Finite Word Length Effects in Digital Filters

Introduction:

Digital Signal Processing algorithms are realized either with special purpose digital hardware or as programs for a general purpose digital computer. In both cases, the coefficients are stored in finite length registers. Therefore, coefficients and numbers are quantized by truncation or rounding off when they are stored.

The following errors arise due to quantization of numbers:

1. Input quantization error
2. Product quantization error
3. Coefficient quantization error

1. The conversion of a continuous time input signal into digital value produces an error, which is known as input quantization error. This error arises due to the representation of input signal by a fixed no of digits in A/D conversion process.
2. Product quantization errors arise at the output of a multiplier. Multiplication of a b-bit data with a b-bit coefficient results in a product having 2b bits. Since a b-bit register is used, the multiplier output must be rounded or truncated to b bits, which produces an error.

3. The filter coefficients are computed to infinite precision in theory. If they are quantized, the freq response of the resulting filter may differ from the desired response and sometimes the filter may fail to meet the desired specification.

The other errors arise from quantization are roundoff noise and limit cycle oscillations.

**Topic I : Number Representation**

**2.1) Defn:** Number $N$ can be represented to any desired accuracy by a finite series

$$N = \sum_{i=n}^{m} c_i r^i$$
where \( r \) is called as radius.

\[
\text{Eqn: } r = 10 \Rightarrow \text{ decimal representation } \Rightarrow 0 \text{ to } 9
\]

\[
N = \sum_{i=-3}^{1} c_i \cdot 10^i
\]

\[
30.285 = 3 \times 10^1 + 0 \times 10^0 + 2 \times 10^{-1} + 8 \times 10^{-2} + 5 \times 10^{-3}
\]

\[
N = 30.285
\]

\[
\text{Eqn: } r = 2 \Rightarrow \text{ binary representation } \Rightarrow 0 \text{ to } 1
\]

\[
N = \sum_{i=-3}^{1} c_i \cdot 2^i
\]

\[
110.010 = 1 \times 2^2 + 1 \times 2^1 + 0 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3}
\]

Problems on Number Representation

1) Convert the decimal no 30.285 to binary form.
### Types of Number Representation

There are three common forms to represent the number in digital form. They are:

- **a) Fixed Point Representation**
- **b) Floating Point Representation**
- **c) Block Floating Point Representation**

### 1.2.1) Fixed Point Representation

In fixed point arithmetic, the position of the binary point is fixed.

- Bit to the right gives fractional part
- Bit to the left gives integer part

**Equation:**
- **Binary:** 01.1100
- **Decimal:** 1.75
There are three different forms for fixed point arithmetic:

(i) Sign-magnitude form.
(ii) Ones Complement form.
(iii) Twos Complement form.

(i) Sign Magnitude form

The most significant bit is set to 1 to represent the negative sign. For eg:
-1.75 is represented as 1.1110000 and 1.75 is represented as 0.1110000.

Here 2^b-1 nos can be represented.

(ii) Ones Complement Form

In an ones complement form, the positive number is represented as in the sign magnitude notation.
But the negative no is obtained by complementing all the bits of the positive no.

Example

+0.875 \Rightarrow (0.1110000)_2

(-0.875) = 1.0001111

\text{Complementing each bit:}

\frac{1}{2} (0.875)_{10} = (0.1110000)_{2}
This is same as subtracting the magnitude from \(2^{-b}\), where \(b\) is the number of bits (without sign bit).

\[2^{-b} = \frac{1}{2^b} = 0.000000 \quad \text{or} \quad 1.11111\]

Now subtract \(0.875 = (0.111000)_{2}\)

\[
\begin{array}{c}
1.11111 \\
0.11100 \\
\hline
1.000111
\end{array}
\]

\(= (-0.875)\) in ones complement form.

In one's complement form, the magnitude of the negative number is given by

\[
1 - \frac{b}{2} (i2^{-i} - 2^{-b})
\]

\(i = 1\)

\[\therefore 1 - (2^{-4} + 2^{-5} + 2^{-6}) - 2^{-b} = 0.875\]

In this type of representation, 0 can be represented as 0.000000 and 1111111. So with \(b\) bits \((2^b - 1)\) nos can be represented exactly.

(iii) Two's complement form:
Here, positive numbers are represented as in sign-magnitude and one's complement. The negative no is obtained by complementing all the bits of the positive no and adding one
to the least significant bit

Example

\[(0.875)_{10} = (0.111000)_{2}\]

\[-(0.875)_{10} = (1.001000)_{2}\]

Then it same as subtracting the magnitude from 2

\[2.0 = 10000000\]

\[-0.111000\]

\[1.001000 = (-0.875)_{10}\] in two's complement form

The Magnitude of the negative no to go by \(1 - 2^{-3} = 0.875\)

\[1 - \frac{1}{2^3} e^{2\pi i}\]

\[i = 1\]

2.2a) Addition of two fixed point numbers

The addition of two fixed point nos is simple. The two numbers are added bit by bit starting from right, with carry bit being added to the next bit.

