Effects of Finite Word length In Digital Filters

Negative Number Representation

Depending on the representation of negative numbers, fixed-point arithmetic can assume three forms

1) Sign-magnitude 2) One's Complement

3) 2's Complement (most common)

In sign-magnitude arithmetic if $|X_n| < 1$ this is stored in binary as

$$X_{n_{sm}} = \underbrace{x_n^0}_{} \cdot \underbrace{x_n^1}_{} \underbrace{x_n^2}_{} \underbrace{x_n^3}_{} \cdots \quad , \quad x_n^k = 0 \text{ or } 1$$

For $X_n \geq 0$ 0

For $X_n < 0$ 1

x_n^k are chosen so that $\quad |X_n| = \sum_{k=1}^{\infty} x_n^k \, 2^{-k}$

The 1's complement representation of $|X_n| < 1$

$$X_n = \pm \cdot x_n^1 \, x_n^2 \, x_n^3 \cdots$$

$$X_{n_{1's}} = \begin{cases} X_n & \text{for } X_n \geq 0 \\ \underbrace{2 - 2^{-B}}_{B+1 \text{ bits with 1's}} - |X_n| & \text{for } X_n \leq 0 \end{cases}$$

B: Word length (No. of bits in the register to the right of the binary point)

$2 - 2^{-B} \equiv$ B+1 locations filled with 1's $\quad 1.111\cdots1$

Thus 1's Complement of a negative number is obtained by representing the number by B+1 bits, including zeros if necessary, and then complementing all bits

e.g. $X_n = -0.11010$

For $B=5$ $\quad X_{n_{1's}} = 1.00101$ $\quad \Rightarrow$ $\begin{array}{r} 1.11111 \\ +\,0.11010 \\ \hline 1.00101 \end{array}$

$B=8$ $\quad X_{n_{1's}} = 1.00101111$

The 2's Complement representation is similar to S-M for $X_n \geq 0$. If $X_n < 0$ then

$$X_{n_{2's}} = \begin{cases} X_n & \text{for } X_n \geq 0 \\ 2 - |X_n| & \text{for } X_n < 0 \end{cases}$$

2's Complement of a negative number can be formed by adding "1" at the least significant position of the 1's Complement

$$X_n = -\sum_{K=1}^{B} x_n^K \, 2^{-K}$$

Start with $\quad 0.\, x_n^1 \, x_n^2 \cdots x_n^B$

1's Complement $\quad 1.\, \bar{x}_n^1 \, \bar{x}_n^2 \cdots \bar{x}_n^B$ $\qquad 2 - 2^{-B} - |X_n|$

Add 2^{-B} $\quad +0.\,0 \;\; 0 \cdots 1$ $\qquad 2 - \cancel{2^{-B}} + 2^{-B} + |X_n| = 2 - |X_n|$

2's Complement

Reason:

$$2-|X_n| = 2 - \sum_{K=1}^{B} x_n^K 2^{-K} = 1 + \underbrace{\sum_{K=1}^{\infty} 2^{-K}}_{2} - \sum_{K=1}^{B} x_n^K 2^{-K}$$

$$= 1 + \sum_{K=1}^{B} \bar{x}_n^K 2^{-K} + \underbrace{\sum_{K=B+1}^{\infty} 2^{-K}}$$

$$= 1 + \sum_{K=1}^{B} \bar{x}_n^K 2^{-K} + 2^{-B} \;//$$

Fact: Can express value represented in 2's Complement in terms of bit values:

$$X_n = -x_n^0 + \sum_{K=1}^{B} x_n^K 2^{-K}, \quad x_n^0 = \text{sign bit}$$

Proof: If $X_n \geq 0$ then $x_n^0 = 0$ the same as S-M rep.

If $X_n < 0$ then store $2-|X_n|$ so that

$$2-|X_n| = \sum_{K=0}^{B} x_n^K 2^{-K}$$

$$|X_n| = 2 - \sum_{K=0}^{B} x_n^K 2^{-K} = \underset{2x_n^0}{2} - x_n^0 - \sum_{K=1}^{B} x_n^K 2^{-K}$$

$$= x_n^0 - \sum_{K=1}^{B} x_n^K 2^{-K}$$

$\therefore$ $X_n < 0$ so $X_n = -x_n^0 + \sum_{K=1}^{B} x_n^K 2^{-K}$

The process

$$\hat{X}_n = Q_B[X_n] = Q_B\left[-x_n^0 + \sum_{K=1}^{\infty} x_n^K 2^{-K}\right] = -x_n^0 + \sum_{K=1}^{B} x_n^K 2^{-K}$$

is quantization.

Arithmetic Operation

a) Addition:

i) 1's Complement : add the 1's complements of two numbers bit by bit. A carry bit at the most significant position (if one) is added at the least significant position (end around carry)

ii) 2's Complement : exactly the same except the carry bit at the most significant bit is ignored

iii) S-M : involve sign checks as well as complementing end around carry

b) Multiplication:

Special algorithms exist for multiplications of two numbers represented by 1's or 2's complements

S-M multiplication : multiply the magnitudes of the two numbers bit by bit and then adjusting the sign bits of the product.

Floating Point Arithmetic

Disadvantages of Fixed point Arithmetic:

(1) The range of numbers that can be handled is small. e.g. in 2's complement the smallest No. is -1 and the largest is $1-2^{-B}$.

(2) The percentage error due to truncation or rounding tends to increase as the magnitude of number is decreased.

