Even when conditions for QT II are not met perfectly, they will be approximately met in most circumstances so that the second-order statistics of the output quantization noise can be estimated with a high degree of reliability. For this purpose, the two quantizers can be replaced by two sources of additive white PQN noise. When this is done, the system is modeled as a linear system with additive independent output noise having a mean of zero and a fixed variance that is not dependent on characteristics of the control signals in the loop. This linear model works almost always.

The linear model does break down however when the amplitudes of the control signals in the loop at the quantizer inputs become small enough. The condition is underload. If and when this happens, the control system exhibits highly nonlinear behavior and analysis is very difficult. Otherwise, the system behavior is almost perfectly linear and easy to analyze.

Thus it is useful to be able to determine command input signal levels that will permit approximate satisfaction of QT II. To explore this question, we will consider two kinds of command inputs, sinusoidal and Gaussian.

For the system of Fig. 17.18, a 5 V_{RMS} sinusoidal command input at approximately half the sampling frequency will cause a full-scale sinusoidal signal of ±5 V at the input of quantizer Q_1, and the same at the input of quantizer Q_2. Recall that Q_1 is part of a 10-bit ADC and that Q_2 is part of an 8-bit DAC. As the command input amplitude is reduced, breakdown in linearity will be caused first by Q_2 since its quantum step size is four times larger than that of Q_1.

Quantizer Q_2 has a total of $2^8 = 256$ steps. It will behave in a manner very close to linear as long as its sinusoidal input covers a few tens of quantum steps or more. A 5 V_{RMS} $\approx$7.07 V_{peak} command input at about $0.01 f_s$ will cause all 256 steps to be covered. Reducing this input by a factor of 25 to about 0.28 V_{peak} will allow coverage of approximately 10 steps. Quantizer Q_2 will behave essentially linearly. So will Q_1 and so will the whole system. The plant output will have additive quantization noise having the mean square value given by Eq. (17.26). Reducing the command input much below 0.4 V will cause underload and nonlinear response.

If the amplitude of the command input is kept at the 5 V_{RMS} level but the frequency is reduced to near zero, the amplitudes of sinusoidal input components to both quantizers will go to near zero and QT II will break down. The only way to have linear behavior at very low frequency or zero frequency would be to inject external dither at the quantizer inputs.

If the command input signal is stochastic, the various transfer functions can once again be used to determine if the conditions for satisfaction of QT II are met. The input could be Gaussian, or it could have some other form of PDF. In most cases, the signals in the loop due to the command input would be essentially Gaussian at the quantizer inputs. This results from filtering. Digitally filtered signals are linear combinations of present and past inputs. If the input signal components are stochastic, linear combinations of stochastic signals regardless of PDF make an approximately Gaussian sum, in accord with the Central Limit Theorem.

Consider the system of Fig. 17.12. Let the command input be a continuous Gaussian signal whose samples taken at the system sampling rate are uncorrelated with each other over time, i.e., whose samples are white. Let these samples have a variance of σ^2. This command input will cause input components at Q_1 and Q_2 that are Gaussian and correlated over time. We need to find the variances of these quantizer input components.

The transfer function $H_{\mathrm{ie}}(z)$ from samples of the command input to the sampled error signal, i.e., the sampled input to Q_1, is given by Eq. (17.15). The inverse z-transform of $H_{\mathrm{ie}}(z)$ was taken to find the corresponding impulse response, and the sum of squares (SOS) of the impulses of this impulse response was calculated to be

$$\mathrm{SOS}_1 = 2.516\,. \tag{17.29}$$

The variance of the component at the input of Q_1 due to the random command input is therefore

$$\sigma^2 \cdot \mathrm{SOS}_1\,. \tag{17.30}$$

The standard deviation is

$$\sigma \cdot \sqrt{\mathrm{SOS}_1}\,. \tag{17.31}$$

For the conditions of QT II to be met with close approximation at the input of quantizer Q_1, it will be sufficient that this standard deviation be equal to or greater than the quantum step size of Q_1, i.e.,

$$\sigma \cdot \sqrt{\mathrm{SOS}_1} \ge q_1\,. \tag{17.32}$$

Meeting the condition of Eq. (17.32) will assure that the quantization noise of Q_1 will behave moment-wise like PQN and that this noise will be essentially white, i.e., uncorrelated over time.

