Digital Signal Processing
Quantization Basics

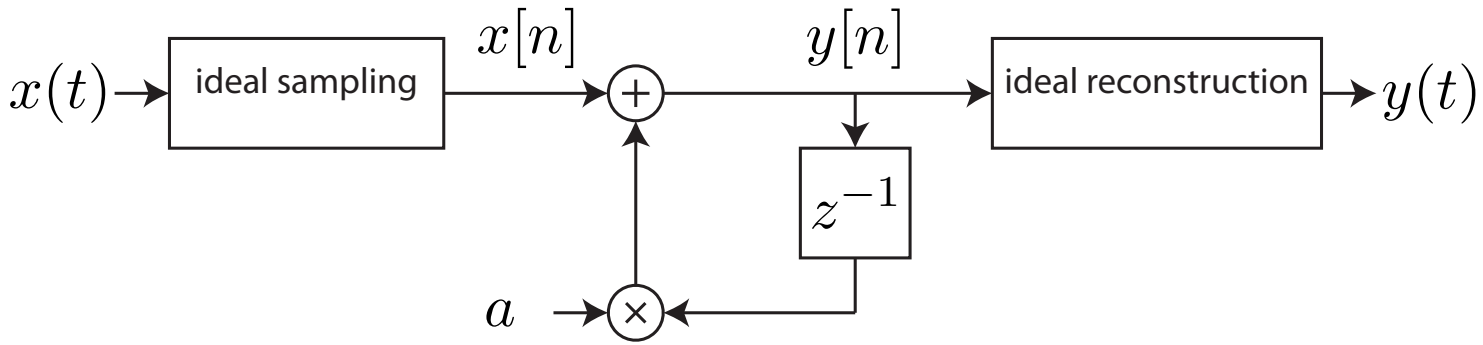
D. Richard Brown III

A Simple DSP System

Suppose we wish to implement the transfer function / difference equation

$$H(z) = \frac{1}{1 - az^{-1}} \qquad \Leftrightarrow \qquad y[n] = x[n] + ay[n-1]$$

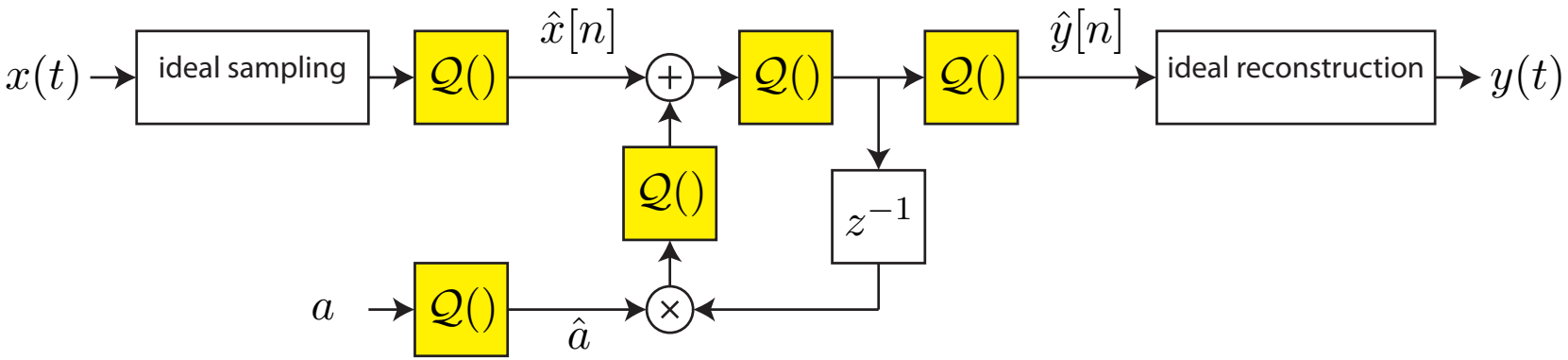
We could draw a block diagram (realization) of a complete system:



Unfortunately, this block diagram represents an idealized view of how the system is actually going to work. Some practical considerations:

- Input/output quantization.
- Filter coefficient quantization.
- Product roundoff.
- Potential overflow in sums.

