Quantization

Jose Krause Perin

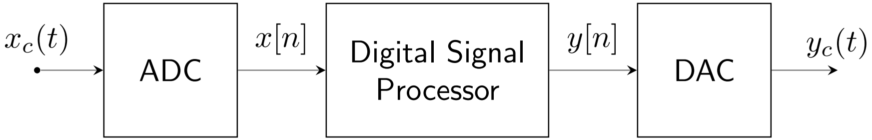
Stanford University

July 25, 2017

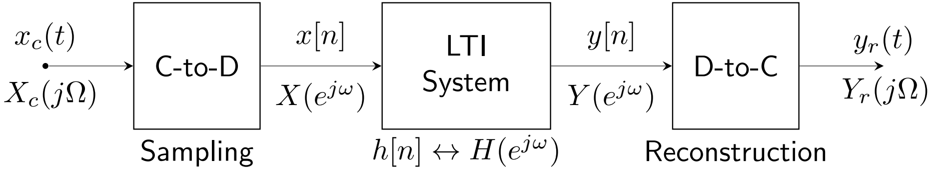
Outline

Practice and theory

In practice



DSP theory

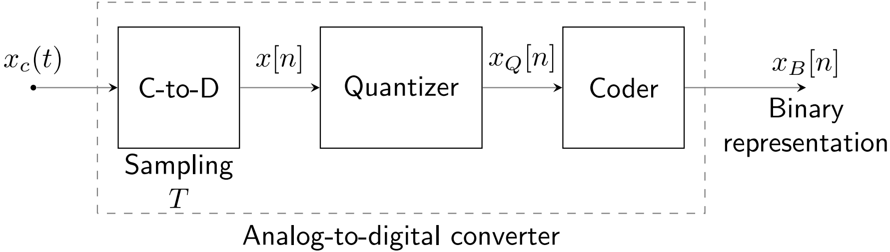


Problem: This simplified model doesn't account for **quantization** (this lecture) or **finite precision arithmetic** (lecture 8).

Including quantization

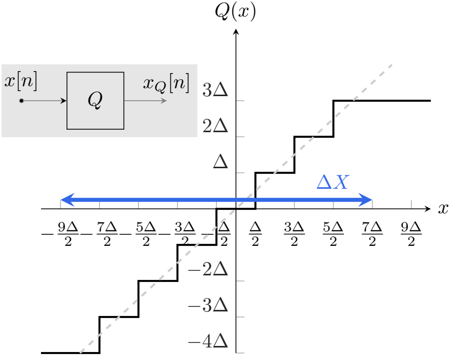
Analog-to-digital converter

A more realistic model

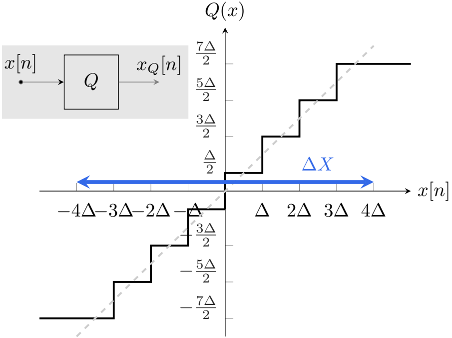


Quantizer

Mid-tread uniform quantizer



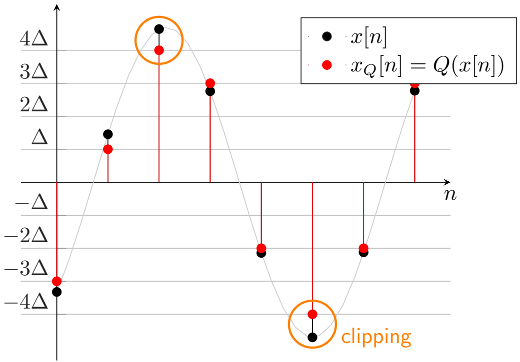
Mid-rise uniform quantizer



Terminology

- The quantizer has B bits of **resolution**
- ΔX is the **dynamic range**
- Δ is the **step size**

Example of quantization



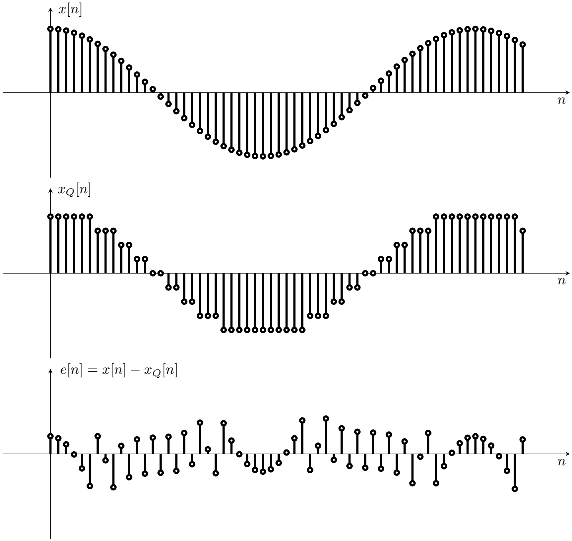
Quantization error:

$$e[n] = x[n] - Q(x[n]) = x[n] - x_Q[n]$$

Note that the quantization error is bounded $-\Delta/2 \leq e[n] \leq \Delta/2$.
Quantization error is deterministic but hard to analyze, so we treat it as noise (random process).

