Quantization

Henry D. Pfister*

May 5, 2017

1 Introduction

Two operations are necessary to transform an analog waveform into a digital signal. The first action, sampling, consists of converting a continuous-time input into a discrete-time sequence. The second operation is the process of approximating continuous-space amplitude values by a discrete set of possible points. This process, termed *quantization*, is also essential to transmit an analog signal over digital media. Quantization invariably induces a loss in signal quality. The distortion between the original and quantized functions is usually unwanted, and cannot be reversed. Yet, for a specific application, the level of signal degradation can be controlled [1].

In this article, we focus primarily on the quantization of real numbers. The techniques described here can easily be extended to complex numbers by quantizing the real and imaginary parts separately. In a more abstract sense, the quantization of a complex number is equivalent to the vector quantization of a pair of real numbers.

2 Random Processes

To understand the effects of quantization, we first have to observe and understand the spectral behavior of white noise and colored noise. The power of a white-noise process is spread uniformly across the full spectrum $[-\pi, \pi]$, whereas the power of a colored noise process is spread non-uniformly across the spectrum. Let $X[n]$ be a stationary discrete-time white-noise process whose sample at time index n is denoted by the random variable $X[n]$. For example, one can generate a length-L sample from such a random process using the Matlab command `x=randn(1,L)`. The autocorrelation function of a random process $X[n]$ is defined to be

$$R_{xx}[\tau] \triangleq \mathbb{E}\Big[X[n]X[n+\tau]\Big] \; ; \; \tau \in \{\ldots, -1, 0, 1, \ldots\}$$

and the power spectral density is related to it as $S_{xx}(e^{j\omega}) = \mathcal{F}\left(R_{xx}[\tau]\right)$, where $\mathcal{F}$ represents the discrete-time Fourier transform operator. Note that, since this process is *stationary*, $R_{xx}[\tau]$ does not depend on n but only on the time difference between the samples, τ.

Now, let us consider the particular example of a white noise process generated from the standard normal distribution $\mathcal{N}(0,1)$. So each $X[n]$ is *independent* and distributed according to $X[n] \sim \mathcal{N}(0,1)$. From the definition of this distribution we know that if $Z \sim \mathcal{N}(0,1)$ then its mean is $\mathbb{E}[Z] = 0$ and its variance is $\mathbb{E}[Z^2] = 1$. Therefore, for our Gaussian white noise process, we have $\mathbb{E}\big[X[n]\big] = 0$ and $\mathbb{E}\big[X[n]^2\big] = 1$. Let us now compute the autocorrelation function and power spectral density of this process. For $\tau = 0$, we get $R_{xx}[0] = \mathbb{E}\big[X[n]^2\big] = 1$. For $\tau > 0$, observe that $X[n]$ and $X[n+\tau]$ are two *independent* standard normal random variables. Since they are independent, the expectation of their product is the product of their expectations,

$$R_{xx}[\tau] = \mathbb{E}\Big[X[n]X[n+\tau]\Big] = \mathbb{E}\Big[X[n]\Big] \cdot \mathbb{E}\Big[X[n+\tau]\Big] = 0 \cdot 0 = 0.$$

Hence we have

$$R_{xx}[\tau] = \begin{cases} 1 & , \tau = 0 \\ 0 & , \tau \neq 0 \end{cases},$$

*Thanks to Narayanan Rengaswamy for typesetting these notes and contributing some new material.

which is a discrete-time impulse function. This means the noise process is completely uncorrelated with itself for any non-zero lag. Computing the DTFT, we see that the power spectral density is given by $S_{xx}(e^{j\omega}) = 1$ and hence the process is "white".

Next, let us run this process $X[n]$ through a low-pass filter whose impulse response is given by $h[n] = 0.5\delta[n] + 0.5\delta[n-1]$. Hence the output process is given by $Y[n] = 0.5X[n] + 0.5X[n-1]$. Let us proceed to compute the autocorrelation function of the process $Y[n]$. We have

$$\begin{aligned}
R_{yy}[\tau] &\triangleq \mathbb{E}\Big[Y[n]Y[n+\tau]\Big] \\
&= \mathbb{E}\Big[\left(0.5X[n] + 0.5X[n-1]\right)\left(0.5X[n+\tau] + 0.5X[n+\tau-1]\right)\Big] \\
&= 0.25\Big[\mathbb{E}\big[X[n]X[n+\tau]\big] + \mathbb{E}\big[X[n]X[n+\tau-1]\big] \\
&\qquad + \mathbb{E}\big[X[n-1]X[n+\tau]\big] + \mathbb{E}\big[X[n-1]X[n+\tau-1]\big]\Big] \\
&= \begin{cases} 0.25\big[1+0+0+1\big] & \text{if } \tau = 0 \\ 0.25\big[0+1+0+0\big] & \text{if } \tau = 1 \\ 0.25\big[0+0+1+0\big] & \text{if } \tau = -1 \\ 0.25\big[0+0+0+0\big] & \text{if } |\tau| > 1 \end{cases} \\
&= \begin{cases} 0.5 & \text{if } \tau = 0 \\ 0.25 & \text{if } \tau = 1 \\ 0.25 & \text{if } \tau = -1 \\ 0 & \text{if } |\tau| > 1. \end{cases}
\end{aligned}$$