A Simple DSP System: A More Realistic View



Our nice simple LTI system is now highly nonlinear.

In general, nonlinear systems like this are difficult to analyze. Hence, we must isolate the sources of quantization error and adopt some approximate approaches to make the analysis tractable.

Quantization Basics

Given a real number x, we denote the quantized value of x as

$$\hat{x} = \mathcal{Q}(x) = x + \epsilon$$

where ϵ is the "quantization error".

There are two main types of quantization:

1. **Truncation**: just discard least significant bits
2. **Rounding**: choose closest value

As an example, suppose we want to quantize $\frac{1}{\sqrt{2}} \approx 0.7071$ to a fixed point number with two fractional bits. If we **truncate**, we have

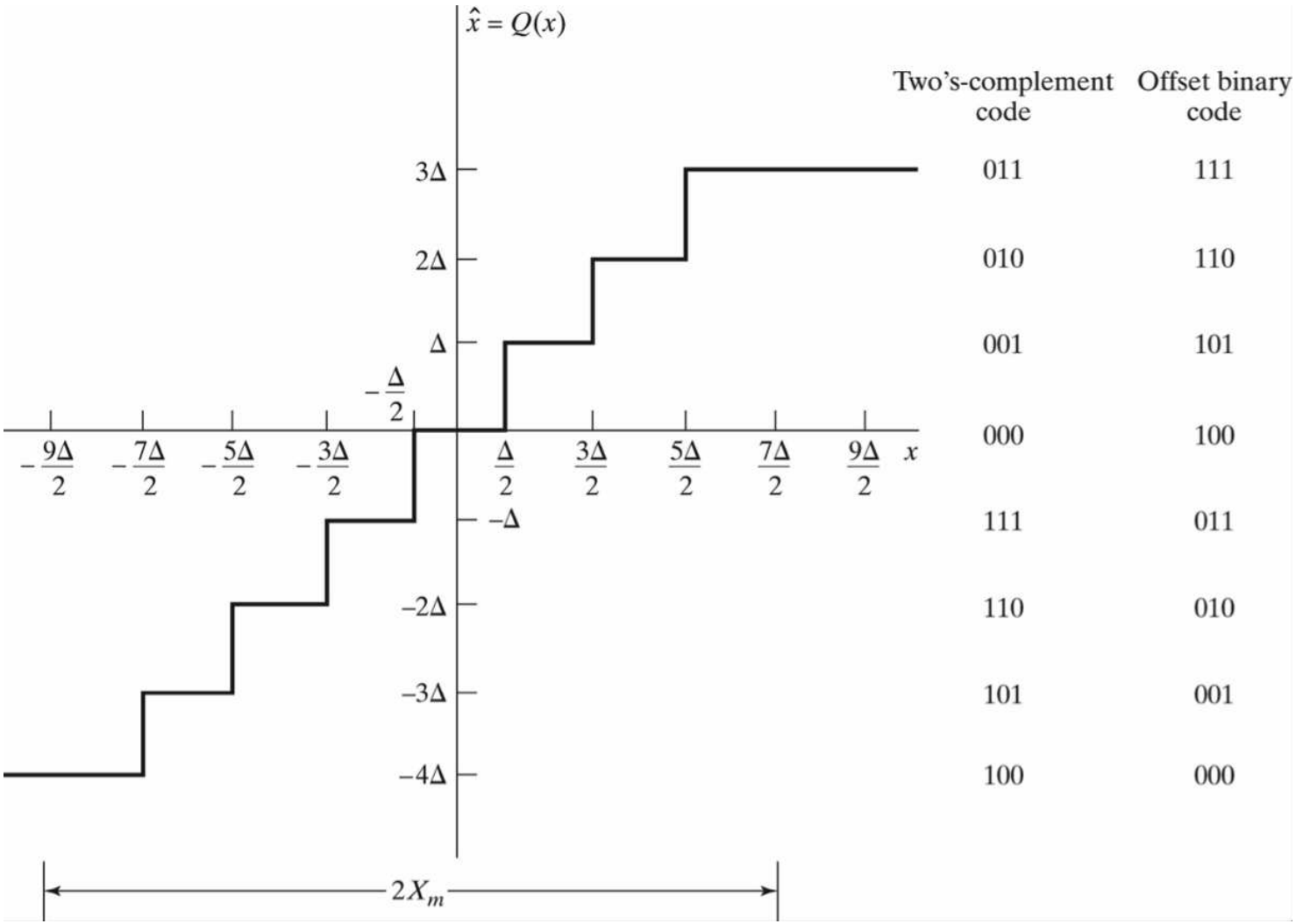
$$\mathcal{Q}(1/\sqrt{2}) = \boxed{0_\Delta 10} 110 \cdots = 0.50$$

