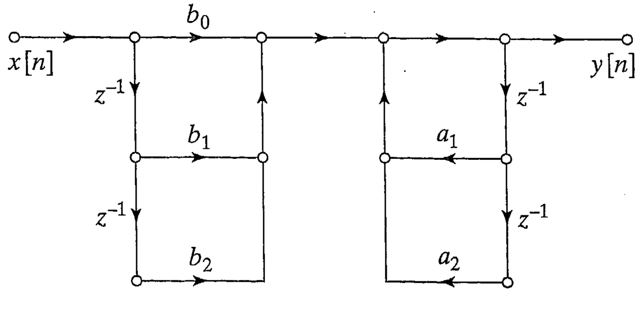
ECE 8440 Unit 14

Effects of Round-Off Noise in Digital Filters-Continued

Consider the linear difference equation associated with a Direct Form I implementation of an IIR digital filter:

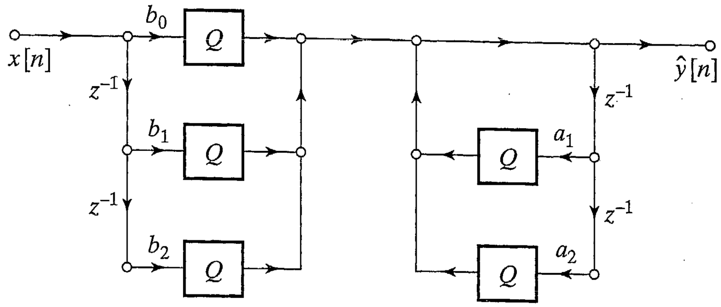
$$y(n) = \sum_{k=0}^{M} b_k x(n-k) + \sum_{k=1}^{N} a_k y(n-k) \qquad \text{(equation 6.90)}$$

A signal flow diagram for the implementation of this system is shown below:



(a)

Round-off error can occur each time a multiplication is implemented in the above diagram. The diagram can be modified to include the effect of round-off error as follows:



(b)

Based on the above figure, we could rewrite the difference equation for the system as

$$\hat{y}(n) = \sum_{k=0}^{M} Q[b_k x(n-k)] + \sum_{k=1}^{N} Q[a_k \hat{y}(n-k)].$$

Note that the system as shown in the above figure is non-linear, due to the properties of the quantizer.

Another way to represent the effects of quantization due to round-off errors is to represent round-off error as independent noise sources which inject noise into the system at each point where round-off error occurs. Each noise source can be formally defined as

$$e(n) = Q[bx(n)] - bx(n).$$

This approach leads to the figure shown below:

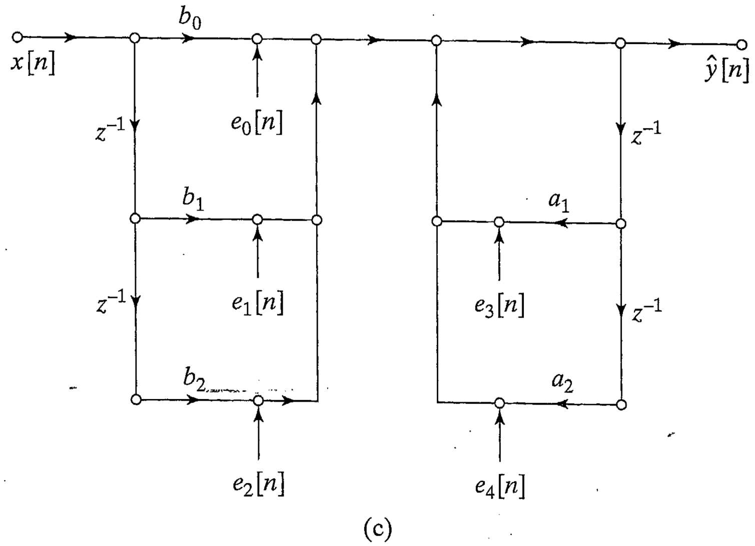


Figure 6.57 Models for direct form I system. (a) Infinite-precision model (b) Nonlinear quantized model (c) Linear-noise model.

In order to obtain a mathematically tractable way to analyze noise, due to round-off errors, <u>we make the following assumptions:</u>

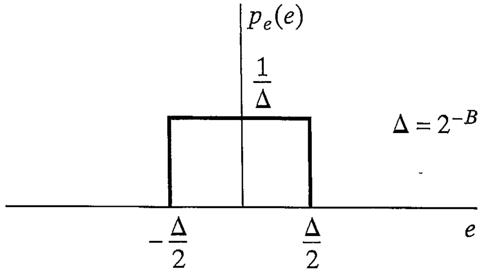
1. **Each noise source e(n) is a wide-sense-stationary white-noise process.**
2. **Each noise source has a uniform density function.**
3. **Each noise source is uncorrelated with the quantizer input, with all other quantization noise sources, and with the input to the system.**

Research has shown that these assumptions are valid if the signal is a "complicated" wideband signal such as speech, in which the signal fluctuates rapidly over all quantization levels and typically traverses over many quantization levels in going from sample to sample.

As seen before, if B +1 bits are available to represent signals that range from -1 to 1, the size range for round-off error for two's complement representation is

$$-\frac{\Delta}{2} < e(n) < \frac{\Delta}{2} \qquad \text{where} \qquad \Delta = 2^{-B}.$$

If the probability density function for e(n) is uniform over this range, as shown in the figure below,



(a)

The variance of the round-off error (noise power) can be found as follows:

$$E\{e^2(n)\} = \sigma_e^2 = \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \frac{1}{\Delta} e^2 de$$

$$= \frac{1}{\Delta}\frac{e^3}{3}\Bigg|_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} = \frac{1}{3\Delta}\left[\frac{\Delta^3}{8} - \frac{(-\Delta)^3}{8}\right] = \frac{\Delta^2}{12}.$$

Expressed in term of B, this becomes

$$\sigma_e^2 = \frac{\left(2^{-B}\right)^2}{12} = \frac{2^{-2B}}{12}.$$

If the round-off error is uncorrelated and has zero mean as assumed, its autocorrelation is

$$\phi_{ee}(n) = \sigma_e^2\delta(n)$$

and its power spectrum is

$$\Phi_{ee}(e^{j\omega}) = \sigma_e^2.$$

Since the various round-off source sources are assumed independent of each other, all the noise sources that inject noise at the same summation node can be combined into a single noise source whose variance (total noise power) is equal to the sum of the variances of the contributing noise sources.

For example, all five noise sources in Figure 6.57(c) can be combined into a single source e(n) where

$$e(n) = e_1(n) + e_2(n) + e_3(n) + e_4(n) + e_5(n)$$

The corresponding signal flow diagram can now be represented as

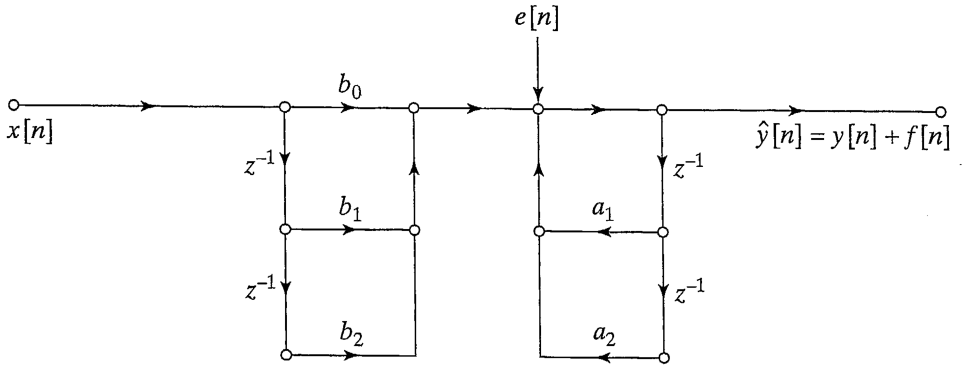


Figure 6.59 Linear-noise model for direct form I with noise sources combined.

The total noise power of the combined round-off noise sources is, since they are assumed to be independent, is

$$\sigma_e^2 = \sigma_{e0}^2 + \sigma_{e1}^2 + \sigma_{e2}^2 + \sigma_{e3}^2 + \sigma_{e4}$$

$$= 5\frac{2^{-2B}}{12}.$$

In general, for a Direct Form I implementation for a filter having M zeros and N poles, this expression for the total noise power of the combined round-off noise sources is

$$\sigma_e^2 = (M + 1 + N)\frac{2^{-2B}}{12}.$$

For Direct Form I, all the round-off noise is injected after the signal passes through the zeros and before it passes through the poles. In other words, the round-off noise passes only though the poles of the filter. Therefore, the transfer function between the point of noise injection and the filter output is

$$H_{ef}(z) = \frac{1}{A(z)}$$

where A(z) is the denominator of the overall system function H(z).

The total output noise power due to internal round-off error is therefore

$$\sigma_f^2 = (M+1+N)\frac{2^{-2B}}{12}\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{d\omega}{|A(e^{j\omega})|^2} \qquad \text{(equation 6.106)}$$

$$= (M+1+N)\frac{2^{-2B}}{12}\sum_{n=-\infty}^{\infty}|h_{ef}(n)|^2$$

where $h_{ef}(n)$ is the unit sample response corresponding to $H_{ef}(z) = \frac{1}{A(z)}$.

Example 6.11 - Round-off Noise in a 1st Order System

Consider an LTI system having system function

$$H(z) = \frac{b}{1-az^{-1}}, \quad |a|<1 \qquad \text{(equation 6.107)}$$

A Direct Form I implementation of this system is shown below:

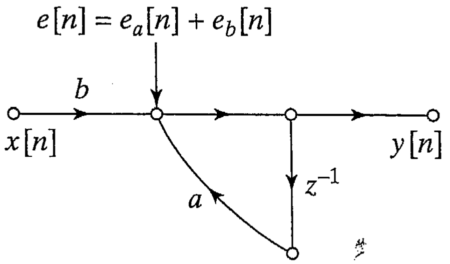


Figure 6.60 1st order linear noise model.

In the Direct Form I implementation, the transfer function that both sources of round-off noise will pass through is

$$H_{ef}(z) = \frac{1}{1 - az^{-1}}$$

with corresponding frequency response of

$$H_{ef}(e^{j\omega}) = \frac{1}{1 - ae^{-j\omega}}.$$

The power spectrum of noise at the output is

$$\Phi_{ff}(e^{j\omega}) = P_{ff}(\omega) = \sigma_e^2 |H_{ef}(e^{j\omega})|^2$$

$$= 2\frac{2^{-2B}}{12}\left(\frac{1}{1 + a^2 - 2a\cos(\omega)}\right).$$

(The leading factor of 2 is because there are two sources of round-off noise in the figure.)

The unit sample response of $H_{ef}(z)$ is

$h_{ef}(n) = a^n u(n)$.

For this simple form of $h_{ef}(n)$ we can find the total power in the output noise using

$$\sigma_f^2 = \sigma_e^2 \sum_{n=0}^{\infty} |\, h(n)\,|^2$$

$$= \sigma_e^2 \sum_{n=0}^{\infty} |\, a\,|^{2n} = 2\frac{2^{-2B}}{12}\left(\frac{1}{1-|\,a\,|^2}\right).$$

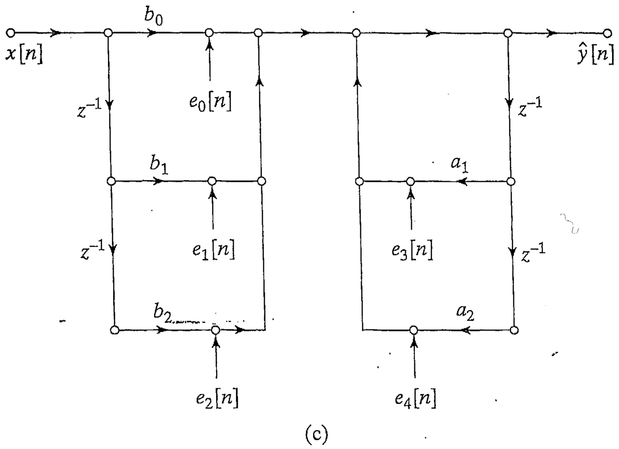
Note that the output noise power increases dramatically as the location of the pole at z = a approaches the unit circle.

Example 6.12 Round-Off Noise in 2nd Order System

Consider the following second order system:

$$H(z)=\frac{b_0+b_1z^{-1}+b_2z^{-2}}{(1-re^{j\theta}z^{-1})(1-re^{-j\theta}z^{-1})}.$$

The Direct Form I implementation of this system is shown below:



(c)

Note that there are five sources of round-off noise in this implementation. The total noise power in the output can be expressed as

$$\sigma_f^2 = 5\sigma_e^2 \int_{-\pi}^{\pi} \left| H_{ef}(e^{j\omega}) \right|^2 d\omega$$

$$= 5\frac{2^{-2B}}{12} \int_{-\pi}^{\pi} \frac{d\omega}{\left|(1 - re^{j\theta}e^{-j\omega})(1 - re^{-j\theta}e^{-j\omega})\right|^2} \qquad \text{(equation 6.111)}$$

As shown previous, we can use a z-transform approach to evaluating σ_f^2. (This was based on material in Appendix A.5). Using the z-transform approach, we have shown that if a white noise source with variance σ_e^2 passed through a second order system with poles at $re^{j\theta}$ and $re^{-j\theta}$ and no zeros, the resulting output noise power is

$$\sigma_f^2 = \gamma_{ff}(0) = \sigma_e^2 \left(\frac{1+r^2}{1-r^2}\right) \frac{1}{r^4 + 1 - 2r^2 \cos 2\theta}.$$

Applying this result to the current example where there are 5 noise sources, each with average power of

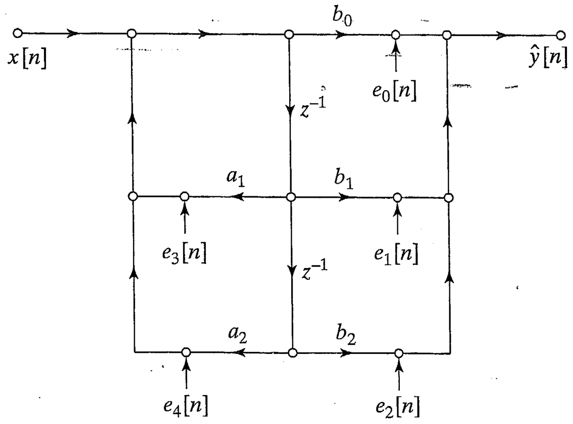
$$\sigma_e^2 = \frac{2^{-2B}}{12},$$

we obtain

$$\sigma_f^2 = 5\frac{2^{-2B}}{12} \left(\frac{1+r^2}{1-r^2}\right) \frac{1}{r^4 + 1 - 2r^2 \cos 2\theta}.$$

Round-Off Noise in Direct Form II

Sources of round-off error in a Direct Form II implementation of a second order system are shown in the figure below:



The corresponding difference equations, also representing the round-off error, as are follows:

$$\hat{w}[n] = \sum_{k=1}^{N} Q[a_k \hat{w}[n-k]] + x[n]$$

and

$$\hat{y}[n] = \sum_{k=1}^{M} Q[b_k \hat{w}[n-k]].$$

The N round-off errors involved in generating $\hat{w}[n]$ in the first equation can be modeled as a combined noise source $e_a(n)$ that enters the system at the system input.

The M + 1 round-off errors involved in generating $\hat{y}[n]$ can be modeled as a combined noise source $e_b(n)$ that enters the system at the system output, as shown in the figure below:

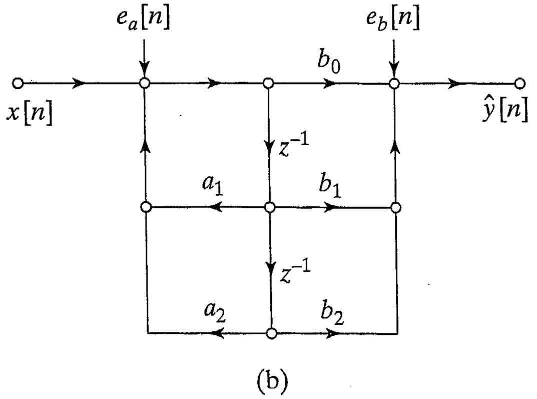


Figure 6.61 Linear-noise models for direct form II. (a) Showing quantization of individual products. (b) With noise sources combined.

The power spectrum of noise in the system output due to round-off error can therefore be represented as

$$\Phi_{ff}(e^{j\omega}) = P_{ff}(\omega) = N\frac{2^{-2B}}{12} \mid H(e^{j\omega}) \mid^2 + (M+1)\frac{2^{-2B}}{12}.$$

The total average noise power in the output can be expressed as

$$\sigma_f^2 = N\frac{2^{-2B}}{12}\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|H(e^{j\omega})\right|^2 d\omega \;+\; (M+1)\frac{2^{-2B}}{12}$$

or as

$$\sigma_f^2 = N\frac{2^{-2B}}{12}\sum_{n=-\infty}^{\infty}\left|h\left[n\right]\right|^2 \;+\; (M+1)\frac{2^{-2B}}{12}.$$

A third option is to use the z-transform based approach to find σ_f^2 as

$$\sigma_f^2 = N\frac{2^{-2B}}{12}\gamma_{ff}(0) + (M+1)\frac{2^{-2B}}{12}.$$

Comparison of Direct Form I and Direct Form II

The question of which Form has the least average noise power in the output due to round-off error depends on the location of the poles and zeros.

Reduction in Quantization Noise

If a double-length adder (having 2B+1 or 2B+2 bits) is used to accumulate sums of products, and if double length registers are used to stored the output of delays (z^{-1} units), then the effect of round-off can be represented for Direct Form I as

$$\hat{y}[n] = Q\left[\sum_{k=1}^{N} a_k \hat{y}[n-k] + \sum_{k=0}^{M} b_k x[n-k]\right].$$

The overall effect is to replace **M+1+N** sources of round-off noise with a single source. This reduces the average power of noise in the output by as factor of **1/ (M+1+N).**

Using a double-length adder/accumulator in a Direct Form II implementation changes the equations to

$$\hat{w}[n] = Q\left[\sum_{k=1}^{N} a_k \hat{w}[n-k] + x[n]\right]$$

(equation 6.117a)

and

$$\hat{y}[n] = Q\left[\sum_{k=0}^{M} b_k \hat{w}[n-k]\right].$$

(equation 6.117b)

Scaling in Fixed-Point Implementations of IIR Systems

To prevent overflow in fixed point implementations, it is often necessary to scale the signal as it passes through the system.

Normally, scaling is based on requiring the signal at each critical node in the implementation to have magnitude less than one. In this way, the two's complement format is assumed to represent a proper fraction.

Let $w_k(n)$ represent the value of the signal at node k within the implementation structure and let $h_k(m)$ denote the unit sample response of that part of the system between the input and the node k.

Then

$$|w_k(n)| = \left| \sum_{m=\infty}^{\infty} x(n-m) h_k(m) \right| .$$

If $|\, x(n) \,| \leq x_{max}$ for all n, then $w_k(n)$ is bounded by

$$|w_k(n)| \leq x_{max} \sum_{m=\infty}^{\infty} |\, h_k(m) \,| .$$

Therefore, a sufficient condition for preventing $|w_k(n)|$ from exceeding a value of 1 is

$$x_{max} \leq \frac{1}{\sum_{m=\infty}^{\infty} |\, h_k(m) \,|} .$$

If x_{max} does not satisfy the above inequality, we can multiply the input x(n) by a scale factor to ensure that the signal at all nodes and at the output satisfies this requirement. That is, we choose the scale factor s so that

$$sx_{max} \leq \frac{1}{\max_k \left[\sum_{m=\infty}^{\infty} |h_k(m)| \right]} .$$

The above approach is often overly conservative. (And multiplying by a scale factor that is smaller than necessary reduces the ratio of signal-to-roundoff noise.)

