Quantization Effects in Digital Filters

Distribution of Truncation Errors

In two's complement representation an exact number would have infinitely many bits (in general). When we limit the number of bits to some finite value we have truncated all the lower-order bits. Suppose the number is 23/64. In two's complement binary this is 0.0101110000... If we now limit the number of digits to 5 we get 0.0101 which represents exactly 20/64 or 5/16. The truncation error is $5/16 - 23/64 = -3/64$. If the exact number is $-15/64$, this is 1.110001000... in two's complement. Truncating to 5 bits yields 1.1100 which exactly represents $-1/4$. The error is $-1/4 - (-15/64) = -1/64$. This illustrates that in two's complement truncation the error is always negative.

Distribution of Truncation Errors

In analysis of truncation errors we make the following assumptions.

1. The distribution of the errors is uniform over a range equivalent to the value of one LSB.
2. The error is a white noise process. That means that one error is not correlated with the next error or any other error.
3. The truncation error and the signal whose values are being truncated are not correlated.

Quantization of Filter Coefficients

Let a digital filter's ideal transfer function be

$$\mathrm{H}(z) = \frac{\sum_{k=0}^{N} b_k z^{-k}}{1 + \sum_{k=1}^{N} a_k z^{-k}}$$

If the *a*'s and *b*'s are quantized, the poles and zeros move, creating a new transfer function

$$\hat{\mathrm{H}}(z) = \frac{\sum_{k=0}^{N} \hat{b}_k z^{-k}}{1 + \sum_{k=1}^{N} \hat{a}_k z^{-k}}$$

where $\hat{a}_k = a_k + \Delta a_k$ and $\hat{b}_k = b_k + \Delta b_k$.

Quantization of Filter Coefficients

The denominator can be factored into the form

$$\mathrm{D}(z) = 1 + \sum_{k=1}^{N} a_k z^{-k} = \prod_{k=1}^{N} \left(1 - p_k z^{-k}\right)$$

and, with quantization,

$$\hat{\mathrm{D}}(z) = 1 + \sum_{k=1}^{N} \hat{a}_k z^{-k} = \prod_{k=1}^{N} \left(1 - \hat{p}_k z^{-1}\right)$$

where $\hat{p}_k = p_k + \Delta p_k$. The errors in the pole locations are related to the errors in the coefficients by

$$\Delta p_i = \Delta a_1 \frac{\partial p_i}{\partial a_1} + \Delta a_2 \frac{\partial p_i}{\partial a_2} + \cdots + \Delta a_N \frac{\partial p_i}{\partial a_N} = \sum_{k=1}^{N} \Delta a_k \frac{\partial p_i}{\partial a_k}$$

Quantization of Filter Coefficients

The partial derivatives can be found from

$$\left[\frac{\partial \mathrm{D}(z)}{\partial a_k}\right]_{z=p_i} = \left[\frac{\partial \mathrm{D}(z)}{\partial z}\right]_{z=p_i} \times \frac{\partial p_i}{\partial a_k}$$

or

$$\frac{\partial p_i}{\partial a_k} = \frac{\left[\partial \mathrm{D}(z)/\partial a_k\right]_{z=p_i}}{\left[\partial \mathrm{D}(z)/\partial z\right]_{z=p_i}}$$

$$\left[\frac{\partial \mathrm{D}(z)}{\partial a_k}\right]_{z=p_i} = \frac{\partial}{\partial a_k}\left[1+\sum_{k=1}^{N} a_k z^{-k}\right]_{z=p_i} = \left[z^{-k}\right]_{z=p_i} = p_i^{-k}$$

$$\left[\frac{\partial \mathrm{D}(z)}{\partial z}\right]_{z=p_i} = \left[\frac{\partial}{\partial z}\prod_{q=1}^{N}\left(1-p_q z^{-1}\right)\right]_{z=p_i}$$

Quantization of Filter Coefficients

The derivative of the qth factor is

$$\left[\frac{\partial}{\partial z}\left(1-p_q z^{-1}\right)\right]=\frac{p_q}{z^2}$$

and the overall derivative is

$$\left[\frac{\partial \mathrm{D}(z)}{\partial z}\right]_{z=p_i}=\left[\sum_{k=1}^{N}\frac{p_k}{z^2}\prod_{\substack{q=1\\q\neq k}}^{N}\left(1-p_q z^{-1}\right)\right]_{z=p_i}=\sum_{k=1}^{N}\frac{p_k}{p_i^2}\prod_{\substack{q=1\\q\neq k}}^{N}\left(1-p_q p_i^{-1}\right)$$

$$\left[\frac{\partial \mathrm{D}(z)}{\partial z}\right]_{z=p_i}=\sum_{k=1}^{N}\frac{p_k}{p_i^2}\,p_i^{-(N-1)}\prod_{\substack{q=1\\q\neq k}}^{N}\left(p_i-p_q\right)=\frac{1}{p_i^{N+1}}\sum_{k=1}^{N}p_k\prod_{\substack{q=1\\q\neq k}}^{N}\left(p_i-p_q\right)$$

Quantization of Filter Coefficients

In the k summation $\sum_{k=1}^{N} p_k \prod_{\substack{q=1 \\ q\neq k}}^{N} \left(p_i - p_q \right)$, for any k other than i,

the iterated product $\prod_{\substack{q=1 \\ q\neq k}}^{N} \left(p_i - p_q \right)$ has a factor $p_i - p_i = 0$. Therefore

the only term in the k summation that survives is the $k = i$ term and

$$\left[\frac{\partial \mathrm{D}(z)}{\partial z} \right]_{z=p_i} = \frac{1}{p_i^{N+1}} p_i \prod_{\substack{q=1 \\ q\neq k}}^{N} \left(p_i - p_q \right) = \frac{1}{p_i^{N}} \prod_{\substack{q=1 \\ q\neq k}}^{N} \left(p_i - p_q \right)$$

Quantization of Filter Coefficients

Then

$$\frac{\partial p_i}{\partial a_k} = \frac{p_i^{-k}}{\frac{1}{p_i^N} \prod_{\substack{q=1 \\ q \neq k}}^{N} \left(p_i - p_q \right)}$$

and

$$\Delta p_i = \sum_{k=1}^{N} \frac{\Delta a_k \, p_i^{N-k}}{\prod_{\substack{q=1 \\ q \neq k}}^{N} \left(p_i - p_q \right)}$$

