COM-202: Signal Processing

Chapter 8.b: Quantization

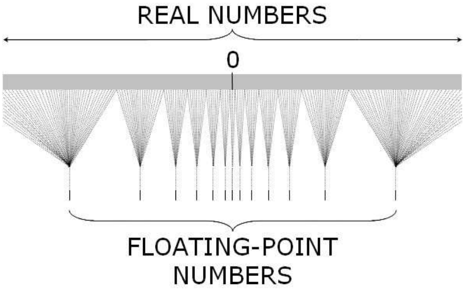
quantization

The digital world

- basic storage unit: the binary digit (bit) with two possible values (0, 1)
- aggregate units: the byte (8 bits), word, dword, etc
- R aggregate bits can hold 2^R distinct integer values

What about floating point?

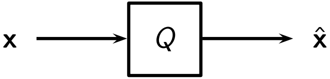
- floating point standards (e.g. IEEE 754) are clever ways of mapping reals to integers
- an R-bit float can represent at most 2^R distinct values
- a floating point representation partitions the real line into intervals of increasing size and maps them to integers



Quantization

- digital devices can only deal with integers (R bits per sample)
- samples of a discrete-time signal must be converted to integers for storage
- the conversion process is called *quantization*
- quantization causes an irreversible loss of information

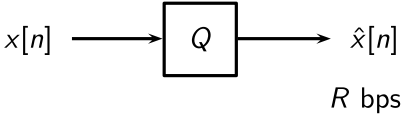
Quantization schemes



Several factors at play:

- storage budget (bits per sample)
- encoding scheme (fixed point, floating point)
- properties of the input
  - dynamic range
  - probability distribution of samples

Scalar, memoryless, fixed-rate quantization

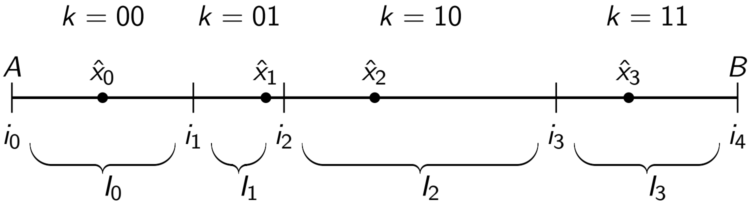


The simplest quantization scheme:

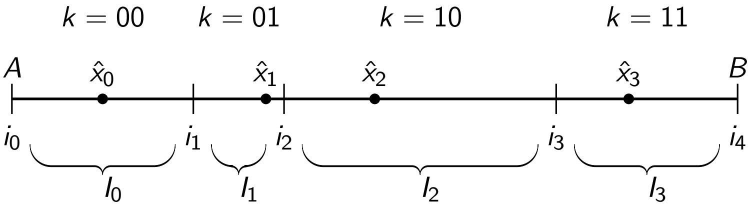
- each sample is encoded individually *(scalar quantization)*
- each sample is quantized independently *(memoryless quantization)*
- each sample is encoded using R bits *(fixed-rate quantization)*

Typical quantization scheme

- input values are within known bounds $A \leq x[n] \leq B$
- with R bits/sample, input range is divided into 2^R intervals $I_k = [i_k, i_{k+1})$
- each interval is associated to a R-bit binary number k
- each interval is associated to a representative value $\hat{x}_k$



Typical quantization scheme



- what are the optimal interval boundaries i_k?
- what are the optimal quantization values $\hat{x}_k$?

