ELEG-636: Statistical Signal Processing

Gonzalo R. Arce

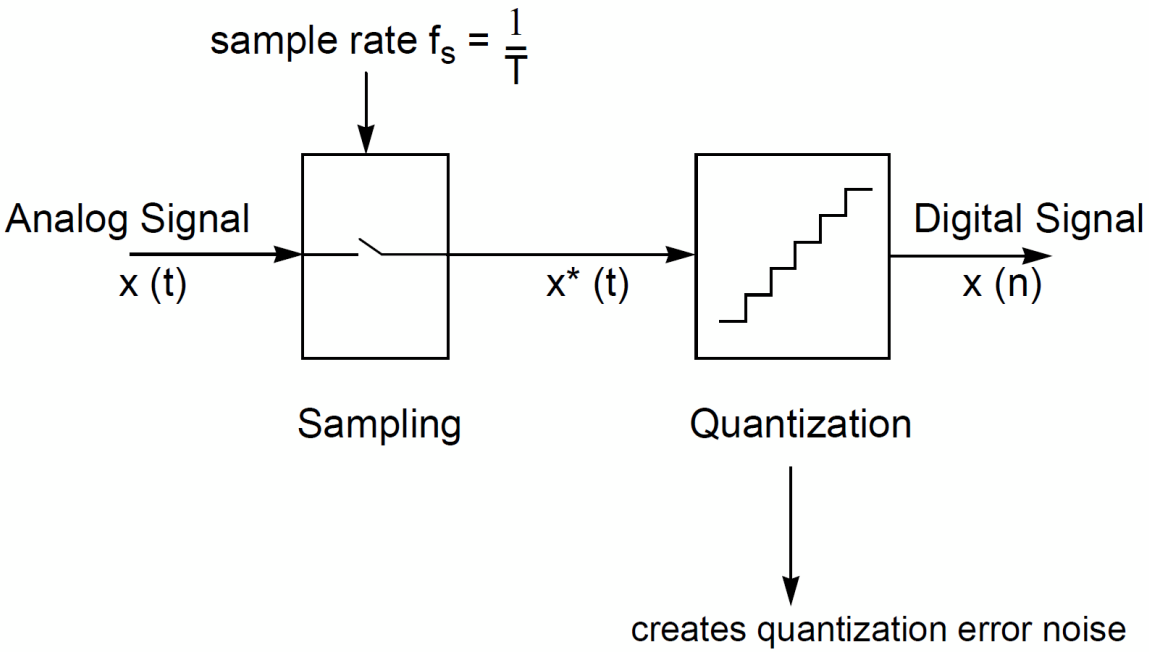
Department of Electrical and Computer Engineering
University of Delaware

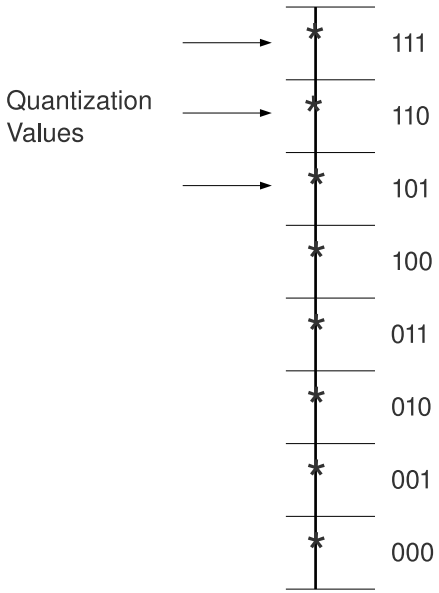
Spring 2010

Quantization

Quantizer

Mapping of the continuous domain of the sample values into a finite number of selected values. A b-bit quantizer can map a signal into 2^b levels.



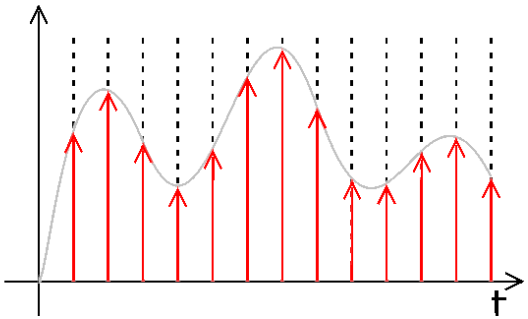


Quantization

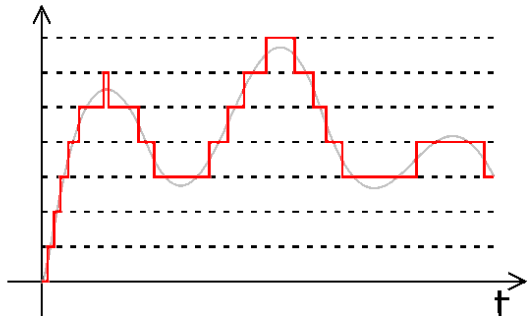
Quantizer

Mapping of the continuous domain of the sample values into a finite number of selected values. A b-bit quantizer can map a signal into 2^b levels.

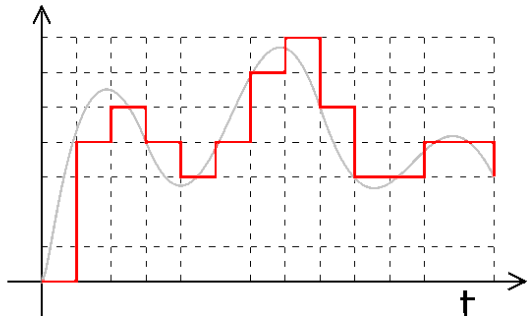
Sampling

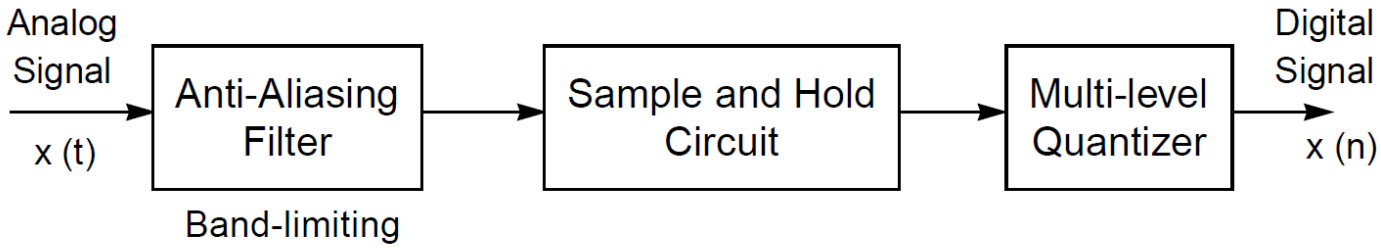


Quantization

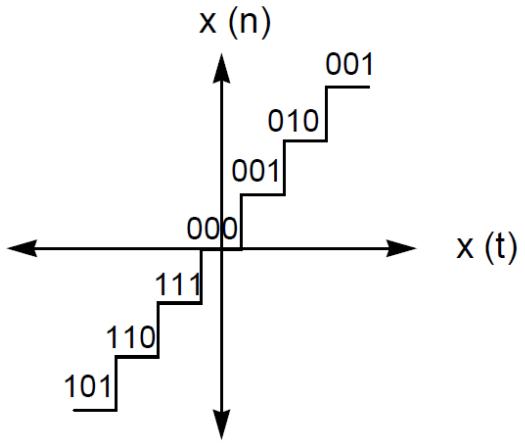


S & Q





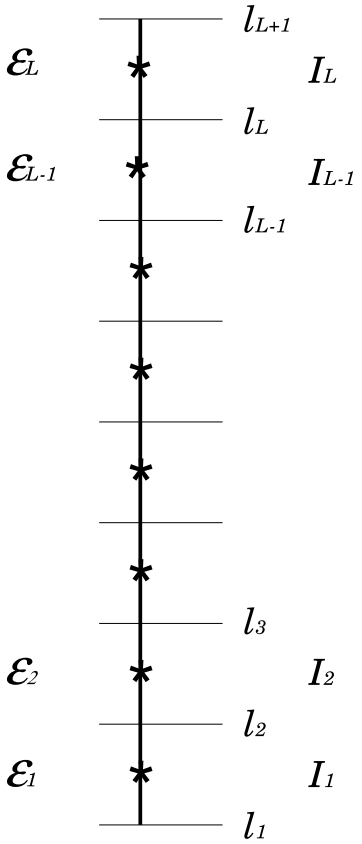
e.g.: Successive Approximation
Flash Conversion
Dual Slope Method



For L intervals we need $[log_2 L]$ bits or nearest larger integer. Sampling at f_s samples/sec., the bit rate is

$$R = f_s \lceil log_2 L \rceil \text{bits/sec.}$$

Consider the quantizer:



Let $MSQE_i$ be the mean squared quantization error when sample is in the i^{th} quantization interval.

$$MSQE = \sum_{i=1}^{L} [MSQE_i] P_i$$

where $P_i = p\{x \in I_i\}$, $p_i = \int_{I_i} f(x)dx$ and $f(x) =$ density function.

$$MSQE_i = \int_{I_i} (x - \varepsilon_i)^2 f(x|i)dx,$$

where $f(x|i)$ is the conditional density of x given that x lies in the i^{th} interval.