whereas if we **round**, we have

$$\mathcal{Q}(1/\sqrt{2}) = \boxed{0_\Delta 11} = 0.75$$

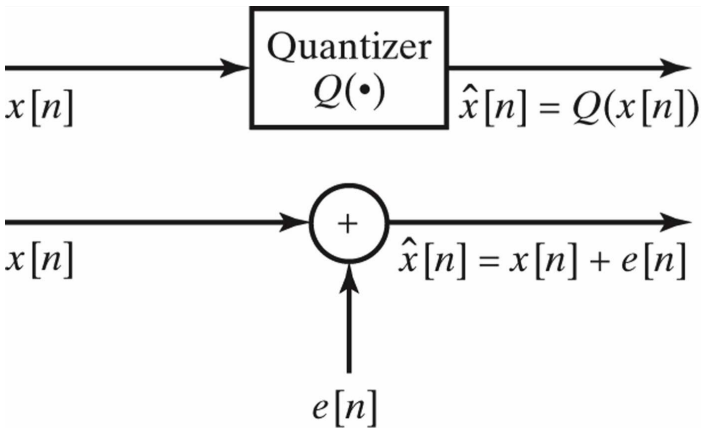
We usually prefer rounding because the quantization errors are zero-mean and bounded $-\frac{\Delta}{2} \leq \epsilon < \frac{\Delta}{2}$, where Δ is the quantizer step size (assuming no overflow).

Bipolar 3 Bit Quantizer Example



Note that $\Delta = \frac{2X_m}{2^{B+1}} = \frac{X_m}{2^B}$ where $2X_m$ is the full-scale range of the quantizer and $B + 1$ is the number of quantizer bits.

Statistical Model of Input Quantization Error



Common assumptions to facilitate analysis:

1. $e[n]$ is a uniformly distributed random variable between $-\frac{\Delta}{2}$ and $\frac{\Delta}{2}$.
2. $e[n]$ is a stationary random process (its statistics don't change over time).
3. $e[n]$ is a white sequence, i.e., $e[n]$ is uncorrelated with $e[m]$ for all $n \neq m$.
4. $e[n]$ is uncorrelated with $x[n]$.

Quantizer Signal-to-Noise Ratio Analysis

The variance of the quantization error of a $B+1$ bit uniform quantizer can be computed as

$$\sigma_e^2 = \int_{-\Delta/2}^{\Delta/2} e^2 \frac{1}{\Delta}\, de = \frac{\Delta^2}{12} = \frac{X_m^2}{12 \cdot 2^{2B}}.$$

The SNR (in dB) of a $B+1$ bit uniform quantizer is then

$$\begin{aligned}
\text{SNR}_Q &= 10 \log_{10}\left(\frac{\sigma_x^2}{\sigma_e^2}\right) \\
&= 10 \log_{10}\left(\frac{12 \cdot 2^{2B}\sigma_x^2}{X_m^2}\right) \\
&= 6.02B + 10.8 - 20 \log_{10}\left(\frac{X_m}{\sigma_x}\right)
\end{aligned}$$

If we assume $x[n]$ is Gaussian distributed and scale the quantizer/input so that $\sigma_x = X_m/4$ (corresponding to a 0.064% chance of saturating the quantizer), this result simplifies to

$$\text{SNR}_Q = 6.02B + 10.8 - 20 \log_{10}(4) = 6.02B - 1.24\,\text{dB}.$$