The transfer function $H_{\mathrm{iQ}_2}(z)$ from samples of the command input to the samples input to Q_2 is given by Eq. (17.18). The sum of squares of the impulses of this impulse response has been computed to be

$$\mathrm{SOS}_2 = 0.1103\,. \tag{17.33}$$

The variance of the component at the input of Q_2 due to the random command input is therefore

$$\sigma^2 \cdot \mathrm{SOS}_2\,. \tag{17.34}$$

The condition for satisfaction of QT II at the input to Q_2 is accordingly:

$$\sigma \cdot \sqrt{\mathrm{SOS}_2} \geq q_2 \,. \tag{17.35}$$

When this condition is met, the quantization noise of Q_2 behaves like PQN from the standpoint of moments. This noise will also be essentially uncorrelated over time. For the overall system to behave linearly, it is necessary for both conditions Eq. (17.32) and Eq. (17.35) to be met. When the samples of the command input are white, testing for these conditions is very easy once SOS_1 and SOS_2 are computed.

In practice it is much more likely that the samples of the command input, when this input is stochastic, will not be white but will be correlated over time. Calculation of the variances of the responses to this input at the inputs to Q_1 and Q_2 will be more complicated. How this can be done will be described next.

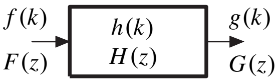

Figure 17.20 A linear digital filter.

Consider the linear digital filter shown in Fig. 17.20. Its input signal is $f(k)$ and its output signal is $g(k)$. The time index or time sample number is k. Its impulse response is $h(k)$. The filter output signal is the convolution of the input signal and the impulse response,

$$\begin{aligned} g(k) &= \sum_{i=0}^{\infty} f(k-i)h(i) \\ &= f(k) \star h(k) \,. \end{aligned} \tag{17.36}$$

The z-transform of the output is the product of the z-transform of the impulse response,

$$G(z) = F(z)H(z) \,. \tag{17.37}$$

The z-transform of the output is $G(z)$. The z-transform of the input is $F(z)$, and the z-transform of the impulse response is $H(z)$, the transfer function of the filter.

Let $f(k)$ be a stochastic input signal. Its autocorrelation function can be defined as

$$r_{ff}(m) \overset{\triangle}{=} \mathrm{E}\{f(k)f(k-m)\} \,, \tag{17.38}$$

where m is a discrete variable, the time lag. The z-transform of this autocorrelation function is

$$R_{ff}(z) = \sum_{m=-\infty}^{\infty} r_{ff}(m) z^{-m} \,. \tag{17.39}$$

With a stochastic input signal, the filter output signal will also be stochastic, and its autocorrelation function will be

$$\begin{aligned} r_{gg}(m) &= \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} h(i)h(i+j) r_{ff}(m-j) \\ &= r_{ff}(m) \star h(m) \star h(-m) \,. \end{aligned} \tag{17.40}$$

The output autocorrelation function is the double convolution of the input autocorrelation function with the impulse response and the impulse response reversed in time. The z-transform of the output autocorrelation function is the product

$$R_{gg}(z) = R_{ff}(z) H(z) H(z^{-1}) \,. \tag{17.41}$$

Relations Eq. (17.38)–(17.41) can be used to find the autocorrelation functions of the input components to Q_1 and Q_2 due to the command input. Let the command input have an autocorrelation function with a z-transform as follows:

$$R_{\mathrm{ii}}(z) = \sum_{m=-\infty}^{\infty} r_{\mathrm{ii}}(m) z^{-m} \,. \tag{17.42}$$

Let the corresponding input to Q_1 have an autocorrelation function with a z-transform as follows:

$$R_{\mathrm{ee}}(z) = \sum_{m=-\infty}^{\infty} r_{\mathrm{ee}}(m) z^{-m} \,. \tag{17.43}$$