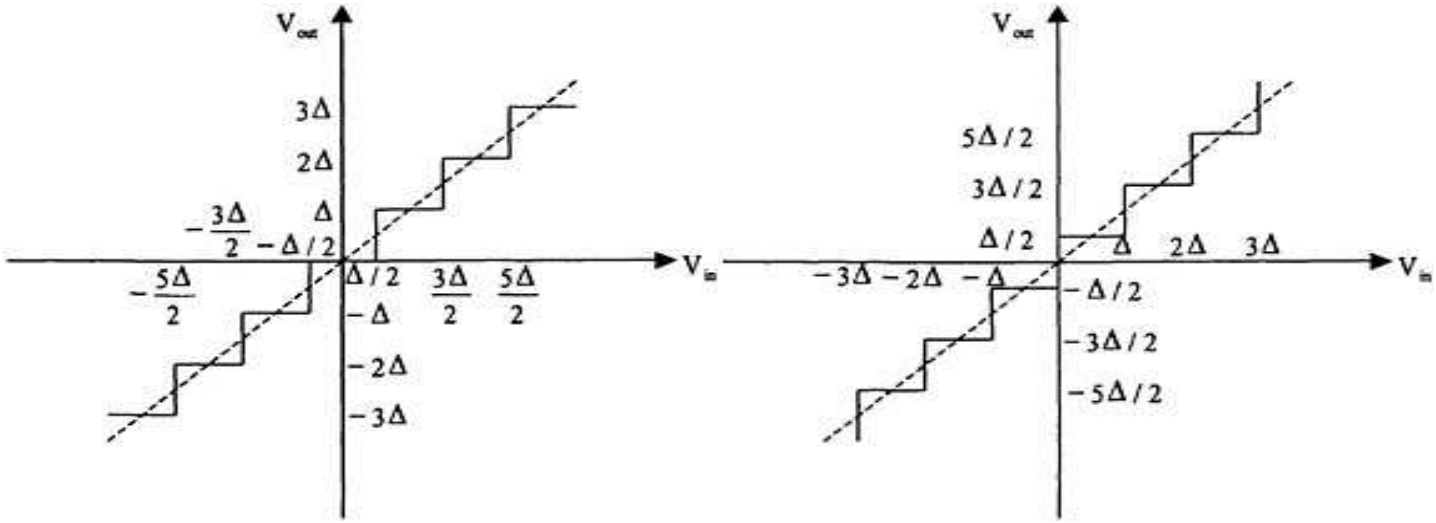
$$f(x|i) = \begin{cases} \frac{f(x)}{P_i} = \frac{f(x)}{\int_{I_i} f(x)dx} & \text{for } x \in I_i \\ 0 & \text{otherwise} \end{cases}$$

So $MSQE$ depends on ε_i, I_i and $f(x)$.

Uniform Quantizer

Generic applications use a *uniform quantizer*. There are two variations called *mid-rise* and *mid-tread* uniform quantizers.
If $x \in [-1, 1]$, an M bit uniform quantizer results in:

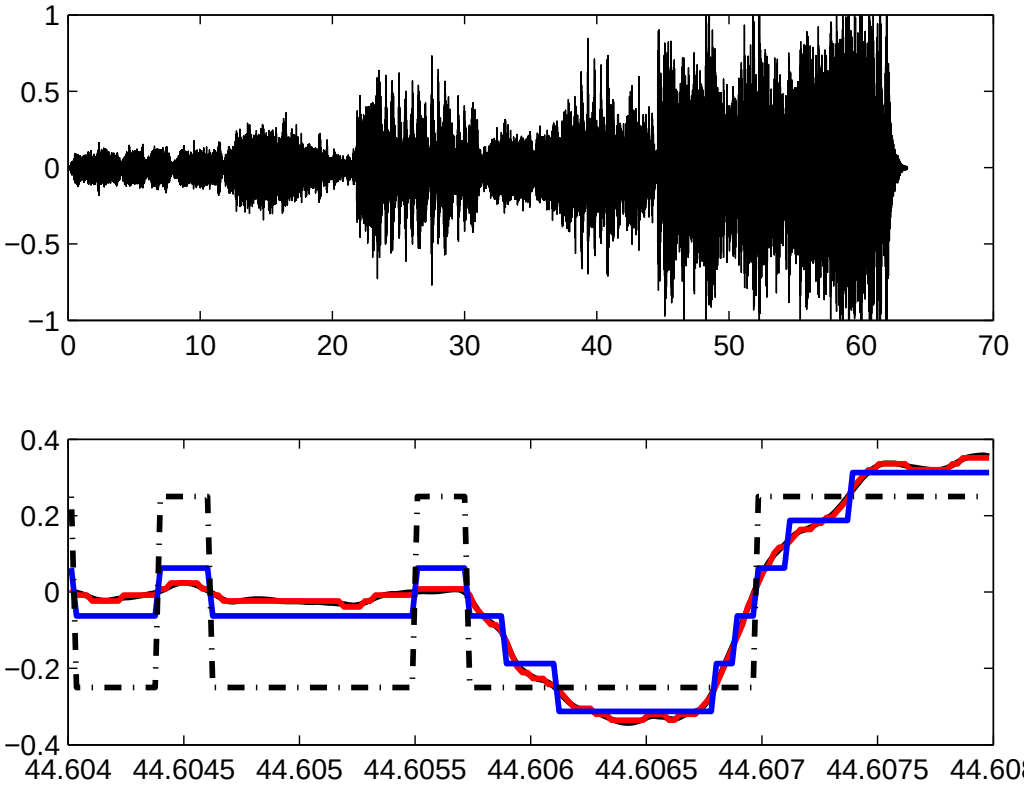
$$Q_{mid-rise}(x) = \frac{\lfloor 2^{M-1}x \rfloor + 0.5}{2^{M-1}} \qquad Q_{mid-tread}(x) = \frac{\lfloor 2^{M-1}x + 0.5 \rfloor}{2^{M-1}}$$



Audio Quantization

Example of Audio quantization:

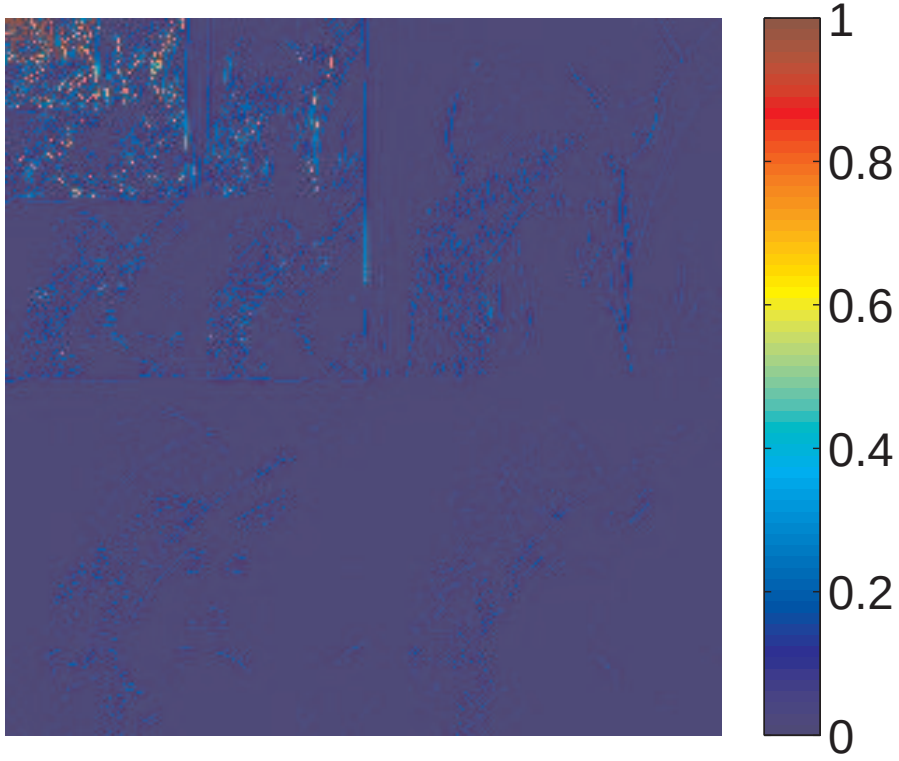
Top: 60 second music recording (16-bit).
Bottom: Signal detail.
Red: 8-bit.
Blue: 4-bit.
Dashed: 2-bit.



Java Applet

Optimal Quantization

Why?



Recall $MSQE = \sum_{i=1}^{L} P_i \int_{l_i} (x - \varepsilon_i)^2 f(x|i) dx$, where $f(x|i) = \frac{f(x)}{P_i}$.

$$MSQE = \sum_{i=1}^{L} \int_{l_i}^{l_{i+1}} (x - \varepsilon_i)^2 f(x) dx$$

. Minimize:

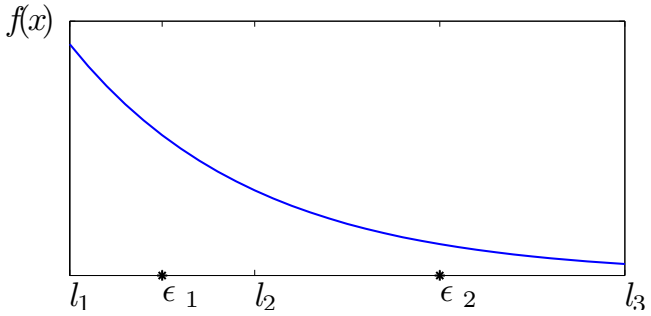
1. $\frac{\partial MSQE}{\partial \varepsilon_i} = 2 \int_{l_i}^{l_{i+1}} (x - \varepsilon_i) f(x) dx = 0, \text{ for } i = 1, \ldots, L$
2. $\frac{\partial MSQE}{\partial l_1} = -(l_1 - \varepsilon_1)^2 f(l_1) = 0$
3. $\frac{\partial MSQE}{\partial l_{L+1}} = (l_{L+1} - \varepsilon_L)^2 f(l_{L+1}) = 0$
4. $\frac{\partial MSQE}{\partial l_i} = (l_i - \varepsilon_{i-1})^2 f(l_i) - (l_i - \varepsilon_i)^2 f(l_i) = 0 \text{ for } i = 1, \ldots, L$

$2L + 1$ Equations.