Optimal Quantization

The optimal quantizer minimizes the energy of the quantization error:

$$e[n] = Q(x[n]) - x[n] = \hat{x}[n] - x[n]$$

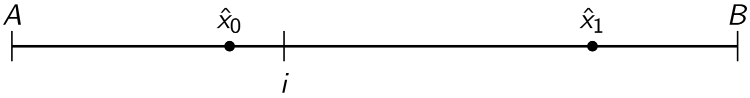
- model $\mathbf{x}$ as a stochastic process
- find the optimal i_k and $\hat{x}_k$ that minimize $\sigma_e^2 = \mathrm{E}\left[e^2[n]\right]$
- optimal quantizer will depend on the input's statistics

Quantization MSE

$$\sigma_e^2 = \mathrm{E}\left[(x - Q(x))^2\right]$$
$$= \int_{-\infty}^{\infty} (x - Q(x))^2 \, f_x(x) \, dx$$
$$= \sum_{k=0}^{2^R - 1} \int_{i_k}^{i_{k+1}} (x - \hat{x}_k)^2 \, f_x(x) \, dx$$

find global minimum wrt i_k, $\hat{x}_k$

Simple example: optimal one-bit quantizer



3 free parameters: $i, \hat{x}_0, \hat{x}_1$

Simple example: optimal one-bit quantizer

$$\sigma_e^2 = \int_A^i (x - \hat{x}_0)^2 \, f_x(x) \, dx + \int_i^B (x - \hat{x}_1)^2 \, f_x(x) \, dx$$

find $i, \hat{x}_0, \hat{x}_1$ such that

$$\frac{\partial \sigma_e^2}{\partial i} = \frac{\partial \sigma_e^2}{\hat{x}_0} = \frac{\partial \sigma_e^2}{\hat{x}_1} = 0$$

Simple example: optimal one-bit quantizer

$$\sigma_e^2 = \int_A^i (x - \hat{x}_0)^2 \, f_x(x) \, dx + \int_i^B (x - \hat{x}_1)^2 \, f_x(x) \, dx$$

find $i, \hat{x}_0, \hat{x}_1$ such that

$$\frac{\partial \sigma_e^2}{\partial i} = \frac{\partial \sigma_e^2}{\hat{x}_0} = \frac{\partial \sigma_e^2}{\hat{x}_1} = 0$$

little calculus reminder

$$\frac{\partial}{\partial t} \int_{\alpha}^{t} f(\tau)\, d\tau = \frac{\partial}{\partial t} \left[F(t) - F(\alpha) \right] = f(t)$$

Optimal one-bit quantizer: threshold

$$\frac{\partial \sigma_e^2}{\partial i} = \frac{\partial}{\partial i}\left[\int_A^i (x-\hat{x}_0)^2 f_x(x)\,dx + \int_i^B (x-\hat{x}_1)^2 f_x(x)\,dx\right]$$

$$= (i-\hat{x}_0)^2 f_x(i) - (i-\hat{x}_1)^2 f_x(i) = 0$$

$$\Rightarrow (i-\hat{x}_0)^2 - (i-\hat{x}_1)^2 = 0$$

$$\Rightarrow i = \frac{\hat{x}_0+\hat{x}_1}{2}$$

Optimal one-bit quantizer: threshold

$$\frac{\partial \sigma_e^2}{\partial i} = \frac{\partial}{\partial i}\left[\int_A^i (x - \hat{x}_0)^2\, f_x(x)\, dx + \int_i^B (x - \hat{x}_1)^2\, f_x(x)\, dx\right]$$

$$= (i - \hat{x}_0)^2\, f_x(i) - (i - \hat{x}_1)^2\, f_x(i) = 0$$

$$\Rightarrow (i - \hat{x}_0)^2 - (i - \hat{x}_1)^2 = 0$$

$$\Rightarrow i = \frac{\hat{x}_0 + \hat{x}_1}{2}$$

Optimal one-bit quantizer: threshold

$$\frac{\partial \sigma_e^2}{\partial i} = \frac{\partial}{\partial i}\left[\int_A^i (x-\hat{x}_0)^2\, f_x(x)\, dx + \int_i^B (x-\hat{x}_1)^2\, f_x(x)\, dx\right]$$