Let the corresponding input to Q_2 have an autocorrelation function with a z-transform as follows:

$$R_{\mathrm{Q}_2\mathrm{Q}_2}(z) = \sum_{m=-\infty}^{\infty} r_{\mathrm{Q}_2\mathrm{Q}_2}(m) z^{-m} \,. \tag{17.44}$$

Accordingly,

$$R_{\mathrm{ee}}(z) = R_{\mathrm{ii}}(z) \cdot H_{\mathrm{ie}}(z) \cdot H_{\mathrm{ie}}(z^{-1}) \,, \tag{17.45}$$

and

$$R_{\mathrm{Q_2Q_2}} = R_{\mathrm{ii}}(z) \cdot H_{\mathrm{iQ_2}}(z) \cdot H_{\mathrm{iQ_2}}(z^{-1})\,. \tag{17.46}$$

The autocorrelation functions may be obtained from Eq. (17.46) by inverse z-transformation:

$$r_{\mathrm{ee}}(m) = \mathcal{Z}^{-1}\{R_{\mathrm{ee}}(z)\}\,, \tag{17.47}$$

and

$$r_{\mathrm{Q_2Q_2}}(m) = \mathcal{Z}^{-1}\{R_{\mathrm{Q_2Q_2}}(z)\}\,. \tag{17.48}$$

By starting with the autocorrelation function of the command input, one can obtain the corresponding autocorrelation functions at the inputs to $\mathrm{Q_1}$ and $\mathrm{Q_2}$ by making use of relations Eq. (17.43), Eq. (17.46), and Eq. (17.48). The mean squares of the inputs to $\mathrm{Q_1}$ and $\mathrm{Q_2}$ can then be obtained as follows:

$$\begin{aligned} E\{(\mathrm{Q_{1\,input}})^2\} &= r_{\mathrm{ee}}(m)\Big|_{m=0} \\ E\{(\mathrm{Q_{2\,input}})^2\} &= r_{\mathrm{Q_2Q_2}}(m)\Big|_{m=0}\,. \end{aligned} \tag{17.49}$$

It is evident that calculations of mean squares and variances at the inputs to $\mathrm{Q_1}$ and $\mathrm{Q_2}$ are much simpler when the command input is white than when this input is correlated over time (colored). The correlated case is the most prevalent however, and this case requires the full use of relations Eq. (17.43), Eq. (17.46), Eq. (17.48), and Eq. (17.49).

There is an additional benefit that comes from the complete calculation of the autocorrelation functions at the quantizer inputs. These inputs will be Gaussian if the command input is Gaussian, and will be close to Gaussian even if the command input is not Gaussian, in accord with the central limit theorem. In the Gaussian case, complete first-order and high-order statistics can be derived from the autocorrelation function.

Knowledge of the standard deviations is therefore sufficient to test for the satisfaction of the first-order QT II at the quantizer inputs. For example, if the command input has a high enough amplitude so that both standard deviations are at least as big as the respective quantization step sizes, then QT II will apply extremely well in a first-order sense. To verify satisfaction of the high-order QT II at both quantizer inputs, we would need the input autocorrelation functions. Since in the Gaussian case the quantizer input samples would need to be extremely highly correlated over time for any correlation over time to appear in the quantization noises, satisfaction of the first-order forms of QT II would almost guarantee that the quantization noises would be not only uniformly distributed but uncorrelated over time. The quantizers could then be replaced, for purposes of analysis, by additive white noises having the statistics of PQN.

Having the autocorrelation functions of the quantizer inputs would allow one to check the correlations over time of the quantizer inputs to verify that the quantization noises will indeed be white. The reader is referred to Chapter 11 for a discussion of the application of the quantizing theorems when the quantizer inputs are Gaussian.

17.7 THE BERTRAM BOUND

Quantization noise is bounded between $\pm q/2$. A quantizer embedded within an otherwise linear system will cause quantization noise at the system output. Assuming that the otherwise linear system is BIBO (bounded input, bounded output) stable, bounded quantization noise injected into the system will cause bounded quantization noise to exist at the system output. The presence of the quantizer will therefore have no effect on stability (except that under certain conditions, it could cause bounded oscillations or limit cycles).