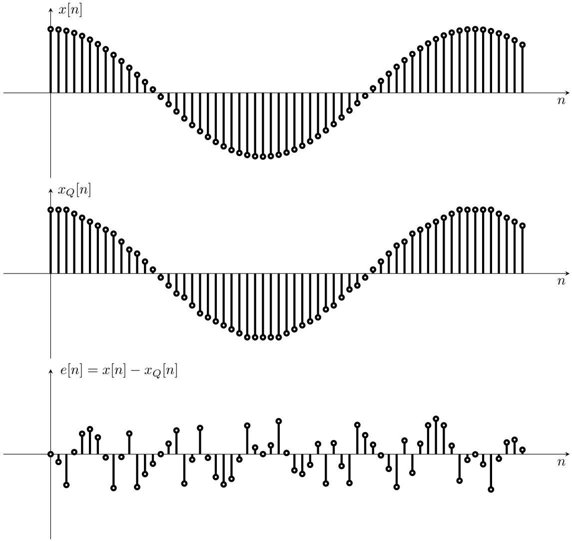
Quantization of a sinusoid

Using a 3-bit quantizer



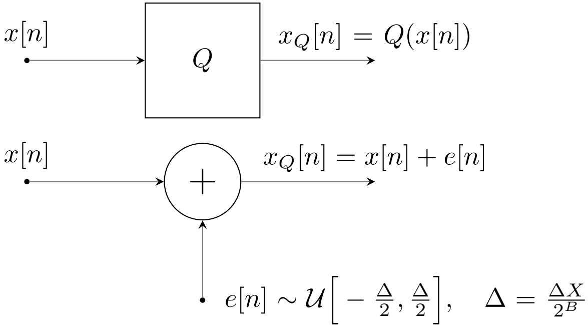
Quantization of a sinusoid

Using an 8-bit quantizer



Linear noise model

We'll model the quantizer as a noise source of a **white uniformly distributed noise** that is independent of the input signal.



From these assumptions:

$$\sigma_e^2 = \frac{\Delta^2}{12} \qquad \text{(average power)}$$

$$\phi_{ee}[n] = \sigma_e^2 \delta[n] \qquad \text{(autocorrelation function)}$$

$$\Phi_{ee}(e^{j\omega}) = \sigma_e^2, |\omega| \leq \pi \qquad \text{(PSD)}$$

Quantizer signal-to-noise ratio (SNR)

It's often convinient to characterize the quantizer in terms of a **signal-to-noise ratio (SNR)**:

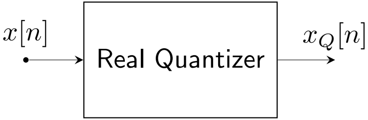
$$
\begin{aligned}
\mathrm{SNR} &= 10\log_{10}\left(\frac{\text{Signal Power}}{\text{Quantization noise power}}\right)\ \text{dB} \\
&= 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_e^2}\right) \\
&= 10\log_{10}\left(\frac{12\sigma_x^2}{\Delta^2}\right) \\
&= 10\log_{10}\left(\frac{12\sigma_x^2(2^{2B})}{\Delta X^2}\right) && \text{(substituting (1))} \\
&= \boxed{6.02B} + 10.79 + 20\log_{10}\left(\frac{\sigma_x}{\Delta X}\right)
\end{aligned}
$$

For every bit in the quantizer we gain 6.02 dB of SNR.
Important: The signal amplitude must be matched to the quantizer dynamic range, otherwise there'll be excessive clipping or some of the bits may not be used.

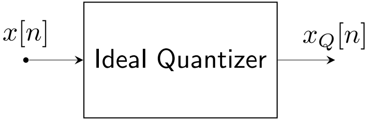
Effective number of bits (ENOB)

Another useful metric to evaluate quantizers is the **effective number of bits (ENOB)**.

- Quantization is not the only source of noise in real quantizers
- Additional noise will consume some bits of resolution
- To continue using the simple linear noise model, we assume that the noisy real quantizer is equal to an ideal quantizer with resolution ENOB $< B$ bits.