These can be solved by using floating-point arithmetic

$$X_n = M \cdot 2^e \qquad e: \text{expanent}$$

e: integer and $\frac{1}{2} \leq M < 1$ $\qquad$ M: mantissa

In floating point arithmetic mantissa can exceed the register length both for multi and additions. But the dynamic range is obsviously much greater than that of the fixed point.

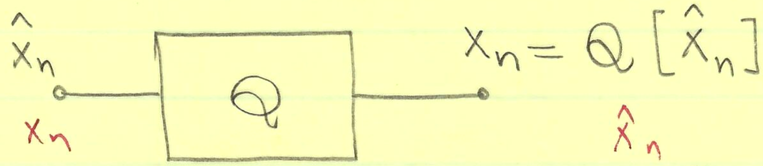
Register for floating point is subdivided into two segments, one for signed mantissa and one for signed exponent.

Disadvantage: Increased cost of hardware and reduced speed of processing

For non-real-time software implementation since neither the cost of hardware nor the speed of processing is a significant factor, floating point arithmetic is preferred.

Quantization:

If the word length of the registers is B, then any number consisting of L bits where L > B must be quantized. This can be accomplished (1) by truncating all the bits that can not be accommodated in the register, and (2) by rounding the number to the nearest machinerepresentable number.



Error produced is $\varepsilon_n = Q[\hat{X}_n] - \hat{X}_n$

Rounding: If B+1 st bit of $\hat{X}_n$ would be "1" then add 2^{-B} and chop off B+1 st, B+2nd ... bits

Truncation: chop off the B+1 st, B+2nd, ... bits

EX $B=2$, Represent $-5/8$

S-M $-5/8$ is represented in ∞ precision as

$$\underbrace{1.10}\underset{B+1}{1}\overset{B+2}{0}00$$

Rounding: $Q_r[-5/8] = 1.11$; value is $-(1/2 + 1/4) = -3/4$
Truncation: $Q_{tr}[-5/8] = 1.10$; value $-(1/2) =$

2's Complement:

$-5/8$ is represented in ∞ precision as $2 - 5/8$: 1.011000

Rounding: 1.10 ; value $-1 + 1/2 = -1/2$
Truncation: 1.01 ; value $-1 + 1/4 = -3/4$

Bounds on $\varepsilon_n = Q[\hat{X}_n] - \hat{X}_n$ for above characteristics

Arithmetic	Error
Sign-magnitude rounding	$\begin{cases} \left(-\frac{2^{-B}}{2}, \frac{2^{-B}}{2}\right] & \hat{X}_n \geq 0 \\ \left[-\frac{2^{-B}}{2}, \frac{2^{-B}}{2}\right) & \hat{X}_n < 0 \end{cases}$
Sign-magnitude Truncation	$\begin{cases} (-2^{-B}, 0] & \hat{X}_n \geq 0 \\ [0, 2^{-B}] & \hat{X}_n < 0 \end{cases}$
2's Complement rounding	$\left(-\frac{2^{-B}}{2}, \frac{2^{-B}}{2}\right] \quad \forall \hat{X}_n$
2's Complement truncation	$(-2^{-B}, 0] \quad \forall \hat{X}_n$

Sources of Error in Digital Filtering

1) Input Quantization (at A/D)
2) Coefficient Quantization (due to this quantization the frequency response of the resulting filter may differ significantly from the desired response).

3) Product Quantization; arises at the output of multipliers which causes roundoff noise

Truncation

1- In SM representation truncation reduces the magnitude of the number. Thus a negative number becomes smaller in magnitude i.e.

$$0 \leq \varepsilon_n \leq (2^{-B} - 2^{-B_1}) \qquad , B < B_1$$

B : No. of bits to the right of point after truncation.
B_1 : No. of bits to the right of point before truncation.

2- For 2's Complement negative number

$$X_{n_{2,s}} = 1.\,x_n^1 x_n^2 \ldots x_n^{B_1}$$

$$|X_n| = 2 - \sum_{i=1}^{B_1} x_n^i \, 2^{-i}$$

$B_1 = \infty$ when Converting an infinite precision number

Truncation to B bits gives

$$|X_n|_{tr} = 2 - \sum_{i=1}^{B} x_n^i \, 2^{-i}$$

$$\Delta |X_n| = \sum_{i=B+1}^{B_1} x_n^i \, 2^{-i} \qquad \text{or} \qquad -(2^{-B} - 2^{-B_1}) \leq \varepsilon_n \leq 0$$

i.e. truncation in 2's complement representation increases the magnitude of the negative number.

3- For 1's complement

$$X_n = 2 - 2^{-B_1} - \sum_{i=1}^{B_1} x_n^i \, 2^{-i}$$

$$X_{n\,tr} = 2 - 2^{-B} - \sum_{i=1}^{B} x_n^i \, 2^{-i}$$

Thus

$$\Delta X_n = \sum_{i=B+1}^{B_1} x_n^i \, 2^{-i} - (2^{-B} - 2^{-B_1})$$

i.e truncation in 1's complement decrease the magnitude of the negative number, the truncation error is positive i.e.

$$0 \leq \varepsilon_n \leq (2^{-B} - 2^{-B_1})$$

Rounding

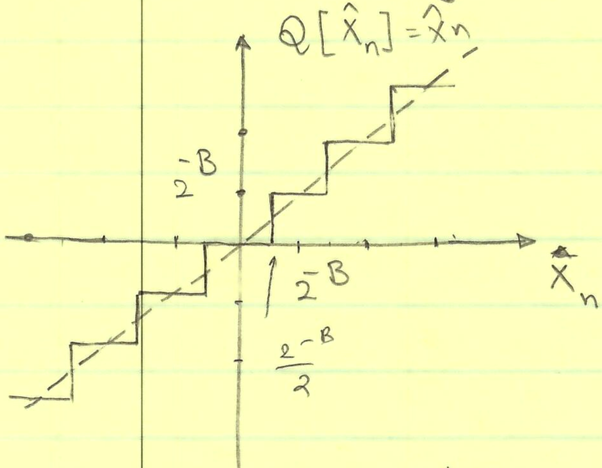
Let B be the number of bits to the right of the point after rounding. The values are quantized in steps 2^{-B} i.e. the smallest nonzero difference between two numbers is 2^{-B}. The max. error has a magnitude of $\frac{1}{2} 2^{-B}$ i.e.

$$-\frac{1}{2}(2^{-B} - 2^{-B_1}) < \varepsilon_n \leq \frac{1}{2}(2^{-B} - 2^{-B_1})$$

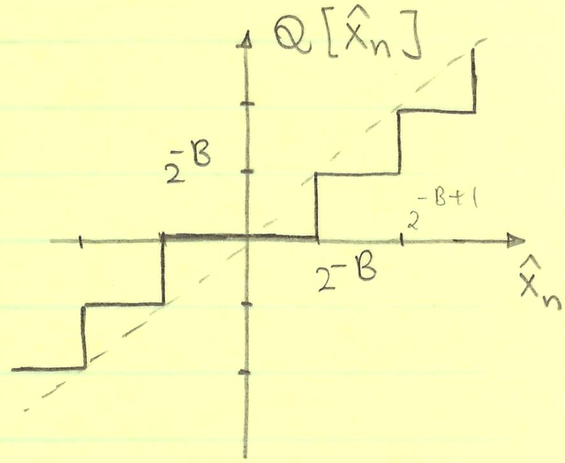
Because rounding is based upon the magnitude of the number, the error is independent of the ways in which negative number are represented. Generally $2^{-B_1} \ll 2^{-B}$ and thus 2^{-B_1} can be neglected.

Quantizer Characteristic

S-M Rounding

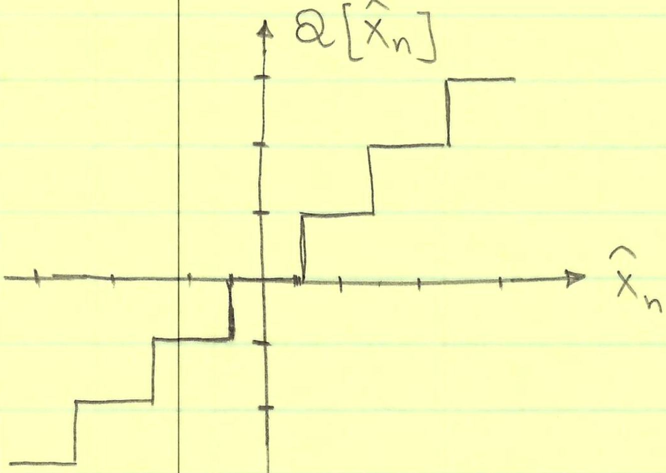


S-M Truncation

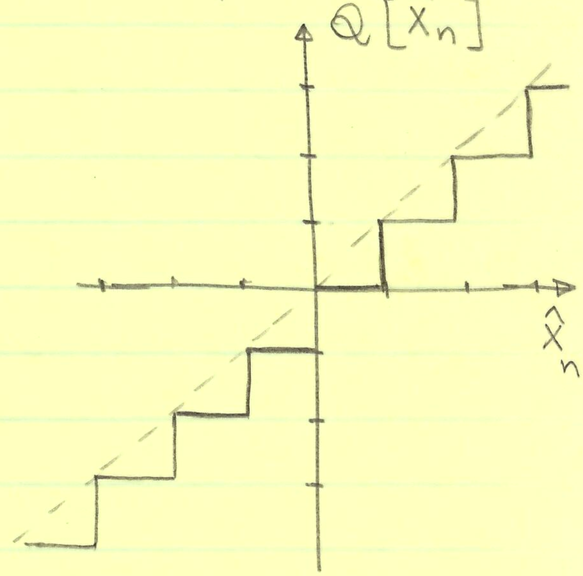


Assumes that round up values at midpoint, operate indep. of sign.

2's Complement Rounding



2's Complement truncation



As can be seen quantization is a nonlinear operation.

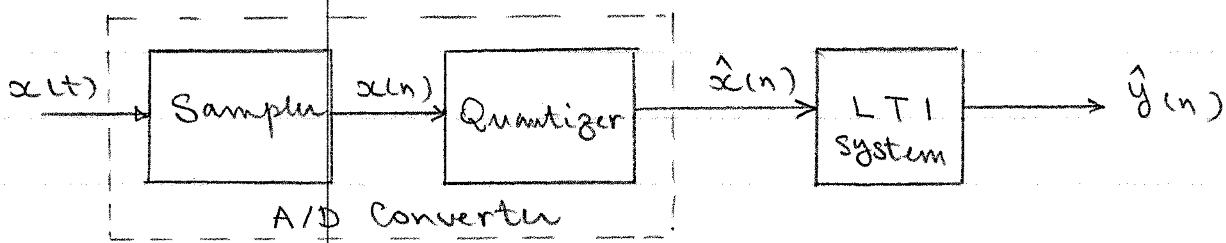
The other Sources are

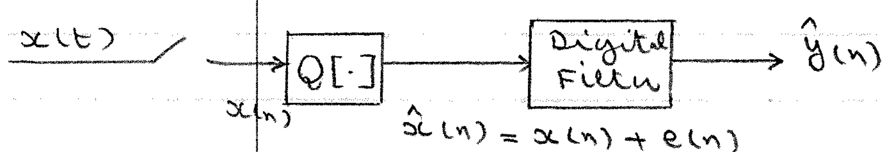
4) Adder overflow.

5) Zero input limit cycles.

(1) Input Quantization

Consider the following system with A/D





$e(n)$: Quantization error

It is assumed that $e(n)$ is

a) a white process

b) uncorrelated with $\{x(n)\}$

c) a stationary process

d) random process with uniform PDF.

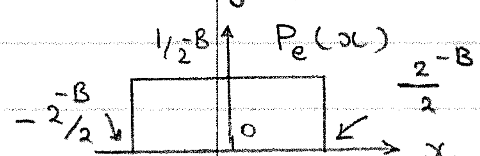
For $x(t) = u_s(t)$, obviously condition (a) is not valid but for complex signals such as speech this condition is valid. In SM and 1's complement representations condition (b) is not valid. At the output of the quantizer we have

$$\hat{x}(n) = x(n) + e(n)$$

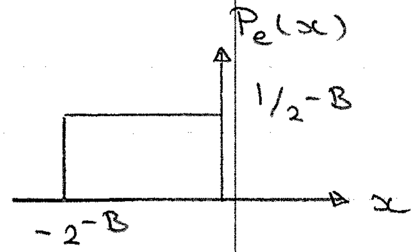
or

$$e(n) = Q[x(n)] - x(n)$$

This input quantization error is uniformly distributed in the ranges shown below



For Rounding (2's complement)



For Truncation (2's Complement)

Thus the mean and variance of $\{e(n)\}$ for rounding are

$$\mu_e = E[e(n)] = \int_{-\infty}^{\infty} x P_e(x)\,dx = \int_{-\frac{2^{-B}}{2}}^{2^{-B}/2} x \cdot \frac{1}{2^{-B}}\,dx$$

$$= 0$$

$$\sigma_e^2 = E[e^2(n)] = \int_{-\frac{2^{-B}}{2}}^{2^{-B}/2} x^2 \frac{1}{2^{-B}}\,dx = \frac{2^{-2B}}{12}$$

Thus

$$r_e(m) = E[e(n)\,e(n-m)] = \frac{2^{-2B}}{12}\,\delta(m)$$

The effect of this error at the output of the digital system is reflected by $\varepsilon(n)$ i.e.

$$\hat{y}(n) = y(n) + \varepsilon(n)$$

where

$$\varepsilon(n) = \sum_{k=0}^{\infty} h(k)\,e(n-k)$$

$h(k)$: impulse response of the LTI system.

The mean of $\varepsilon(n)$ is

$$\mu_\varepsilon = E[\varepsilon(n)] = \left(\sum_{k=0}^{\infty} h(k)\right)\mu_e = 0$$

and

$$\sigma_\varepsilon^2 = E[\varepsilon^2(n)] = E\left[\sum_k h(k)\,e(n-k) \sum_e h(e)\,e(n-e)\right]$$

$$= \sum_k \sum_e h(k)\,h(e)\,E[e(n-k)\,e(n-e)]$$

$$= \frac{2^{-2B}}{12} \sum_{k=0}^{\infty} h^2(k) = \sigma_e^2 \sum_{k=0}^{\infty} h^2(k)$$

In frequency domain

$$S_{\varepsilon}(e^{j\Omega}) = S_e(e^{j\Omega})\,|H(e^{j\Omega})|^2$$

$$= \frac{2^{-2B}}{12}\,|H(e^{j\Omega})|^2$$

Using Parseval's theorem

$$\sigma_{\varepsilon}^2 = \frac{2^{-2B}}{12}\sum_{K=0}^{\infty} h^2(K) = \frac{2^{-2B}}{12}\cdot\frac{1}{2\pi}\int_{-\pi}^{\pi}|H(e^{j\Omega})|^2\,d\Omega$$

(2) Coefficient Quantization

When the coefficients of the transfer function or corresponding difference equation are quantized, the poles and the zeros of the system move to new positions in the z-plane. These changes obviously perturb the frequency response and the resulting system may no longer meet the desired specs. In case of IIR systems they may even become unstable. Let $\{a_K\}$ and $\{b_K\}$ be the ideal infinite precision coefficients, then the ideal transfer function is

$$H(z) = \frac{\sum_{K=0}^{M} b_K z^{-K}}{\sum_{K=0}^{N} a_K z^{-K}}, \quad a_0 = 1$$

If we quantize the coefficients the actual transfer function becomes

$$\hat{H}(z) = \frac{\sum_{K=0}^{M} \hat{b}_K z^{-K}}{1 + \sum_{K=1}^{N} \hat{a}_K z^{-K}}$$

where $\hat{a}_K = a_K + \Delta a_K$, $\quad \hat{b}_K = b_K + \Delta b_K$

$\qquad = Q[a_K] \qquad\qquad\qquad = Q[b_K]$

Δa_K, Δb_K are quantization errors.

To see the effect of quantization on the poles, assume that all poles are distinct, then

$$1 + \sum_{k=1}^{N} a_k z^{-k} = \prod_{j=1}^{N} (1 - z_j z^{-1})$$

where $z = z_j$ is a pole, $j = 1, \ldots N$. Now the poles of $\hat{H}(z)$ will be $z_j + \Delta z_j$, $j = 1, \ldots N$. The error in the location of the ith pole is

$$\Delta z_i = \sum_{k=1}^{N} \frac{\partial z_i}{\partial a_k} \cdot \Delta a_k \quad , \quad i = 1, \ldots N$$

Using the above Eq.

$$\frac{\partial z_i}{\partial a_k} = \frac{z_i^{N-k}}{\prod_{j=1,\, j \neq i}^{N} (z_i - z_j)} \quad , \quad \begin{array}{l} i = 1, 2, \ldots N \\ k = 1, 2, \ldots N \end{array}$$

$\frac{\partial z_i}{\partial a_k}$: Sensitivity of the ith pole to the quantization error in the kth coeff. i.e. a_k.

A similar result can be obtained for the zeros.

<u>Remark</u>

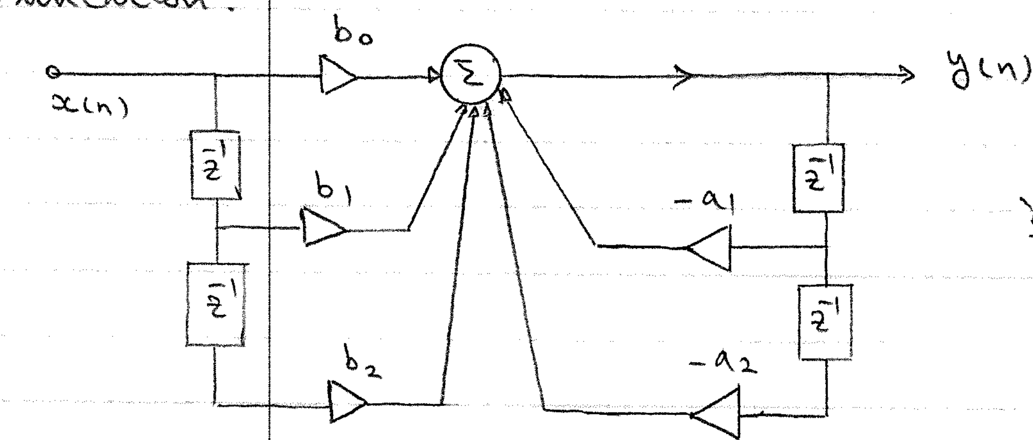
If the poles (or zeros) are tightly clustered, small errors in the denominator (or numerator) coefficients may cause large shifts of the poles (zeros) for the direct form structures. The cascade form is generally much less sensitive to coefficient quantization. This also applies to parallel forms.

(3) Roundoff Noise

Consider the direct form structure for difference equation

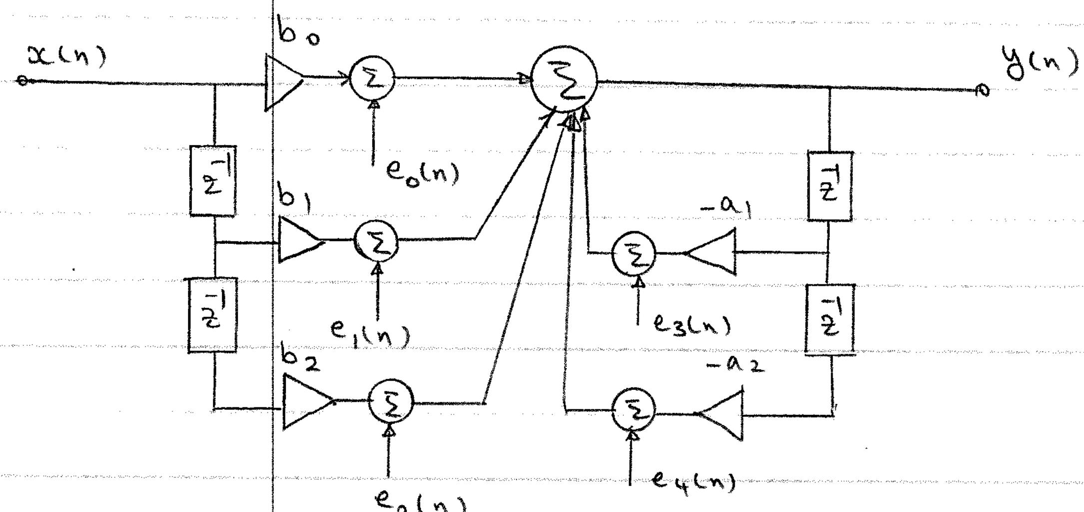
$$y(n) = -\sum_{k=1}^{N} a_k\, y(n-k) + \sum_{k=0}^{M} b_k\, x(n-k)$$

All the signal samples and coefficients are represented by $(B+1)$-bit fixed point binary numbers. Then implementing the difference equation with a $(B+1)$-bit adder, it is necessary to reduce the length of the $(2B+1)$-bit products to $(B+1)$-bits. i.e we discard the last B bits by either rounding or truncation.



$$\frac{Y(z)}{X(z)} = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1+\sum_{k=1}^{N} a_k z^{-k}}$$

$$H_{xy}(z) = \frac{B(z)}{A(z)}$$



The difference equation taking into account the roundoff becomes

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

Let

$$e_i(n) = Q[b_i x(n-i)] - b_i x(n-i) \qquad , i = 0, 1 \ldots M$$

$$e_i(n) = Q[a_i \hat{y}(n-i)] - a_i y(n-i) \qquad , i = M+1, \ldots M+N$$

It is assumed that $e_i(n)$, $\forall i$ satisfies the following properties

(a) $e_i(n)$ is a wide-sense stationary white noise process.

(b) $e_i(n)$ has a uniform distribution over one quantization interval.

(c) $e_i(n)$ is uncorrelated with $e_j(n)$, $i \neq j$, the input $x(n-i)$ to that quantizer and the input to the system.

For $(B+1)$-bit quantization (rounding)

$$-\frac{1}{2} 2^{-B} < e_i(n) \leq \frac{1}{2} 2^{-B}$$

and for 2's complement truncation

$$-2^{-B} < e_i(n) < 0.$$

From assumption (b)

$$\mu_e = 0 \quad \text{and} \quad \sigma_e^2 = \frac{2^{-2B}}{12} \quad \text{for rounding.}$$

and

$$\mu_e = -\frac{2^{-B}}{2} \quad \text{and} \quad \sigma_e^2 = \frac{2^{-2B}}{12} \quad \text{for truncation.}$$

Since $e_i(n)$ are additive, an overall roundoff noise can be generated

$$e(n) = e_0(n) + e_1(n) + \cdots + e_4(n)$$

The variance of $e(n)$ is

$$\sigma_e^2 = \sigma_{e_0}^2 + \cdots + \sigma_{e_4}^2 = 5 \cdot \frac{2^{-2B}}{12}$$

or for the general direct form

$$\sigma_e^2 = (M+N+1)\,\frac{2^{-28}}{12}$$

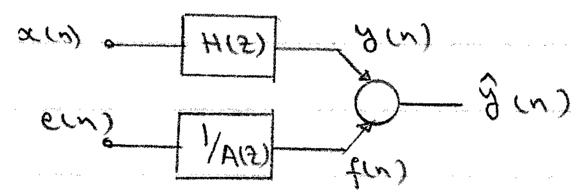
The output taking into account the quantization is

$$\hat{y}(n) = y(n) + f(n)$$

$y(n)$: output of the ideal unquantized system

$f(n)$: output due to input $e(n)$

Note that since $e(n)$ is injected after the zeros and before the poles, thus



$$f(n) = -\sum_{k=1}^{N} a_k f(n-k) + e(n)$$

and

$$\frac{F(z)}{E(z)} = H_{ef}(z) = \frac{1}{A(z)}$$ Transfer function between $e(n)$ and $f(n)$.

The statistics of the output noise $f(n)$ can be obtained using

$$f(n) = \sum_{k=-\infty}^{\infty} h_{ef}(k)\, e(n-k)$$ Convolution sum

Thus

$$\mu_f = \mu_e \sum_{k=-\infty}^{\infty} h_{ef}(k) = \mu_e H_{ef}(e^{j0})$$

The variance is

$$\sigma_f^2 = \sigma_e^2 \sum_{k=-\infty}^{\infty} |h_{ef}(n)|^2$$

$$= \sigma_e^2 \frac{1}{2\pi} \int_{-\pi}^{\pi} |H_{ef}(e^{j\Omega})|^2 \, d\Omega$$

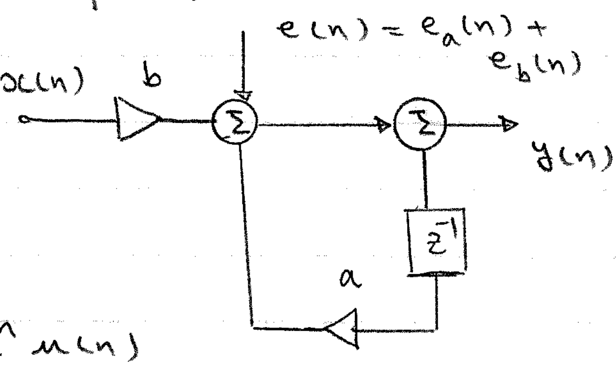
The power spectrum is

$$S_{ff}(e^{j\Omega}) = \sigma_e^2 \, |H_{ef}(e^{j\Omega})|^2 + 2\pi \mu_f^2 \, \delta(\Omega)$$

__Example__

Consider a stable system having transfer function

$$H(z) = \frac{b}{1 - a z^{-1}}$$



From the above analysis

$$H_{ef}(z) = \frac{1}{1 - a z^{-1}} \quad \text{or} \quad h_{ef}(n) = a^n u(n)$$

The SDF is

$$S_{ff}(e^{j\Omega}) = 2 \cdot \frac{2^{-2B}}{12} \, \frac{1}{1 + a^2 - 2a \cos\Omega} + \frac{2\pi \mu_e^2 \, b}{1-a} \, \delta(\Omega)$$

Where $\mu_e = 0$ for rounding

$= -\frac{2^{-B}}{2}$ for 2's complement truncation

$$\sigma_f^2 = 2 \, \frac{2^{-2B}}{12} \sum_{n=0}^{\infty} a^{2n} = 2 \, \frac{2^{-2B}}{12} \, \frac{1}{1 - a^2}$$