Example

Design an optimal one bit quantizer for samples obeying an exponential density,

$$f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$



From (1),

$$\int_{l_1}^{l_2} x e^{-x} dx = \varepsilon_1 \int_{l_1}^{l_2} e^{-x} dx$$

or,

$$\varepsilon_1 = \frac{\int_0^{l_2} x e^{-x} dx}{P_1} = \frac{1 - e^{-l_2}(1 - l_2)}{(1 - e^{-l_2})}$$

$$\varepsilon_2 = \frac{\int_{l_2}^{\infty} x e^{-x} dx}{P_2} = \frac{e^{-l_2}(1 + l_2)}{(1 - P_1)} = (1 + l_2)$$

From (4) and $i = 2$,

$$\left[(l_2 - \varepsilon_1)^2 - (l_2 - \varepsilon_2)^2\right] f(l_2) = 0$$

or

$$(l_2 - \varepsilon_1)^2 - (l_2 - [1 + l_2])^2 = 0$$

$$(l_2 - \varepsilon_1)^2 = 1$$

$$l_2 = 1 + \varepsilon_1$$

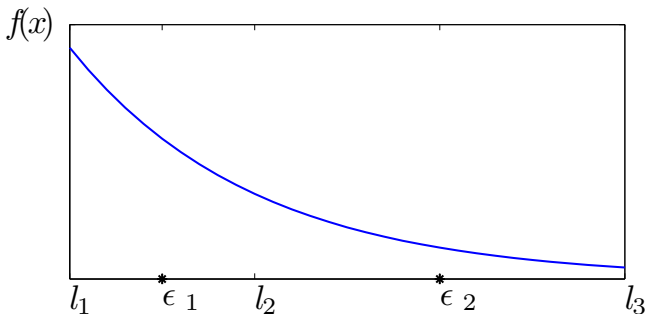
Solve for the 3 eqs.:

$1 - e^{-l_2}(1 - l_2) = \varepsilon_1(1 - e^{-l_2})$,

$\varepsilon_2 = l_2 + 1$

and

$\varepsilon_1 = l_2 - 1$



$l_2 = 2.265$, $\varepsilon_2 = 3.265$ and $\varepsilon_1 = 1.265$

Dithering

The quantization is a non-linear operation, which has been approximately described by:

$$y(t) = Q[x(t)] \approx x(t) + n(t, x)$$

Where $x(t)$ is the input signal, $y(t)$ is the quantized signal, and the effect of the quantizing operation is reflected in a noise signal, $n(t, x)$, which depends on the input signal x.
A Dither signal is added to the input of the quantizer and then subtracted after the quantizing operation.
Thus,

$$y(t) = Q[x(t) + d(t)] - d(t) \approx x(t) + n(t, x + d)$$

Substraction, is normally done at the receiver.

Minimum loss of statistical data due to the quantizer operation occurs when the quantization noise is independent of the input signal.

$$N(t,x) = N(t)$$

A metric used to determine wether the noise is independent of the signal is the second-order statistic (referred to in the literature as the $''D''$ factor), which is defined as follows:

$$D == \int (x - \bar{y}_x)^2 f(x) dx,$$

where

$$\bar{y}_x \triangleq \int f(y|x) y dy$$

and

$$y = n + x + d$$

The quantizer output $\tilde{y}$ is related to the system output y by

$$\tilde{y} - d = y$$

Revealing that the following relations exist

$$E(\tilde{y} - d) = E(y)$$

$$E(\tilde{y} - d|x = X) = E(y|x = X)$$

$$E(\tilde{y}|x = X) - E(d|x = X) = E(y|x = X)$$

Assuming

$$f(d|x = X) = f(d) \textbf{ and } E(d) = 0$$

Then,

$$E(\tilde{y}|x = X) = E(y|x = X)$$

And

$$E(\tilde{y}|x = X) = \int [n(x + d) + x + d]f(d)dd$$

$$E(y|x = X) = x + \int n(x + d)f(d)dd$$

Thus,

$$E(\tilde{y}|x = X) = \int [n(x + d) + x + d]f(d)dd$$

$$E(y|x = X) = x + \int n(x + d)f(d)dd$$

Substituting into D's definition:

$$D = \int [E(n|x = X)]^2 f(x)dx$$

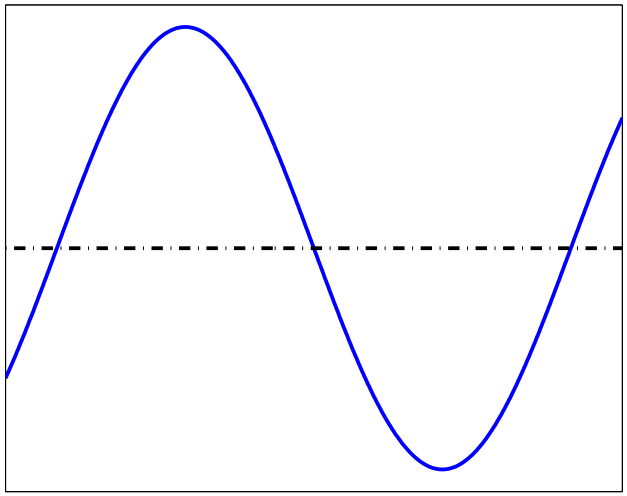
Since $f(x)$ and $[E(n|x)]^2$ are non-negative, $D = 0$ if, and only if,

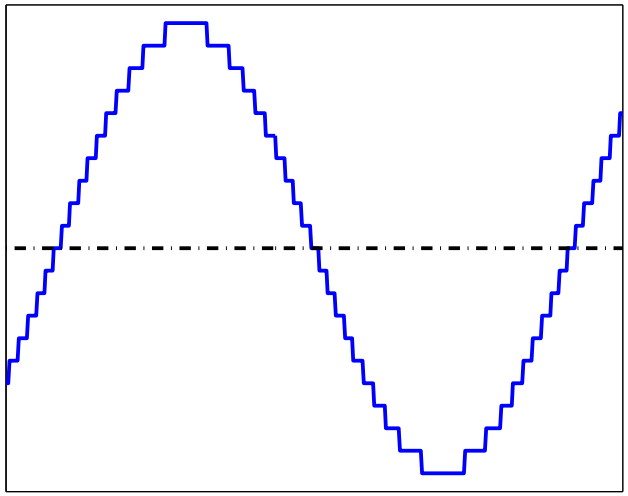
$$E(n|x) = 0$$

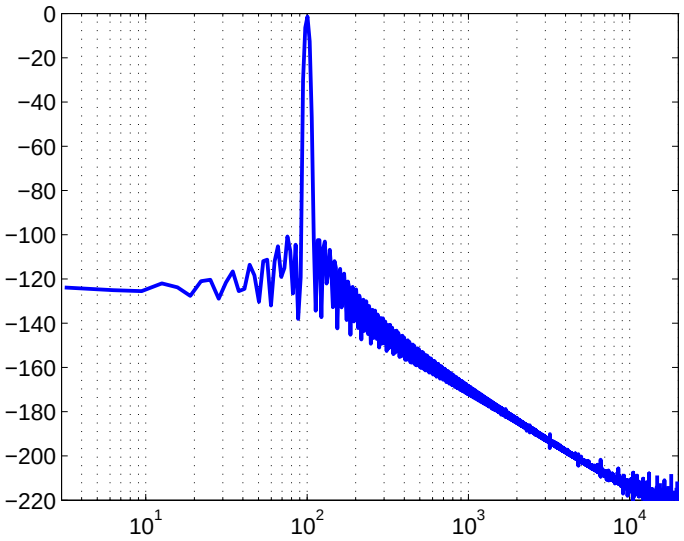
Using the moment generating function it's possible to conclude that $D = 0$ if, and only if,

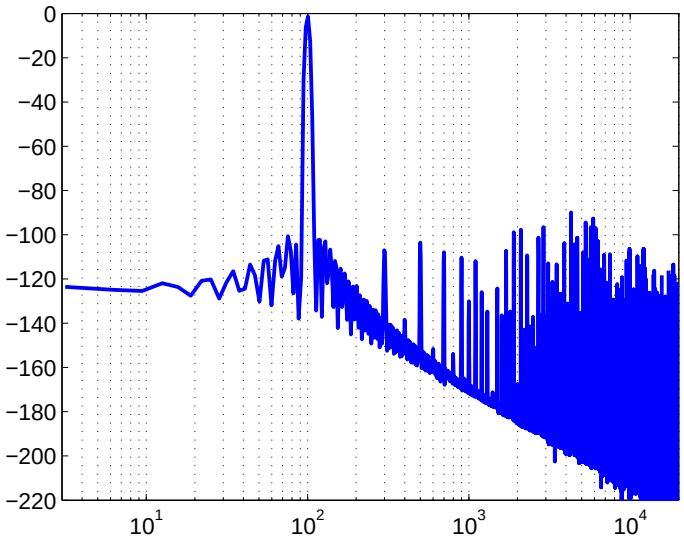
$$f(n|x = X) = f(n)$$

Sinusoidal 100Hz, 44.1kHz sampling. (24-bit, 16-bit)