Example: Assume total no of bits \(b+1 = 4\) (including sign bit).
\((0.5)_{10} = (0.100)_2\)

\((0.125)_{10} = \frac{0.001}{0.101}_2 = (0.0101)_2\)

**Sign bit**

Here if two numbers of b bits are added, the sum cannot be represented by b bits, an overflow will occur.

**Example 2**

\((0.5)_{10} \times (0.625)_{10}\), if \(b = 3\), addition of two nos \(b\)

\((0.5)_{10} = (0.100)_2\)

\((0.625)_{10} = (0.101)_2\)

\[1.001_2 = (-0.125)_{10}\] in Sign Magnitude

**Sign bit**

\((1.125)_{10}\)

Thus, here overflow occurs since the result \(\neq 1\)

\(\Rightarrow\) Addition of two nos (fixed point) have overflow.
Problems on Fixed Point Number Representation

1) Subtract 0.25 from 0.5

\[ \begin{align*}
\text{Toa's complement representation of } &\quad 0.25 \\
0.25 &= 0.100 \quad \text{add} \\
-0.25 &= 0.110 \\
(0.100)_{10} &= 0.25 \\
= 0.0110 \\
\end{align*} \]

Carry is generated after the addition. So the result is negative.

Neglect the carry bit to get the result in decimal:

\[ (0.25)_{10} = (0.0100)_{2} \]

2) Subtract 0.5 from 0.25

1.2.2 Multiplication in fixed point arithmetic

In multiplication of two fixed point numbers, first the sign and magnitude components are separated. \( b \times b = 2b \) bits are produced.
Example

\[ 0.1001 \times 0.0011 = 0.00011011 \]

4 bits 4 bits 8 bits

Thus it implies, magnitude of the res are multiplied first, then sign of the product is determined and applied to the result.

In fixed point arithmetic, multiplication of two fractions result in a fraction. For multiplication with fraction, overflow can never occur.

1.a) Floating point representation

In floating point representation, a positive number is represented as \( F = 2^c H \), where \( H \) called mantissa, is a fraction such that \( \frac{1}{2} \leq H \leq 1 \) and \( c \), the exponent can be either positive or negative.

The decimal numbers 4.5, 1.5, 6.5 and 0.625 have floating point representation as

\[ 2^3 \times 0.5625, \quad 2^1 \times 0.75, \quad 2^8 \times 0.0125, \quad 2^9 \times 0.0625 \text{ resp.} \]

Equivalently,

\[
\begin{align*}
2^3 \times 0.5625 &= 2^{11} \times 0.1001 \\
2^1 \times 0.75 &= 2^{10} \times 0.1100 \\
0.5625 &= 0.1001 \\
(0.5625) &= (0.1001)_{10}
\end{align*}
\]
Negative floating point numbers are generally represented by considering the mantissa as a fixed point number. The sign of the floating point number is obtained from the first bit of mantissa.

**1.3.b) Multiplication in Floating Point Arithmetic**

\[
P_1 = 2^{c_1} \times H_1 \times x_1
\]

\[
P_2 = 2^{c_2} \times H_2
\]

\[
P_3 = P_1 \times P_2 = (H_1 \times H_2) \times 2^{c_1 + c_2} \rightarrow (1)
\]

From the eqn (1) if the mantissa are multiplied using fixed point arithmetic.

(ii) Exponents are added.

(iii) Product \((H_1 \times H_2)\) must be in the range of 0.25 to 1.

(iv) Exponent \((c_1 + c_2)\) must be

---

**Example**

\[
(1.5)_{10} = 2^1 \times 0.75 = 2^0.01 \times 0.1100
\]

\[
(1.25)_{10} = 2^1 \times 0.625 = 2^0.01 \times 0.1010
\]

Now, \((1.5)_{10} \times (1.25)_{10} = (2^0.01 \times 0.1100) \times (2^0.01 \times 0.1010)\)

\[
= 2^{0.01} \times 0.110111
\]
I.2.b.2) Addition and Subtraction of Floating Point Numbers

This is different than fixed point representation.

1. Adjust the exponent of small numbers until it matches the exponent of larger numbers.
2. The mantissa are then added or subtracted.
3. The resulting representation is rescaled (so that mantissa lies in the range of 0 to 1).

Example

Add 3.0 and 0.125

3.0 = 2^10 × 0.110000 → ①
0.125 = 2^000 × 0.0001000 → ②

Adjust the exponent of smaller no, both exponents are equal.

0.125 = 2^0 × 0.0000100

① + ② = sum is equal to 2^10 × 0.110010.

I.2.c) Block Floating Point Numbers

A compromise between floating point and fixed point is called Block Floating Point Arithmetic. Set of signals to floating point arithmetic. Set of signals to be handled is divided into blocks. Each block have the same exponent. The memory is caused by...
using one exponent per block and fixed
point arithmetic operation is used in arithmetic block. This representation of numbers is most suitable
in certain FFT flow graphs and in digital audio applications.

