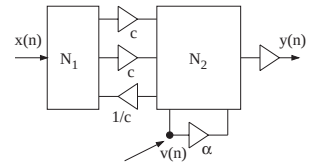
Scaling of Signal Levels

"Every little bit helps"

Measures must also be taken to prevent overflow from occurring too often, since overflows cause large distortion.

The probability of overflow can be reduced by inserting scaling multipliers that only affect signal levels inside the filter and not the poles and zeros.

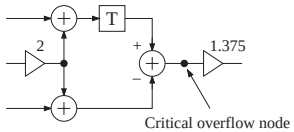


Critical overflow node

Scaling is not required in floating-point arithmetic since the exponent is adjusted so that the mantissa always represents the signal value with full precision.

An important advantage of using two's-complement representation for negative numbers is that temporary overflows in repeated additions can be accepted if the final sum is within the proper signal range.

The incident signal to a multiplier with a noninteger coefficient must not overflow, since that would cause large errors.



Critical overflow node

Safe Scaling

One strategy used to choose the scaling coefficient can be derived in the following way. The signal in the scaling node is given by

$$v(n) = f(n) * x(n)$$

where $f(n)$ is the impulse response from the input of the filter to the critical

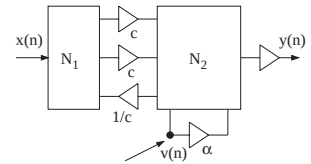
overflow node. The magnitude of the output signal is bounded by

$$|v(n)| = \left| \sum_{k=0}^{\infty} f(k)x(n-k) \right| \le \sum_{k=0}^{\infty} |f(k)||x(n-k)| \le M \sum_{k=0}^{\infty} |f(k)|$$

where

$$|x(n)| \le M$$

In this scaling approach, we insert a scaling multiplier(s), c, between the input and the critical overflow node.



Critical overflow node

The resulting impulse response becomes

$$f'(n) = cf(n)$$

Now, we choose c so that

$$\sum_{n=0}^{\infty} |f'(n)| \le 1$$

The magnitude of the scaled input signal to the multiplier will be equal to, or less than, the magnitude of the input signal of the filter.

The input to the multiplier will never overflow if the input to the filter does not overflow.

This scaling policy is therefore called *safe scaling*.

The safe scaling method is generally too pessimistic since it uses the available signal range inefficiently.

The safe scaling method is suitable for short FIR filters because the probability for overflow is high for a filter with a short impulse response.

It is sometimes argued that parasitic oscillations caused by overflow can not occur if the filter is scaled according to the safe scaling criterion.

However, this is not correct since abnormal signal values can occur due to malfunctioning hardware—for example, in the memory (delay) elements due to external disturbances (e.g., ionic radiation and disturbances from the supply lines).

L_p-Norms

For a sequence, $x(n)$, with Fourier transform $X(e^{j\omega T})$, the L_p-norm is defined as

$$\left\|X(e^{j\omega T})\right\|_p \triangleq \sqrt[p]{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|X(e^{j\omega T})\right|^p d\omega T}$$

for any real $p \geq 1$ such that the integral exists. It can be shown that

$$\|X\|_p \geq \|X\|_q \qquad \text{for } p \geq q$$

L_1-Norm

For $p = 1$ we have

$$\left\|X(e^{j\omega T})\right\|_1 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|X(e^{j\omega T})\right| d\omega T$$

L_2-Norm

$$\left\|X(e^{j\omega T})\right\|_2 = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|X(e^{j\omega T})\right|^2 d\omega T}$$

The L_2-norm is simple to compute by using Parseval's relation which states that the power can be expressed either in the time domain or in the frequency domain. We get from Parseval's relation

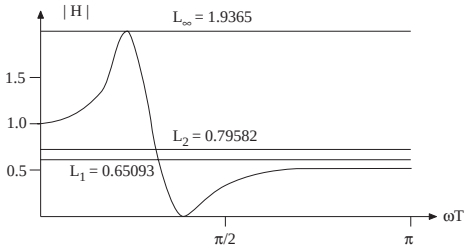
$$\|X\|_2 = \sqrt{\sum_{n=-\infty}^{\infty} x(n)^2}$$

L_∞-Norm

$$\left\|X(e^{j\omega T})\right\|_\infty = \lim_{p\to\infty}\left\|X(e^{j\omega T})\right\|_p = \max_{\omega T}\{\left|X(e^{j\omega T})\right|\}$$

The values of L_p-norms are illustrated below for the second-order section having the transfer function

$$H(z) = \frac{0.5z^2 - 0.4z + 0.5}{z^2 - 1,2z + 0.8}, \quad |z| > \sqrt{0.8}$$



Scaling

The first step in scaling a filter is to determine the appropriate L_p-norm that characterizes the input signal.

Generally, we distinguish between wide-band and narrow-band input signals.

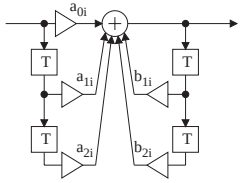
Jackson has derived the following bound on the variance of the signal in the critical node v

$$\sigma_v^2 \leq \left\|F_v(e^{j\omega T})\right\|_{2p}^2 \left\|S_x(e^{j\omega T})\right\|_q, \quad \frac{1}{p}+\frac{1}{q} = 1, \text{ for } p, q \geq 1$$

where $F_v(e^{j\omega T})$ is the frequency response to the critical node v and $S_x(e^{j\omega T})$ is the power spectrum of the input signal.

Scaling of Wide-Band Signals

A wide-band input signal is characterized by the L_∞-norm, $\|S_x\|_\infty$.



$$\sigma_v^2 \le \left\|F_v(e^{j\omega T})\right\|_{2p}^2 \left\|S_x(e^{j\omega T})\right\|_q \quad , \quad \frac{1}{p}+\frac{1}{q} = 1, \text{ for } p, q \ge 1$$

Hence, $q = \infty \Rightarrow p = 1$ and the filter should therefore be scaled such that $\|F_v\|_2 = c$, where $c \le 1$ for all critical nodes.

Scaling of Narrow Band Signals

The L_1-norm, $\|S_x\|_1$ characterizes a sinusoidal or narrow-band signal.

From

$$\sigma_v^2 \le \left\|F_v(e^{j\omega T})\right\|_{2p}^2 \left\|S_x(e^{j\omega T})\right\|_q \quad , \quad \frac{1}{p}+\frac{1}{q} = 1, \text{ for } p, q \ge 1$$

we find, with $q = 1$ and $p = \infty$, that the upper bound for the variance of the signal in the critical node is determined by $\|F_v\|_\infty$.

Thus, the filter should be scaled such that the maximum value of $|F_v|_{max} = \|F_v\|_\infty \le 1$ for all critical nodes.

Round-Off Noise

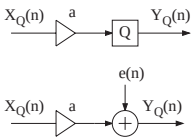
A simple linear model of the quantization operation in fixed-point arithmetic can be used if the signal varies over several quantization levels, from sample to sample, in an irregular way.

The quantization of a product

$$y_Q(n) = [ax_Q(n)]_Q$$

is modeled by an additive error

$$y_Q(n) = ax_Q(n) + e(n)$$



where $e(n)$ is a stochastic process.

Normally, $e(n)$ can be assumed to be white noise and independent of the signal.

The density function for the errors is often approximated by a rectangular function.

However, the density function is a discrete function if both the signal value and the coefficient value are binary.

The difference is only significant if only a few bits are discarded by the quantization. The average value and variance for the noise source are

$$m = \begin{cases} \dfrac{Q_c}{2}Q & \text{rounding} \\ \dfrac{Q_c - 1}{2}Q & \text{truncation} \end{cases}$$

$$\sigma_e^2 = k_e(1 - Q_c^2)\sigma^2$$

where

$$k_e = \begin{cases} 1 & \text{rounding or truncation} \\ 4 - \dfrac{6}{\pi} & \text{magnitude truncation} \end{cases}$$

$\sigma^2 = Q^2/12$, Q is the data quantization step, and Q_c is the coefficient quantization step.

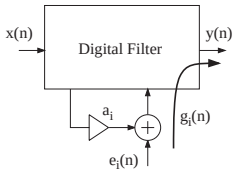
For long coefficient wordlengths—the average value is close to zero for rounding and $-Q/2$ for truncation.

Correction of the average value and variance is only necessary for short coefficient wordlengths, for example, for the scaling coefficients.

A digital filter with M quantization points has a DC offset of

$$m_y = \sum_{i=1}^{M} G_i(1)\, m_i = m \sum_{i=1}^{M} \sum_{n=0}^{\infty} g_i(n)$$

where $g_i(n)$ are the impulse responses measured from the noise sources to the output of the filter.



The noise sources contribute to the noise at the output of the filter. The variance at the output, from source i, is

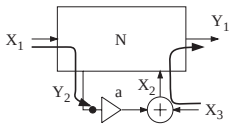
$$\sigma_{y_i}^2 = \sigma_i^2 \sum_{n=0}^{\infty} g_i(n)^2 = \sigma_i^2 \left\| G_i(e^{j\omega T}) \right\|_2^2$$

The variance of the round-off noise at the output is equal to the sum of the contributions from all the uncorrelated noise sources

$$\sigma_{y_{tot}}^2 = \sum_{i=1}^{M} \sigma_{y_i}^2$$

Coefficient Sensitivity

Consider the LSI network which is described by



$$Y_1 = AX_1 + BX_2$$

$$Y_2 = CX_1 + DX_2$$

$$X_2 = aY_2 + X_3$$

We get the transfer function

$$H = \left.\frac{Y_1}{X_1}\right|_{X_3 = 0} = A + \frac{aCB}{1 - aD}$$

The transfer function from the input to the multiplier is

$$F = \left.\frac{Y_2}{X_1}\right|_{X_3 = 0} = \frac{C}{1 - aD}$$

and the transfer function from the output of the multiplier to the output of

the network is

$$G = \left.\frac{Y_1}{X_3}\right|_{X_1 = 0} = \frac{B}{1 - aD}$$

Taking the derivative of the transfer function, H, with respect to the coefficient, a, leads to the main result

$$\frac{\partial H}{\partial a} = \frac{CB}{(1 - aD)^2} = FG$$

where F is the transfer function from the input of the filter to the input of the multiplier and G is the transfer function from the output of the multiplier to the output of the filter.

Sensitivity and Round-Off Noise

Fettweis has shown that coefficient sensitivity and round-off noise are closely related. Jackson has derived the following lower bounds on round-off noise in terms of the sensitivity.

Let F_i be the scaled frequency response from the input of the filter to the input of the multiplier a_i, and G_i the frequency response from the output of the multiplier to the output of the filter. For a scaled filter we have

$\|F_i\|_p = 1$ for all critical nodes $i = 0, 1, \ldots, n$

The round-off noise variance at the output of the filter is

$$\sigma_{y_e}^2 = \sum_{i=0}^{n} \sigma_i^2 \|G_i\|_2^2 = \sum_{i=0}^{n} \sigma_i^2 \|F_i\|_p \|G_i\|_2^2$$

We get, using Hölder's inequality[1], for $p = 2$

1. Hölders inequality: $\|FG\|_1 \le \|F\|_p \|G\|_q \quad , \frac{1}{p} + \frac{1}{q} = 1 \quad , \quad p, q \ge 1$

$$\sigma_{y_e}^2 \ge \sum_{i=0}^{n} \sigma_i^2 \|F_i G_i\|_1^2$$

and

$$\sigma_{y_e}^2 \ge \sum_{i=0}^{n} \sigma_i^2 \left\| \frac{\partial H}{\partial a_i} \right\|_1^2$$

This is a lower bound of the noise variance for a filter scaled for wide-band input signals.

Another lower bound that is valid instead for L_∞-norm scaling

$$\sigma_{y_e}^2 \ge \sum_{i=0}^{n} \sigma_1^2 \left\| \frac{\partial H}{\partial a_i} \right\|_2^2$$

These two bounds are important, since they show that a structure with high sensitivity will always have high round-off noise and requires a longer data wordlength.