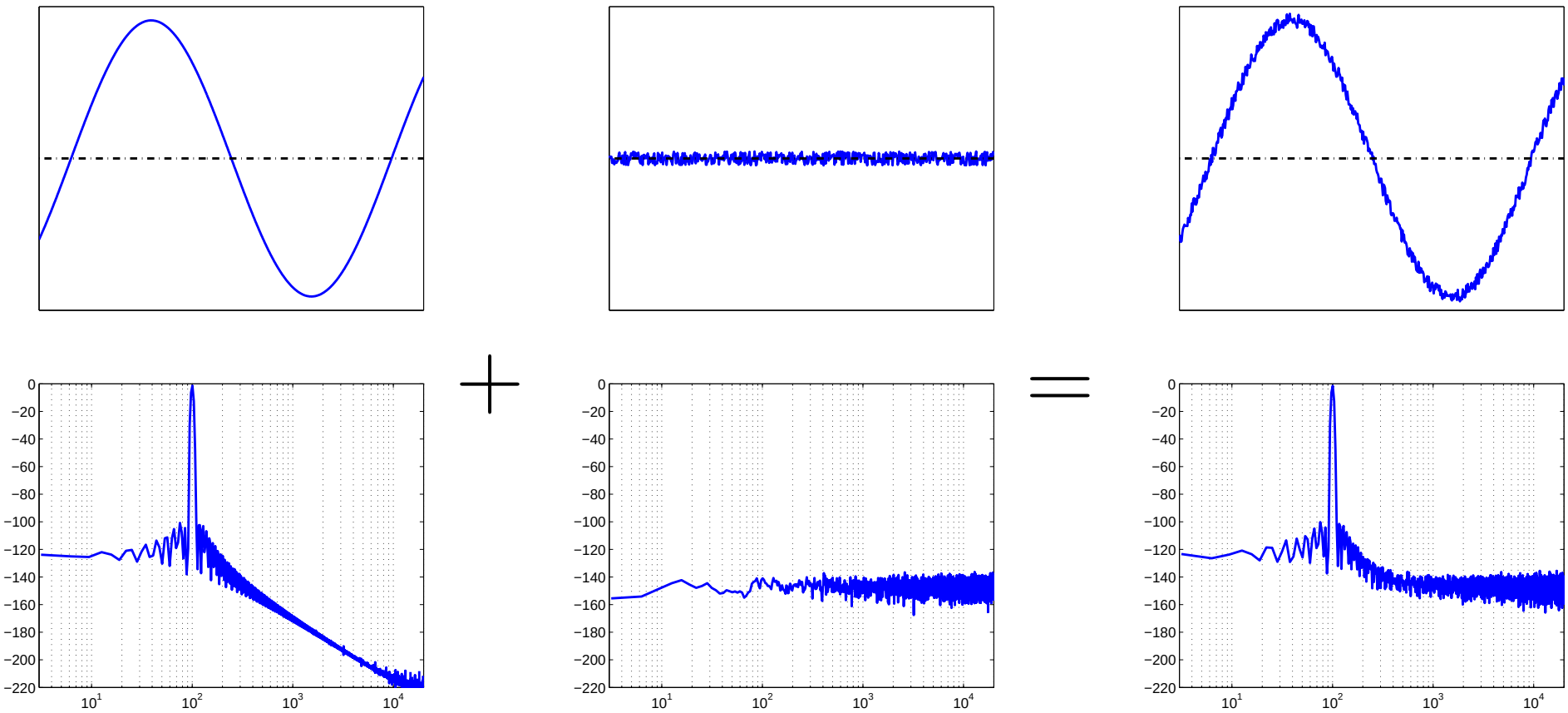






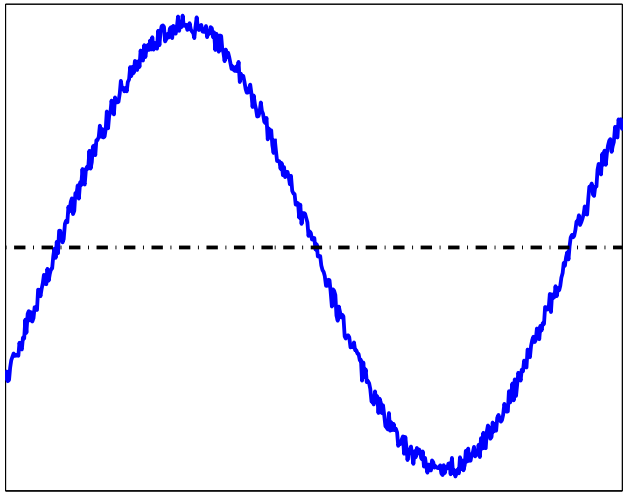
Dithering

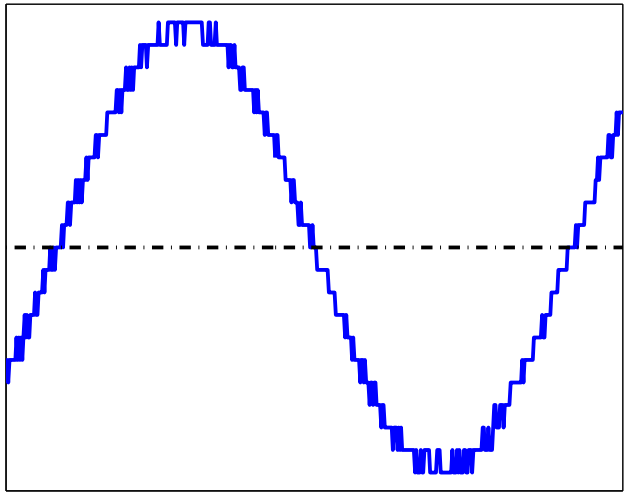
Sinusoidal 100Hz, 44.1kHz sampling + Dither

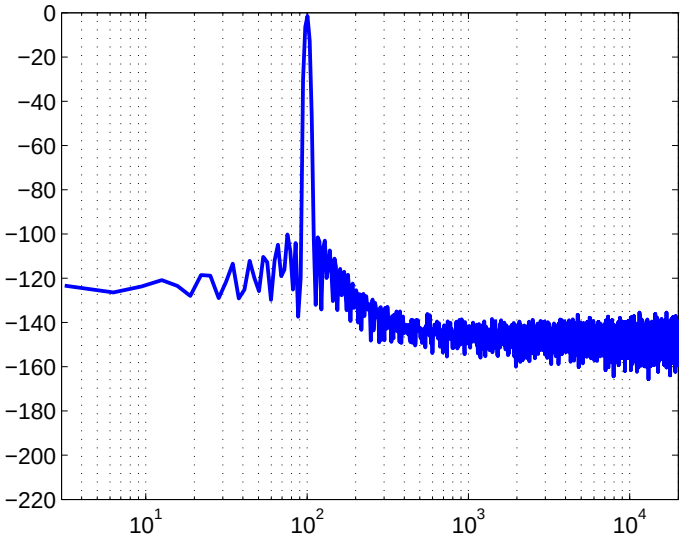


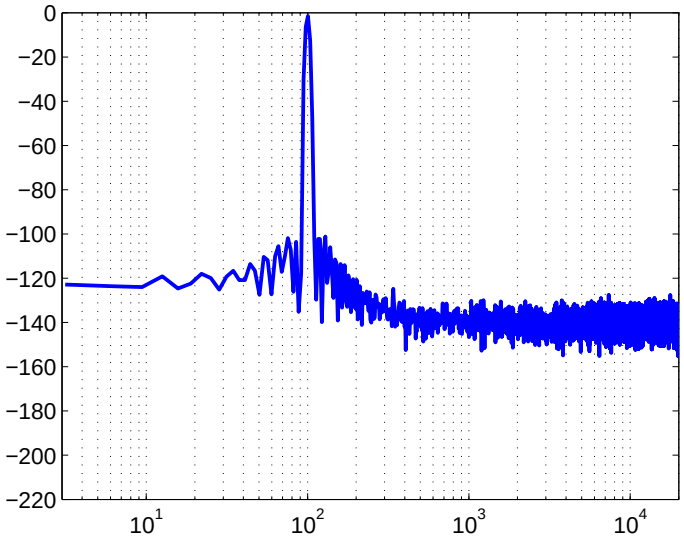
Dithering

Sinusoidal 100Hz, 44.1kHz sampling. (24-bit + dither, 16-bit dithered)