B bits of resolution



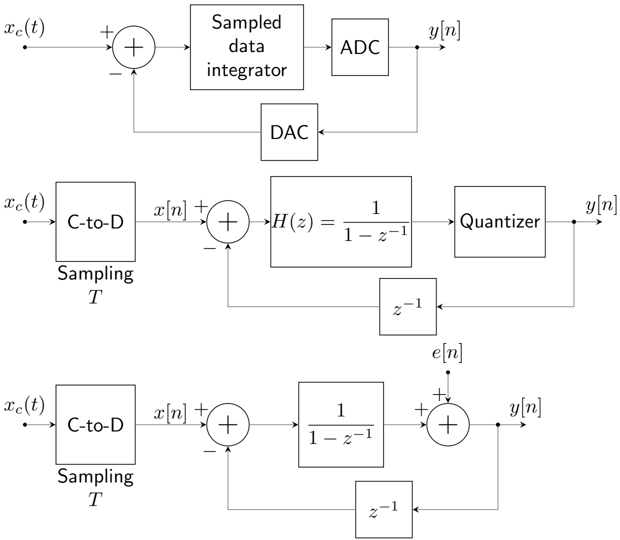
ENOB bits of resolution

Datasheets of ADCs will typically give you the ENOB at a certain frequency.

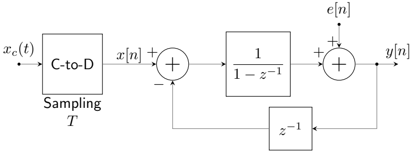
Noise shaping

- Quantization noise is unavoidable, but there are strategies to mitigate it
- One example is **noise shaping**. The goal is to shape the quantization noise PSD, so that most of the noise power falls outside the signal band
- To perform noise shaping the signal must be **oversampled**, otherwise noise aliasing would make most of the noise power fall in the signal band.
- Noise shaping can be used in both A-to-D and D-to-A converters

Noise shaping in A-to-D conversion



Noise shaping in A-to-D conversion



Using superposition, we can separately study the effect of the system on the signal $x[n]$ and on quantization noise $e[n]$.
For the signal

$$
\begin{aligned}
Y(z) &= (X(z) - Y(z)z^{-1})\frac{1}{1-z^{-1}} \\
Y(z)(1-z^{-1}) &= (X(z) - Y(z)z^{-1}) \\
Y(z) &= X(z) && \text{(signal is unaffected)}
\end{aligned}
$$

For the noise

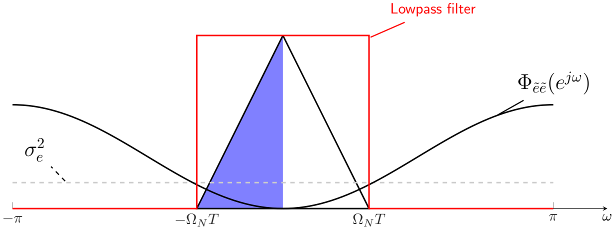
$$
\begin{aligned}
Y(z) &= E(z) - Y(z)\frac{z^{-1}}{1-z^{-1}} \\
&= E(z)(1-z^{-1}) && \text{(noise is filtered)}
\end{aligned}
$$

The noise is filtered by

$$\frac{Y(z)}{E(z)} = 1 - z^{-1}$$

Therefore, the noise PSD at the output will be

$$\begin{aligned}\Phi_{\tilde{e}\tilde{e}}(e^{j\omega}) &= |1 - e^{-j\omega}|^2 \sigma_e^2 \qquad \text{(since } e[n] \text{ is white)} \\ &= 4\sigma_e^2 \sin^2(\omega/2)\end{aligned}$$



After noise shaping most of the noise power falls out of the signal band, so we can use a simple lowpass filter to minimize quantization noise.

Important: This strategy of noise shaping only works when oversampling is sufficiently high. Otherwise, quantization noise would still fall in the signal band due to aliasing.

Noise shaping in A-to-D conversion

Table 4.1 of the textbook:

TABLE 4.1 EQUIVALENT SAVINGS IN QUANTIZER BITS RELATIVE TO $M = 1$ FOR DIRECT QUANTIZATION AND FIRST-ORDER NOISE SHAPING

M	Direct quantization	Noise shaping
4	1	2.2
8	1.5	3.7
16	2	5.1
32	2.5	6.6
64	3	8.1

M denotes the amount of oversampling. That is $M = \frac{\text{Sampling frequency}}{\text{Nyquist frequency}}$.

Summary

- Quantization is unavoidable in DSP systems
- Although quantization is a nonlinear operation on a signal, we can model the quantization error as a uniformly distributed random process (linear noise model)
- Using this linear noise model, we simply replace quantizers by noise sources of average power $\sigma_e^2 = \Delta^2/12$
- Quantization noise is assumed white (samples are uncorrelated)
- Every extra bit of resolution in a quantizer improves the SNR by 6.02 dB
- The signal amplitude must be matched to the dynamic range of the quantizer, otherwise there'll be excessive clipping or some bits won't be used
- Noise shaping is a strategy that minimizes quantization noise in A-to-D and D-to-A converters. The goal is to shape the quantization noise PSD, so that most of the noise power falls outside the signal band
- Noise shaping requires oversampling to minimize noise aliasing