$$= (i-\hat{x}_0)^2\, f_x(i) - (i-\hat{x}_1)^2\, f_x(i) = 0$$

$$\Rightarrow (i-\hat{x}_0)^2 - (i-\hat{x}_1)^2 = 0$$

$$\Rightarrow i = \frac{\hat{x}_0 + \hat{x}_1}{2}$$

Optimal one-bit quantizer: threshold

$$\frac{\partial \sigma_e^2}{\partial i} = \frac{\partial}{\partial i}\left[\int_A^i (x - \hat{x}_0)^2 \, f_x(x)\, dx + \int_i^B (x - \hat{x}_1)^2 \, f_x(x)\, dx\right]$$

$$= (i - \hat{x}_0)^2 \, f_x(i) - (i - \hat{x}_1)^2 \, f_x(i) = 0$$

$$\Rightarrow (i - \hat{x}_0)^2 - (i - \hat{x}_1)^2 = 0$$

$$\Rightarrow i = \frac{\hat{x}_0 + \hat{x}_1}{2}$$

Optimal one-bit quantizer: values

$$\frac{\partial \sigma_e^2}{\partial \hat{x}_0} = \frac{\partial}{\partial x_0} \int_A^i (x - \hat{x}_0)^2 \, f_x(x) \, dx$$

$$= \int_A^i 2(\hat{x}_0 - x) \, f_x(x) \, dx = 0$$

$$\Rightarrow \hat{x}_0 = \frac{\int_A^i x \, f_x(x) \, dx}{\int_A^i f_x(x) \, dx} \qquad \text{(center of mass)}$$

$$\Rightarrow \hat{x}_1 = \frac{\int_i^B x \, f_x(x) \, dx}{\int_i^B f_x(x) \, dx}$$

Optimal one-bit quantizer: values

$$\frac{\partial \sigma_e^2}{\partial \hat{x}_0} = \frac{\partial}{\partial x_0} \int_A^i (x - \hat{x}_0)^2 \, f_x(x) \, dx$$

$$= \int_A^i 2(\hat{x}_0 - x) \, f_x(x) \, dx = 0$$

$$\Rightarrow \hat{x}_0 = \frac{\int_A^i x \, f_x(x) \, dx}{\int_A^i f_x(x) \, dx} \qquad \textit{(center of mass)}$$

$$\Rightarrow \hat{x}_1 = \frac{\int_i^B x \, f_x(x) \, dx}{\int_i^B f_x(x) \, dx}$$

Optimal one-bit quantizer: values

$$\frac{\partial \sigma_e^2}{\partial \hat{x}_0} = \frac{\partial}{\partial x_0} \int_A^i (x - \hat{x}_0)^2 \, f_x(x) \, dx$$

$$= \int_A^i 2(\hat{x}_0 - x) \, f_x(x) \, dx = 0$$

$$\Rightarrow \hat{x}_0 = \frac{\int_A^i x \, f_x(x) \, dx}{\int_A^i f_x(x) \, dx} \qquad \textit{(center of mass)}$$

$$\Rightarrow \hat{x}_1 = \frac{\int_i^B x \, f_x(x) \, dx}{\int_i^B f_x(x) \, dx}$$