1.3) Comparison of Fixed Point and Floating

Point Arithmetic

<table>
<thead>
<tr>
<th>Fixed Point Arithmetic</th>
<th>Floating Point Arithmetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Fast operation</td>
<td>Slow operation</td>
</tr>
</tbody>
</table>
| 2) Relatively economical| More expensive because of
easier hardware           |
| 3) Small dynamic range  | Increased dynamic range   |
| 4) Round off errors occur only in
addition                | Round off errors can occur with both addition
|                        | Multiplication            |
| 5) Overflow occurs in
addition                | Overflow does not arise   |
| 6) Used in small
computers              | Used in larger, general
purpose computers         |
Quantization Noise

The input signal is continuous in time or analog waveform for engineering applications.
The signal is converted to digital by A/D converter.
and the conversion is shown below in Fig1.

Fig1: Block diagram of A/D Converter

→ The signal $x(t)$ is sampled at regular intervals $t = nT$, $n = 0, 1, 2, \ldots$ to create a sequence $x(n)$.
→ Sampler converts $x(t)$ to $x(n)$.
→ $x(n)$ is expressed by a finite number of bits.
→ $x(n)$ is quantized to $Q$ levels.
→ Difference signal $e(n) = x_q(n) - x(n)$ is called quantization noise or A/D conversion noise.

$Q = 2^{b+1}$ levels available for quantizing.
$m$ is $2^b$. The interval $2^{-b}$ successively.

$Q = \frac{2}{2^{b+1}} = 2^{-b} \rightarrow (A)$
where \( q \) is called an quantization step size.

Example \( b = 3, \) \( q = 2^{-3} = 0.125 \)

\[ \Rightarrow \text{The common methods of quantization are} \]

1. Truncation
2. Rounding

II.1) Truncation

Truncation is a process of discarding all bits less significant than least significant bit that is retained.

Example

Truncation of binary nos from 8 bits to 4 bits

\[ \Rightarrow 0.00110011 \quad 8\text{bits} \quad \text{to} \quad 0.0011 \quad 4\text{bits} \]

\[ 1.01001001 \quad 8\text{bits} \quad \text{to} \quad 1.0100 \quad 4\text{bits} \]

Here signal value is approximated by the highest quantization level that is not greater than the signal.

II.2) Rounding

Rounding of a number of \( b \) bits is accomplished by choosing the rounded result as the \( b \) bit number closest to the original number rounded.
Example
0.11010 rounded to three bits as
0.110 or 0.111

(ii.3) Quantization Error ranges

<table>
<thead>
<tr>
<th>Type of quantization</th>
<th>Type of Arithmetic</th>
<th>Fixed Point Number range</th>
<th>Floating Point no-Relative Error range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rounding</td>
<td>Sign Magnitude</td>
<td>$-\frac{2^b}{2} \leq x &lt; \frac{2^b}{2}$</td>
<td>$-2^b &lt; x &lt; 2^b$</td>
</tr>
<tr>
<td></td>
<td>Ones complement</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Twos complement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Truncation</td>
<td>Twos Complement</td>
<td>$-\frac{2^b}{2} &lt; x \leq 0$</td>
<td>$-2^b &lt; x &lt; 2^b$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sign Magnitude</td>
<td>Ones complement</td>
<td>$-2^b &lt; x \leq 0, x &gt; 0$</td>
<td>$-2^b &lt; x &lt; 2^b$</td>
</tr>
<tr>
<td>truncation</td>
<td>Sign-magnitude</td>
<td>$0 \leq x &lt; 2^b, x &lt; 0$</td>
<td></td>
</tr>
</tbody>
</table>

Topic III: Quantization noise power (Types of Quantization Error)

Topic III.1: Input Quantization Error

Quantization error arises when a continuous signal is converted to digital value.

:. quantization error is given by
\[ e(n) = x_q(n) - x(n) \]  \(\Rightarrow (1)\)

Where \(x_q(n)\) → sampled quantized value is \(x(n)\) → sampled unquantized value.

In the way in which \(x(n)\) in quantized, different distributions of quantization noise may be obtained. If rounding of a number is used to get \(x_q(n)\), then the error signal satisfies the relation

\[-\frac{a}{2} \leq e(n) \leq \frac{a}{2} \]  \(\Rightarrow (2)\)

For example

Let \(x(n) = (0.1010)_{10} = (0.101100110\ldots)_{2}\)

After rounding \(x(n)\) to 3 bits, we have

\[ x_q(n) = 0.101 + 0.0 \overset{add}{\rightarrow} 0.110 = (0.75)_{10} \]

Now the error,

\[ e(n) = x_q(n) - x(n) = 0.05 \]

which satisfies the inequality

\[ \Rightarrow \text{The probability density function } p(e) \text{ for roundoff error and quantization characters with rounding is shown in fig below.} \]
Fig 1. a) Quantizer characteristics with rounding

The other type of quantization can be obtained by truncation. In two’s complement truncation, the error $e(n)$ is always negative and satisfies the inequality $-q < e(n) < 0$. Corresponding to quantizer in Fig below.