Quantization of Filter Coefficients

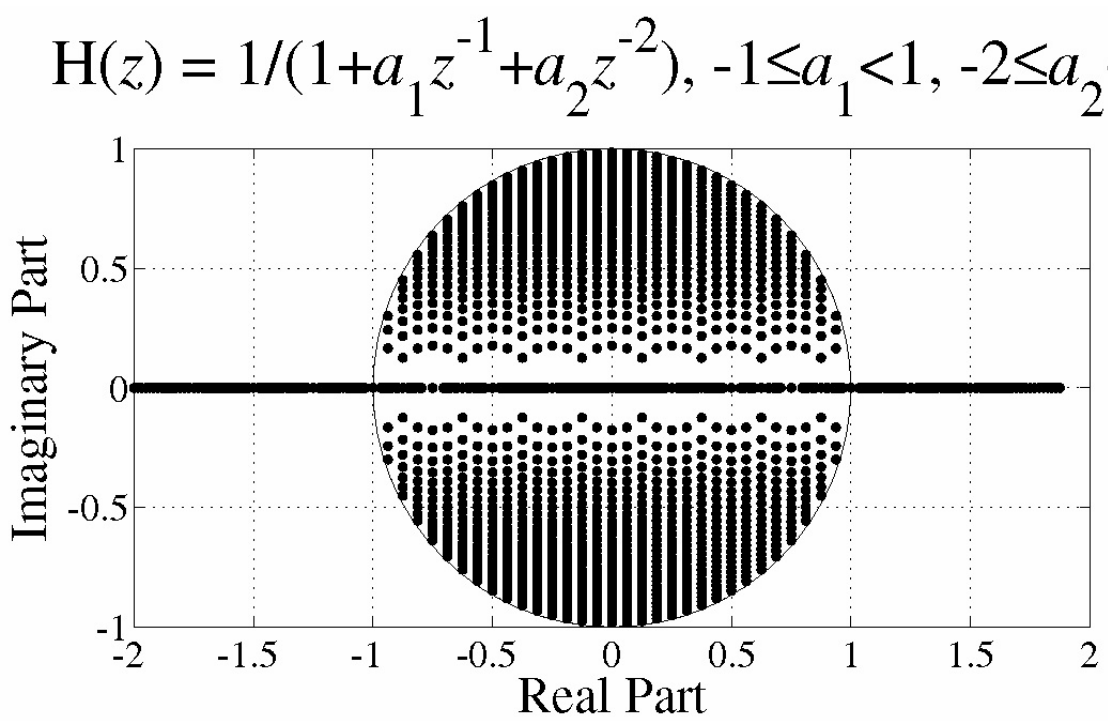
The error in a pole location caused by errors in the coefficients is strongly affected by the denominator factor $\prod_{\substack{q=1 \\ q \neq k}}^{N} \left(p_i - p_q \right)$ which is a product of differences between pole locations. So if the poles are closely spaced, the distances are small and the error in the pole location is large. Therefore systems with widely-spaced poles are less sensitive to errors in the coefficients.

Quantization of Pole Locations

Consider a 2nd order system with transfer function

$$\mathrm{H}(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}}$$

and quantize a_1 and a_2 with 5 bits. If we quantize a_1 in the range $-1 \le a_1 < 1$ and $-2 \le a_2 < 2$ we get a finite set of possible pole locations.

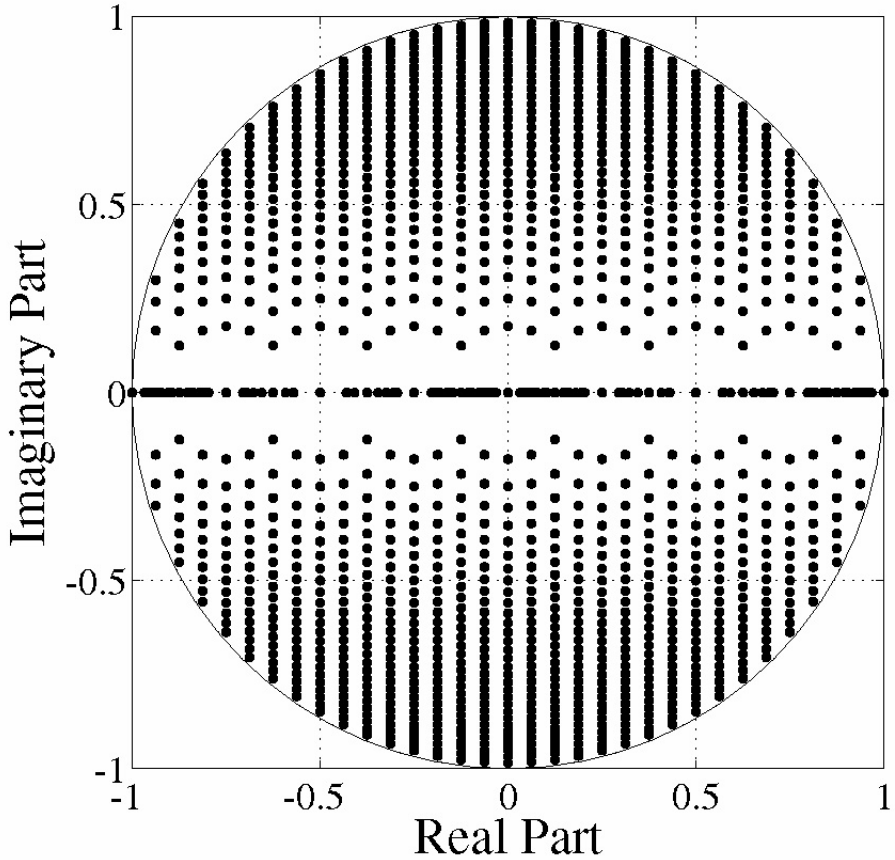


Quantization of Pole Locations

If we look at only pole locations for stable systems,

$$H(z) = 1/(1+a_1 z^{-1}+a_2 z^{-2}),\ -1 \le a_1 < 1,\ -2 \le a_2 < 2$$

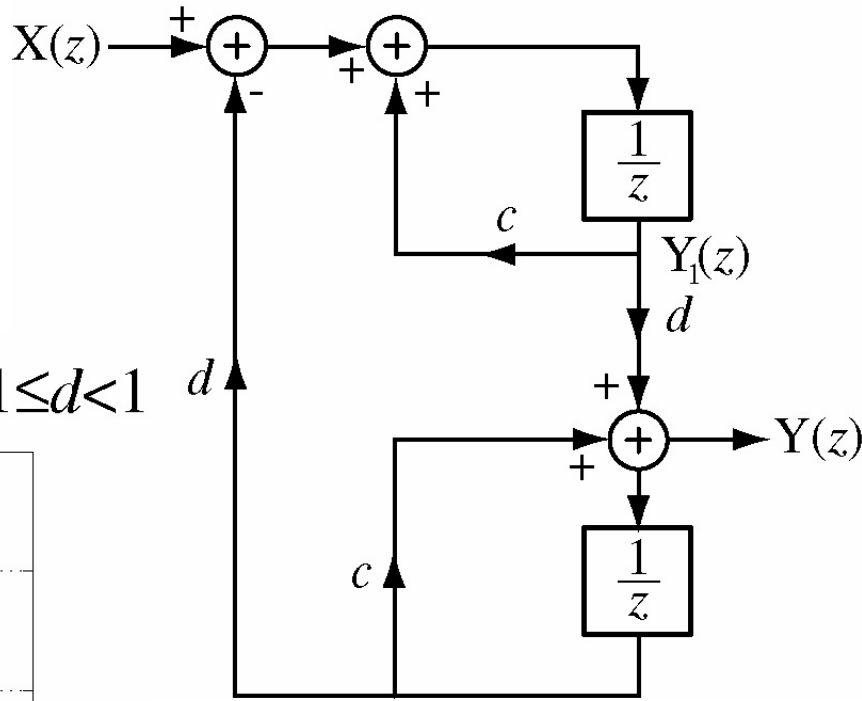
Obviously the pole locations are non-uniformly distributed.



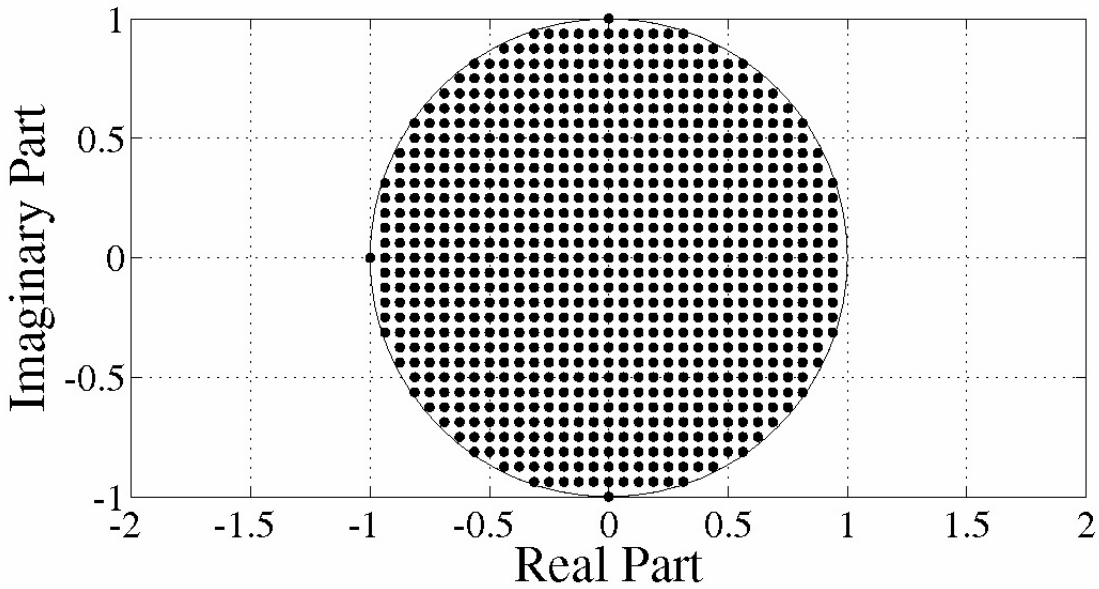
Quantization of Pole Locations

Consider this alternative 2nd order system design.

$$H(z) = \frac{dz^{-1}}{\left(1-(c+jd)z^{-1}\right)\left(1-(c-jd)z^{-1}\right)}$$



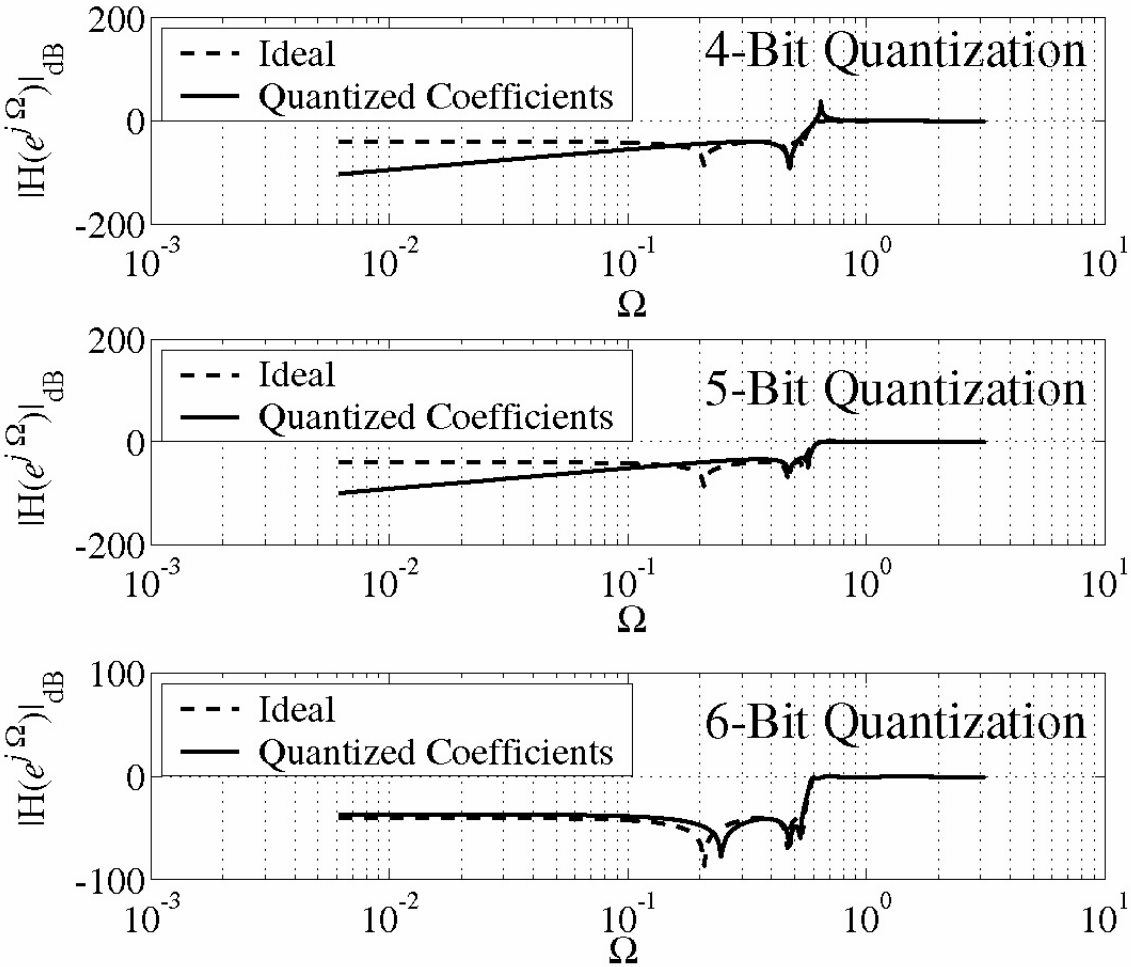
$$H(z) = 1/(1-2cz^{-1}+(c^2+d^2)z^{-2}),\ -1 \le c < 1,\ -1 \le d < 1$$