For uniformly-distributed input

$$f_x(x) = \frac{1}{B - A}$$

$$\hat{x}_0 = \frac{\int_A^i x\,dx}{\int_A^i dx} = \frac{A + i}{2}$$

$$\hat{x}_1 = \frac{\int_i^B x\,dx}{\int_i^B dx} = \frac{i + B}{2}$$

$$i = \frac{\hat{x}_0 + \hat{x}_1}{2} = \frac{A + B}{2}$$

For uniformly-distributed input

$$f_x(x) = \frac{1}{B - A}$$

$$\hat{x}_0 = \frac{\int_A^i x\,dx}{\int_A^i dx} = \frac{A + i}{2}$$

$$\hat{x}_1 = \frac{\int_i^B x\,dx}{\int_i^B dx} = \frac{i + B}{2}$$

$$i = \frac{\hat{x}_0 + \hat{x}_1}{2} = \frac{A + B}{2}$$

For uniformly-distributed input

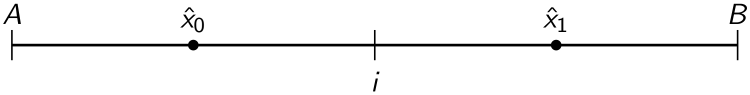
$$f_x(x) = \frac{1}{B - A}$$

$$\hat{x}_0 = \frac{\int_A^i x\,dx}{\int_A^i dx} = \frac{A + i}{2}$$

$$\hat{x}_1 = \frac{\int_i^B x\,dx}{\int_i^B dx} = \frac{i + B}{2}$$

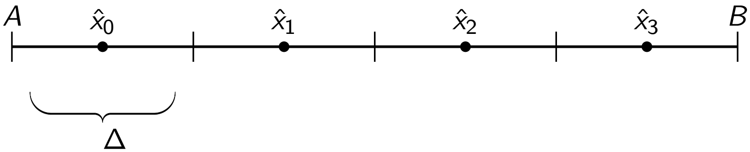
$$i = \frac{\hat{x}_0 + \hat{x}_1}{2} = \frac{A + B}{2}$$

Optimal one-bit quantizer

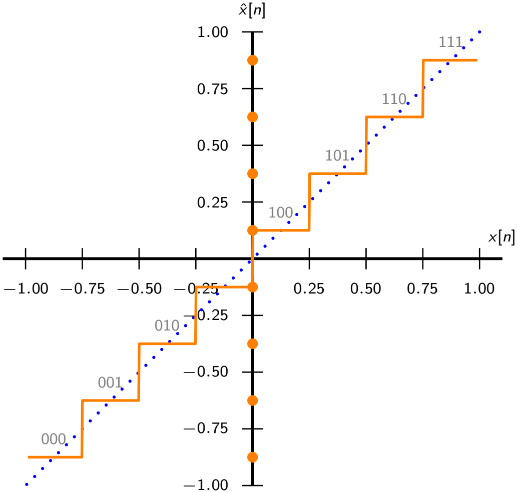


Uniform quantization of uniform input

- for uniformly-distributed input values, optimal quantizer is uniform
- optimal subdivision: 2^R *equal* intervals of width $\Delta = (B - A)2^{-R}$
- optimal quantization values are the midpoints of each interval



Uniform 3-Bit quantization function



Uniform quantization of uniform input: error analysis

$$\sigma_e^2 = \int_A^B f_x(x)(Q(x) - x)^2 \, dx$$
$$= \sum_{k=0}^{2^R-1} \int_{I_k} f_x(x)(\hat{x}_k - x)^2 \, dx$$

$$f_x(s) = \frac{1}{B - A}$$
$$\Delta = \frac{B - A}{2^R}$$
$$I_k = [A + k\Delta, A + (k+1)\Delta]$$
$$\hat{x}_k = A + (k + 1/2)\Delta$$

Uniform quantization of uniform input: error analysis

$$\sigma_e^2 = \int_A^B f_x(x)(Q(x) - x)^2 \, dx$$

$$= \sum_{k=0}^{2^R - 1} \int_{I_k} f_x(x)(\hat{x}_k - x)^2 \, dx$$

$$f_x(s) = \frac{1}{B - A}$$

$$\Delta = \frac{B - A}{2^R}$$

$$I_k = [A + k\Delta, A + (k+1)\Delta]$$

$$\hat{x}_k = A + (k + 1/2)\Delta$$

Uniform quantization of uniform input: error analysis

$$\sigma_e^2 = \sum_{k=0}^{2^R-1} \int_{A+k\Delta}^{A+(k+1)\Delta} \frac{(A+(k+1/2)\Delta - x)^2}{B-A}\, dx$$

$$= \sum_{k=0}^{2^R-1} \int_{-\Delta/2}^{\Delta/2} \frac{x^2}{B-A}\, dx \qquad x \leftarrow x + A + k(+1/2)\Delta$$