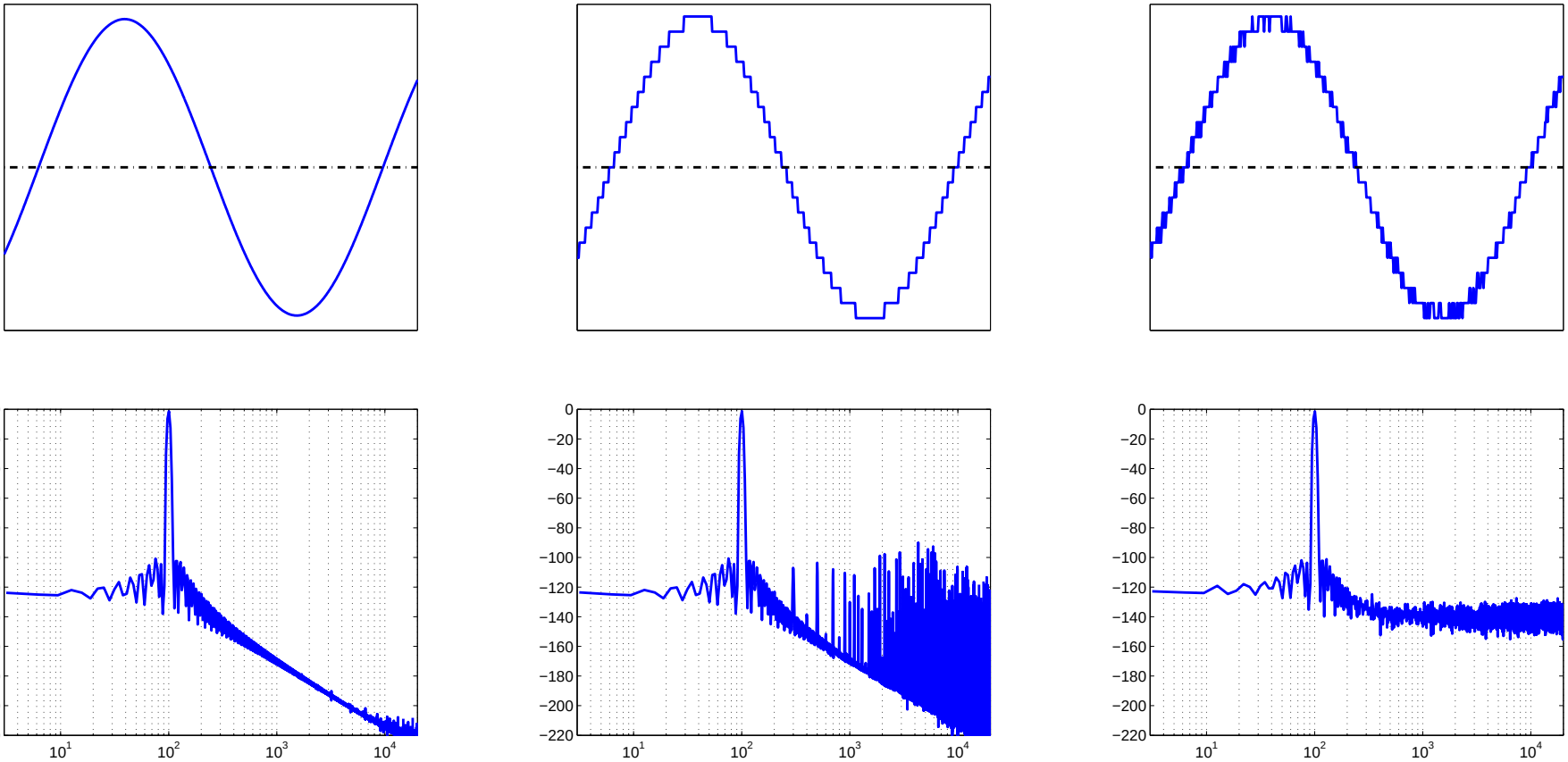






Dithering

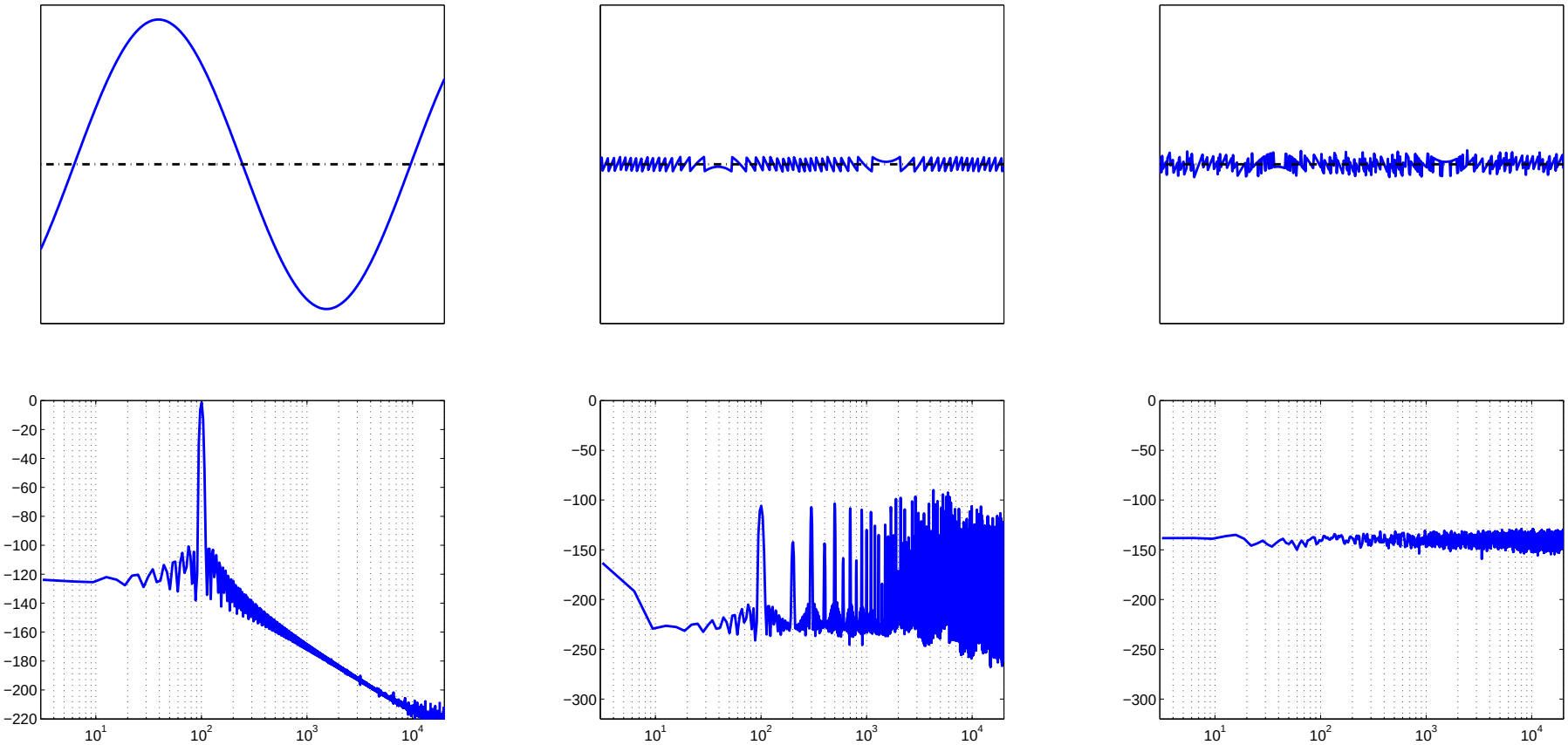
Sinusoidal 100Hz, 44.1kHz sampling (24-bit, 16-bit, 16-bit dithered).



Java Applet

Dithering - Quantization Noise

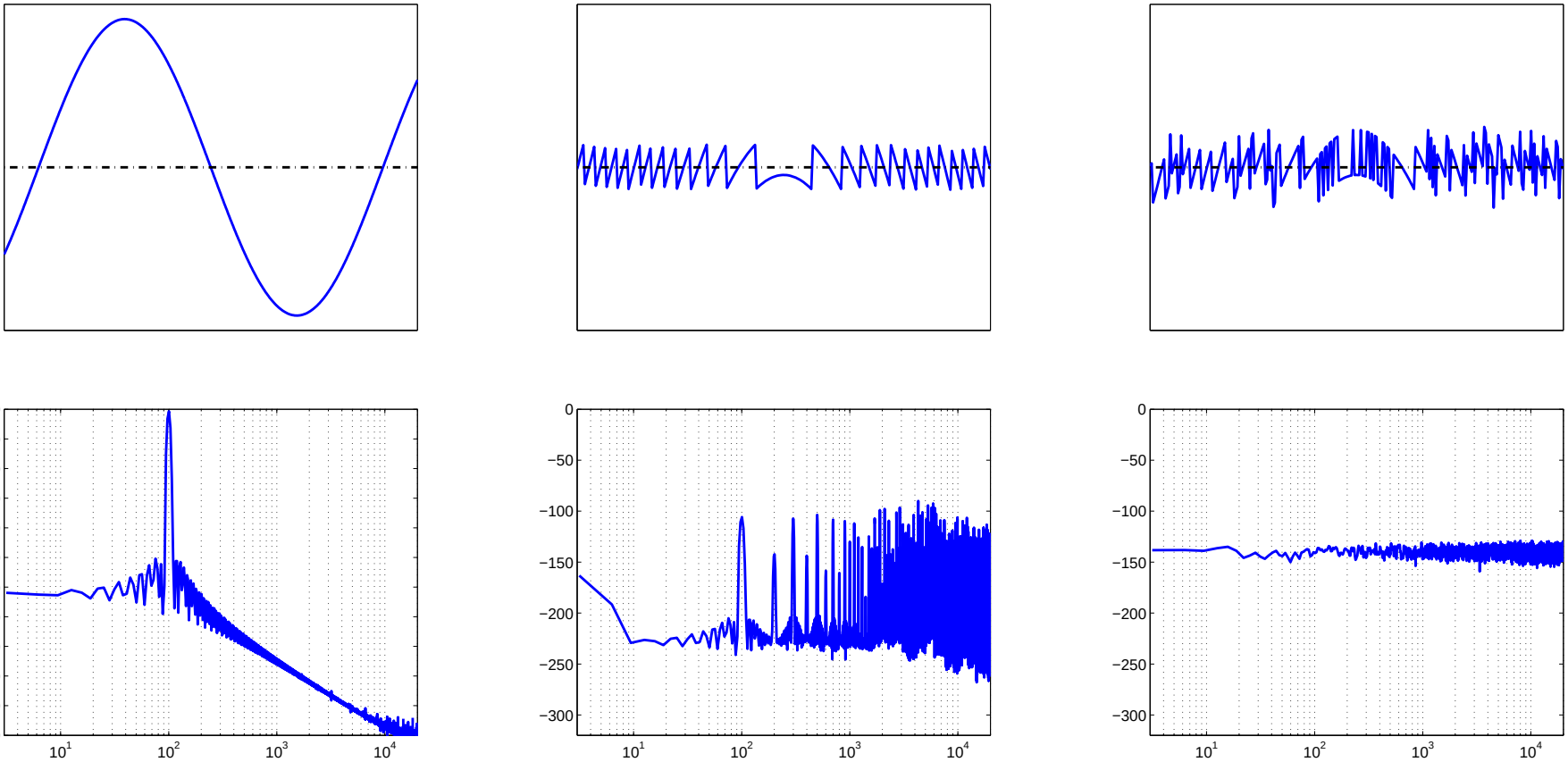
Sinusoidal 100Hz, 44.1kHz sampling (24-bit, 16-bit, 16-bit dithered).



Java Applet

Dithering - Quantization Noise

Sinusoidal 100Hz, 44.1kHz sampling (24-bit, 16-bit, 16-bit dithered).