The bound on the quantization noise at the system output is called the Bertram bound. It was discovered by John E. Bertram (Bertram, 1958).

The simplest way to study this bound would be to place a quantizer at the input to a linear filter as shown in Fig. 17.21(a), and find a bound for the quantization noise at the filter output. It is useful to compare this bound to the standard deviation of the output quantization noise. The standard deviation can be computed by placing a source of additive PQN at the linear filter input as shown in Fig. 17.21(b). Since PQN is white, the variance of the quantization noise at the filter output would be equal to $q^2/12$ multiplied by the sum of squares of the impulses of the impulse response, and the standard deviation would be obtained as the square root of this.

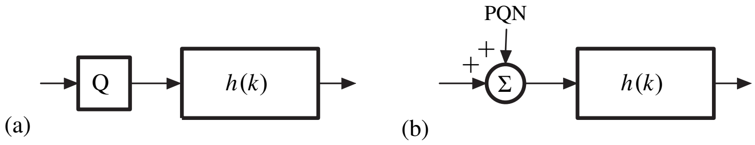

Figure 17.21 A linear digital filter with noise input, (a) quantization noise; (b) pseudo-quantization noise.

The bound can be computed by referring to Fig. 17.21. The impulse response of the digital filter is $h(k)$. The filter output is the convolution of the filter input with this impulse response. The convolution process can be illustrated by turning the impulse response around backward in time, multiplying the samples of the input signal with the backward impulse response samples, and forming a sum of the products.

To calculate the quantization noise in the output of the filter of Fig. 17.21(a), the quantization noise is convolved with the impulse response. The quantization noise is highly variable. If one could control the quantization noise, a maximum quantization noise component at the filter output could be generated by choosing the appropriate quantization noise sequence. Since the quantization noise is bounded between $\pm q/2$, the maximum noise at the filter output is generated when the filter input is a sequence of quantization noise samples of individual amplitude $+q/2$ or $-q/2$. The convolution is illustrated in Fig. 17.22. The quantization noise input sequence has been chosen to cause a maximum noise output at time sample k.

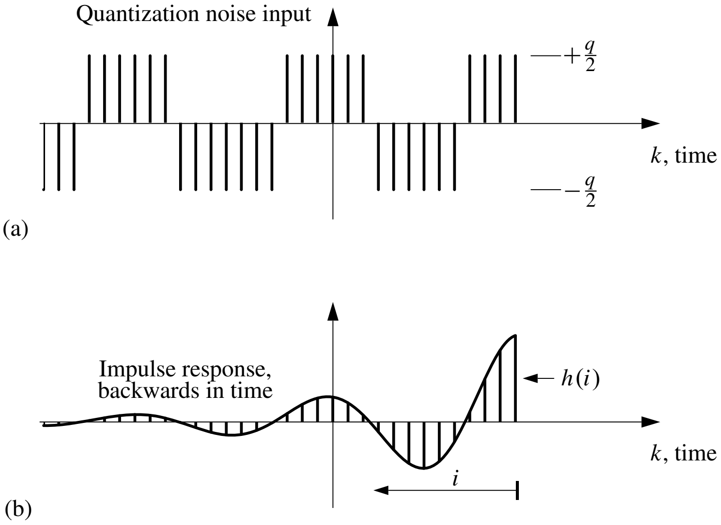

Figure 17.22 Convolution of worst-case quantization noise at the filter input with the impulse response, (a) sequence of $\pm q/2$ quantization noise samples; (b) impulse response plotted backward in time.

The convolution is given by expression Eq. (17.50).

$$g(k) = \sum_{i=0}^{\infty} f(k-i)h(i)\,. \tag{17.50}$$

From it, we obtain the largest possible output, the maximum output at time k as

$$g(k)\Big|_{\max} = \frac{q}{2}\sum_{i=0}^{\infty} |h(i)|\,. \tag{17.51}$$

Thus, at any moment of time, the quantization noise component at the filter output will be bounded between

$$\pm q/2 \sum_{i=0}^{\infty} |h(i)| \,. \tag{17.52}$$

This is the Bertram bound. Its magnitude is $q/2$ multiplied by the sum of magnitudes of the impulses of the impulse response.