$$= \frac{2^R}{B-A} \frac{2(\Delta/2)^3}{3}$$

$$= \frac{\Delta^2}{12}$$

Uniform quantization of uniform input: error analysis

$$\sigma_e^2 = \sum_{k=0}^{2^R-1} \int_{A+k\Delta}^{A+(k+1)\Delta} \frac{(A+(k+1/2)\Delta - x)^2}{B-A}\, dx$$

$$= \sum_{k=0}^{2^R-1} \int_{-\Delta/2}^{\Delta/2} \frac{x^2}{B-A}\, dx \qquad\qquad x \leftarrow x + A + k(+1/2)\Delta$$

$$= \frac{2^R}{B-A} \frac{2(\Delta/2)^3}{3}$$

$$= \frac{\Delta^2}{12}$$

Uniform quantization of uniform input: error analysis

$$
\begin{aligned}
\sigma_e^2 &= \sum_{k=0}^{2^R-1} \int_{A+k\Delta}^{A+(k+1)\Delta} \frac{(A+(k+1/2)\Delta - x)^2}{B-A}\, dx \\
&= \sum_{k=0}^{2^R-1} \int_{-\Delta/2}^{\Delta/2} \frac{x^2}{B-A}\, dx \qquad\qquad x \leftarrow x + A + k(+1/2)\Delta \\
&= \frac{2^R}{B-A} \frac{2(\Delta/2)^3}{3} \\
&= \frac{\Delta^2}{12}
\end{aligned}
$$

Error analysis

- error energy

$$\sigma_e^2 = \Delta^2/12, \qquad \Delta = (B - A)/2^R$$

- signal energy

$$\sigma_x^2 = (B - A)^2/12$$

- signal to noise ratio

$$\mathrm{SNR} = 2^{2R}$$

- in dB

$$\mathrm{SNR}_{\mathrm{dB}} = 10 \log_{10} 2^{2R} \approx 6R \text{ dB}$$

Error analysis

- error energy

$$\sigma_e^2 = \Delta^2/12, \qquad \Delta = (B - A)/2^R$$

- signal energy

$$\sigma_x^2 = (B - A)^2/12$$

- signal to noise ratio

$$\mathrm{SNR} = 2^{2R}$$

- in dB

$$\mathrm{SNR_{dB}} = 10 \log_{10} 2^{2R} \approx 6R \text{ dB}$$

Error analysis

- error energy

$$\sigma_e^2 = \Delta^2/12, \qquad \Delta = (B - A)/2^R$$

- signal energy

$$\sigma_x^2 = (B - A)^2/12$$

- signal to noise ratio

$$\text{SNR} = 2^{2R}$$

- in dB

$$\text{SNR}_{\text{dB}} = 10 \log_{10} 2^{2R} \approx 6R \text{ dB}$$

Error analysis

- error energy

$$\sigma_e^2 = \Delta^2/12, \qquad \Delta = (B - A)/2^R$$

- signal energy

$$\sigma_x^2 = (B - A)^2/12$$

- signal to noise ratio

$$\text{SNR} = 2^{2R}$$

- in dB

$$\text{SNR}_{\text{dB}} = 10 \log_{10} 2^{2R} \approx 6R \text{ dB}$$

The "6dB/bit" rule of thumb

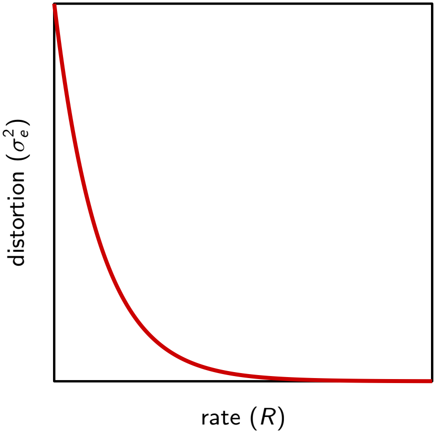
- a compact disk has 16 bits/sample:

$$\max \text{SNR} = 96\text{dB}$$

- a DVD has 24 bits/sample:

$$\max \text{SNR} = 144\text{dB}$$

Rate/Distortion Curve

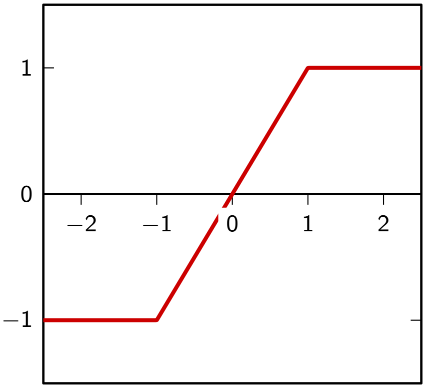


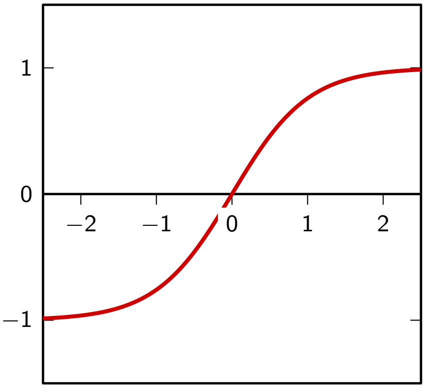
Other quantization errors

If input is not bounded to $[A, B]$ several options; eg:

- clip samples to $[A, B]$: linear distortion (can be put to good use in guitar effects!)
- smoothly saturate input: this simulates the saturation curves of analog electronics

Clipping vs saturation





Analysis of the quantization error

- so far we have only a *quantitative* result on the error (its power)
- to understand the distortion we need the error's spectrum
- quantizer is nonlinear: impossible to compute the spectrum exactly
- the common approach is to make *assumptions* on the error statistics

High-resolution hypothesis

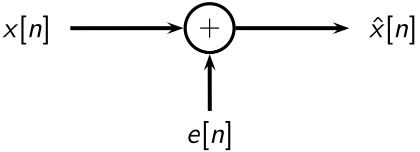
drastic simplification of the problem: if

- input samples are iid (they are not)
- R is relatively large

then we can try to use the following model:

- error samples are iid
- error is uncorrelated to the signal
- quantization error eqivalent to additive white noise with $P_e(\omega) = \Delta^2/12$

High-resolution hypothesis



problems with this model:

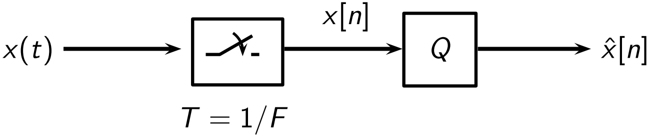
- error is not random!
- error is not white or uncorrelated to the input

common approaches:

- use *dithering* to whiten the noise spectrum
- use *feedback* in the quantization loop to perform *noise shaping*

oversampled A/D conversion

A/D conversion



$$\hat{\mathbf{x}} = \mathbf{x} + \mathbf{e}$$

Oversampled A/D

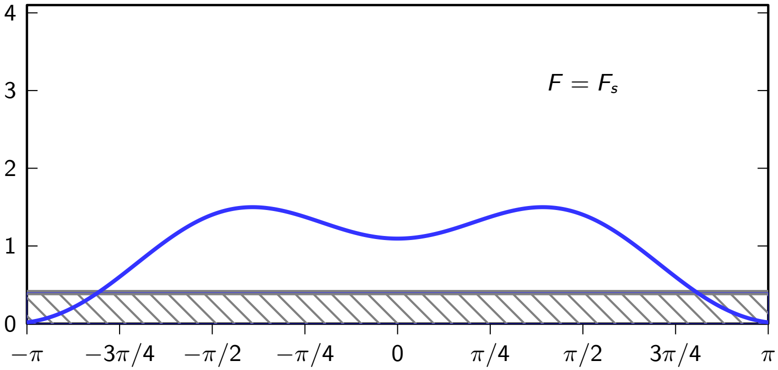
Key assumptions on quantization error:

- **e** is a white noise process, independent of **x**
- PSD of quantization noise is flat, $P_e(\omega) = \frac{\Delta^2}{12}$
- PSD of quantization noise is independent of sampling rate F

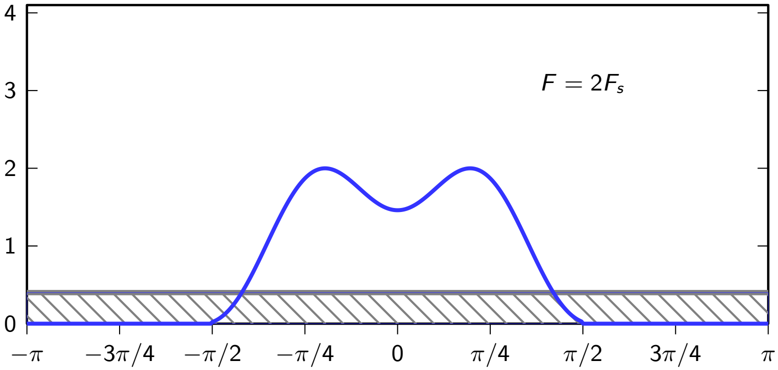
Key observations:

- $x(t)$ is F_s-BL
- spectrum of sampled signal is $X(\omega) = FX(\frac{\omega}{2\pi}F)$
- with N-times oversampling, spectral support is $[-\pi/N, \pi/N]$