Java Applet

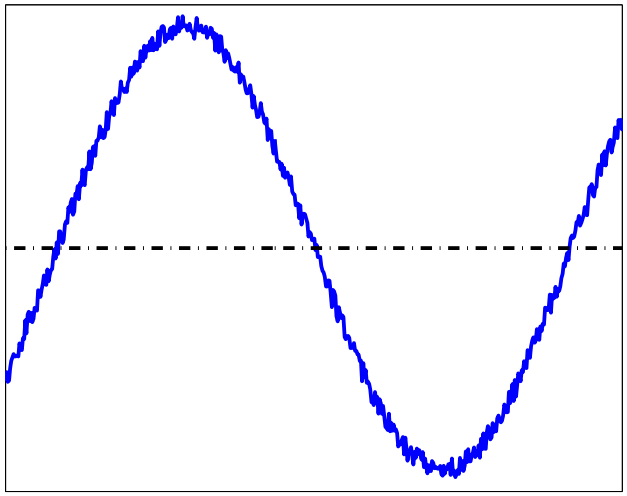
Oversampling

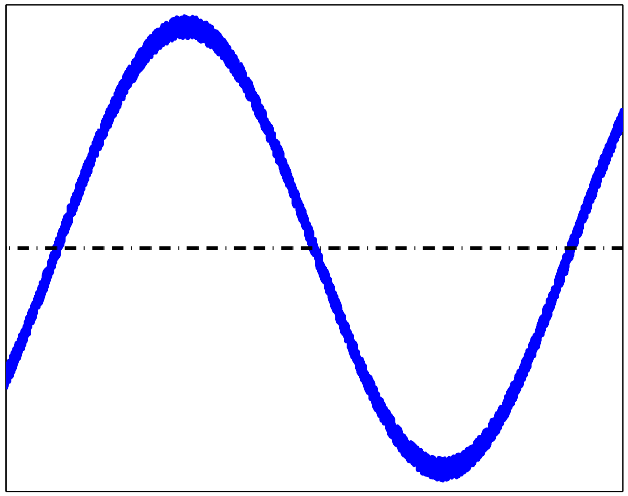
Signals are typically sampled at Nyquist rate. Oversampling refers to the process of sampling at frequencies significantly higher.
Advantages:

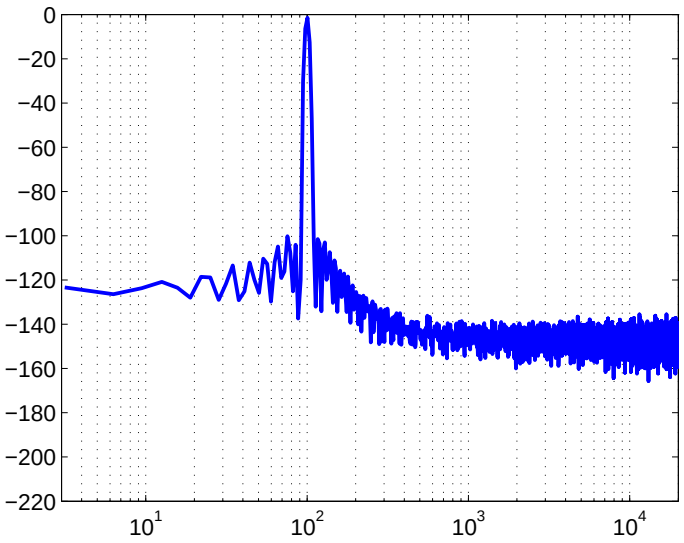
1. Quantization noise can be reduced.
2. Allows better filtering.
3. Since noise spreads out, the SNR increases. ADC resolution can be effectively increased in the digital domain.

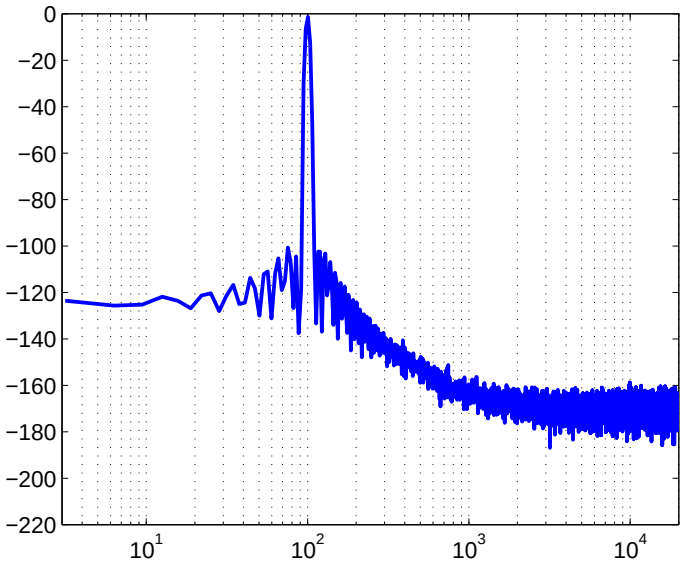
Note that while Nyquist ADCs require sample and hold mechanisms to achieve higher resolution. Oversampled converters use difference between samples to obtain better estimates of the signal.

Sinusoidal 100Hz + Dither, 44.1kHz/441kHz sampling.