(a) Quantizer characteristics with two's complement

(b) Probability density function of truncation error.
In digital processing of analog signals, the quantization error is commonly viewed as an additive noise signal that:

\[ x_q(n) = x(n) + e(n) \]

- **Fig 2: Quantization Noise Model**

The above figure says about A/D converter and output is sum of input signal \( x(n) \) and the error signal \( e(n) \).

- If rounding is used for quantization then the quantization error \( e(n) = x_q(n) - x(n) \) is bounded by \(-q/2 \leq e(n) \leq q/2\).

- The analog to digital conversion error \( e(n) \) has the following properties:
  - Error sequence \( e(n) \) is a sample sequence of a stationary random process.
b) Error sequence is uncorrelated with \( x(n) \) and other signals in the system.

c) Error is a white noise process with uniform amplitude probability distribution over the range of quantization error.

In case of rounding the error is 0 or \( \frac{\pm 1}{2} \) with equal probability. The variance of the error is given by

\[
\sigma_e^2 = E[e^2(n)] - E^2[e(n)] \quad (4)
\]

where \( E[e^2(n)] \) is the average of \( e^2(n) \) and \( E[e(n)] \) is the mean value of \( e(n) \).

From Eqn (4) \( E(e^2(n)) \) to find \( \sigma_e^2 \)

\[
E[e^2(n)] = \int_{-\infty}^{\infty} e^2(n) p(e) \, de.
\]

\[
p(e) = \frac{1}{q} \text{ for } -q/2 \leq e(n) \leq q/2
\]

from the fig (b)

\[
E(e^2(n)) = \frac{q/2}{-q/2} \int_{-q/2}^{q/2} e^2(n) \, de - \frac{q^2}{12} \quad \rightarrow (5)
\]

\[
E[e^2(n)] = \int_{-q/2}^{q/2} e^2(n) \, de = \frac{1}{q} \int_{-q/2}^{q/2} e^2(n) \, de \quad \rightarrow (5)
\]
\[
\sigma_e^2 = \frac{1}{q^2} \int_{-q/2}^{q/2} e^2(n) \, dn - \frac{q^2}{2} = \sigma_e^2 - \frac{q^2}{2} \rightarrow (6)
\]

Sub Eqn (5) in (6)

\[
\sigma_e^2 = \frac{(a-b)^2}{12} = \frac{q^2}{12} \rightarrow (7)
\]

In case of two's complement truncation, the mean value of \( e(n) \) lies between 0 and \(-q/2\), having mean value of \(-q/2\). The variance or power of the error \( e(n) \) is given by

\[
\sigma_e^2 = \frac{1}{q^2} \int_{-q/2}^{q/2} e^2(n) \, dn - \left[ \frac{-q/2}{2} \right]^2 \int_{-q/2}^{q/2} e^2(n) \, dn
\]

\[= q^2/12 - q^2/12 = \frac{q^2}{12} \rightarrow (8)\]

We have \( \sigma_e^2 = \frac{q^2}{12} \rightarrow (9)\)

In both cases, the value \( \sigma_e^2 = \frac{q^2}{12} \) which is also known as the steady state noise power due to input quantization.
Signal to noise ratio for rounding is

\[ \frac{\sigma_n^2}{\sigma_x^2} = \frac{\sigma_x^2}{2 \cdot 2^b \cdot 1/2} = 12 \cdot \left(2^2 \cdot \sigma_x^2 \right) \]

where \( \sigma_x^2 \) is variance
\( k(n) \) is input signal

This can be given by dB

\[ 10 \log \frac{\sigma_n^2}{\sigma_x^2} = 6 \cdot 0.2 \cdot b + 10 \cdot 1.9 + 10 \log \sigma_x^2 \]

3) Steady State Output Noise Power

Due to A/D conversion noise, one can represent the quantized input to a digital system with impulse response \( h(n) \) as shown in fig. below.

![Diagram](image)

Fig 3: Representation of A/D Conversion Noise

Let \( e(n) \) be the output noise due to quantization of the input.
\[ e(n) = e(n) * h(n) \]

\[ e(n) = \sum_{k=0}^{\infty} h(k^n) e(n-k) \]

The variance of the sum of independent random variables is the sum of their variances. If the quantization errors are assumed to be independent at different sampling instances, then the variance of the output (steady state output noise power)

\[ \sigma_{e^2}(n) = \sigma_{e^2} \sum_{k=0}^{\infty} h^2(k^n) \]

To find the steady state variance, extend the limit \( k \) up to infinity.

\[ \sigma_{e^2} = \sigma_{e^2} \sum_{n=0}^{\infty} h^2(n) \]

Using Parseval's theorem, the steady state output noise variance due to the quantization error is given by

\[ \sigma_{e^2} = \sigma_{e^2} \sum_{n=0}^{\infty} h^2(n) = \frac{\sigma_{e^2}^2}{2\pi j} \int_{-\infty}^{\infty} H(z) H(z^{-1}) z^{-1}dz \]

Condition: 
Closed contour integration around unit circle \(|z|=1\) (only poles that lie inside the unit circle are evaluated using residue theorem)
1.2) Product Quantization Error

In fixed point arithmetic, the product of two b-bit numbers results in numbers 2b bits long. In digital signal processing applications, it is necessary to round this product to a b-bit number, which produces an error known as product quantization error or product roundoff noise.