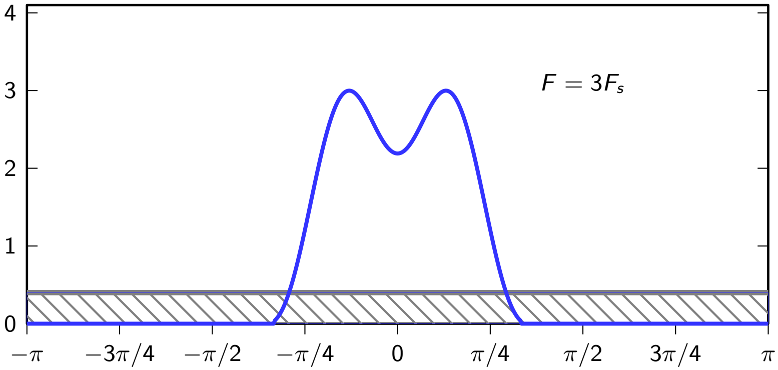
Oversampled A/D



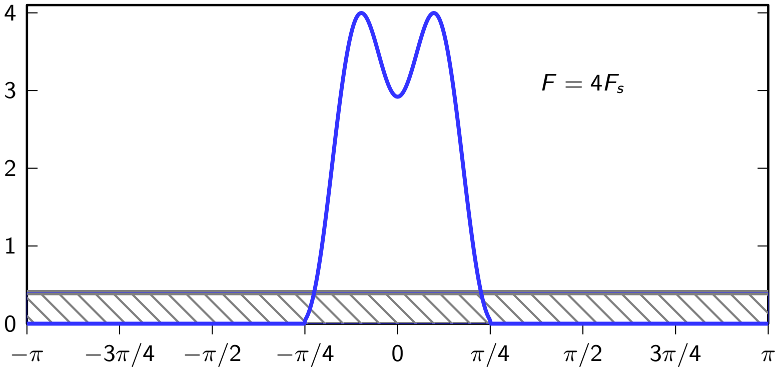
Oversampled A/D



Oversampled A/D



Oversampled A/D

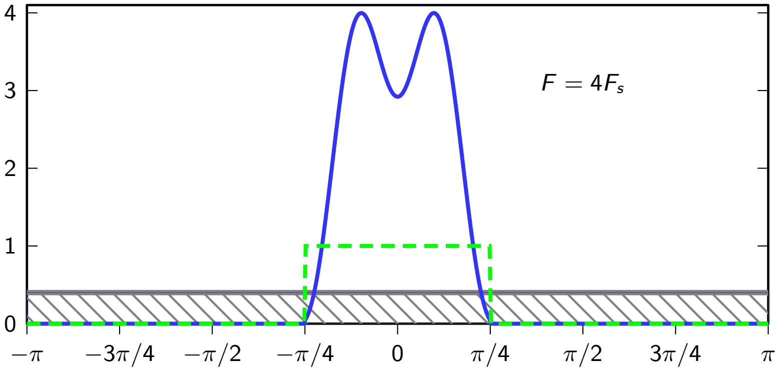


Oversampled A/D

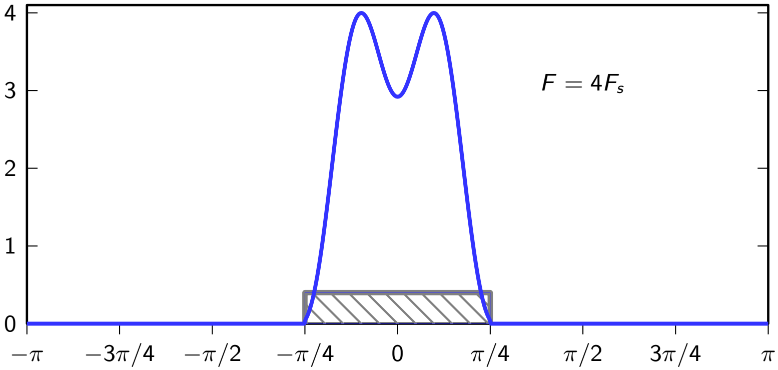
Idea:

- oversample by a factor of N
- signal's spectral support shrinks
- if quantization noise remains independent, its PSD remains flat
- filter out the quantization noise out of band
- downsample back to F_s

Oversampled A/D

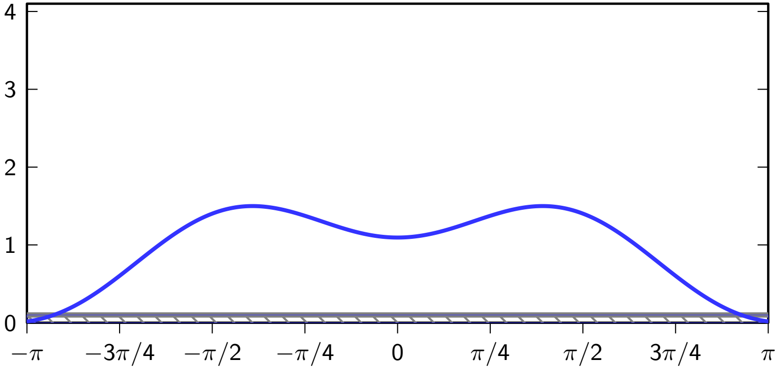


Oversampled A/D

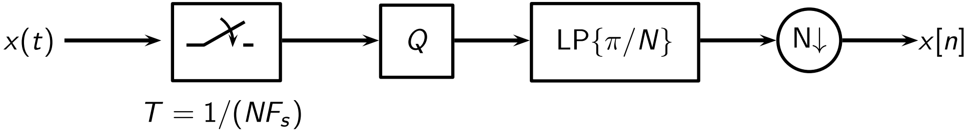


Oversampled A/D

after downsampling by N, $X_o(\omega) = (1/N)X(\omega/N)$



Oversampled A/D



- in theory, SNR at the output is N times better
- 3dB gain per octave (i.e. per doubling of the sampling rate)
- but key assumptions (independence of error) breaks down fast...