Example 17.1 Bertram Bound and Standard Deviation
We can make some calculations of quantization noise bound and standard deviation with a simple example. Let the filter impulse response be a geometric sequence (like an exponential decay),

$$h(k) = \begin{cases} \left(\frac{1}{2}\right)^k \,, & k \geq 0 \,, \\ 0 \,, & k < 0 \,. \end{cases} \tag{17.53}$$

The sum of magnitudes of the impulse response is

$$\begin{aligned} \begin{pmatrix} \text{sum of} \\ \text{magnitudes} \end{pmatrix} = \sum_{k=0}^{\infty} h(k) &= 1 + 1/2 + 1/4 + \cdots \\ &= \frac{1}{1-\frac{1}{2}} = 2 \,. \end{aligned} \tag{17.54}$$

The sum of squares of the impulses of the impulse response is

$$\begin{aligned} \begin{pmatrix} \text{sum of} \\ \text{squares} \end{pmatrix} = \sum_{k=0}^{\infty} h^2(k) &= 1 + 1/4 + 1/16 + \cdots \\ &= \frac{1}{1-\frac{1}{4}} = \frac{4}{3} \,. \end{aligned} \tag{17.55}$$

The bound on the quantization noise magnitude at the filter output is

$$\frac{q}{2} \begin{pmatrix} \text{sum of} \\ \text{magnitudes} \end{pmatrix} = \frac{q}{2} \cdot 2 = q \,. \tag{17.56}$$

The variance of the quantization noise at the filter output is

$$\frac{q^2}{12} \begin{pmatrix} \text{sum of} \\ \text{squares} \end{pmatrix} = \frac{q^2}{12} \cdot \frac{4}{3} = \frac{q^2}{9} \,. \tag{17.57}$$

The standard deviation of the quantization noise at the filter output is

$$\sqrt{\frac{q^2}{9}} = \frac{q}{3} \,. \tag{17.58}$$

A computer simulation was made of a digital filter having the impulse response given by Eq. (17.53). The filter input was white, uniformly distributed quantization noise. The probability density of the output noise was calculated, and is plotted in Fig. 17.23. It is symmetrical and resembles a Gaussian density with its "tails" cut off. The bound and standard deviation of the output quantization noise are indicated in the plot. The bound magnitude equals three standard deviations for this case.

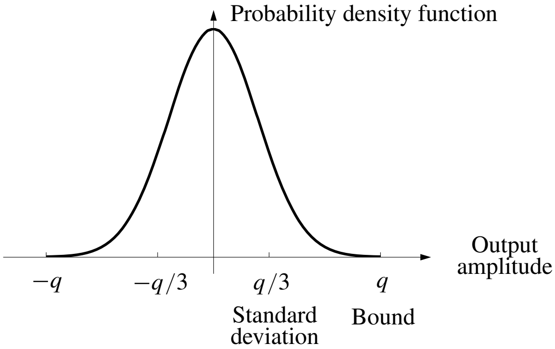

Figure 17.23 Probability density of quantization noise at filter output.

The methods employed above in finding the bound and standard deviation of the quantization noise at the filter output can be used to find the bound and standard deviation of the quantization noise at the output of a control system.

Example 17.2 Output Noise of a Feedback System
A good example would be the control system of Fig. 17.18, with two quantizers.

For the system of Fig. 17.18, the total output quantization noise power is given by Eqs. (17.25) and (17.26) as

$$\begin{aligned}\begin{pmatrix}\text{mean square}\\ \text{of output noise}\\ \text{due to } \mathrm{Q}_1 \text{ and } \mathrm{Q}_2\end{pmatrix} &= \frac{0.1386q_1^2 + 0.1415q_2^2}{12}\\ &= \frac{0.1386 \cdot 0.00977^2 + 0.1415 \cdot 0.0391^2}{12} \qquad (17.59)\\ &= 1.91 \cdot 10^{-5}\,\mathrm{V}^2\,.\end{aligned}$$

The standard deviation of the output quantization noise is the square root,

$$\begin{pmatrix}\text{standard deviation of}\\ \text{output quantization noise}\end{pmatrix} = 4.4 \cdot 10^{-3}\,\mathrm{V}\,. \qquad (17.60)$$

The step sizes of the quantizers are $q_1 = 0.00977$ V and $q_2 = 0.0391$ V.

Calculation of the output quantization noise bound requires knowledge of the sum of magnitudes of the impulse response from the output of quantizer Q$_1$ to the control system output, and the same for the impulse response from the output of Q$_2$ to the control system output. Obtaining the sum of magnitudes is generally not as easily done as the calculation of Eq. (17.54). This is often achieved by simulation of the system, and the sum of magnitudes is calculated numerically. This was done for the system of Fig. 17.18. Unit impulse inputs were applied separately, in place of the noise sources n_1 and n_2 in Fig. 17.19. Each impulse response was obtained by taking samples at the plant output. The results are

$$\begin{pmatrix} \text{sum of magnitudes} \\ \text{of impulse response} \\ \text{from Q}_1 \text{ to output} \end{pmatrix} = 2.56\,, \tag{17.61}$$

and

$$\begin{pmatrix} \text{sum of magnitudes} \\ \text{of impulse response} \\ \text{from Q}_2 \text{ to output} \end{pmatrix} = 2.572\,. \tag{17.62}$$

Finding the magnitude of the bound of the output quantization noise when the system contains more than one quantizer requires some thought. Since the bound comes from a worst-case analysis, let the worst-case quantization noise sequences (like that of Fig. 17.21(a)) be chosen independently for the two quantizers. The magnitude of the bound can be seen to be the sum of the bound magnitudes that would result from each quantizer as if acting alone. The magnitude of the bound is

$$\begin{aligned} \begin{pmatrix} \text{bound magnitude} \\ \text{of output} \\ \text{quantization noise} \end{pmatrix} &= \frac{q_1}{2}2.56 + \frac{q_2}{2}2.572 \\ &= \frac{0.00977}{2}2.56 + \frac{0.0391}{2}2.572 \\ &= 62.8 \cdot 10^{-3}\,\text{V}\,. \end{aligned} \tag{17.63}$$

Thus, the output quantization noise is bounded between $\pm 62.8 \cdot 10^{-3}$ V. The bound magnitude turns out to be equal to 4.24 standard deviations. Since the output quantization noise has a PDF that is similar to Gaussian, most of the probability exists between ±(a few standard deviations). The output quantization noise reaching the bound is therefore improbable for this case.

> For the system of Fig. 17.18, the magnitude of the bound for the output quantization noise would always be the same, given by (17.63), whether the command input is sinusoidal, stochastic, or some other type of signal. Satisfaction of quantizing theorems would not be an issue. On the other hand, the standard deviation of the output quantization noise would be given by (17.60) as long as the quantization noises of Q_1 and Q_2 behave like PQN, regardless of the characteristics of the command input signal.

17.8 SUMMARY

Digital feedback control systems have analog filtering, digital filtering, analog-to-digital conversion, and digital-to-analog conversion within one or more feedback loops. Analog-to-digital conversion is a nonlinear process that is a combination of sampling and quantization. Digital-to-analog conversion is a nonlinear process that is a combination of quantization and interpolation. Analysis of the effects of quantization in digital feedback systems is simple when the entire system is linear except for quantization in the ADC and DAC units. Quantization noise or quantization error at the system output is bounded and is most often random, and its statistical properties are quite predictable.

To analyze quantization effects at the system output, the quantizers are replaced with additive noises. Each quantizer causes noise in the system output. Impulse responses are found from the quantization points to the system output.

Since the noise injected by each quantizer is bounded between $\pm q/2$, the output noise or output error is bounded by $\pm q/2$ multiplied by the sum of the magnitudes of the impulses of the impulse response from quantizer to system output. With multiple quantizers, the magnitude of the output noise or output error is bounded by the sum of the magnitudes of the individual noise or error bounds. This is the Bertram bound for the system.