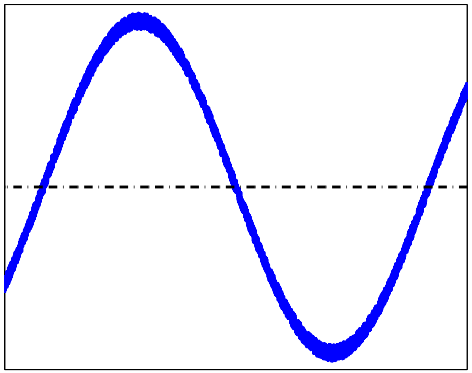


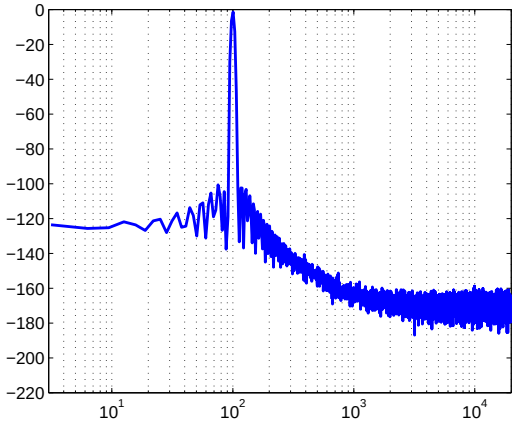




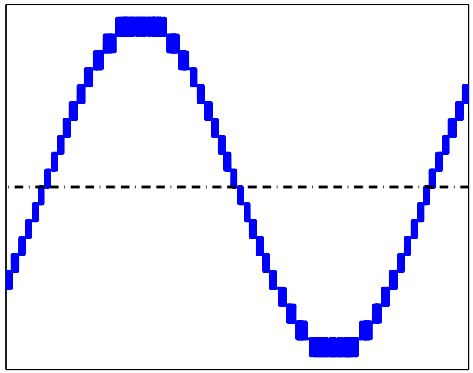
Sinusoidal 100Hz, 44.1kHz sampling

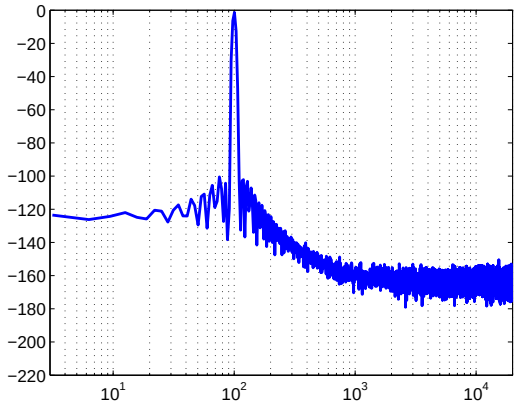
24-bit OS +dither



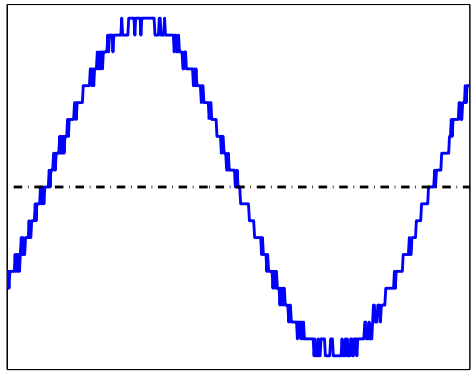


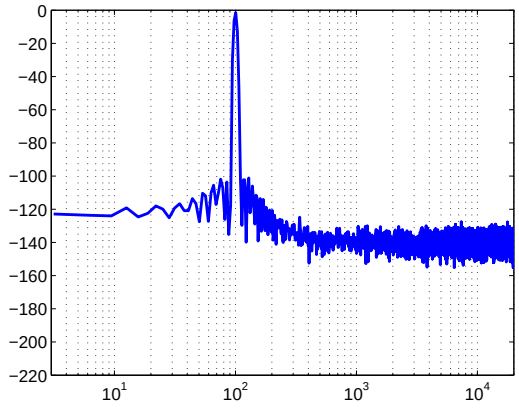
16-bit OS dithered



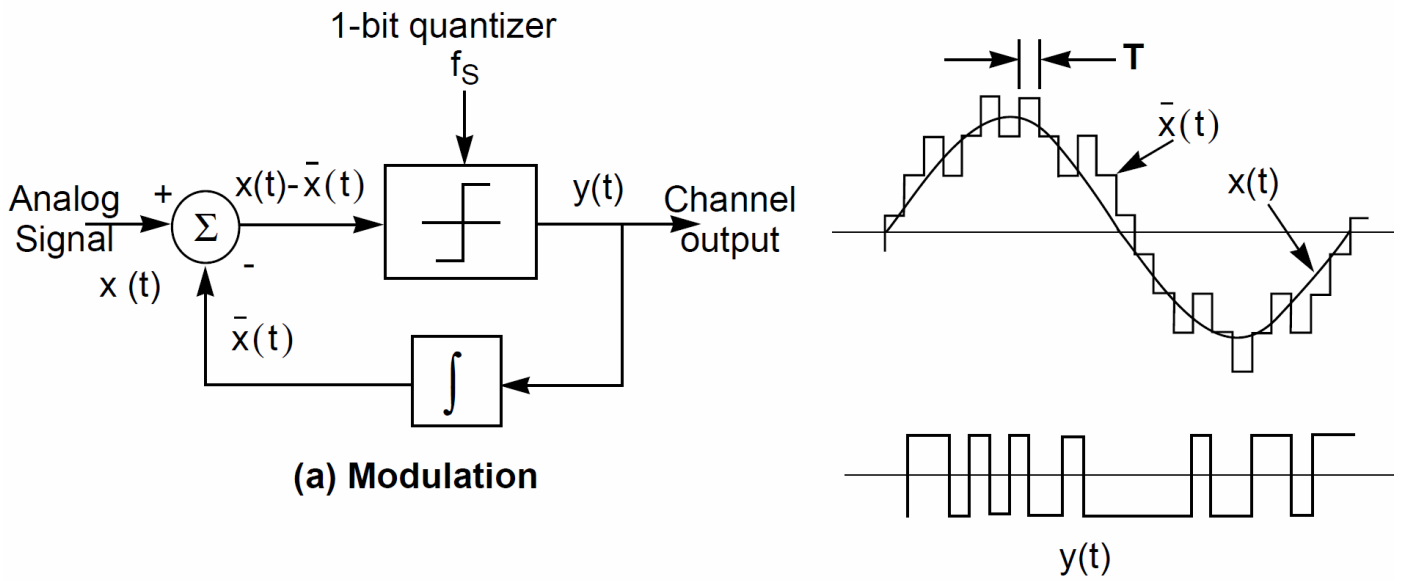


16-bit dithered





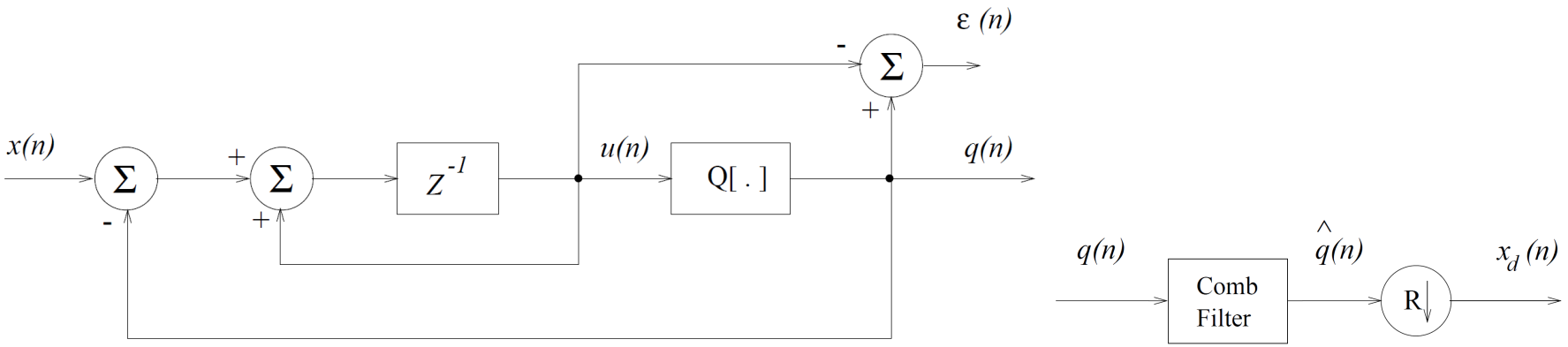
Delta Sigma (Δ − Σ) ADC



Delta Sigma converters are oversampled converters with noise shaping capabilities. The goal is to push the noise power out of the band of interest. This technique furthers increase the SNR.

This technique, allows for this type of converters to achieve the higher resolution.

Delta Sigma (Δ − Σ) ADC



Delta Sigma converters are oversampled converters with noise shaping capabilities. The goal is to push the noise power out of the band of interest. This technique furthers increase the SNR.

This technique, allows for this type of converters to achieve the higher resolution.

A b-bit uniform quantizer $Q[x(t)]$ maps a real-valued input signal $x(t)$ into one of 2^b quantization values; the interval between successive levels is $q = \frac{X}{2^{b+1}}$, where X is the range of the quantizer.
Define the Signal-to-Quantization Noise Ratio,

$$SQNR = 10 \log_{10} \frac{P_x}{P_n}$$

where $P_x = \sigma_x^2$ is the signal power and $P_n = \sigma_e^2$ is the power of the quantization noise.

If the quantization noise is uniformly distributed in $(\frac{q}{2}, -\frac{q}{2})$. The mean value of the error is zero and the variance (quantization error power) is

$$P_n = \sigma_e^2 = \frac{1}{q}\int_{-q/2}^{q/2} e^2 de = \frac{q^2}{12}$$

Thus

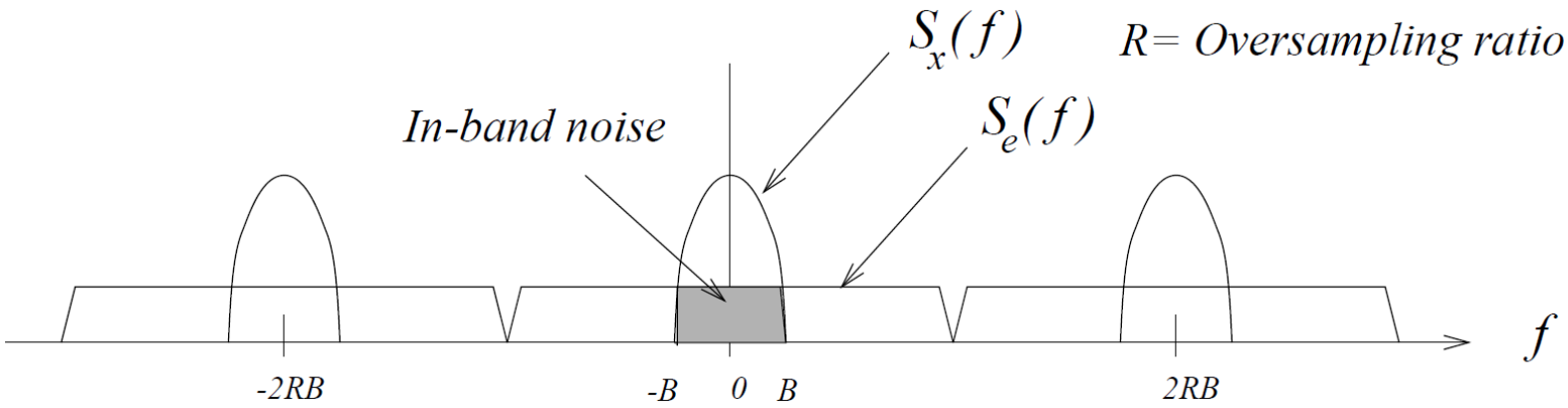
$$\begin{aligned} SQNR &= 10\log\frac{P_x}{P_n} = 10\log\frac{\sigma_x^2}{\sigma_e^2} = 10\log\frac{\sigma_x^2}{q^2/12} \\ &= 10\log\left(\frac{\sigma_x^2}{X^2}12 * (2^{b+1})^2\right) \\ &= 6.02b + 16.81 - 20\log\frac{X}{\sigma_x} \quad dB \end{aligned}$$

This formula is used to specify the precision needed in an ADC: each additional bit in the quantizer increases the SQNR by 6dB.

If the input signal is sampled at $2RB$ (oversampled by R) and a uniform quantizer of b bits is applied, the quantization error in $S_e(f)$ is now spread over a larger region ($\sigma^2_{ose} = \sigma^2_e/R$), but the signal power contained in $S_x(f)$ is still concentrated within its bandwidth:

$$SQNR = 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_e^2/R}\right) = 10\log_{10}(R) + 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_e^2}\right)$$

Thus, for PCM, doubling the sampling frequency decreases the in-band noise by 3dB, increasing the resolution by half a bit.



Let f_s and f_N be the sampling frequency and the Nyquist frequency, respectively; the oversampling ratio is defined as $R = f_s/F_N$. Let $x_c(t)$ and $x(n) = x^c(nT_R)$ be the input analog signal and the oversampled discrete-time sequence fed to the SDM encoder. T_R is the oversampled sampling period.

The sequence $u(n)$ is the integrator state describing the output of the integrator. It is assumed that $x(n) \in [-b, b]$. The binary quantizer is defined by

$$Q[u(n)] = \begin{cases} +b & \text{if } u(n) \geq 0 \\ -b & \text{otherwise.} \end{cases}$$

Define the binary quantizer error sequence as $\epsilon(n) = Q[u(n)] - u(n)$, where $\epsilon(n)$ is approximated as i.i.d. Observe that,

$$u(n) = x(n-1) - (q(n-1) - u(n-1)) = x(n-1) - \epsilon(n-1)$$

and $u(n) = q(n) - \epsilon(n)$. Thus, the SDM output can be written as $q(n) = x(n-1) + \epsilon(n) - \epsilon(n-1)$, which is the one-step delayed input sequence plus a first-order difference of the binary quantizer error.

The power spectral density the noise, $N(n) = e(n) - e(n-1)$, is

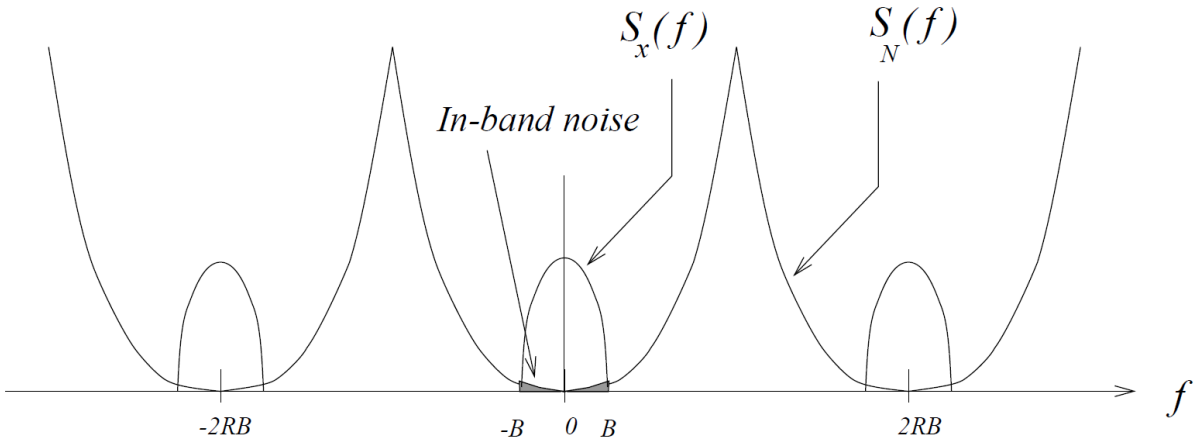
$$\begin{aligned} N(z) &= e(z) - e(z)z^{-1} = e(z)(1 - z^{-1}) \\ H(z) &= \frac{N(z)}{e(z)} = 1 - z^{-1} \\ |H(f)|^2 &= |1 - e^{-j\frac{2\pi f}{f_s}}|^2 = 4\sin^2\left(\frac{\pi f}{f_s}\right) \approx 4\pi^2\left(\frac{f}{f_s}\right)^2 \end{aligned}$$

$$\sigma_n^2 = \int_{-B}^{B} |H(f)|^2 S_e(f) df \approx \int_{-B}^{B} \frac{4\pi^2}{f_s^2} f^2 \left(\frac{\sigma_e^2}{f_s}\right) df = \frac{8\pi^2 \sigma_e^2}{3} \left(\frac{1}{R}\right)^3$$