Fixed point roundoff noise model for multiplication is shown in Fig. 1.

\[ y(n) = ax_q(n) + e(n) \]

**Fig. 1:** Fixed point roundoff noise model

The roundoff noise is a zero mean random variable and variance is \( \sigma_e^2 = \frac{2^b}{12} \), where \( b \) is the number of bits to represent the variables.

To model the effects of rounding due to multiplication in digital filters, certain assumptions are made:

1. For any \( n \), error sequence \( e(n) \) is uniformly distributed over the range \(-\frac{1}{2} \) and \( \frac{1}{2} \).

This implies the mean value of \( e(n) \) is 0.
and its variance is \( \sigma^2 = \frac{\sigma^2 - b}{12} \)

2. Error sequence \( e(n) \) is a stationary white noise sequence.

3. Error sequence \( e(n) \) is uncorrelated with signal sequence \( x(n) \). Power density spectrum of sequence \( x(n) \) is \( \frac{2-2b}{12} \).

Example 1

Consider an first order quantization noise model in Fig. 5.

\[
y(n) = 2y(n-1) + x(n) + e(n)
\]

By Fig. 5, the eqn can be represented as:

Here in the Fig, finite precision multiplier has been replaced by an ideal multiplier and an additional round off noise \( e(n) \).

Example 2

Let us consider second order quantization noise model in Fig. 6 with five noise sources.
Figure 6: Second order filter with five noise sources

Here, since noise sources are added at the same point in the filter, all these sources can be replaced by a single noise source with

\[ e(n) = e_1(n) + e_2(n) + \ldots + e_5(n) \]

zero mean and variance

\[ \sigma^2 = \sigma_0^2 + \sigma_1^2 + \ldots + \sigma_5^2 \]

below.

Figure 7: Single noise source quantization model
Fig 6 can be realized as cascade form of two first order s tim e as in Fig 8.

Fig 8: Quantization noise model of a cascade system.

In the above figure, there are five noise sources and they are added at different transfer functions.

The total variance can be obtained by adding the individual variances of the noise sources.

Response from the noise source can be obtained by convolution

\[ E_k(n) = \sum_{m=0}^{n} h_k(m) e_k(n-m) \]  

where \( h_k(n) \) is the impulse response of the filter.

The variance of \( E_k(n) \) is
\[ \sigma_{ok}^2 = E \left[ \sum_{m=0}^{n} \frac{1}{2} h_k(m) e_k(m+n) \cdot \frac{1}{2} h_k(l) e_k(n-l) \right] \]

\[ = \frac{n}{2} \sum_{m=0}^{n} h_k(m) h_k(l) E \left[ e_k(n-m) e_k(n-l) \right] \]

\[ = \frac{n}{2} \sum_{m=0}^{n} h_k(m) h_k(l) \delta(l-m) \delta(l-m) \]

\[ \delta(l-m) = 1 \text{ for } l=m \]

\[ = 0 \text{ for } l \neq m \]

\[ \sigma_{ok}^2 = \sigma_{e2}^2 \sum_{m=0}^{n} h_k^2(m) \]

Where \[ \sigma_{e2}^2 = \frac{a-2b}{12} \]

For all better the impulse response approach to zero as \( m \) tends to infinity. Then the steady state noise variance can be obtained by tending \( m \) to infinity.

\[ \sigma_{ok}^2 = \sigma_{e2}^2 \sum_{m=0}^{\infty} h_k^2(m) \]

.. Total steady state noise variance \( \sigma_{e2}^2 \) is

\[ \sigma_{e2} = \sum_{k} \sigma_{ok}^2 \]
Eqn (8) can be written as

\[ e_{ok}^2 = \sigma_e^2 \frac{1}{2\pi j} \int_{C} H_k(z) H_k(z^{-1}) z^{-1} dz \]

where \( H_k(z) \) is defined as the noise transfer function.

### Coefficient Quantization Error

Thus, in the design of digital filters, the coefficients are evaluated with infinite precision. But when they are quantized, the frequency response of the actual filter deviates from that which would have been obtained with an infinite word length representation, and the filter may actually fail to meet desired specifications.

If the poles of the desired filter are close to the unit circle, then those of the filter with quantized coefficients may lie just outside the unit circle leading to instability.
Problems on Quantization Noise Power

(Quantization Error)

1. Consider the transfer function \( H(z) = H_1(z)H_2(z) \)

where

\[
H_1(z) = \frac{1}{1-a_1z^{-1}} \quad \text{and} \quad H_2(z) = \frac{1}{1-a_2z^{-1}}
\]

Find the output roundoff noise power.

Assume \( a_1 = 0.5 \) and \( a_2 = 0.6 \) and find output roundoff noise power.

The roundoff noise model for \( H(z) = H_1(z)H_2(z) \) is given by

\[
\text{Fig. 1: Roundoff noise model}
\]

From the realization we can find that the noise transfer function seen by new source \( e_1(n) \) is \( H(z) \), where
\[ H(z) = \frac{1}{(1-a_1 z^{-1} - a_2 z^{-2})} \rightarrow 0 \]

Whereas the noise transfer function seen by \( e_2(n) \) is

\[ H_2(z) = \frac{1}{(1-a_2 z^{-1})} \]

The total steady state noise variance can be obtained from eqn \((8)\)

\[ \sigma_0^2 = \sigma_{e_1}^2 + \sigma_{e_2}^2 \]

From eqn \((7)\)

\[ \sigma_{e_1}^2 = \frac{1}{2\pi j} \oint \frac{1}{H(z) H(z^{-1})} \frac{1}{z} \frac{1}{1-a_1 z^{-1}} \frac{1}{1-a_2 z^{-2}} \frac{1}{z^{-1}} \frac{1}{z^{-2}} \frac{dz}{z} \]

\[ = \sigma_e^2 \left[ \text{residue of } H(z) H(z^{-1}) z^{-1} \right] \]

At poles \( z = a_1, z = a_2, z = \frac{1}{a_1}, z = \frac{1}{a_2} \)

If \( a_1 \times a_2 \) are less than 1 then the poles

\[ z = \frac{1}{a_1} \times \frac{z = \frac{1}{a_2}}{z \text{ lie outside of the circle } \left| z \right| = 1} \]

so \( H(z) H(z^{-1}) z^{-1} \) at \( z = \frac{1}{a_1}, z = \frac{1}{a_2} \) are zero.
\[
\sigma_{01}^2 = \int \text{residue of } H(z)H(z^{-1})z^{-1}dz \\
\text{at } z = a_1, a_2,
\]

\[
= \left[ \frac{z^{-1}}{(1-a_1 z^{-1})(1-a_2 z^{-1})} \right] \left[ \frac{(1-a_1 z^{-1})}{(1-a_2 z^{-1})} \right] \left[ \frac{(1-a_2 z^{-1})}{(1-a_1 z^{-1})} \right] z^{-1} + \left[ \frac{z^{-1}}{(1-a_1 z^{-1})(1-a_2 z^{-1})} \right] \left[ \frac{(1-a_2 z^{-1})}{(1-a_1 z^{-1})} \right] z^{-1}
\]

\[
= \sigma_e^2 \left[ \frac{1}{(1-a_1 a_2)(1-a_2 a_1)} + \frac{1}{(1-a_1 a_2)} \right] + \frac{1}{a_1 - a_2}
\]

\[
\sigma_{01}^2 = \sigma_e^2 \left[ \frac{a_1}{a_1 - a_2} \frac{1}{1-a_1^2} + \frac{a_2}{a_2 - a_1} \frac{1}{1-a_2^2} \right]
\]

In the same way,

\[
\sigma_{02}^2 = \sigma_e^2 \left[ \frac{H(z)}{H(z)} \right] \frac{1}{1-a_2 z^{-1}} \frac{1}{1-a_2 z^{-1}} z^{-1} dz
\]
\[ \sigma_e^2 = \sigma_e^2 \left[ \frac{1}{1-a_2^2} + \frac{a_1}{a_1-a_2} \times \frac{1}{1-a_1^2} + \frac{a_2}{a_2-a_1} \times \frac{1}{1-a_2^2} \right] \]

\[ \implies \sigma_e^2 = \sigma_e^2 \left[ \frac{1}{1-a_2^2} + \frac{a_1 (1-a_2^2) - a_2 (1-a_2)}{(1-a_1^2)(1-a_2^2)(1-a_1-a_2)} \right] \]

\[ \sigma_e^2 = \sigma_e^2 \left[ \frac{1}{1-a_2^2} + \frac{(a_1-a_2)(1+a_1a_2)}{(1-a_1^2)(1-a_2^2)(1-a_1)(1-a_2)} \right] \]

\[ \sigma_e^2 = \frac{\sigma_e^2}{12} \left[ \frac{1}{1-a_2^2} + \frac{(1+a_1a_2)}{(1-a_1^2)(1-a_2^2)(1-a_1)(1-a_2)} \right] \]

The steady state noise power for \( a_1 = 0.5 \), \( a_2 = 0.6 \)

\[ \sigma_e^2 = \frac{\sigma_e^2}{12} \left[ \frac{1}{1-(0.6)^2} + \frac{1+(0.5)(0.6)}{(1-0.5)^2(1-0.6)^2(1-0.6)(0.5)} \right] \]
\[ f_0^2 = \frac{g_1 - 2b}{12} \] 

2) Draw the quantization noise model for a second order filter \[ H(z) = \frac{1}{1 - 2r \cos \theta z^{-1} + r^2 z^{-2}} \] and find the steady state output noise variance.