When the noise injected by each quantizer into the system is white, uniformly distributed between $\pm q/2$, and uncorrelated from quantizer to quantizer, the variance of the quantization noise at the system output due to an individual quantizer is equal to $q^2/12$ multiplied by the sum of squares of the impulses of the impulse response from quantizer to system output. With multiple quantizers, the variance of the total quantization noise at the system output is the sum of the individual variances. The standard deviation of the total noise is the square root of the variance. Knowing the standard deviation, one could create a Gaussian PDF that would closely approximate the true output noise PDF, at least over plus or minus several standard deviations. The true output PDF cannot be perfectly Gaussian however, since it is bounded.

17.9 EXERCISES

17.1 Regarding the diagram in Fig. E17.1.1, the input to the ADC is white, zero mean, and Gaussian with a standard deviation 1. Quantization within the ADC has a step size of $q = 0.001$.

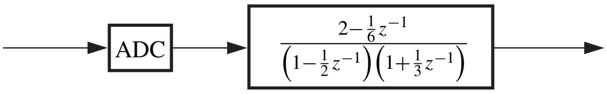
ADC
$\frac{2-\frac{1}{6}z^{-1}}{\left(1-\frac{1}{2}z^{-1}\right)\left(1+\frac{1}{3}z^{-1}\right)}$

Figure E17.1.1 A discrete dynamic system.

(**a**) Calculate the standard deviation of the quantization noise at the output of the digital filter.

(**b**) Calculate the theoretical bound of the quantization noise at the output of the digital filter.

17.2 The impulse response of a one-pole digital filter is exponential, like in Eq. (17.53). Let the geometric ratio be λ (the geometric ratio is the ratio between an impulse response sample and the previous sample). Determine the ratio of the Bertram bound to the standard deviation for the system in Fig. 16.2, by a calculation similar to Eqs. (17.56) and (17.58).

17.3 For the diagram in Fig. E17.1.1, the input to the ADC is a discrete sinusoid whose frequency is one hundredth of the sampling frequency and whose RMS amplitude is unity. The ADC has a step size of $q = 0.001$.

(**a**) Calculate the standard deviation of the quantization noise at the output of the digital filter.

(**b**) Calculate the theoretical bound of the quantization noise at the output of the digital filter.

Note that the magnitude of the quantization error does not depend on the input signal, if it is large enough. Therefore, the answer to this exercise is the same as for Exercise 17.1.

17.4 Regarding the diagram in Fig. E17.4.1, the input to the ADC is white, zero mean, and Gaussian with a unit standard deviation. Quantization within the ADC has a step size of $q_1 = 0.001$, and quantization within the DAC has a step size of $q_2 = 0.001$.

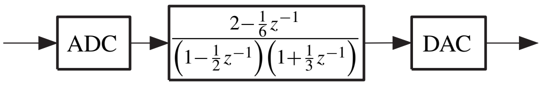
ADC
$\frac{2-\frac{1}{6}z^{-1}}{\left(1-\frac{1}{2}z^{-1}\right)\left(1+\frac{1}{3}z^{-1}\right)}$
DAC

Figure E17.4.1 A discrete dynamic system with DAC at the output.

(a) Calculate the standard deviation of the total quantization noise at the DAC output.

(b) Calculate the bound of the total quantization noise at the DAC output.

17.5 For the diagram in Fig. E17.4.1, the input to the ADC is a discrete sinusoid whose frequency is one hundredth of the sampling frequency and whose RMS amplitude is unity. The ADC has a step size of $q_1 = 0.001$, the DAC has a step size of $q_2 = 0.001$.

(a) Calculate the standard deviation of the total quantization noise at the output of the DAC.

(b) Calculate the bound of the total quantization noise at the output of the DAC.

Note that the magnitude of the quantization error does not depend on the input signal, if it is large enough. Therefore, the answer to this exercise is the same as for Exercise 17.4.

