Notes on Uniform Quantization

Quantization Signal Model

The process of converting a discrete–time signal $x[n]$ with infinite precision to a discrete–time signal with finite–bit precision $x_q[n]$ is referred to as *quantization*. This will be necessary when the objective is to implement a DSP algorithm using a fixed point DSP processor. In mathematical terms this is denoted as :

$$x_q[n] = Q(x[n]).$$

We define the quantization error $e_q[n]$ as the error introduced in the quantization process via:

$$e_q[n] = x_q[n] - x[n] = Q(x[n]) - x[n].$$

The process of quantization is therefore modeled conveniently as a linear noise addition process. Of course this assumption is valid provided that the quantizer is not clipping the input source.

In this section we will be specifically looking at two modes of quantization. The first mode is the *round-off* mode where the source sequence value is rounded to the nearest value in the output alphabet. The second mode is the *saturation* mode where the source sample is ceiled up to the nearest value in the output alphabet. For applications in communications such as equalization, where a rational signal power spectrum is needed the round–off mode is more appropriate.

Uniform Noise Source Model

For the purposes of this discussion we will assume that our source $x[n]$ is uniformly distributed, i.e.,

$$x[n] \sim U([-X_{\max}, X_{\max}]).$$

For a B bit quantizer with 1 bit allocated for the sign of the source, the step-size of the desired uniform quantizer is given by:

$$\Delta = \frac{2X_{\max}}{2^{B+1}} = \frac{X_{\max}}{2^B}.$$

Specifically if the source produced a random signal $x[n]$ that was uniformly distributed, i.e, $x[n] \sim U([-X_{\max}, X_{\max}])$ then the mean and average power of the source would be:

$$\mu_x[n] = 0, \quad P^x_{\text{ave}} = \sigma^2_x = \frac{X^2_{\max}}{3}.$$

In any case in the design of the quantizer we need to match the range of the quantizer data, i.e, $x[n] \in [-X_{\max}, X_{\max}]$ to the variance of the signal. Otherwise if we do get data that is not in this range then the quantizer will clip the signal. In this region the additive noise model described before is no longer valid.

Intuitively since the input source does not show preference to a particular quantizer interval, i.e., the source is uniform the quantization error $e_q[n]$ is also expected to be uniform. Indeed it can be shown that in the case of the round-off and saturation options:

$$e^{(1)}_q[n] \sim U\left(\left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right]\right) \quad , \quad e^{(2)}_q[n] \sim U\left([0, \Delta]\right).$$

The difference in the quantization noise model between the “round–off” mode and the “saturation mode” is the presence of a non–zero mean in the first–order PDF of the noise in the “saturation mode” that represents the bias towards the upper output alphabet.

Signal to Noise Ratio

The average power in the quantization error for either quantization type is therefore, given by:

$$P^e_{\text{ave}} = \sigma^2_e = \frac{\Delta^2}{12} = \frac{X^2_{\max}}{2^{2B}12}.$$

The *signal to noise ratio* (SNR) of the output of the quantizer is defined via:

$$\text{SNR} = 10\log_{10}\left(\frac{\sigma^2_x}{\sigma^2_e}\right) \ dB.$$

For the specific case of the uniform quantizer this reduces to

$$\text{SNR} = 6.02B + 10.8 - 20\log_{10}\left(\frac{X_{\max}}{\sigma_x}\right) \ dB.$$

There is approximately a 6 dB increase in the SNR of the quantized signal for every bit increase in resolution provided that we are in the linear range of operation of the quantizer and not clipping the signal, i.e, $\sigma_x \leq X_{\max}$.