Therefore the power spectral density is

$$\begin{aligned}
S_{yy}(e^{j\omega}) &= \sum_{\tau=-\infty}^{\infty} R_{yy}[\tau]e^{-j\omega\tau} \\
&= 0.5 \cdot e^{-j\omega(0)} + (0.25) \cdot \left(e^{-j\omega(1)} + e^{+j\omega(1)}\right) \\
&= 0.5(1 + \cos(\omega)),
\end{aligned}$$

which equals the magnitude squared $\left|H(e^{j\omega})\right|^2$ of the low-pass filter $h[n]$ and is clearly not uniform on $[-\pi, \pi]$. Since this spectrum has more power at low frequencies than high frequencies, the noise process $Y[n]$ is called "colored" noise.

3 Scalar Quantizers

Quantizers can generally be designed to be very robust for a large class of signals. In scalar quantization, each source value is processed individually; the input value is mapped to an output taking one of finitely many values. The number of quantization levels is typically chosen to be a power-of-2 because the outputs are usually represented using binary strings. Mathematically, a quantizer is a function that maps its input to a value from a finite set. Hence, one can generally define the quantizer as a function $Q : \mathbb{R} \mapsto \mathcal{Q}$ with output

$$\hat{x} = Q(x).$$

A quantizer can separate its input space into intervals of either uniform or non-uniform lengths and map all points within one interval to a particular output level from the set $\mathcal{Q}$. Consider uniform quantization and let Δ be the quantization step size. Then we will use the following quantizer function for our purposes of audio processing:

$$Q(x) = \left\lfloor \frac{x}{\Delta} + \frac{1}{2} \right\rfloor \Delta = \text{round}\left(\frac{x}{\Delta}\right)\Delta. \tag{1}$$

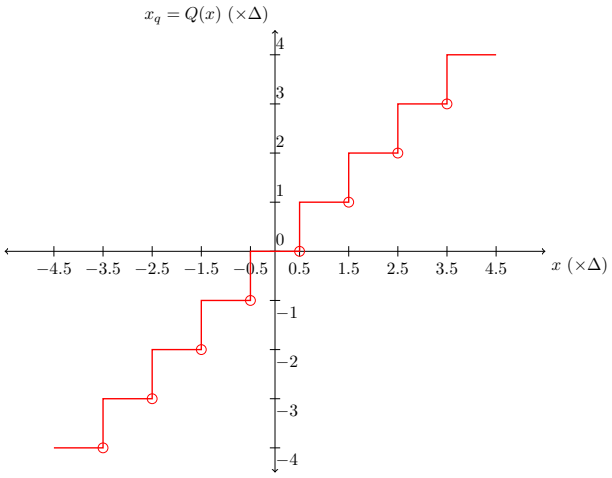

Figure 1: Graphical representation of the quantizer given by eqn. (1). The circles indicate points not included on the curve.

Figure 1 depicts this uniform quantizer graphically.

3.1 Quantization Noise

Let us now analyze our quantization scheme. If $x[n]$ is our input signal with continuous amplitudes, we model the quantized signal $\hat{x}[n]$ with an additive error model:

$$\hat{x}[n] \triangleq Q(x[n]) = x[n] + e[n],$$

where $e[n]$ is the quantization error or quantization noise signal. Error due to rounding is called *granular distortion* and error due to saturation is called *overload distortion*. To quantify the amount of distortion introduced by quantization, we can measure the power in $e[n]$. The average power in $e[n]$ is defined as

$$P_e \triangleq \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} |e[n]|^2 = \int_{-\infty}^{\infty} f_X(x)|x - Q(x)|^2 dx,$$

where $f_X(x)$ is the probability density function of the signal amplitude and the second term is the squared error due to quantization. Note that this relationship between the time average of the noise power and the statistical average of the noise power holds under mild conditions, such as if the process is wide sense stationary.

In general, the average power of the quantization noise depends on the distribution of the signal amplitude. A very important observation is that this dependence is very weak when Δ is small and $f_X(x)$ is continuous. Under these conditions, the signal amplitude, conditioned on $Q(X) = k\Delta$ for integer k, is approximately uniform over the interval $\left[(k - \frac{1}{2})\Delta, (k + \frac{1}{2})\Delta\right]$ *irrespective* of the true distribution of the signal amplitude. Thus, the quantization noise power, conditioned on $Q(X) = k\Delta$ for integer k, satisfies

$$\begin{aligned} P_e &\approx \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} \frac{1}{\Delta}|x - Q(x)|^2 dx \\ &= \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} \frac{1}{\Delta}|x - k\Delta|^2 dx \\ &= \frac{1}{\Delta}\int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} x^2 dx \end{aligned}$$

$$= \frac{1}{\Delta} \frac{x^3}{3}\Big|_{-\Delta/2}^{\Delta/2}$$
$$= \frac{\Delta^2}{12}.$$

For the quantization of a signal with amplitudes in $[-1, 1]$ to m bits per sample, we have 2^m levels that uniformly divide the interval $[-1, 1]$ into subintervals of length Δ. Hence, we have

$$\frac{2}{\Delta} = 2^m \Rightarrow \Delta = 2^{-m+1}$$

and the quantization noise power is

$$P_e = \frac{\Delta^2}{12} = \frac{2^{-2m+2}}{12} = \frac{2^{-2m}}{3}.$$

The signal-to-quantization-noise-ration (SQNR) is defined as

$$\text{SQNR} = \frac{P_x}{P_e} = \frac{\int f_X(x)|x|^2 dx}{\int f_X(x)|x - Q(x)|^2 dx}. \tag{2}$$

If $x[n]$ is a full-scale sinusoid $x[n] = \cos(\omega n + \phi)$, then $P_x = \frac{1}{2}$ and hence $\text{SQNR} = \frac{3}{2} \cdot 2^{2m}$, which on the log-scale gives

$$\text{SQNR} = 10 \log_{10} \left(\frac{3}{2} \cdot 2^{2m} \right) = 6.02m + 1.76 \text{ (dB)}.$$

Hence every bit increase in the quantization scheme improves the SQNR by approximately 6 dB and the additional factor is a constant dependent on the signal power.

3.2 Dithering

While the SQNR accurately measures the increase in noise power caused by quantization, it does not give any information about the spectral content of the quantization noise. For many applications, it is also important that the quantization noise be white (i.e., uncorrelated in time). To see the problem with standard quantization, consider the periodic signal x_n that satisfies $x_{n+N} = x_n$. In this case, the quantized version $y_n = Q(x_n)$ and the quantization error $e_n = y_n - x_n$ are also periodic. Therefore, the spectral energy of the quantization noise is concentrated in the harmonics of the fundamental frequency, which introduces audible distortions.

Since the quantizer affects only one value at a time, one may wonder how the quantization noise becomes correlated. The mechanism for this phenomenon can be explained through the fact that the quantization noise is correlated with the input value. For example, one can compute this correlation for our quantizer and obtain

$$\mathbb{E}_X \left[X \left(X - Q(X) \right) | Q(X) = k\Delta \right] = \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} \frac{1}{\Delta} x(x - Q(x)) dx$$
$$= \frac{1}{\Delta} \left[\frac{x^3}{3}\Big|_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} - \frac{k\Delta}{\Delta} \frac{x^2}{2}\Big|_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} \right]$$
$$= \frac{1}{3\Delta} \left[\left(\frac{\Delta}{2} + k\Delta \right)^3 - \left(-\frac{\Delta}{2} + k\Delta \right)^3 \right] - \frac{k}{2} \left[\left(\frac{\Delta}{2} + k\Delta \right)^2 - \left(-\frac{\Delta}{2} + k\Delta \right)^2 \right]$$
$$= \frac{1}{3\Delta} \left[\Delta \left\{ \left(\frac{\Delta}{2} + k\Delta \right)^2 + \left(\frac{\Delta}{2} + k\Delta \right) \left(-\frac{\Delta}{2} + k\Delta \right) + \left(-\frac{\Delta}{2} + k\Delta \right)^2 \right\} \right] - \frac{k}{2} \cdot \frac{4k\Delta^2}{2}$$

$$
\begin{aligned}
&= \frac{1}{3\Delta}\left[\Delta\left\{2\left(\frac{\Delta}{2}\right)^2 + 2\left(k\Delta\right)^2 + \left(k\Delta\right)^2 - \left(\frac{\Delta}{2}\right)^2\right\}\right] - k^2\Delta^2 \\
&= \frac{1}{3\Delta}\left[\Delta\left\{\frac{\Delta^2}{4} + 3k^2\Delta^2\right\}\right] - k^2\Delta^2 \\
&= \frac{\Delta^2}{12} + k^2\Delta^2 - k^2\Delta^2 \\
&= \frac{\Delta^2}{12}.
\end{aligned}
$$

From this, we see that the correlation is the same as the mean squared error computed previously as P_e. This means that quantizing a pure sinusoid will typically create new spurious harmonics whose powers are proportional to the power in the original sinusoid.

The process of adding a small amount of noise before quantization is called dithering. Of course, the added noise increases the overall noise power in the system by a small amount. But, if the noise sequence is chosen to be independent and uniformly distributed over one quantization interval, then the above correlation becomes exactly zero. To see this, we use our quantizer $Q(x)$ and let Z be a uniform random variable on $\left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right]$. In this case, we get

$$
\begin{aligned}
\mathbb{E}_Z[Q(X+Z)|X=x] &= \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \frac{1}{\Delta} Q(x+z)dz \\
&= \frac{1}{\Delta}\int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \left\lfloor \frac{x+z}{\Delta} + \frac{1}{2} \right\rfloor \Delta dz \\
&= \int_{\frac{x}{\Delta}}^{\frac{x}{\Delta}+1} \lfloor y \rfloor \Delta dy \quad \left(y \triangleq \frac{x+z}{\Delta} + \frac{1}{2}\right) \\
&= \Delta\left[\left\lfloor \frac{x}{\Delta} \right\rfloor \left(1 - \left\{\frac{x}{\Delta}\right\}\right) + \left(\left\lfloor \frac{x}{\Delta} \right\rfloor + 1\right)\left\{\frac{x}{\Delta}\right\}\right] \\
&= \Delta\left[\left\lfloor \frac{x}{\Delta} \right\rfloor + \left\{\frac{x}{\Delta}\right\}\right] \\
&= \Delta \cdot \frac{x}{\Delta} \\
&= x,
\end{aligned}
$$

where $\{x\} \triangleq x - \lfloor x \rfloor$ is the fractional part of x. Hence, the added uniform noise ensures that the signal values are preserved on average by the quantization process. Using this, we can compute

$$
\begin{aligned}
\mathbb{E}_Z\left[X\left(X - Q(X+Z)\right)|Q(X) = k\Delta\right] &= \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} f_X(x) \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} f_Z(z) x(x - Q(x+z))\, dz\, dx \\
&= \frac{1}{\Delta}\int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} x(x-x)\, dx = 0.
\end{aligned}
$$

Even more generaly, we observe that

$$
\begin{aligned}
\mathbb{E}_{\mathrm{X,Z}}\left[X\left(X - Q(X+Z)\right)\right] &= \int_{-\infty}^{\infty} f_X(x) \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} f_Z(z) x(x - Q(x+z))\, dz\, dx \\
&= \int_{-\infty}^{\infty} f_X(x) x(x-x)\, dx = 0.
\end{aligned}
$$

This implies that the quantization noise is uncorrelated with the signal. With a little more work, one can also show that it is white (i.e., uncorrelated with time-shifts of itself). However, the addition of uniform

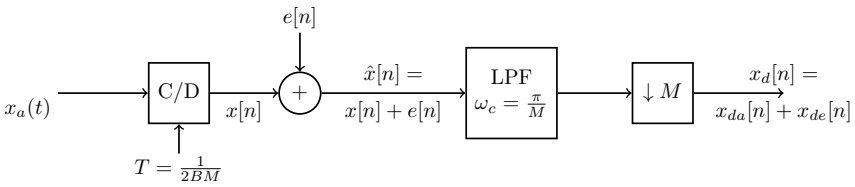

Figure 2: The oversampled A/D conversion process diagram.

dither increases the average quantization noise power as shown in Section A. In high-end audio applications, it is more common to add dither with a triangular PDF over one quantization interval since it prevents noise modulation, thereby making the noise almost exactly white. In that case, the noise power is increased by $\frac{\Delta^2}{6}$.

4 Analog-to-Digital (A/D) Conversion

As mentioned earlier, in order to convert an analog signal $x_a(t)$ into a digital signal $\hat{x}[n]$, we have to first sample it in time to produce $x[n] = x_a(nT)$, where T is our sampling period, and then quantize the sample amplitudes to obtain $\hat{x}[n] = Q(x[n])$, where $Q(\cdot)$ is the quantizer. If the signal of interest is not band-limited to $\frac{1}{2T}$ Hz, then energy above the Nyquist frequency could be aliased down into the band of interest. In order to prevent this, we assume the signal $x_a(t)$ is pre-filtered using an *anti-aliasing filter* designed to bandlimit the signal to $\frac{F_s}{2} = \frac{1}{2T}$ Hz. For example, an ideal anti-aliasing filter would have the frequency response

$$H_{lp}(e^{j\omega}) = \begin{cases} 1 & \text{if } |\omega| \leq \frac{\pi}{T} \\ 0 & \text{if } \frac{\pi}{T} < |\omega| \leq \pi. \end{cases}$$

After prefiltering, the signal is denoted $x_a(t)$ and passed through a *sample-and-hold* circuit that "samples" the signal every T seconds and "holds" the value for the same interval of time. During the hold time, the A/D converter quantizes the signal value. The signal obtained at the output of the sample-and-hold circuit is given by

$$x_{sh}(t) = \sum_{n=-\infty}^{\infty} x_a(nT) h_{sh}(t - nT) = \left(\sum_{n=-\infty}^{\infty} x_a(nT)\delta(t - nT) \right) * h_{sh}(t),$$

where

$$h_{sh}(t) = \begin{cases} 1 & \text{if } 0 \leq t < T \\ 0 & \text{otherwise} \end{cases}$$

is the sample-and-hold filter response. The discrete-time signal is obtained from this signal as $x[n] = x_{sh}(nT) = x_a(nT)$. Next, we need to convert the continuous amplitudes of $x[n]$ into a sequence of values obtained from a discrete set of values $\mathcal{Q}$. This is precisely the purpose of the quantizer and hence we obtain the fully digital signal $\hat{x}[n] = Q(x[n])$. For a more detailed discussion of A/D conversion, see [2, Section 4.8].

4.1 Oversampled A/D Conversion

Consider a signal $x_{sh}(t)$ with bandwidth B Hz which is sampled at $F_s = 2BM$ Hz, where M is the oversampling factor, and then quantized to obtain $\hat{x}[n] = Q(x[n] + z[n])$. Recollect that oversampling compresses the spectrum in $[-\pi, \pi]$ to $\left[-\frac{\pi}{M}, \frac{\pi}{M}\right]$ and also introduces periodic repetitions of this shrinked spectrum with period $\frac{2\pi}{M}$. Hence we need to pass this signal through a low-pass filter with cutoff frequency $\frac{\pi}{M}$ and then quantize to obtain $\hat{x}[n]$. The process is shown in Figure 2.

The quantization noise signal is defined to be $e[n] = x[n] - \hat{x}[n]$. Under mild conditions, the noise can be modeled as white noise with noise power $\sigma_e^2 = \frac{1}{3}2^{-2m}$, where m is the number of bits allocated per quantized

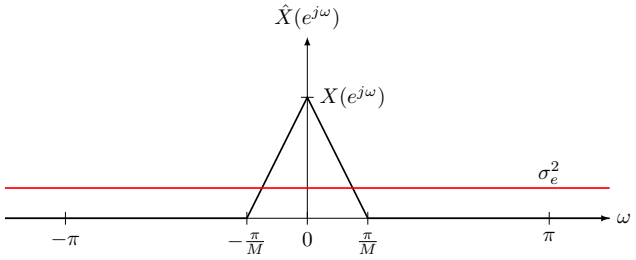

Figure 3: The DTFT of oversampled quantized signal $\hat{x}[n]$.

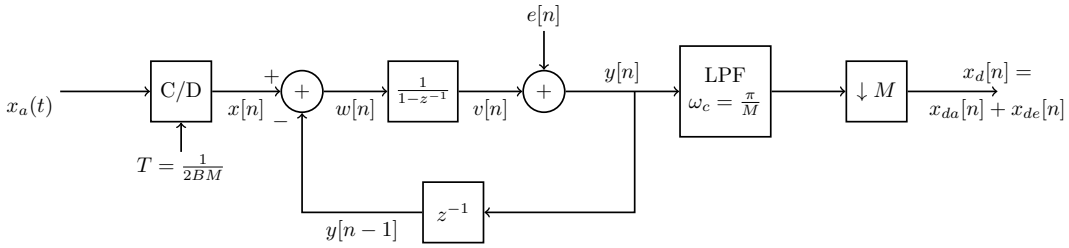

Figure 4: Block diagram for oversampled A/D conversion with noise shaping. $e[n]$ represents quantization error signal.

sample. In that case, an example of the power spectrum (i.e., expected value of the DTFT magnitude-squared) of $\hat{x}[n]$ is shown in Figure 3. Observe that oversampling reduces the overlap between the signal spectrum and the quantization noise spectrum. Using Parseval's theorem, we can compute the noise power in the signal bandwidth as

$$P_e = \frac{1}{2\pi}\int_{-\frac{\pi}{M}}^{\frac{\pi}{M}} \frac{1}{3}2^{-2m}d\omega = \frac{1}{3M}2^{-2m}.$$

Hence, after low-pass filtering, the SQNR has increased by a factor of M due to oversampling! For more details on oversampled A/D conversion, see [2, Section 4.9.1].

4.2 Oversampled A/D Conversion with Noise Shaping

We can rewrite the quantization noise power under oversampling as

$$\log_2(3MP_e) = -2m \Rightarrow m = -\frac{1}{2}\log_2 M - \frac{1}{2}\log_2 P_e - \frac{1}{2}\log_2 3.$$

Now observe that, for a fixed quantization noise power P_e, doubling the oversampling factor M reduces the quantization resolution by half-a-bit. So if we want to reduce the resolution of the system by 1 bit, we need to oversample by a factor of 4. This means to reduce 16 bit resolution to 12 bit resolution, we need to oversample by a factor of $4^4 = 256$! Instead, we can achieve the same noise power (in the signal spectrum) if we "shape" the uniform noise in Figure 3 so that most of the noise energy is outside $\left[-\frac{\pi}{M}, \frac{\pi}{M}\right]$. The block diagram for this procedure is given in Figure 4.

Let us observe the operations of this procedure and try to understand how noise shaping is achieved. We have

$$w[n] = x[n] - y[n-1], \qquad v[n] = v[n-1] + w[n],$$
$$e[n] = Q(v[n]) - v[n], \qquad y[n] = Q(v[n]).$$

By applying the z-transform to all equations we obtain

$$W(z) = X(z) - z^{-1}Y(z)$$

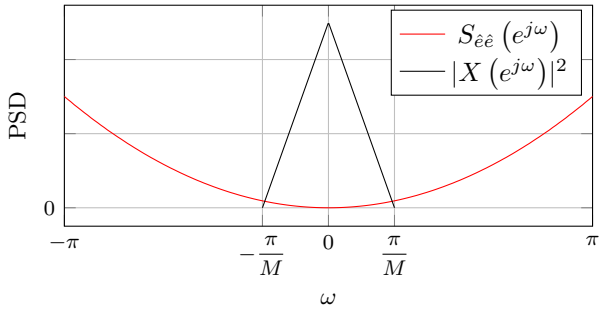

Figure 5: Power spectral density of signal and noise after noise shaping.

$$\begin{aligned} V(z) &= z^{-1}V(z) + W(z) = z^{-1}V(z) + X(z) - z^{-1}Y(z) \\ V(z) &= Y(z) - E(z) \\ \Rightarrow Y(z) - E(z) &= z^{-1}\left(Y(z) - E(z)\right) + X(z) - z^{-1}Y(z) \\ \Rightarrow Y(z) &= E(z)\left(1 - z^{-1}\right) + X(z). \end{aligned}$$

Now we see that the quantization noise is filtered by $H_e(z) = 1 - z^{-1}$, which is an differentiator or, in other words, a first order high-pass filter. The difference equation is $\hat{e}[n] = e[n] - e[n-1]$. We know that the power spectral density of this filter is $\left|H_e(e^{j\omega})\right|^2 = 4\sin^2(\omega/2)$. Therefore, the power spectral density of the filtered noise is given by

$$S_{\hat{e}\hat{e}}\left(e^{j\omega}\right) = \sigma_e^2\left|H_e(e^{j\omega})\right|^2 = \sigma_e^2 \cdot 4\sin^2\left(\frac{\omega}{2}\right).$$

This is shown in Figure 5. Observe that now most of the noise energy is outside the signal bandwidth. Let us compute the quantization noise power in $\left[-\frac{\pi}{M}, \frac{\pi}{M}\right]$.

$$\begin{aligned} P_e &= \frac{1}{2\pi}\int_{-\frac{\pi}{M}}^{\frac{\pi}{M}} S_{\hat{e}\hat{e}}\left(e^{j\omega}\right) d\omega \\ &= \frac{1}{2\pi}\int_{-\frac{\pi}{M}}^{\frac{\pi}{M}} \frac{\Delta^2}{12} 4\sin^2\left(\frac{\omega}{2}\right) d\omega \\ &\leq \frac{1}{2\pi}\frac{\Delta^2}{12}\int_{-\frac{\pi}{M}}^{\frac{\pi}{M}} 4 \cdot \frac{\omega^2}{4} d\omega \\ &= \frac{\Delta^2}{12}\frac{1}{2\pi}\frac{1}{3} \cdot 2 \cdot \frac{\pi^3}{M^3} \\ &= \frac{\Delta^2}{12} \cdot \frac{\pi^2}{3M^3}. \end{aligned}$$

Therefore, we get a SQNR gain of $10\log_{10}\left(\frac{3M^3}{\pi^2}\right) \approx 9\log_2 M - 5.17$ dB! Also, while doubling M resulted in just a 1/2-bit increase in resolution without noise shaping, now doubling M results in a 1.5-bit increase in resolution.