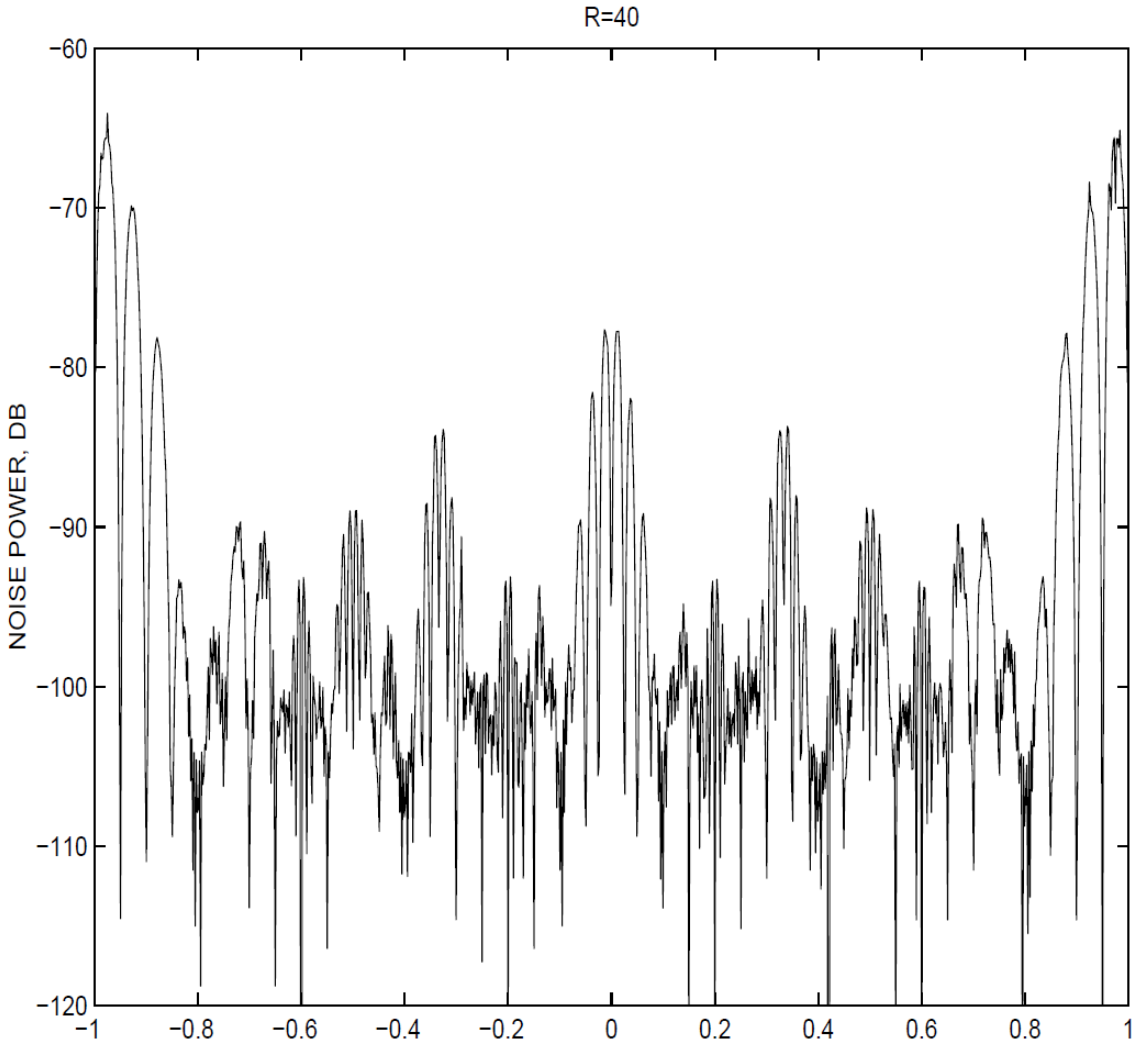
The in-band SQNR can be computed in this case as

$$SQNR = 30 \log_{10}(R) - 10 \log_{10}(8\pi^2/3) + 20 \log_{10}(\sigma_x/\sigma_e)$$

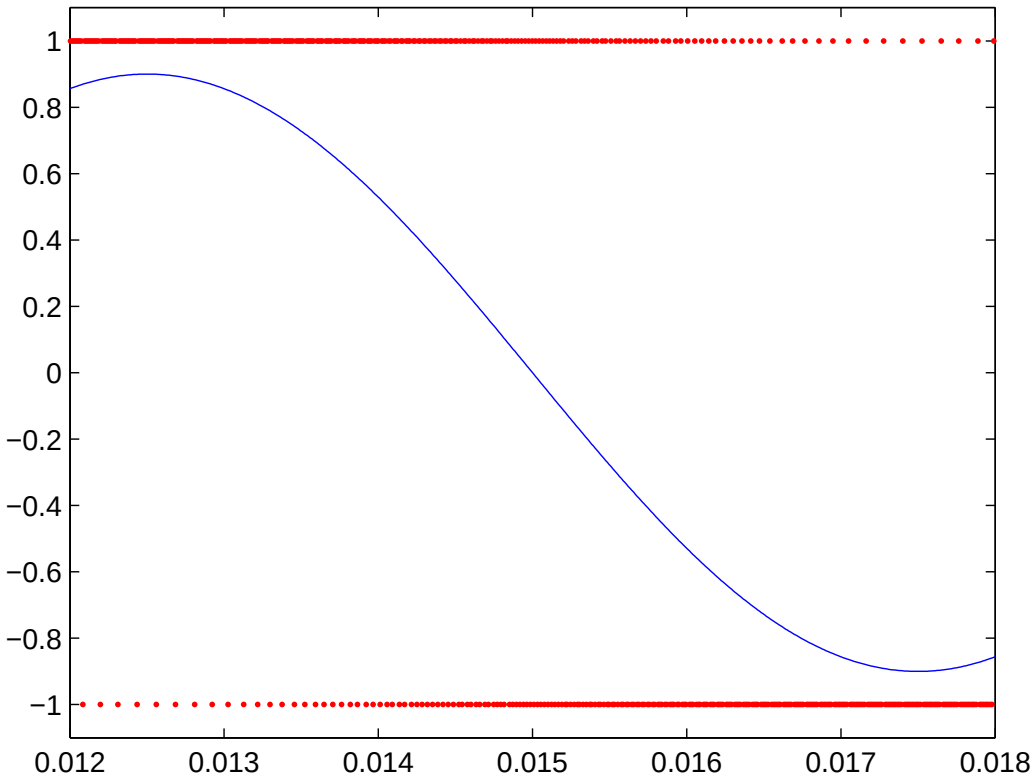
For single-loop $\Delta - \Sigma$ modulation, doubling the sampling frequency increases the SNR by 9 dB, increasing the resolution by one and a half bits.



Error from sigma-delta acquired DC inputs using a linear decimating filter:

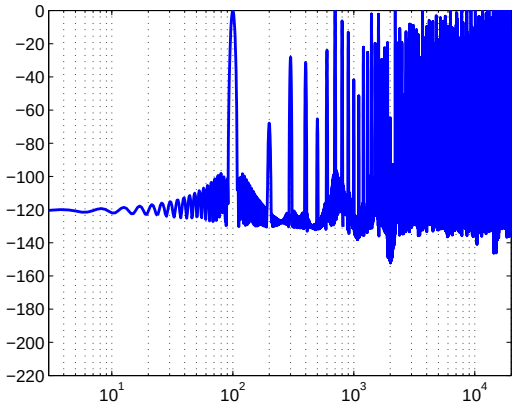


Delta-Sigma 1-bit quantizer:

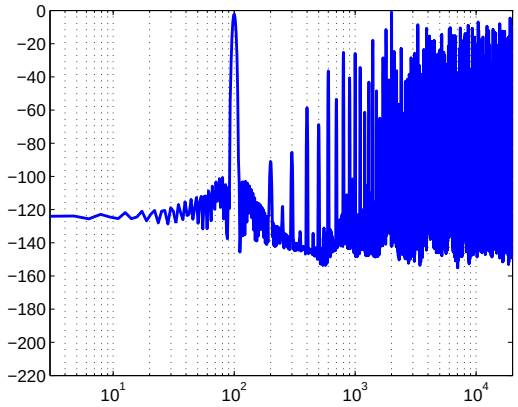


Delta-Sigma 1-bit quantizer (Spectral):

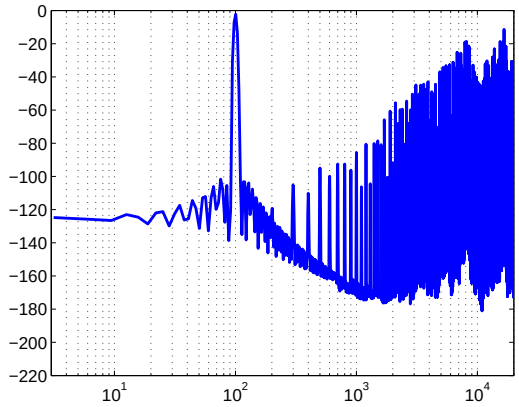
2x Over-sampling



5x Over-sampling



20x Over-sampling



Adding Dither: