Overview of Effects of Finite Register Length in DSP

1) Quantization in A/D

2) Coefficient inaccuracy

3) Arithmetic "roundoff"

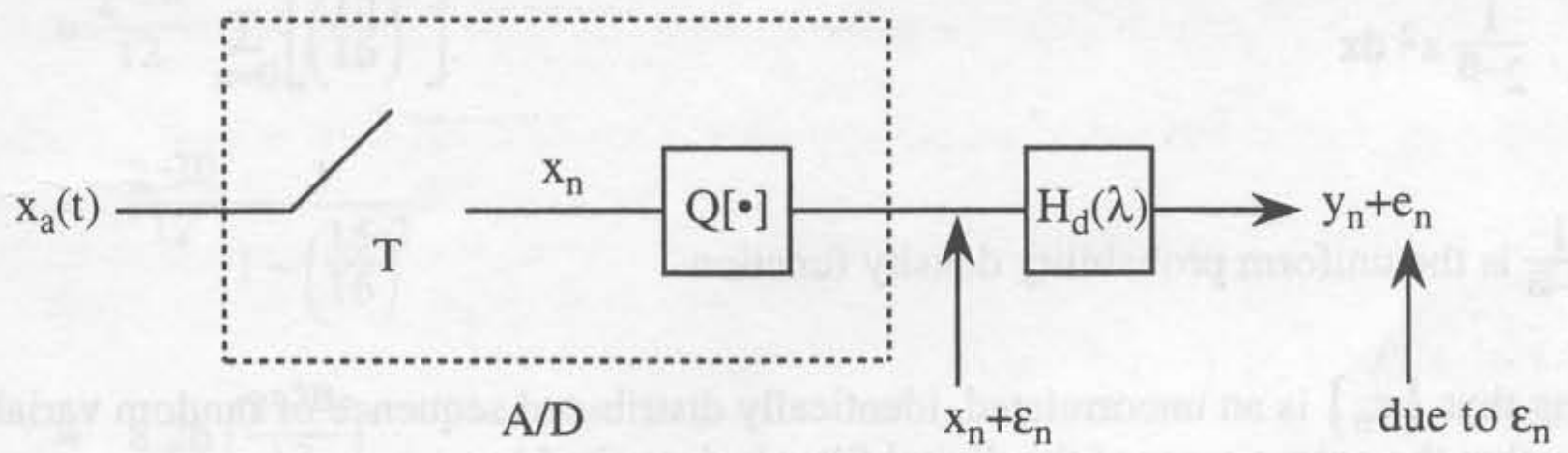
   a) "Roundoff noise"

   b) Adder overflow

   c) Zero-input limit cycles

1) Quantization in A/D

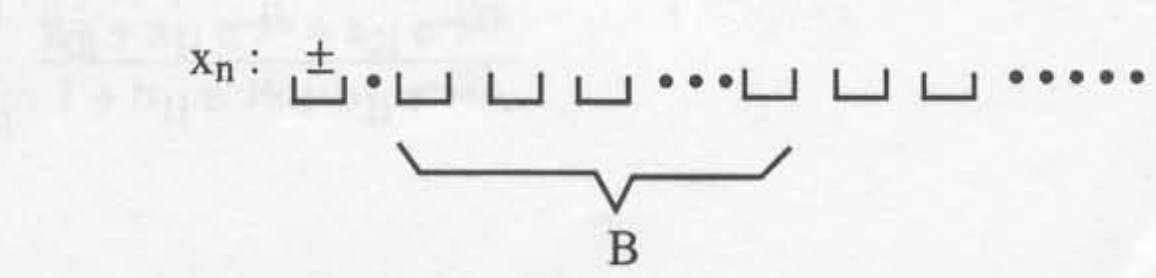
The A/D employs a quantizer to round off the value of x_n to fit into the available register length. The rounding operation creates an error ε_n that propagates to the output of the DSP system. As an example of this, consider an A/D followed by a digital filter:



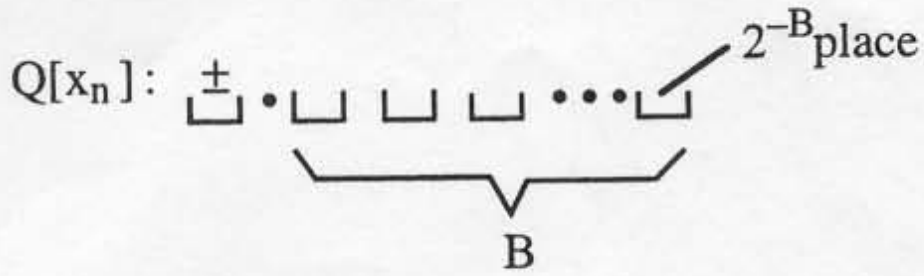
The quantizer Q is a nonlinearity. We will model it by the additive error source $\varepsilon_n = Q[x_n] - x_n$

.

Assume a fixed-point binary arithmetic representation with B bits plus sign. Also assume $|x_a(t)| < 1$ so that x_n is a fraction. An exact binary representation of x_n would generally require an infinite number of bits:



where each bit is a zero or a one. The quantizer Q rounds or truncates this representation to B bits plus sign:



Suppose Q[•] rounds to B bits plus sign. Then

$$-\frac{2^{-B}}{2} \le \varepsilon_n \le \frac{2^{-B}}{2}$$

Assuming ε_n is a random variable, uniformly distributed on its range, the mean-squared value of ε_n is

$$\boxed{E\left\{\varepsilon_n^2\right\} = \frac{2^{-2B}}{12}}$$

Here, E denotes the mean or probabilistic "average" and the above formula is computed as

$$\int_{\frac{-2^{-B}}{2}}^{\frac{2^{-B}}{2}} \frac{1}{2^{-B}} x^2 \, dx$$

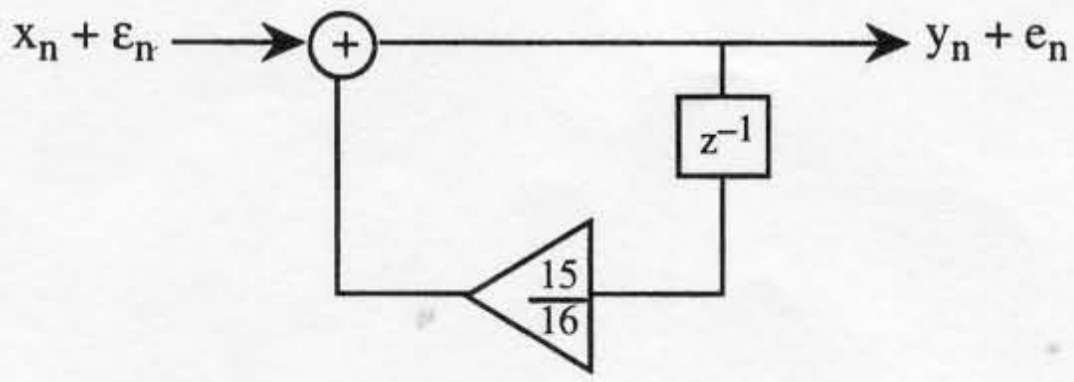
where $\frac{1}{2^{-B}}$ is the uniform probability density function.

Assuming that $\{\varepsilon_n\}$ is an uncorrelated, identically distributed sequence of random variables, we can show that the output error of the digital filter is described by

$$\boxed{E\left\{e_n^2\right\} = \frac{2^{-2B}}{12} \cdot \frac{1}{2\pi} \int_{-\pi}^{\pi} |H_d(\lambda)|^2 \, d\lambda}$$

Example

Given the digital filter below, with A/D quantization noise at its input, find the resulting mean-squared error at the filter output.



Solution

$$E\left\{e_n^2\right\} = \frac{2^{-B}}{12}\,\frac{1}{2\pi}\int_{-\pi}^{\pi} |H_d(\lambda)|^2\, d\lambda$$

$$\overset{\uparrow}{=} \frac{2^{-2B}}{12}\sum_{n=0}^{\infty} |h_n|^2$$

Parseval

$$= \frac{2^{-2B}}{12}\sum_{n=0}^{\infty}\left[\left(\frac{15}{16}\right)^n\right]^2$$

$$= \frac{2^{-2B}}{12}\,\frac{1}{1-\left(\frac{15}{16}\right)^2}$$

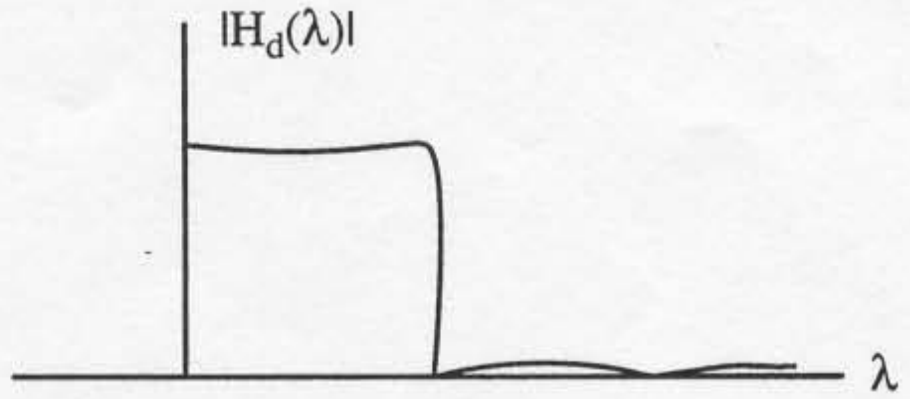
$$\approx 8.26\left(\frac{2^{-2B}}{12}\right)$$

So, the filter increases the mean-squared value of the A/D quantization noise by a factor of about eight.

2) Coefficient Inaccuracy

Can be treated deterministically. Suppose the desired infinite-precision frequency response is

$$H_d(\lambda) = \prod_{i=1}^{M} \frac{a_{0i} + a_{1i}\, e^{-j\lambda} + a_{2i}\, e^{-j2\lambda}}{1 + b_{1i}\, e^{-j\lambda} + b_{2i}\, e^{-j2\lambda}}$$

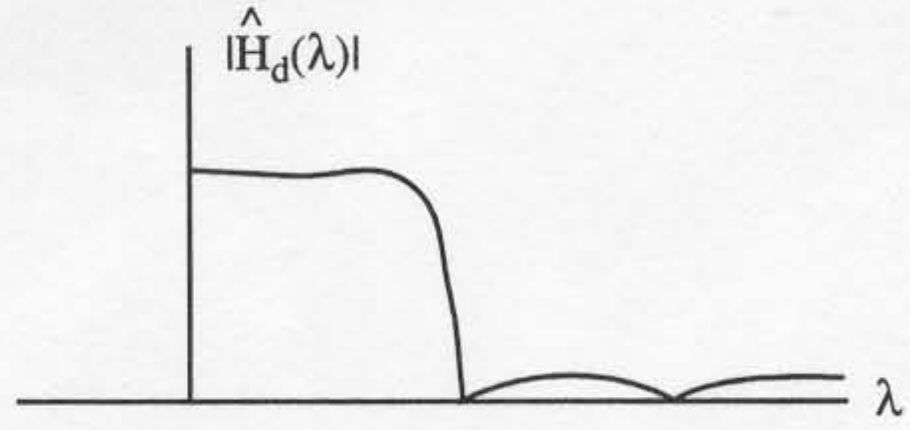


Then the actual frequency response after quantization of the filter coefficients will be

$$\hat{H}_d(\lambda) = \prod_{i=1}^{M} \frac{\hat{a}_{0i} + \hat{a}_{1i}\, e^{-j\lambda} + \hat{a}_{2i}\, e^{-j2\lambda}}{1 + \hat{b}_{1i}\, e^{-j\lambda} + \hat{b}_{2i}\, e^{-j2\lambda}}$$

where $\hat{a}_i = Q[a_i]$.

The result of coefficient inaccuracy is a deterministic degradation in the frequency response:



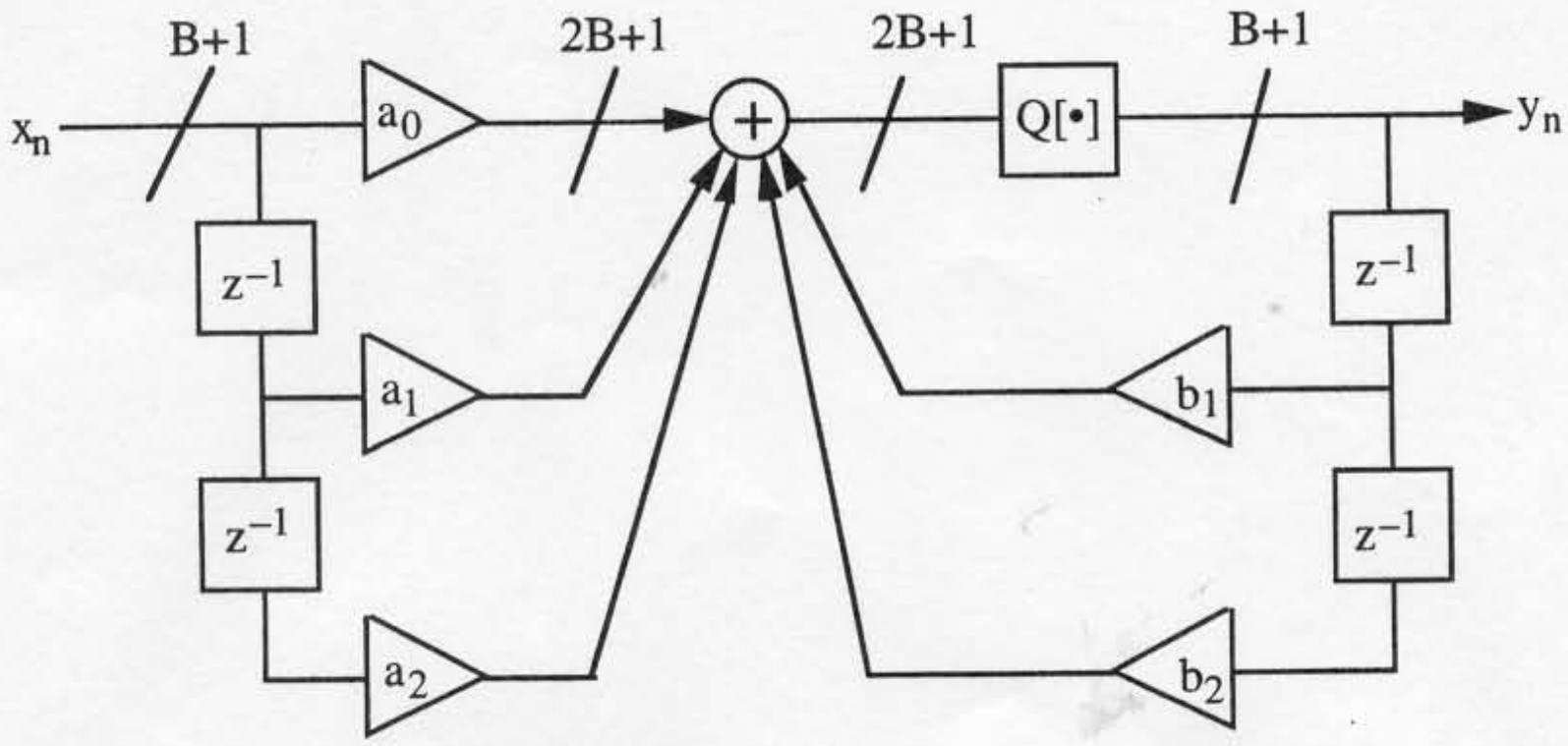
A sensitivity analysis can be helpful in determining, for a given filter structure, how badly the frequency response will be perturbed by a small change in the filter coefficients. Study:

$$\frac{\partial}{\partial a_k} |H_d(\lambda)|, \qquad \frac{\partial}{\partial b_k} |H_d(\lambda)|,$$

$$\frac{\partial}{\partial a_k} \angle H_d(\lambda), \qquad \frac{\partial}{\partial b_k} \angle H_d(\lambda)$$

3a) Multiplication Roundoff Noise

Consider the filter structure



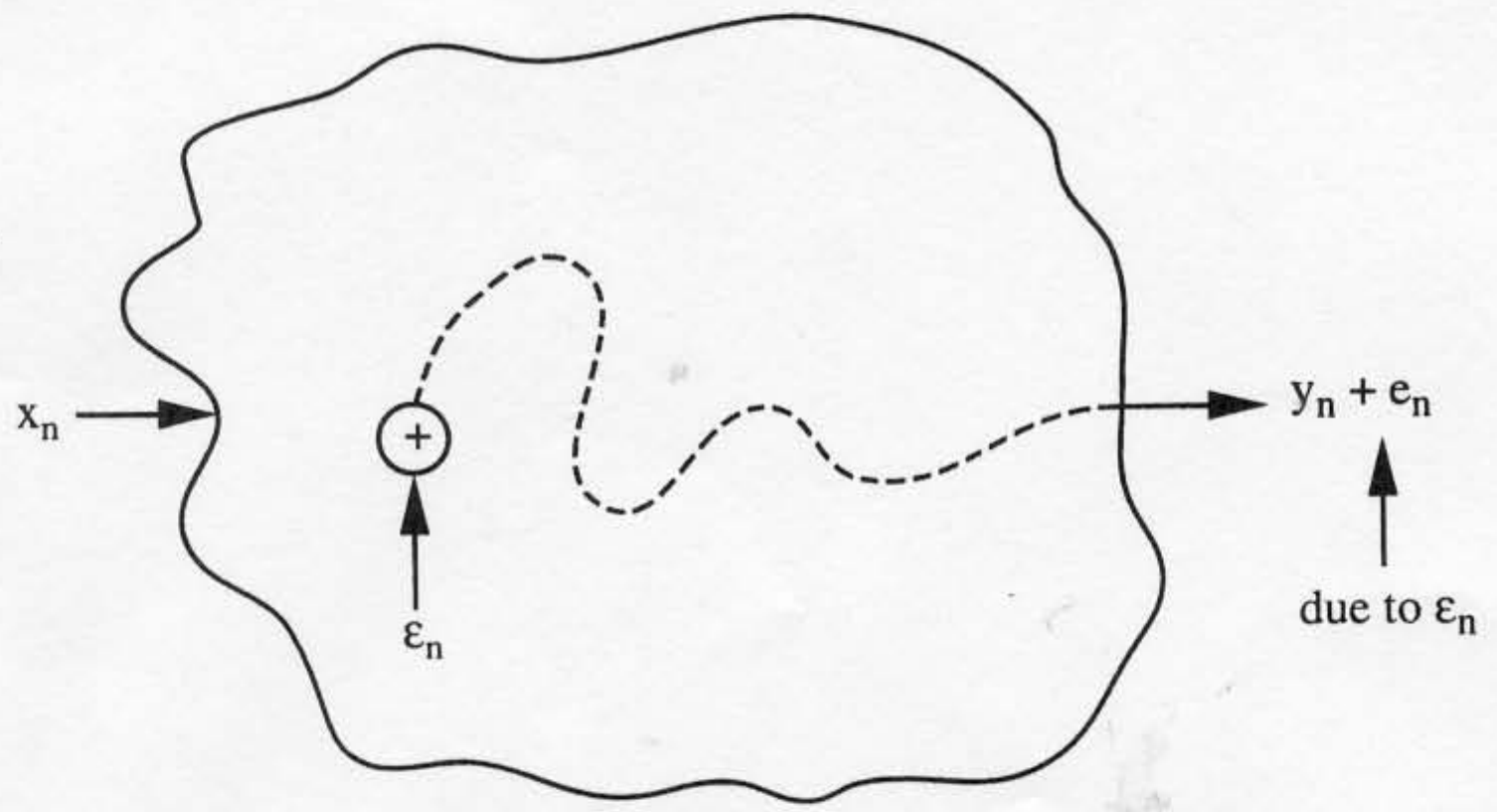
where the register lengths are indicated at various points in the filter structure. Note that the multiplier outputs have length 2B + 1 and these must be quantized back to B + 1 bits before they enter a feedback loop. Otherwise, the required register length will grow without bound.

Here we assume a double-length accumulator, in which case a single quantization occurs after the accumulator. Alternatively, we could quantize at each multiplier output, individually.

For a typical x_n, we can treat the error $\{\varepsilon_n\}$ at the internal quantizer as additive, uncorrelated noise with

$$E\left\{\varepsilon_n^2\right\} = \frac{2^{-2B}}{12} \qquad \text{(for rounding).}$$

General conceptual setting:



Let F(z) be the transfer function from the error source (quantizer) to y_n. Then for fixed-point rounding

$$\text{MSE} = E\left\{e_n^2\right\} = \frac{2^{-2B}}{12}\,\frac{1}{2\pi}\int_{-\pi}^{\pi} |F_d(\lambda)|^2\, d\lambda$$

For the above second-order section:

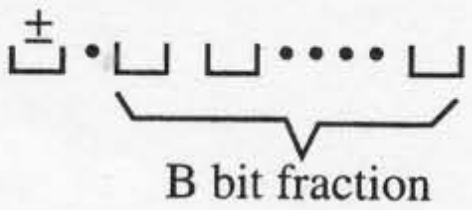
$$F(z) = \frac{1}{1 - b_1 z^{-1} - b_2 z^{-2}}$$

In general, $F(z) \neq H(z)$.

There may be several quantization points in a filter. We generally can assume all error sources are uncorrelated so that the mean-squared error at the output is the sum of the individual output mean-squared errors.

3b) Adder Overflow

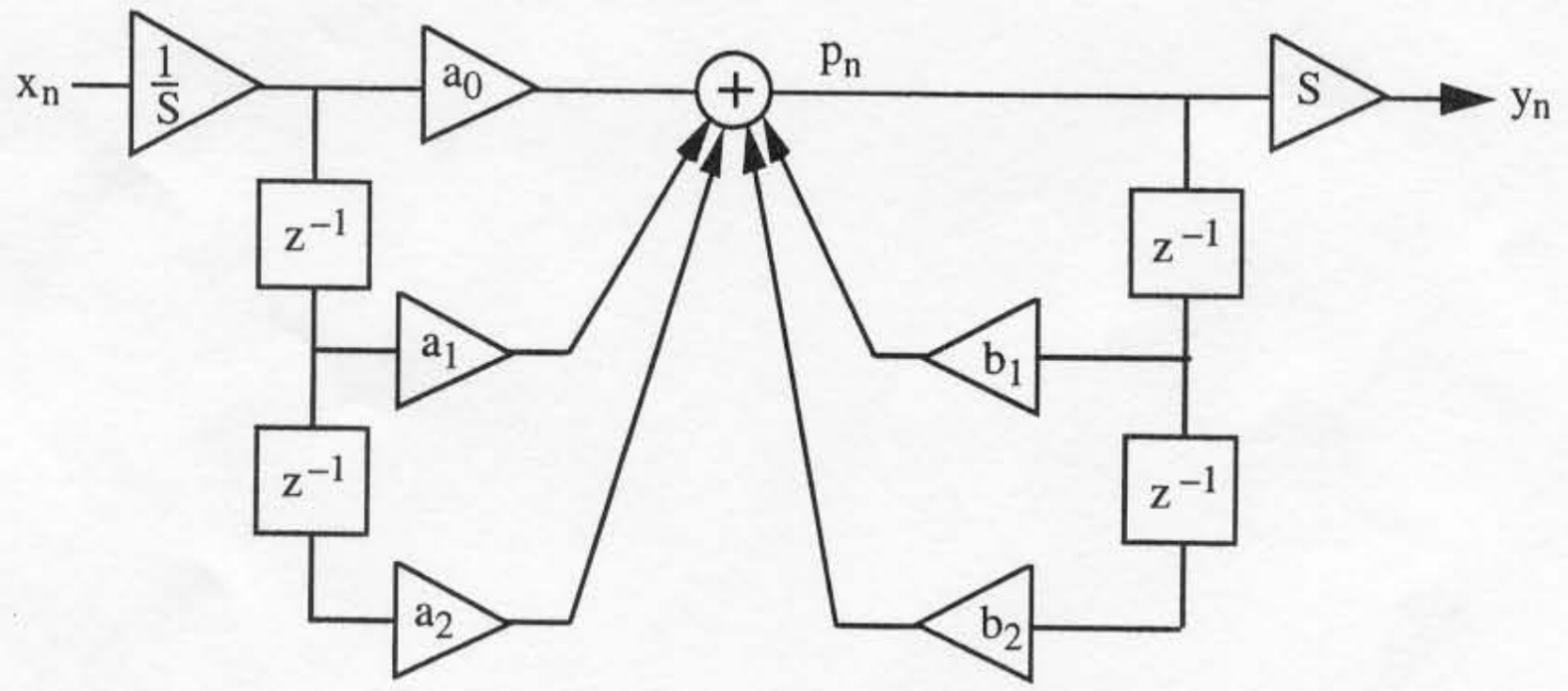
Without loss of generality assume all quantities are stored as



Even if the input satisfies $|x_n| < 1$ for all n, we may have |adder output| > 1. If this happens, we will have overflow (large error) at the adder output.

Solution: Scale down adder inputs by S to prevent overflow.

Picture:



In practice, the scaling operation does not necessitate additional multiplications. S may be a power of 2, or alternatively, $\frac{1}{S}$ can be incorporated into the a_i and you may not care whether the output is renormalized by S.

We want $|p_n| < 1$. How do we choose S?

Let G(z) be the transfer function from x_n to the <u>adder</u> output <u>prior</u> to scaling. Let g_n be the corresponding unit pulse response. Then:

$$p_n = \frac{1}{S} \sum_{m=0}^{\infty} g_m \, x_{n-m}$$

$$\Rightarrow \quad |p_n| \le \frac{1}{S} \sum_{m=0}^{\infty} |g_m| \underbrace{|x_{n-m}|}_{<1}$$

$$< \frac{1}{S} \sum_{m=0}^{\infty} |g_m|.$$

$\Rightarrow$ Choose

$$\boxed{S = \sum_m |g_m|} \qquad (*)$$

(*) is called ℓ_1 scaling and it will guarantee no overflow at the adder.

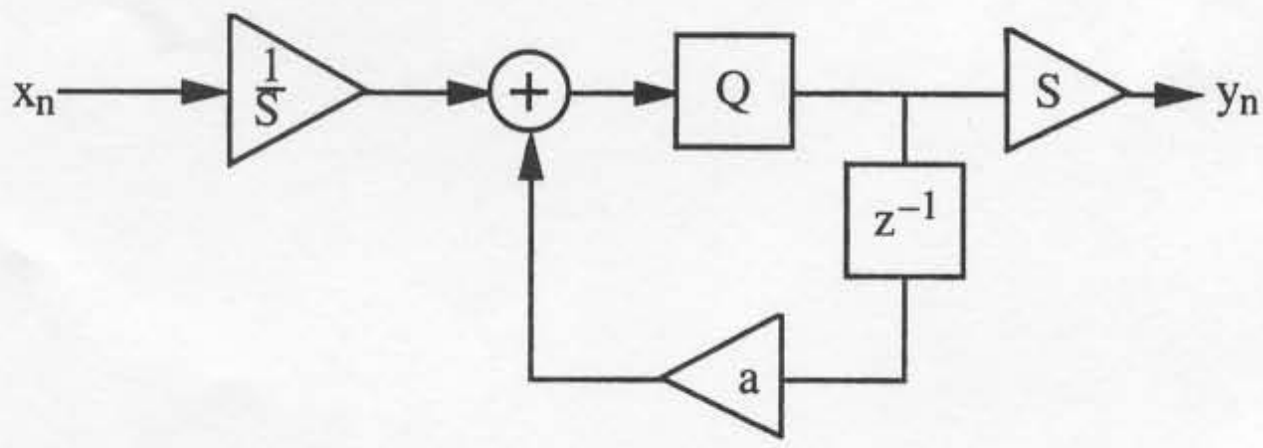
Problem: Large S increases roundoff noise at the filter output. There exist alternative scaling policies. One of these is called ℓ_2 scaling:

$$S_{\ell_2} = \left[\sum_m g_m^2\right]^{1/2}$$

It can be shown that $S_{\ell_2} \le S_{\ell_1}$ (and often <u>much</u> less).

With ℓ_2 scaling, overflow <u>can occur</u> (since S_{ℓ_2} is generally less than S_{ℓ_1}) but it is <u>unlikely</u>. If your system can afford to have very rare overflows, you may choose a less stringent scaling policy than ℓ_1, to keep roundoff noise lower at the filter output.

Example Combined effects of scaling and roundoff noise.



Assume fixed point, rounding arithmetic with B bits plus sign. Do these things:

a) Scale the filter using ℓ_1 scaling.

b) Find the MSE due to rounding in terms of 'a' and B.

c) Examine what happens to the MSE as a function of pole location.

Solution

a) With no scaling, the unit-pulse response from the input x_n to the adder output is

$$g_n = a^n u_n \qquad \left(TF = \frac{1}{1 - az^{-1}}\right)$$

In this example, $g_n = h_n$. In general, however, this will not be the case.

To scale the filter, we use (*) to select S as

$$S = \sum_{n=0}^{\infty} |a^n| = \sum_{n=0}^{\infty} |a|^n = \frac{1}{1 - |a|}$$

b) The transfer function from the quantizer to the filter output is $F(z) = S \bullet H(z)$. Thus,

$$MSE = \frac{2^{-2B}}{12} \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{S}{1 - ae^{-j\lambda}} \right|^2 d\lambda$$

$$= \frac{2^{-2B}}{12} S^2 \sum_{n=0}^{\infty} |a^n|^2$$

$$= \frac{2^{-2B}}{12} S^2 \sum_{n=0}^{\infty} (a^2)^n$$

$$= \frac{2^{-2B}}{12} \left(\frac{1}{1 - |a|}\right)^2 \frac{1}{1 - a^2}$$

As an example, if a = .95, then $MSE = 4013 \frac{2^{-2B}}{12}$.

Since $4013 \sim 2^{12}$, we see that roughly the six lowest bits in the output register would be filled with roundoff noise.

Comments:

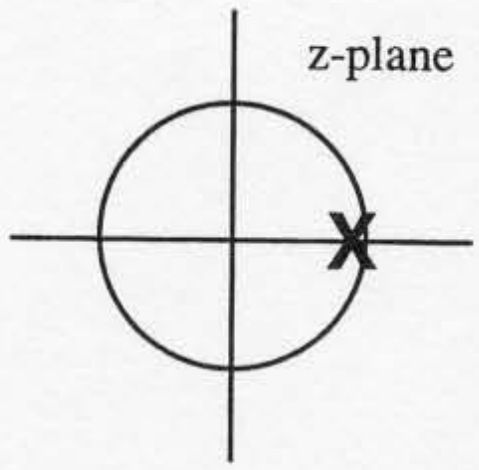
i) An actual filter implementation often uses longer internal register lengths (e.g., by 6 bits) than output register lengths, since there is no reason to present noise to the output.

ii) Scaling is a problem only for fixed-point arithmetic, not for floating point.

Now, examine what happens to the MSE as a function of pole location.

c) $H(z) = \dfrac{1}{1 - az^{-1}} \Rightarrow$ pole at $z = a$

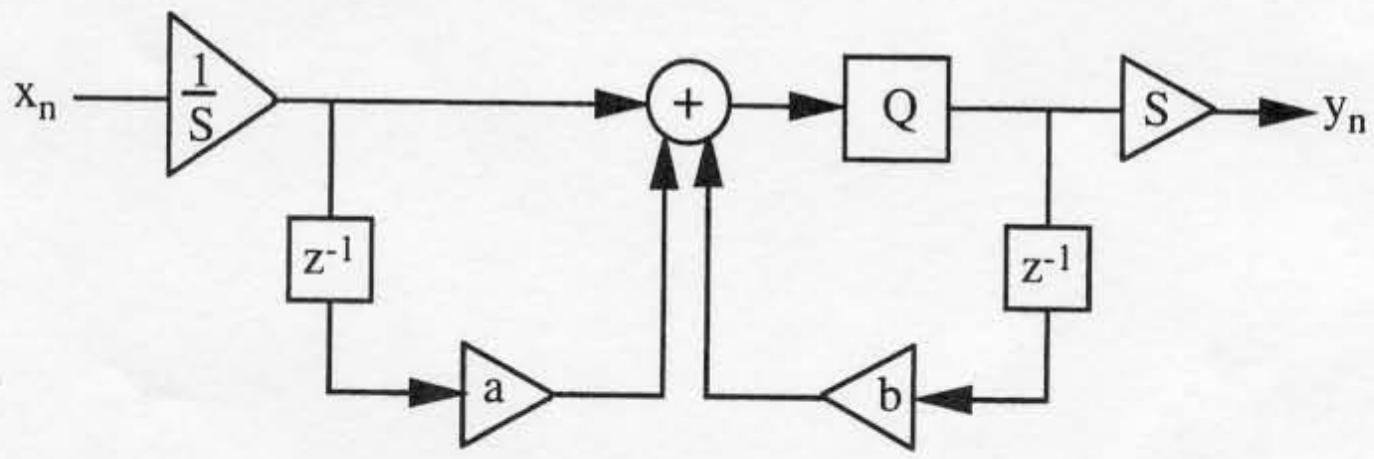
Pole location:



Notice that as the pole approaches the unit circle (in this example, as $|a| \to 1$), the MSE $\to \infty$. Thus, longer register lengths will be needed if the pole is closer to the unit circle. This observation is not unique to this example. Recursive filters having sharp transitions in their frequency response tend to have poles located very near the unit circle. Such filters require long register lengths.

Example

Find an expression for the mean-squared roundoff noise in the following ℓ_1-scaled filter.



Assume $a > 0$ and $b > 0$.

The transfer function from the input to the adder output, with no scaling, is

$$G(z) = \frac{1 + az^{-1}}{1 - bz^{-1}}$$

$$= \frac{z + a}{z - b}$$

$$\Rightarrow g_n = b^n u_n + ab^{n-1} u_{n-1}$$

Using the ℓ_1 policy, we choose

$$S = \sum_n |g_n| = \sum_{n=0}^{\infty} |b^n + a\, b^{n-1} u_{n-1}|$$

$$= \sum_{n=0}^{\infty} b^n + \frac{a}{b} \sum_{n=1}^{\infty} b^n$$

$\uparrow$ since
$a, b > 0$

$$= \frac{1}{1-b} + \frac{a}{b}\left[-1 + \frac{1}{1-b}\right]$$

$$= \frac{1}{1-b} + \frac{a}{b}\left[\frac{b}{1-b}\right] = \boxed{\frac{1+a}{1-b}}$$

Now, to compute the MSE due to roundoff noise, we note that the transfer function from the quantizer to the filter output is

$$F(z) = \frac{S}{1 - b\,z^{-1}} = \frac{S\,z}{z-b}$$

$$\Rightarrow f_n = S\,b^n\,u_n$$

So,

$$MSE = \frac{2^{-2B}}{12} \frac{1}{2\pi} \int_{-\pi}^{\pi} |F_d(\lambda)|^2\, d\lambda$$

$$= \frac{2^{-2B}}{12} \sum_n |f_n|^2$$

$$= \frac{2^{-2B}}{12} \sum_{n=0}^{\infty} S^2\,(b^n)^2$$

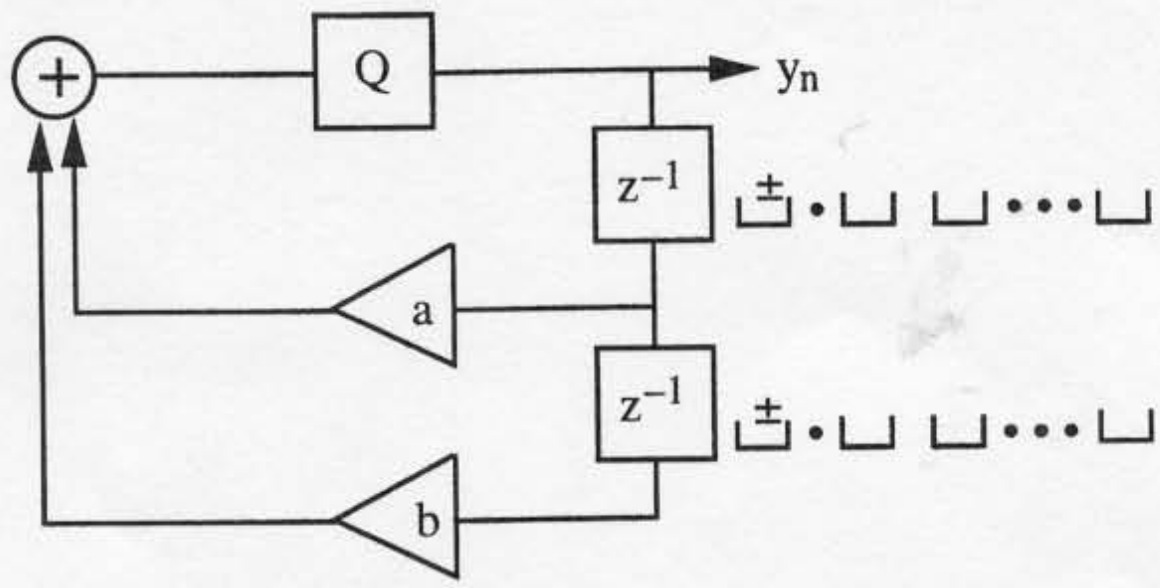
$$= \frac{2^{-2B}}{12} \left[\frac{1+a}{1-b}\right]^2 \sum_{n=0}^{\infty} (b^2)^n$$

$$= \boxed{\frac{2^{-2B}}{12} \left[\frac{1+a}{1-b}\right]^2 \frac{1}{1-b^2}}$$

3c) **Zero-Input Limit Cycles** (only occur in IIR filters)

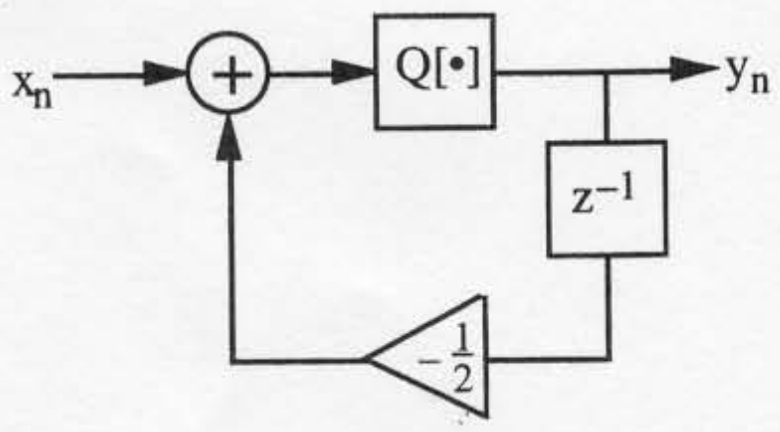
The output of a stable linear filter must decay to zero if the input drops to zero. However, in an IIR filter, the quantizer or quantizers (nonlinear!) may cause a nonzero periodic output, which is called a limit cycle.

Why periodic? Consider the following recursive filter, with zero input and finite-length registers:



The output must be periodic in this situation because the feedback delay registers can contain only a finite number of possible pairs of values, so that at some point they must repeat. Once the contents of the delay registers repeats, the output will become periodic because the contents of the delay registers is a state of the system, which determines all future outputs.

Example 1



$$y_n = Q\left[x_n - \frac{1}{2}y_{n-1}\right]$$

Assume sign-magnitude rounding, with B = 3, $x_n = 0$, $n \geq 0$, and $y_{-1} = \frac{1}{2}$. (Here, B = 3 would be unrealistically small for a digital filter, but this choice serves perfectly well as a simple example.) Also, assume that the quantizer rounds up at the midpoint and that the quantizer operates only on the bits to the right of the binary point (positive and negative numbers get quantized in a similar way). Then:

$$y_0 = Q\left[-\frac{1}{2}y_{-1}\right] = Q\left[-\frac{1}{4}\right] = -\frac{1}{4}$$

$$y_1 = Q\left[-\frac{1}{2}y_0\right] = Q\left[\frac{1}{8}\right] = \frac{1}{8}$$

$$y_2 = Q\left[-\frac{1}{2}y_1\right] = Q\left[-\frac{1}{16}\right] = -\frac{1}{8}$$

$$y_3 = Q\left[-\frac{1}{2}y_2\right] = Q\left[\frac{1}{16}\right] = \frac{1}{8}$$

$$\vdots$$

So, y_n enters the limit cycle $\frac{1}{8}, -\frac{1}{8}, \frac{1}{8}, -\frac{1}{8}, \ldots$

In this example, the limit cycle occupies only the least significant bit.