17.6 By computer simulation of the system given in Fig. E17.1.1, measure the standard deviation and bound for the quantization noise at the digital filter output,

(a) for the Gaussian input of Exercise 17.1,

(b) for the sinusoidal input of Exercise 17.3.

Explain your measurement techniques.

17.7 By computer simulation of the system given in Fig. E17.4.1, measure the standard deviation and bound for the total quantization noise at the output of the DAC,

(a) for the Gaussian input of Exercise 17.4,

(b) for the sinusoidal input of Exercise 17.5.

Explain your measurement techniques.

17.8 For the system diagrammed in Fig. 17.18 (page 446), verify by computer simulation the mean square of the output quantization noise as predicted by Eq. (17.26).

17.9 For the system diagrammed in Fig. 17.18, let the command input be a sinusoid whose frequency is one hundredth of the sampling frequency. By computer simulation, determine the minimum amplitude of this input that permits satisfaction of QT II at both quantizers. What are the amplitudes of the sinusoidal components at the quantizer inputs under this condition?

17.10 Regarding Exercise 17.9, reduce the amplitude of the sinusoidal command input to 0.02 V.

(a) Plot the sinusoidal command input and the plant output.

(b) Add independent white dither noises, uniformly distributed between $\pm q/2$, to the inputs of the quantizers. Once again, plot the command input and plant output signals. Notice the change in the plant output. Calculate the variance of the quantization noises and the dither noises in the plant output. Sum the variances and calculate the standard deviation of the total noise in the plant output. Compare this with the RMS amplitude of the sinusoidal component of the plant output.

(**c**) Repeat Exercise 17.10(b), but use white zero mean Gaussian noises as dither signals, make the standard deviations of the dither noises equal to one third the quantum step sizes of Q_1 and Q_2.

(**d**) Repeat Exercise 17.10(b), but use high frequency sine waves as dither signals. Let the frequencies of the sine waves be 100.31 times the sampling frequency, and let the RMS amplitudes of the sine waves be equal to the quantum step sizes for Q_1 and Q_2.

17.11 Consider the system illustrated in Fig. 17.18 (page 446), with a 10-bit ADC with input range ±5 V, and a 8-bit DAC with output range ±5 V.

(**a**) Verify with simulations that a static error is present for a step input.

(**b**) Determine the maximum value of this error.

(**c**) Compare this maximum error to the Bertram bound.

(**d**) Check the system for limit cycles (look for oscillating responses with different previous excitations and different constant inputs).

(**e**) Inject uniform or Gaussian or sinusoidal dither with carefully chosen parameters before the ADC and/or the DAC, and check the static error again.

17.12 The control system of Fig. 17.18 (page 446) has a 10-bit ADC with input range ±5 V, and an 8-bit DAC with output range ±5 V. At the input of the plant, let there be a hysteresis block, illustrated in Fig. E17.12.1.

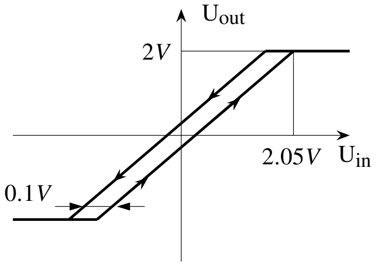

Figure E17.12.1 Hysteresis block.

(**a**) By simulation, verify the existence of limit cycles.

(**b**) Can you determine the Bertram bound?

(**c**) Inject Gaussian dither with carefully chosen parameters before the hysteresis loop, and check the presence of limit cycles again.

17.13 We have scaled the signal at the input of Q_2 in the system shown in Fig. 17.18, to be ±5 V when a sinusoidal input signal with peak value 7.07 V is applied.

(**a**) With a square wave input signal, determine the proper scaling (change the two coefficients in the feedforward link while maintaining their product), to have maximum ±5 at the input of the DAC.

(b) For any signal, determine the scaling to have maximum ±5 at the input of the DAC.

17.14 Change the scaling of the system shown in Fig. 17.18, to avoid underload of the ADC and the DAC, with a sine wave input with 0.005 V peak value. Scale so that the peak-to-peak sine amplitude for each quantizer covers 5 quantum steps. How large is the maximum input then without overload?