The quantization noise model is shown below.

Fig: Quantization noise model for a second order filter both noise sources give the transfer function as
\[ H(z) = \frac{1}{1 - 2\cos(\theta) z^{-1} + \sin^2(\theta) z^{-2}} \]

The impulse response of the transfer function is given by
\[ h(n) = \gamma(n) \frac{\sin(\theta n) e^{j\theta n}}{\sin(\theta)} \]

Now steady state output noise variance is
\[ \sigma_e^2 = \sigma_1^2 + \sigma_2^2 \]

but \[ \sigma_1^2 = \sigma_2^2 = \sigma_e^2 \]
\[ h(n) \text{ which gives} \]
\[ n = -\infty \]
\[ \sigma_e^2 = 2 - 2b \left[ \frac{2}{n=0} \frac{\gamma^2 n \sin^2(\pi n) e^{j\theta}}{\sin^2(\theta)} \right] \]
\[ = 2 - 2b \left[ \frac{1}{12 \sin^2(\theta)} \sum_{n=0}^{\infty} \gamma^2 n \cos(\pi n) e^{j\theta} \right] \]
\[ = 2 - 2b \left[ \frac{1}{12 \sin^2(\theta)} \sum_{n=0}^{\infty} \gamma^2 n (1 - \cos(\pi n) e^{j\theta}) \right] \]
\[ = 2 - 2b \left[ \frac{1}{12 \sin^2(\theta)} \sum_{n=0}^{\infty} \gamma^2 n \cos(\pi n) e^{j\theta} \right] \]
\[ = 2 - 2b \left[ \frac{1}{12 \sin^2(\theta)} \left( \frac{1 - \gamma^2}{2} \sum_{n=0}^{\infty} \gamma^2 n \cos(\pi n) e^{j\theta} \right) \right] \]
\[ \frac{2 - 2b}{b} \cdot \frac{1}{2^{2m^2b}} \cdot \left[ \frac{1}{1 - \gamma^2} - \frac{1}{2} \cdot \frac{e^{j2\theta}}{1 - \gamma^2 e^{j\theta}} + \frac{e^{-2\theta}}{1 - \gamma^2 e^{-j\theta}} \right] \]

\[ = \frac{2 - 2b}{b} \cdot \frac{1}{2^{2m^2b}} \cdot \left[ \frac{1}{1 - \gamma^2} - \frac{1}{2} \cdot \frac{\cos 2\theta - \gamma^2}{1 - 2\gamma^2 \cos 2\theta + \gamma^4} \right] \]

\[ = \frac{2 - 2b}{b} \cdot \frac{1}{2^{2m^2b}} \cdot \left[ \frac{(1 + \gamma^2)(1 - \cos 2\theta)}{(1 - \gamma^2)(1 - 2\gamma^2 \cos 2\theta + \gamma^4)} \right] \]

\[ \sigma^2 = \frac{2 - 2b}{6} \cdot \frac{(1 + \gamma^2)}{(1 - \gamma^2)(1 - 2\gamma^2 \cos 2\theta + \gamma^4)} \]

3) The output signal of an A/D converter is passed through a first-order low-pass filter with transfer function given by

\[ H(z) = \frac{(1 - a)z}{z - a} \quad \text{for} \quad 0 < a < 1 \]

Find the steady-state output noise power due to quantization at the output of the digital filter.
From Eqn (11)

\[
\sigma_e^2 = \sigma_e^2 \frac{1}{\pi} \int_{-\pi}^{\pi} \frac{H(z) H(z^{-1}) z^{-1} dz}{(z^{-1} - a)}
\]

\[
H(z) = \frac{(1-a)z}{(z-a)}
\]

\[
H(z^{-1}) = \frac{(1-a)z^{-1}}{(z^{-1} - a)}
\]

Substituting \(H(z)\) and \(H(z^{-1})\) in Eqn (10)

\[
\Rightarrow \sigma_e^2 = \sigma_e^2 \frac{1}{\pi} \int_{-\pi}^{\pi} \frac{(1-a)^2 z^{-1} dz}{(z^{-1} - a)(z^{-1} - a)}
\]

\[
= \sigma_e^2 \left[ \text{residue of } H(z) H(z^{-1}) z^{-1} \text{ at } z=a 
+ \text{residue of } H(z) H(z^{-1}) z^{-1} \text{ at } z=1/a \right]
\]

\[
= \sigma_e^2 \left[ \frac{(z-a)(1-a)^2 z^{-1}}{(z-a)(z^{-1} - a)} \right]_{z=a}
\]

\[
= \sigma_e^2 \left[ \frac{(1-a)^2}{1-a^2} \right] = \sigma_e^2 \left[ \frac{1-a}{1+a} \right]
\]
where \( \sigma_e^2 = \frac{\sigma^2 - 2b}{12} \)