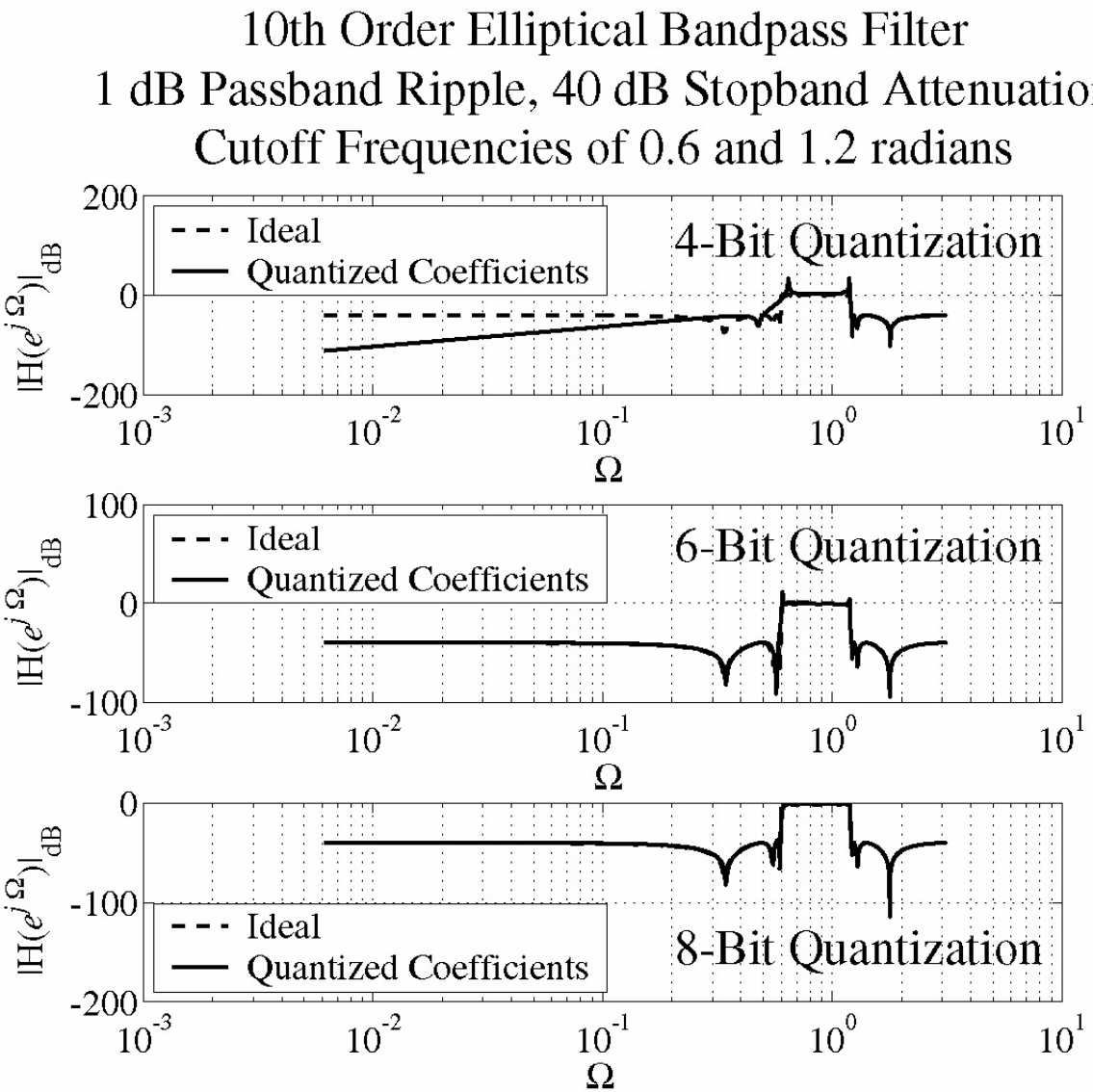
Pole locations are uniformly distributed.

Quantization of Pole Locations

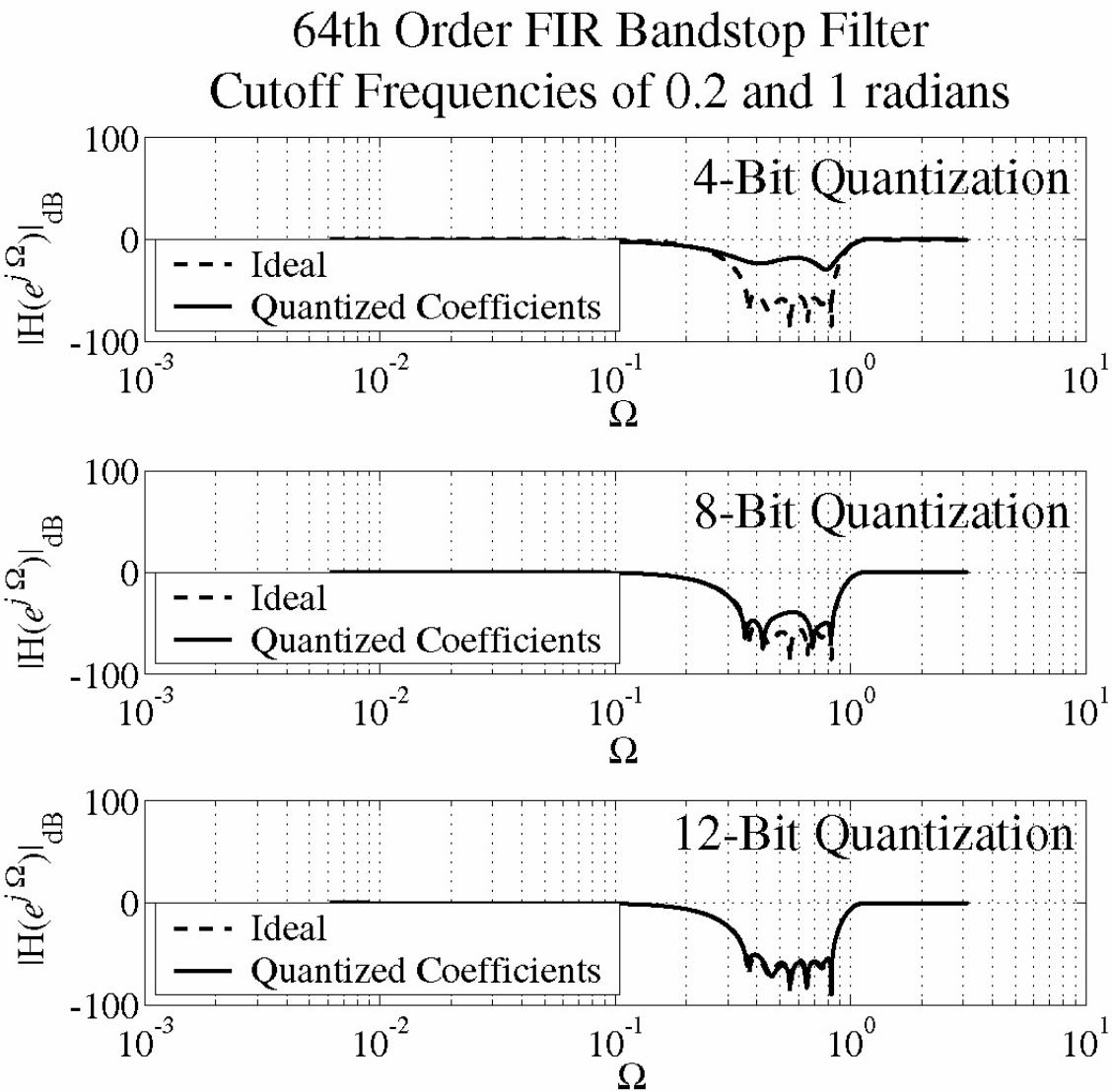
6th Order Elliptical Highpass Filter
1 dB Passband Ripple, 40 dB Stopband Attenuation
Cutoff Frequency of 0.6 radians