A less conservative approach to scaling is based on assuming that the input is a narrowband signal that can be approximated as

$$x(n) = x_{max} \cos \omega_0 n.$$

The signal at the various nodes in the implementation can then be represented as

$$w_k(n) = |H_k(e^{j\omega_0})|\, x_{max} \cos(\omega_0 n + \sphericalangle H_k(e^{j\omega_0})).$$

Overflow is avoided for all sinusoidal frequencies if

$$\max_{k, |\omega| < \pi} |H_k(e^{j\omega})| x_{max} < 1.$$

If this condition is not naturally satisfied, we can force it to be satisfied by multiplying the input by a scale factor s so that

$$sx_{max} \leq \frac{1}{\max_{k, |\omega| < \pi} |H_k(e^{j\omega})|}.$$

A third possible scaling approach is based on the total energy of the input instead of x_{max}. The signal at node k can be expressed as

$$w_k(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W_k(e^{j\omega})e^{j\omega n}d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} H_k(e^{j\omega})X(e^{j\omega})e^{j\omega n}d\omega.$$

Then

$$|\, w_k(n)\, |^2 \;=\; \left|\frac{1}{2\pi}\int_{-\pi}^{\pi} H_k(e^{j\omega})X(e^{j\omega})e^{j\omega n}d\omega\right|^2 .$$

Schwarz's inequality says that for square-integrable complex-valued functions f(x) and g(x) ,

$$\left|\int f(x)g(x)\, dx\right|^2 \le \int |f(x)\,|^2\, dx \bullet \int |g(x)\,|^2\, dx.$$

Applying this inequality to the previous equation and using $f(\omega) = H_k(e^{j\omega})$ and $g(\omega) = X(e^{j\omega})e^{j\omega n}$ provides the following upper bound for $|\, w_k(n)\, |^2$:

$$|\, w_k(n)\, |^2 \;=\; \left|\frac{1}{2\pi}\int_{-\pi}^{\pi} H_k(e^{j\omega})X(e^{j\omega})e^{j\omega n}d\omega\right|^2$$

$$\le \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} |H_k(e^{j\omega})\,|^2\, d\omega\right)\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} |X_k(e^{j\omega})\,|^2\, d\omega\right). \qquad \text{(equation 6.125)}$$

By applying Parseval's rule to both terms on the right-hand side above, we can write the upper bound on $|w_k(n)|^2$ as

$$|w_k(n)|^2 \;\le\; \left(\sum_{n=-\infty}^{\infty} |h_k(n)|^2\right)\left(\sum_{n=-\infty}^{\infty} |x(n)|^2\right).$$

Therefore, to ensure that $|w_k(n)| \le 1$ for all nodes, we can multiply x(n) by the scale factor s where s satisfies

$$s^2\left(\sum_{n=-\infty}^{\infty} |x(n)|^2\right) \;=\; s^2E \;\le\; \frac{1}{\max\limits_{k} \sum\limits_{n=\infty}^{\infty} |h_k(n)|^2}. \qquad \text{(equation 6.126)}$$

The corresponding bound on s is

$$s \;\le\; \frac{1}{\sqrt{\left(\sum\limits_{n=\infty}^{\infty} |x(n)|^2\right) \max\limits_{k} \sum\limits_{n=\infty}^{\infty} |h_k(n)|^2}}.$$

To summarize, the three bounds proposed for scaling are listed below:

First method considered:

$$s \le \frac{1}{\left(x_{max}\right) \max\limits_{k}\left[\sum\limits_{m=\infty}^{\infty} |h_k(m)|\right]}$$

Second method considered:

$$s \le \frac{1}{\left(x_{max}\right) \max\limits_{k,|\omega|<\pi} |H_k(e^{j\omega})|}$$

Third method considered:

$$s \leq \frac{1}{\sqrt{\left(\sum_{n=\infty}^{\infty} | x(n) |^2\right) \max_k \sum_{n=\infty}^{\infty} | h_k(n) |^2}}.$$

The first bound is more conservative that the second bound since

$$\left\{\sum_{n=\infty}^{\infty} | h_k(n) |^2\right\}^{1/2} \leq \max_{k,|\omega|<\pi} |H_k(e^{j\omega})| \leq \sum_{n=\infty}^{\infty} | h_k(n) | . \qquad \text{(equation 6.127)}$$

For most signals, the third bound is the least conservative of the three founds. The third bound is usually the easiest to evaluate analytically, since it can be evaluated using the z-transform method of Appendix A5, finding

$$\sum_{n=\infty}^{\infty} | h(n) |^2 = \gamma_{yy}(0)$$ where $\gamma_{yy}(n)$ is the inverse z-transform of

$$\Gamma_{yy}(z) = \sigma_x^2 H(z) H^*(\frac{1}{z})$$ for the case where $\sigma_x^2 = 1$.

<u>Note 1</u>: If a scale factor s < 1 is used, the signal-to-noise ratio at the system output is reduced.

<u>Note 2</u>: If non-saturation two-complement arithmetic is used, it is not necessary to examine every node in the system for possible overflow. Only nodes that represent "complete sums" (not "partial sums") must be considered.

Non-saturation property of two-complement arithmetic:

If N values are added using (N-1) steps of pair-wise additions, overflow in intermediate sums does not affect the accuracy of the overall sum, if the overall sum can be represented correctly using the available number of bits.

Example (using 4-bit two's complement format):

		partial sum	decimal value of partial sum
0111	(7)		
+ 0111	(7)	1110	(-2)
+ 1010	(-6)	1000	(-8)
+ 1111	(-1)	0111	(7) = correct sum of 7 + 7 - 6 – 1

The figure below shows the nodes for which scaling must be examined for Direct Form I and Direct Form II.

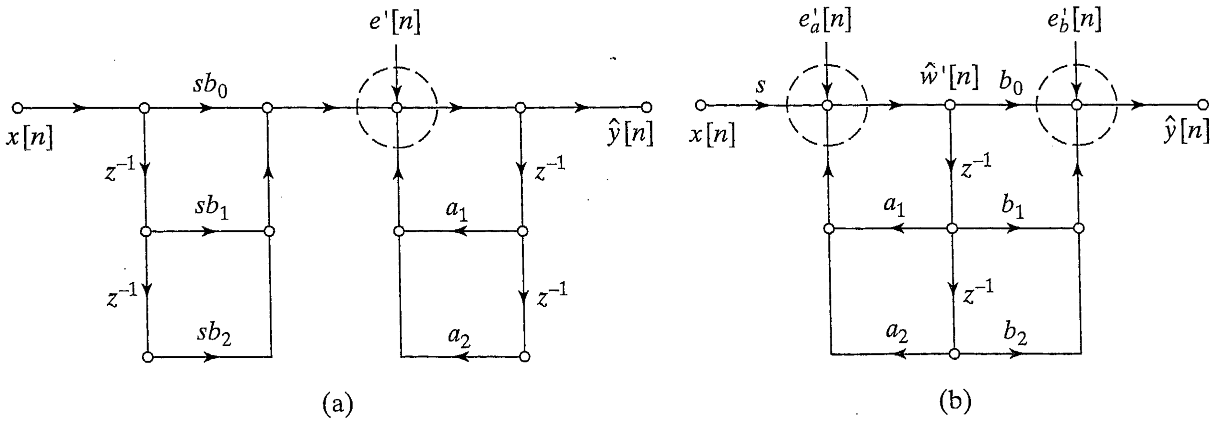


Figure 6.62 Scaling of direct form systems. (a) Direct form I. (b) Direct form II.

If the input x(n) is multiplied by a scale factor s , the overall transfer function for both Direct Form I and Direct Form II is now sH(z) instead of H(z).

Therefore, the "good signal" part of the output is sy(n) instead of y(n).

The round-off noise magnitude is not affected by scaling, since round-off occurs after scaling for both forms of implementation.

Therefore, the ratio of signal power to noise power in the output is multiplied by s^2 which is typically less than 1.

Note that for Direct Form II, the scaling step itself introduces another source of round-off noise, so the noise power in the input is increased from $N\frac{2^{-2B}}{12}$ to $(N+1)\frac{2^{-2B}}{12}$.

In Direct Form I, scaling can be combined with the b_i multipliers, so that no additional sources of round-off error is required to implement scaling.
See figure 6.62 on the previous slide.