This is called first-order noise shaping and we can get the transfer function $H_e(z) = \left(1 - z^{-1}\right)^p$ by adding p stages of noise shaping. The resulting PSD would be $S_{\hat{e}\hat{e}}\left(e^{j\omega}\right) = \sigma_e^2\left(2\sin^2\left(\frac{\omega}{2}\right)\right)^p$. However, in practice circuit limitations dominate if order is larger than 2 or 3. For more details on oversampled A/D conversion with noise shaping, see [2, Section 4.9.2].

4.3 Oversampled D/A Conversion

Now we extend the ideas discussed above to convert a digital signal into an analog signal. A simple oversampled D/A conversion can be realized using the process shown in Figure 6.

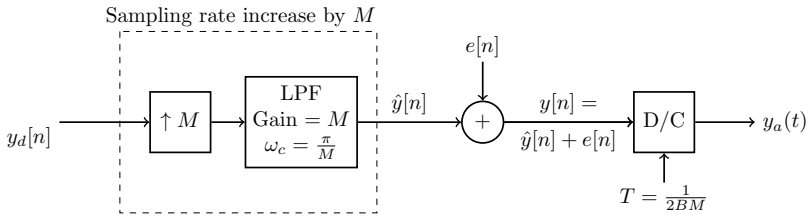

Figure 6: Oversampled D/A conversion without noise shaping. $e[n]$ represents quantization error signal.

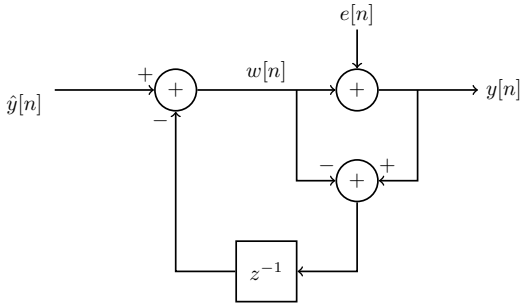

Figure 7: Oversampled D/A conversion with noise shaping.

However, if it is not guaranteed the quantization noise energy is mostly outside the signal spectrum, then we have to resort to noise shaping. This can be implemented as shown in Figure 7. Let us again look at the relationships between different signals in the diagram. We have

$$\begin{aligned} w[n] &= \hat{y}[n] - y[n-1] + w[n-1] \\ y[n] &= Q(w[n]) \\ e[n] &= Q(w[n]) - w[n] \\ \Rightarrow w[n] &= y[n] - e[n]. \end{aligned}$$

Now taking the z-transform, we obtain

$$\begin{aligned} W(z) &= \hat{Y}(z) - z^{-1}Y(z) + z^{-1}W(z) \\ W(z) &= Y(z) - E(z) \\ \Rightarrow Y(z) - E(z) &= \hat{Y}(z) - z^{-1}Y(z) + z^{-1}\left(Y(z) - E(z)\right) \\ \Rightarrow Y(z) &= \hat{Y}(z) + E(z)\left(1 - z^{-1}\right). \end{aligned}$$

Hence we again see that the noise is shaped by the high-pass filter $H_e(z) = 1 - z^{-1}$. Therefore the noise power spectral density is again $S_{\hat{e}\hat{e}}\left(e^{j\omega}\right) = \sigma_e^2 \left|H_e(e^{j\omega})\right|^2 = \sigma_e^2 \cdot 4\sin^2\left(\frac{\omega}{2}\right)$.

References

[1] J. G. Proakis and D. K. Manolakis, *Digital Signal Processing.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 4th ed., 2006.

[2] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time Signal Processing.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2nd ed., 1999.

A Appendix: Dither Increases Noise Power

In this appendix, we will show that the addition of uniform dither before quantization increases the average noise power. First let us compute the squared error for a particular value of X, averaged over the noise PDF.

$$\begin{aligned}
\mathbb{E}_Z\Big[\,(X-Q(X+Z))^2\,\Big|\,X=x\Big] &= \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \frac{1}{\Delta}\left(x-\left\lfloor\frac{x+z}{\Delta}+\frac{1}{2}\right\rfloor\Delta\right)^2 dz \\
&= \frac{1}{\Delta}x^2\cdot\Delta - 2x\cdot x + \Delta\int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}}\left\lfloor\frac{x+z}{\Delta}+\frac{1}{2}\right\rfloor^2 dz \\
&= -x^2+\Delta^2\int_{\frac{x}{\Delta}}^{\frac{x}{\Delta}+1}\lfloor y\rfloor^2\,dy \quad \left(y\triangleq\frac{x+z}{\Delta}+\frac{1}{2}\right) \\
&= -x^2+\Delta^2\left[\left\lfloor\frac{x}{\Delta}\right\rfloor^2\left(1-\left\{\frac{x}{\Delta}\right\}\right)+\left(\left\lfloor\frac{x}{\Delta}\right\rfloor+1\right)^2\left\{\frac{x}{\Delta}\right\}\right] \\
&= -x^2+\Delta^2\left[\left\lfloor\frac{x}{\Delta}\right\rfloor^2-\left\lfloor\frac{x}{\Delta}\right\rfloor^2\left\{\frac{x}{\Delta}\right\}+\left\lfloor\frac{x}{\Delta}\right\rfloor^2\left\{\frac{x}{\Delta}\right\}+2\left\lfloor\frac{x}{\Delta}\right\rfloor\left\{\frac{x}{\Delta}\right\}+\left\{\frac{x}{\Delta}\right\}\right] \\
&= -x^2+\Delta^2\left[\left(\left\lfloor\frac{x}{\Delta}\right\rfloor+\left\{\frac{x}{\Delta}\right\}\right)^2-\left\{\frac{x}{\Delta}\right\}^2+\left\{\frac{x}{\Delta}\right\}\right] \\
&= -x^2+\Delta^2\cdot\frac{x^2}{\Delta^2}+\Delta^2\cdot\left\{\frac{x}{\Delta}\right\}\left[1-\left\{\frac{x}{\Delta}\right\}\right] \\
&= \Delta^2\left\{\frac{x}{\Delta}\right\}\left[1-\left\{\frac{x}{\Delta}\right\}\right].
\end{aligned}$$

$$\begin{aligned}
P_e &= \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} f_X(x)\mathbb{E}_Z\left[(X-Q(X+Z))^2\,|X=x\right]dx \\
&= \frac{1}{\Delta}\int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta}\Delta^2\left\{\frac{x}{\Delta}\right\}\left[1-\left\{\frac{x}{\Delta}\right\}\right]dx \\
&= \Delta\int_{-\frac{\Delta}{2}+k\Delta}^{k\Delta}\left(k-\frac{x}{\Delta}\right)\left(1-k+\frac{x}{\Delta}\right)dx+\Delta\int_{k\Delta}^{\frac{\Delta}{2}+k\Delta}\left(\frac{x}{\Delta}-k\right)\left(1-\frac{x}{\Delta}+k\right)dx \\
&= \Delta\int_{-\frac{\Delta}{2}+k\Delta}^{k\Delta}\left(k-k^2+\frac{kx}{\Delta}-\frac{x}{\Delta}+\frac{kx}{\Delta}-\frac{x^2}{\Delta^2}\right)dx+\Delta\int_{k\Delta}^{\frac{\Delta}{2}+k\Delta}\left(\frac{x}{\Delta}-\frac{x^2}{\Delta^2}+\frac{kx}{\Delta}-k+\frac{kx}{\Delta}-k^2\right)dx \\
&= \Delta k\cdot\frac{\Delta}{2}-\Delta k^2\cdot\frac{\Delta}{2}+(2k-1)\frac{x^2}{2}\Big|_{-\frac{\Delta}{2}+k\Delta}^{k\Delta}-\frac{1}{\Delta}\cdot\frac{x^3}{3}\Big|_{-\frac{\Delta}{2}+k\Delta}^{k\Delta}-\Delta k\cdot\frac{\Delta}{2}-\Delta k^2\cdot\frac{\Delta}{2}+(2k+1)\frac{x^2}{2}\Big|_{k\Delta}^{\frac{\Delta}{2}+k\Delta}-\frac{1}{\Delta}\cdot\frac{x^3}{3}\Big|_{k\Delta}^{\frac{\Delta}{2}+k\Delta} \\
&= -\Delta^2k^2+k\left[\left(\frac{\Delta}{2}+k\Delta\right)^2-\left(-\frac{\Delta}{2}+k\Delta\right)^2\right]+\frac{1}{2}\left(\frac{\Delta}{2}+k\Delta\right)^2-\frac{1}{2}k^2\Delta^2 \\
&\qquad -\frac{1}{2}k^2\Delta^2+\frac{1}{2}\left(-\frac{\Delta}{2}+k\Delta\right)^2-\frac{1}{3\Delta}\left[\left(\frac{\Delta}{2}+k\Delta\right)^3-\left(-\frac{\Delta}{2}+k\Delta\right)^3\right] \\
&= -2\Delta^2k^2+k\cdot 4\frac{\Delta}{2}\cdot k\Delta+\frac{\Delta^2}{4}+k^2\Delta^2-\frac{1}{3\Delta}\Delta\left[\frac{2\Delta^2}{4}+2k^2\Delta^2+k^2\Delta^2-\frac{\Delta^2}{4}\right] \\
&= \frac{\Delta^2}{4}+k^2\Delta^2-k^2\Delta^2-\frac{\Delta^2}{12} \\
&= \frac{\Delta^2}{6}.
\end{aligned}$$

Hence the addition of a uniform dither increases the average noise power by $\frac{\Delta^2}{12}$.