Quantization of Pole Locations



Quantization of Pole Locations



Scaling to Prevent Overflow

Let $y_k[n]$ be the response at node k to an input signal $x[n]$ and let $h_k[n]$ be the impulse response at node k.

$$\left|y_k[n]\right| = \left|\sum_{m=-\infty}^{\infty} h_k[m]\,x_k[n-m]\right| \le \sum_{m=-\infty}^{\infty} \left|h_k[m]\right|\left|x_k[n-m]\right|$$

Let the upper bound on $x[n]$ be A_x.

$$\left|y_k[n]\right| \le A_x \sum_{m=-\infty}^{\infty} \left|h_k[m]\right|$$

If we want $y_k[n]$ to lie in the range $(-1,1)$ then

$$A_x < \frac{1}{\sum_{n=-\infty}^{\infty} \left|h_k[n]\right|}$$

guarantees that oveflow will not occur.

Scaling to Prevent Overflow

The condition

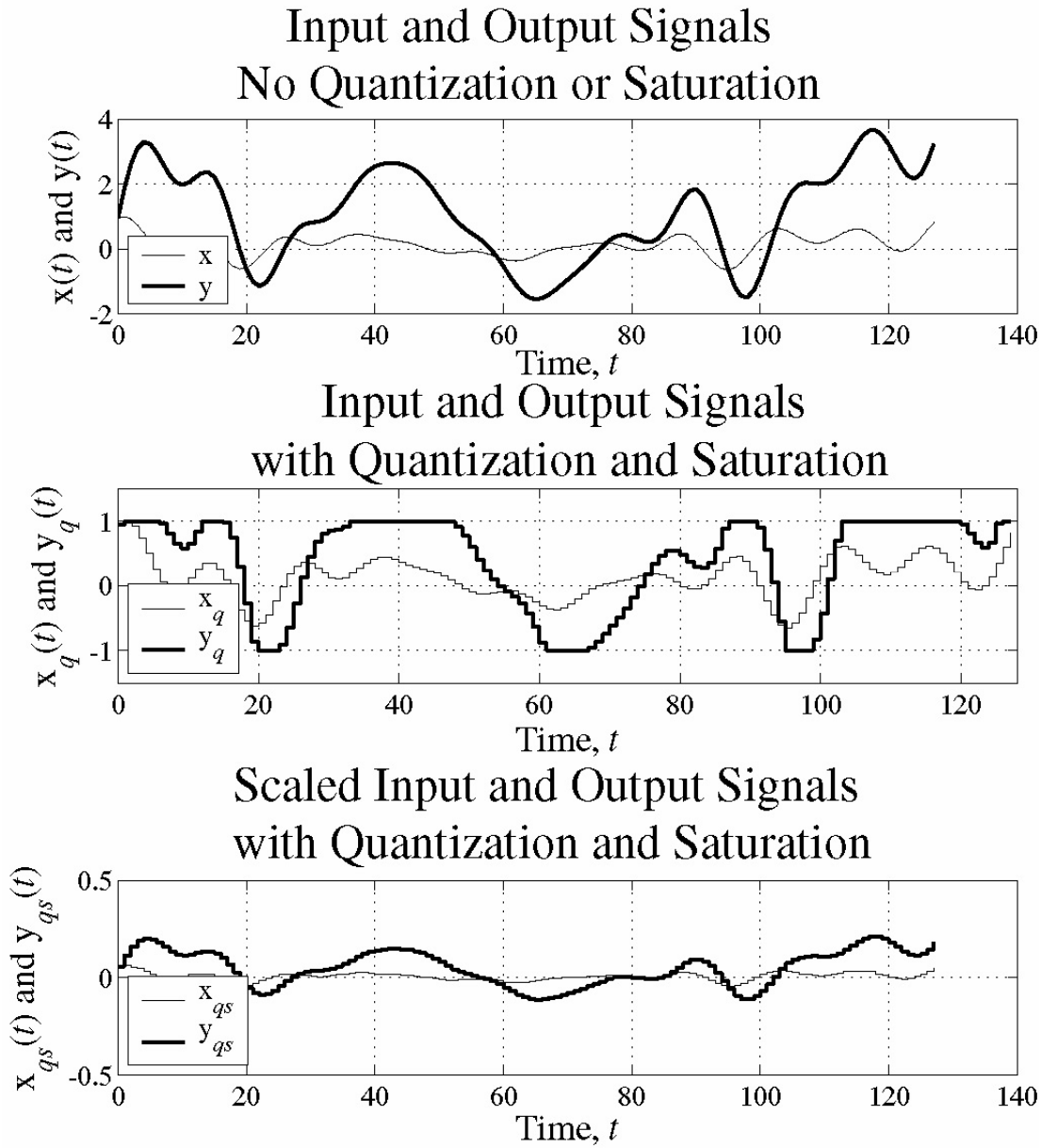
$$A_x < \frac{1}{\sum_{n=-\infty}^{\infty} \left| \mathrm{h}_k[n] \right|}$$

may be too severe in some cases. Another type of scaling that may be more appropriate in some cases is