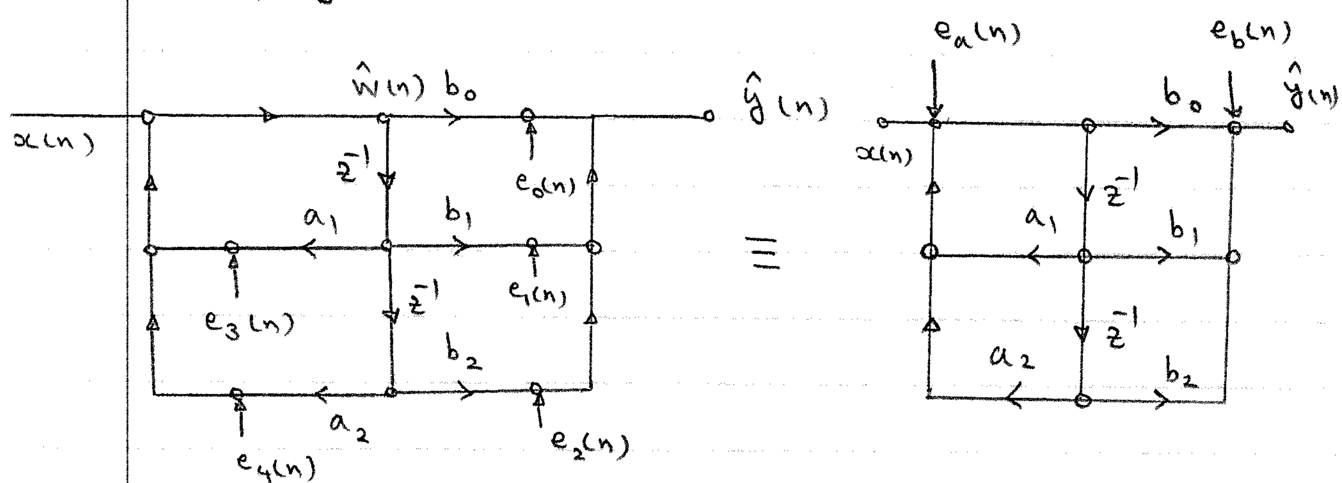
Thus as the pole $z = a$ approaches the unit circle the output error increases in variance.

For the canonical realization the difference equation taking into account the quantization effects is

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

and

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



The statistics for the output noise are

$$\sigma_f^2 = \underbrace{N \frac{2^{-2B}}{12} \sum_{n=-\infty}^{\infty} |h(n)|^2}_{\text{Contribution from } e_a(n) \text{ with transfer function } H(z)} + \underbrace{(M+1)\frac{2^{-2B}}{12}}_{\text{contribution from } e_b(n) \text{ directly added to the output}}$$

or

$$\sigma_f^2 = N \frac{2^{-2B}}{12} \frac{1}{2\pi j} \oint_c H(z) H(z^{-1}) z^{-1} dz + (M+1)\frac{2^{-2B}}{12}$$

The SDF is

$$S_{ff}(e^{j\Omega}) = N \frac{2^{-2B}}{12} |H(e^{j\Omega})|^2 + (M+1)\frac{2^{-2B}}{12}$$

<u>Remark</u>

A Comparison of the results for the direct and canonical realization shows that the choice of filter realization has an important impact on the roundoff error performance of the system. The choice of the best realization which leads to min. output noise variance depends on the particular system in hand and no general conclusion can be made.

(4) Adder overflow and Scaling

The possibility of overflow is another important consideration in the implementation of IIR systems utilizing fixed-point arithmetic. The roundoff error models above assume no overflow at the adders i.e. $|\hat{y}(n)| < 1$. To prevent the overflow we can scale down the input. To see this let $w_K(n)$ be the Kth node variable and $h_K(n)$ be the impulse response from the input $x(n)$ to the node variable $w_K(n)$, then

$$|w_K(n)| = \left| \sum_{m=-\infty}^{\infty} x(n-m)\, h_K(m) \right|$$

$$\leqslant x_{max} \sum_{m=-\infty}^{\infty} |h_K(m)|$$

A sufficient condition that $|w_K(n)| < 1$ is

$$x_{max} < \frac{1}{\sum_{m=-\infty}^{\infty} |h_K(m)|}, \quad \forall K .$$

If x_{max} does not satisfy the above inequality, then we can multiply $x(n)$ by a scaling multiplier s at the input to the system so that $s\,x_{max}$ satisfies the above inequality for all the nodes. i.e.

$$s\,x_{max} < \frac{1}{\max_K \left[\sum_{m=-\infty}^{\infty} |h_K(m)| \right]} \qquad \text{upper bound}$$

However, this is a very conservative scaling of the input for most signals. The signal levels are quite restricted, leaving much of the filter dynamic range unused.

l_p - Norm Scaling

If $x(n)$ is deterministic sequence with z-transform $X(z)$ then

$$w_K(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H_K(e^{j\Omega})\, X(e^{j\Omega})\, e^{j\Omega n}\, d\Omega$$

If we define the l_p norm $(p \geq 1)$ of a Fourier transform say $A(e^{j\Omega})$ as

$$\|A\|_p = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} |A(e^{j\Omega})|^p\, d\Omega\right]^{1/p}$$

when the integral is finite, then as $p \to \infty$ the limit is

$$\|A\|_\infty = \max_{-\pi \leq \Omega \leq \pi} |A(e^{j\Omega})|$$

i.e l_∞ norm is the peak value of $|A(e^{j\Omega})|$ over all Ω.

Now if $\|X\|_\infty$ is bounded then

$$|w_K(n)| \leq \|H_K\|_1 \, \|X\|_\infty$$

Similarly if $\|H_K\|_\infty$ is bounded then

$$|w_K(n)| \leq \|H_K\|_\infty \, \|X\|_1$$

Using Schwarz inequality on the 1st equation

$$|w_K(n)|^2 \leq \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} |H_K(e^{j\Omega})|^2 d\Omega\right]\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} |X(e^{j\Omega})|^2 d\Omega\right]$$

or $\quad |w_K(n)| \leq \|H_K\|_2 \, \|X\|_2$

In general it can be shown that

$$|w_K(n)| \leq \|H_K\|_p \, \|X\|_q$$

with $\quad 1/p + 1/q = 1 \quad , \quad p, q \geq 1$

The case $p = q = 2$ corresponds to placing an energy constraint on both input and transfer function. For example let the energy in input be E i.e.

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} |X(e^{j\Omega})|^2 \, d\Omega$$

Thus $\quad \|X\|_2 = \sqrt{E}$

and $\quad |w_k(n)| \leq \sqrt{E} \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\Omega})|^2 \, d\Omega \right]^{1/2}$

As a result the scaling factor is

$$S = 1 \Big/ \sqrt{E} \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\Omega})|^2 \, d\Omega \right]^{1/2}$$

The case $p = \infty$ and $q = 1$, on the other hand, corresponds to bounding the peak spectrum level of $H_K(e^{j\Omega})$. If the input is unit amplitude sinusoid the scaling factor in this case is

$$S = 1 \Big/ \max_{-\pi \leq \Omega \leq \pi} |H_k(e^{j\Omega})|$$

The case $p = 1$, $q = \infty$ corresponds to knowing the peak magnitude of the input spectrum and bounding the L_1 norm of $H_K(e^{j\Omega})$.

5) Zero - Input Limit Cycles

The output of stable linear systems should decay to zero when no excitation is applied to the system. However, owing to the finite word length used in the implementation of IIR filters, a nonzero periodic output is possible under zero-input conditions. Called "limit cycles" these non-linearities may be produced by

(a) internal register overflow

(b) internal product quantization

overflow limit cycles can always be eliminated by using "saturation arithmetic". In this section only the 2nd limit cycle type will be studied.

Example

Consider the 1st order system

$$y(n) = x(n) - 0.5\, y(n-1)$$

Assume SM rounding and $B = 3$, $y(0) = 0.5$ (0.1 in SM) and $x(n) = 0$

Then $y(1) = -(0.5)(0.5) = -0.25 = 1.01$ SM

$y(2) = 0.125 = 0.001$ SM

$y(3) = -0.5 \times 0.125 = -Q[0.0001] = 1.001$ sm

$= -0.125$

$y(4) = 0.125 = 0.001$ SM

$\vdots$

i.e output oscillates from 0.125 to -0.125 (deadbands).

Remark

Limit cycles can only occur if the result of rounding effectively leads to poles on the unit circle.

Consider a 1st order filter given by

$$\hat{y}(n) = -Q[a\,\hat{y}(n-1)] + x(n)$$

By the definition of rounding

$$|Q[a\,\hat{y}(n-1)] - a\,\hat{y}(n-1)| \leq \tfrac{1}{2}(2^{-B})$$

Furthermore, for values of n in the limit cycle

$$|Q[a\,\hat{y}(n-1)]| = |\hat{y}(n-1)|$$

i.e. the effective value of a is 1, corresponding to the pole of the filter being on the unit circle. Thus

$$|\hat{y}(n-1)| - |a\,\hat{y}(n-1)| \leq \tfrac{1}{2}(2^{-B})$$

or $$|\hat{y}(n-1)| \leq \frac{1/2\,(2^{-B})}{1-|a|}$$ Deadband for 1st order filters.

For $\alpha < 0$, the limit cycle is of constant magnitude and sign.
For $\alpha > 0$, the limit cycle is of constant magnitude and alternating sign. As a result of rounding values within the deadband are quantized in steps of 2^{-B}.

A 2nd order demonstrates a larger variety of limit cycle behavior. Consider

$$y(n) = x(n) - a_1\,y(n-1) - a_2\,y(n-2)$$

with $a_1^2 < 4a_2$, the poles are complex conjugate and with $a_2 = 1$, the poles are on the unit circle. Due to product quantization

$$\hat{y}(n) = x(n) - Q[a_1\,\hat{y}(n-1)] - Q[a_2\,\hat{y}(n-2)]$$

Then

$$|Q[a_2\hat{y}(n-2)] - a_2\hat{y}(n-2)| \leq \tfrac{1}{2}(2^{-B})$$

With $x(n)=0$, the poles of the system are on the unit circle if

$$Q[a_2\hat{y}(n-2)] = \hat{y}(n-2)$$

Then $$|\hat{y}(n-2)|\,|1-a_2| \leq \tfrac{1}{2}(2^{-B})$$

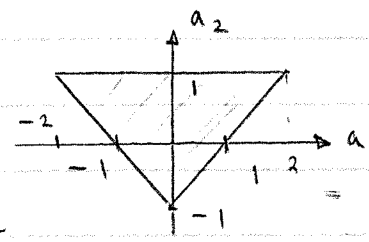
or $$|\hat{y}(n-2)| \leq \frac{\tfrac{1}{2}(2^{-B})}{|1-a_2|}$$

Note that in this case the value of a_1 controls the frequency of oscillation. In a 2nd order system limit cycle can also occur when the effective poles are at $z=+1$ and $z=-1$. The deadband corresponding to this case is bounded by $\dfrac{1}{1-|a_1|-a_2}$.

For a 2nd order system

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}$$

the stability triangle is shown below. The system is stable iff a_1 and a_2 lie in this triangle.



For effective poles at $z=\pm 1$, the condition for a limit cycle of amplitude K to occur is

$$Q[a_2 K] \pm Q[a_1 K] = K$$

The regions for (a_1, a_2) plane within the triangle for various of K can readily be determined.