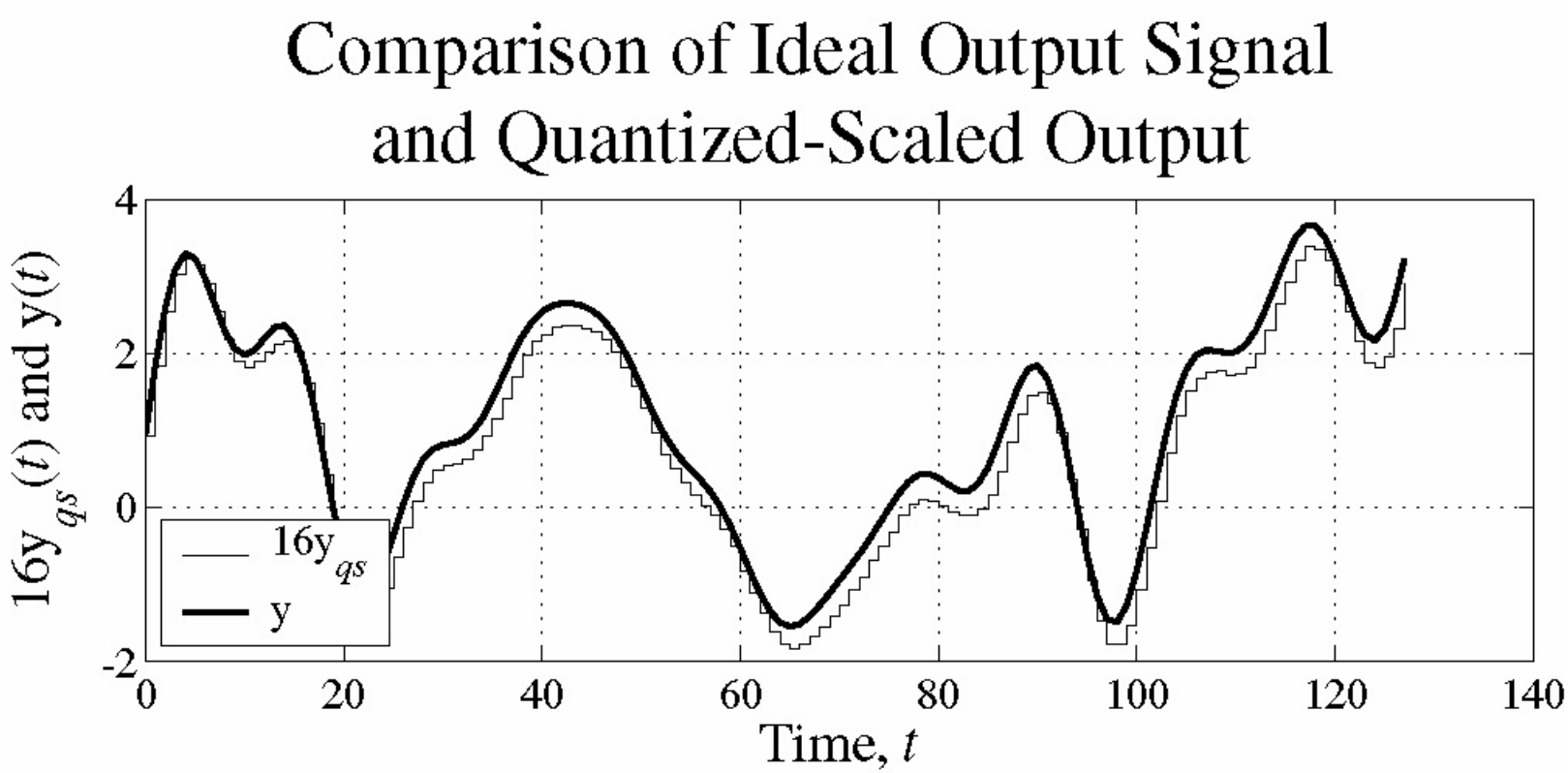
$$A_x = \frac{1}{\max\limits_{0 \le \Omega \le \pi} \left| \mathrm{H}_k\left(e^{j\Omega}\right) \right|}$$

where $\mathrm{h}_k[n] \overset{\mathrm{F}}{\longleftrightarrow} \mathrm{H}_k\left(e^{j\Omega}\right)$.

Scaling to Prevent Overflow



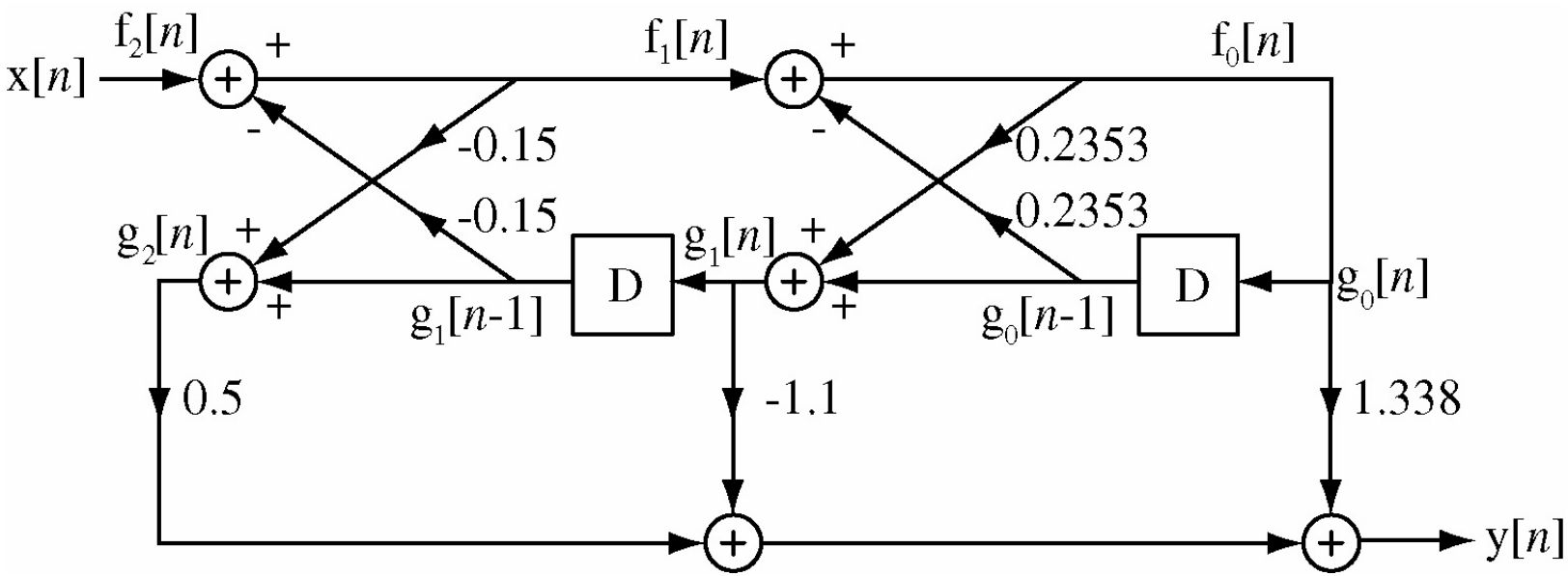
Scaling to Prevent Overflow



Scaling Example

Assume that the input signal x is white noise bounded between -1 and +1 and that all additions and multiplications are done using 8-bit two's complement arithmetic. Also let all numbers (signals and coefficients) be in the range -1 to 1.

$$H(z) = \frac{1 - z^{-1} + 0.5z^{-2}}{1 + 0.2z^{-1} - 0.15z^{-2}} = \frac{z^2 - z + 0.5}{z^2 + 0.2z - 0.15}$$



Scaling Example

Quantize the *K*'s to 8 bits.

$$K_1 = 0.2353 \rightarrow [Q] \rightarrow 0.2344 \ (0.0011110)$$

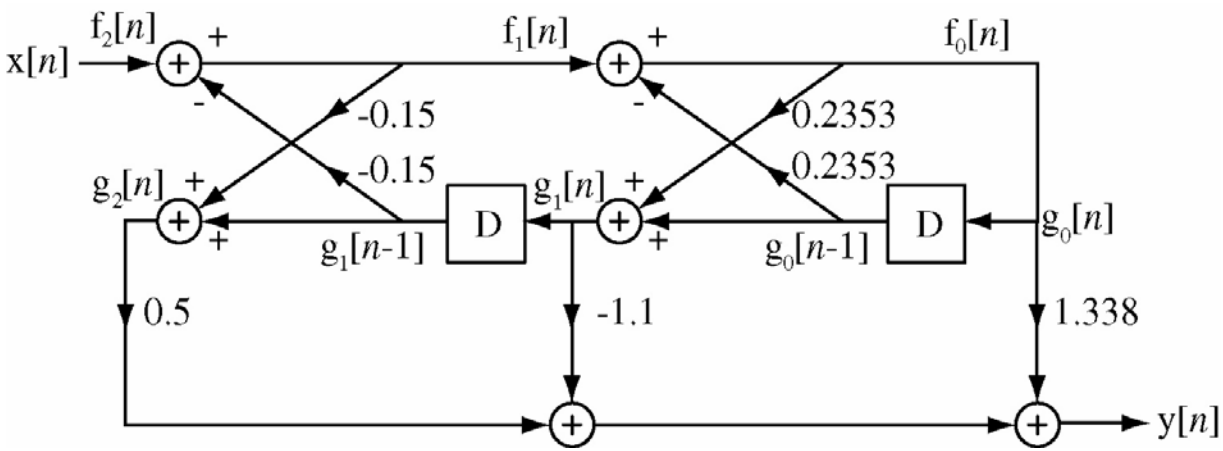
$$K_2 = -0.15 \rightarrow [Q] \rightarrow -0.1562 \ (1.1101100)$$

Scale and quantize the *v*'s.

$$v_0 = 1.338 \rightarrow 1.338/1.338 = 1 \rightarrow [Q] \rightarrow 0.9922 \ \ (0.1111111)$$

$$v_1 = -1.1 \rightarrow -1.1/1.338 = -0.8221 \rightarrow [Q] \rightarrow -0.8281 \ \ (1.0010110)$$

$$v_2 = 0.5 \rightarrow 0.5/1.338 = 0.3737 \rightarrow [Q] \rightarrow 0.3672 \ \ (0.0101111)$$



Scaling Example

Scaling the *v*'s scales the output signal but the input signal may still need scaling to prevent overflow.

The exact transfer function is $\mathrm{H}(z)=\frac{1}{\mathrm{A}_2(z)}\sum_{m=0}^{N} v_m \mathrm{B}_m(z)$ where

$\mathrm{A}_2(z)=1+0.2z^{-1}-0.15z^{-2}$. The actual $\mathrm{A}_2(z)$, after quantizing the *K*'s is $\mathrm{A}_2(z)=1+0.1978z^{-1}-0.1562z^{-2}$. $\mathrm{B}_0(z)=1$,

$\mathrm{B}_1(z)=0.2344+z^{-1}$ and $\mathrm{B}_2(z)=-0.1562+0.1978z^{-1}+z^{-2}$.

Therefore the actual transfer function is

$$\mathrm{H}(z)=\frac{0.7407-0.7555z^{-1}+0.3672z^{-2}}{1+0.1978z^{-1}-0.1562z^{-2}}.$$

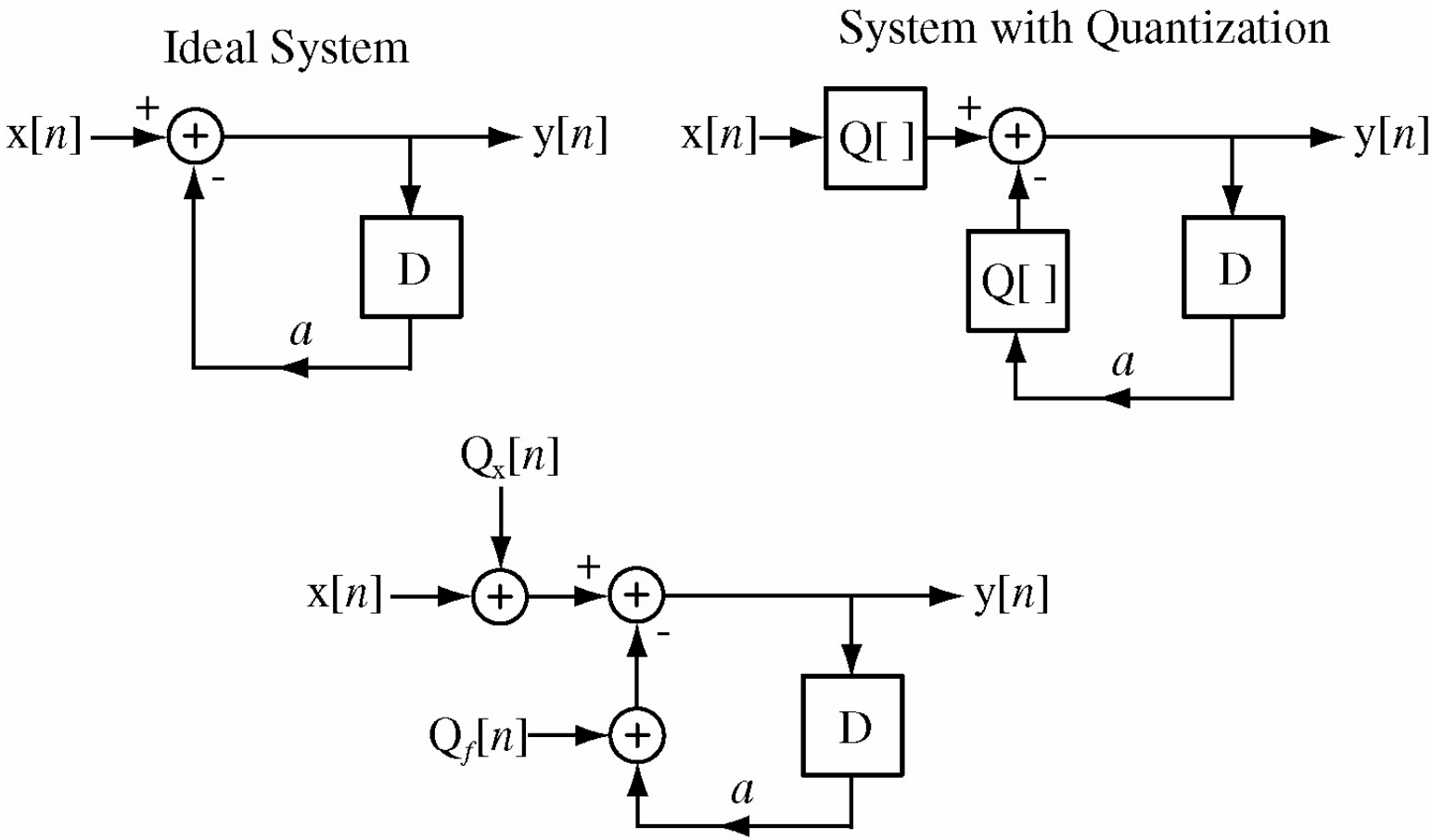
Scaling Example

The maximum magnitude of $\mathrm{H}\left(e^{j\Omega}\right)$ in this example is 2.882 and the maximum magnitude of $\mathrm{H}_2\left(e^{j\Omega}\right)$ is 1.5455. So the scaling factor should be $1/2.882 = 0.347$. The simplest scaling is by shifting right. In this case that would require a shift of two bits to the right to scale by 0.25. This is significanly less that 0.347 so a more complicated scaling might be preferred. We could instead scale by $0.25 + 0.0625 + 0.03125 = 0.34375$ by shifting two bits, then four bits and then 5 bits and adding the results.

Statistical Analysis of Quantization Effects

The exact estimation of quantization effects requires numerical simulation and is not amenable to exact analytical methods. But an approach that has proven useful is to treat the quantization noise effects as a random process problem. In doing this we get an approximate analytical estimate instead of an exact numerical simulation. We do this by treating quantization noise as an added noise signal in the system everywhere quantization occurs.

Statistical Analysis of Quantization Effects



System for Analysis of Quantization Effects using Statistical Methods

Statistical Analysis of Quantization Effects

The probability density function for quantization noise using two's complement representation is

$$\mathrm{f}_Q(q) = \frac{1}{\mathrm{LSB}} \begin{cases} 1\ , & -\mathrm{LSB} < q < 0 \\ 0\ , & \text{otherwise} \end{cases} .$$

The expected value of the quantization noise is $\mathrm{E}(Q) = -\mathrm{LSB}/2.$

The variance of the quantization noise is $\sigma_Q^2 = \mathrm{LSB}^2/12.$ Therefore the mean-squared quantization noise is

$$\mathrm{E}\left(Q^2\right) = \sigma_Q^2 + \mathrm{E}^2(Q) = \mathrm{LSB}^2/12 + \left(-\mathrm{LSB}/2\right)^2 = \mathrm{LSB}^2/3$$

Statistical Analysis of Quantization Effects

The power spectral density function for quantization noise using two's complement representation is $\mathrm{G}_Q(F)$. It has an impulse of strength $\mathrm{LSB}^2/4$ at $F=0$ and is otherwise flat with a value K in the range $-1/2 < F < 1/2$. It repeats periodically outside this range.

$$\mathrm{E}\left(Q^2\right) = \int_{-1/2}^{1/2} \mathrm{G}_Q(F)\,dF = \int_{-1/2}^{1/2} \left[\left(\mathrm{LSB}^2/4\right) \sum_{k=-\infty}^{\infty} \delta(F-k) + K \right] dF$$

$$= \mathrm{LSB}^2/4 + K = \mathrm{LSB}^2/3$$

Solving, $K = \mathrm{LSB}^2/12$ and

$$\mathrm{G}_Q(F) = \mathrm{LSB}^2/12 + \left(\mathrm{LSB}^2/4\right) \sum_{k=-\infty}^{\infty} \delta(F-k)$$

Statistical Analysis of Quantization Effects

$G_Q(F) = \mathrm{LSB}^2/12 + \left(\mathrm{LSB}^2/4\right)\sum_{k=-\infty}^{\infty}\delta(F-k)$ is the power spectral density of both quantization noise sources. The power spectral density of the quantization noise effect on the output signal is

$$G_y(F) = G_{y,x}(F) + G_{y,f}(F)$$

where $G_{y,x}(F) = G_x\left|H_{xy}(F)\right|^2 = G_Q\left|H_{xy}(F)\right|^2$

and $G_{y,f}(F) = G_y\left|H_{fy}(F)\right|^2 = G_Q\left|H_{fy}(F)\right|^2$

and $H_{xy}(F) = \dfrac{Y(F)}{Q_x(F)} = \dfrac{e^{j2\pi F}}{e^{j2\pi F} - a}$ and $H_{fy}(F) = \dfrac{Y(F)}{Q_f(F)} = \dfrac{e^{j2\pi F}}{e^{j2\pi F} - a}$

Statistical Analysis of Quantization Effects

The two quantization noise sources have the same power spectral density and they are independent so the output quantization noise is just twice what would be produced by one of them acting alone.

$$P_{Q,\mathrm{y}} = 2\int_{-1/2}^{1/2} \mathrm{G}_Q(F)\left|\frac{e^{j2\pi F}}{e^{j2\pi F}-a}\right|^2 dF$$

$$= \frac{\mathrm{LSB}^2}{2}\left(\frac{1}{1-a}\right)^2 + \frac{\mathrm{LSB}^2}{6}\int_{-1/2}^{1/2}\left|\frac{e^{j2\pi F}}{e^{j2\pi F}-a}\right|^2 dF$$

Evaluating the integral, $P_{Q,\mathrm{y}} = \dfrac{\mathrm{LSB}^2}{2}\left(\dfrac{1}{1-a}\right)^2 + \dfrac{\mathrm{LSB}^2}{6}\dfrac{1}{1-a^2}.$

Statistical Analysis of Quantization Effects

$$\text{In } P_{Q,\text{y}} = \frac{\text{LSB}^2}{2}\left(\frac{1}{1-a}\right)^2 + \frac{\text{LSB}^2}{6}\frac{1}{1-a^2}$$

$\frac{\text{LSB}^2}{2}\left(\frac{1}{1-a}\right)^2$ is the effect of the expected value of the input quantization noise and

$\frac{\text{LSB}^2}{6}\frac{1}{1-a^2}$ is the effect of the variance of the input quantization noise.

Both are dependent on a. The closer a is to one, the larger the output quantization noise.