4) Draw the quantization noise model for the following system (remains).

\[ x(n) \quad b_0 \quad y(n) \quad k(n) \quad b_1 \quad A \]

\[ z^{-1} \]

\[ a^2 \quad b_1 \quad A \quad b_1 \]

\[ \text{Fig. A} \]

\[ \text{Fig. B} \]

- Quantization Noise Model

5) Find the steady state variance of the noise in the output due to quantization of input for the first order filter

\[ y(n) = ay(n-1) + x(n) \]

So

The impulse response for the above filter is given by

\[ h(n) = a^n u(n) \]

From Eqn 18

\[ \sigma_e^2 = \sigma_e^2 \sum_{k=0}^{\infty} h^2(k) \]
\[
\sigma_e^2 = \sigma_0^2 \sum_{k=0}^\infty a^{2k} = \sigma_0^2 \left[ 1 + a^2 + a^4 + \ldots \right]
\]

\[
= \sigma_0^2 \frac{1}{1-a^2} = \frac{2}{1-a^2} \left[ \frac{1}{1-a^2} \right] (\text{OR})
\]

\[
G_n
\]

\[
y(n) = ay(n-1) + x(n)
\]

Taking Z transform on both sides we have:

\[
y(z) = az^{-1}y(z) + x(z)
\]

\[
H(z) = \frac{y(z)}{x(z)} = \frac{1}{1-a z^{-1}} = \frac{z}{z-a}
\]

\[
H(z^{-1}) = \frac{z^{-1}}{z^{-1}-a}
\]

We know:

\[
\sigma_e^2 = \sigma_0^2 \cdot \frac{1}{2\pi j} \oint H(z) \cdot H(z^{-1}) \cdot z^{-1} dz
\]

Substituting \(H(z)\) & \(H(z^{-1})\) values in the above equation we get:

\[
\sigma_e^2 = \sigma_0^2 \cdot \frac{1}{2\pi j} \oint \frac{z}{z-a} \cdot \frac{z^{-1}}{z^{-1}-a} \cdot z^{-1} \cdot dz
\]

\[
\sigma_e^2 = \sigma_0^2 \cdot \frac{1}{2\pi j} \oint \frac{z^{-1}}{(z-a)(z^{-1}-a)} \cdot dz
\]
\[
\varepsilon_e^2 \left[ \text{residue of } \frac{z^{-1}}{(z-a)(z^{-1}-a)} \text{ at } z = a \right] \\
+ \text{residue of } \frac{z^{-1}}{(z-a)(z^{-1}-a)} \text{ at } z = \frac{1}{a} \\
\varepsilon_e^2 \left[ \left( z-a \right) \frac{z^{-1}}{(z-a)(z^{-1}-a)} \right]_{z=a} \\
\varepsilon_e^2 \left[ \frac{a^{-1}}{a^{-1}-a} \right] = \varepsilon_e^2 \frac{a^{-1}}{1-a^2} \\
\]

**HW**

6) Realize the first order transfer function
   \[ H(z) = \frac{1}{1-a^2 z^{-1}} \]
   and draw its quantization noise model. Find the steady state noise power due to product roundoff.

7) The output signal of an A/D converter is passed through a first order low pass filter with transfer function given by
   \[ H(z) = \frac{(1-a)z}{z-a} \text{ for } 0 < a < 1 \]
   Find the steady state output noise power due to quantization at the output of the digital filter.
8) Determine the output noise power for the 
SAR shown in Fig. (Quantization noise model)

Express your answer in terms of $\sigma_e^2$.

\[ a) \quad 2\sigma_e^2 + \sigma_e^2 \left( b_0^2 + \frac{(a_0 + b_1)^2}{1-a^2} \right) \]

\[ b) \quad \frac{3\sigma_e^2}{1-a^2} \]

9) Consider a second-order IIR filter with

\[ H(z) = \frac{1.0}{(1-0.5^2)(1-0.45^2)} \]

Find the effect on quantization on pole location 
of the given SAR function in direct form II in 
cascade form. Take $b = 3$ bits

Solve

**Direct Form II**

\[ H(z) = \frac{1}{(1 - 0.95z^{-1} + 0.225z^{-2})} \]

\[ (0.95)_{10} = (0.111\ldots)_2 \]

\[ (-0.95)_{10} = (1.111\ldots)_2 \]

After truncation

we have \( (1.111)_2 = -0.875 \)

\[ (0.225)_{10} = (0.001\ldots)_2 \]
After truncation we have \((0.001)_2 = 0.125\)

\[ H(z) = \frac{1}{(1 - 0.875z^{-1}) + 0.125z^{-2}} \]

Cascade form

\[ H(z) = \frac{1}{(1-0.5z^{-1})(1-0.45z^{-1})} \]

\[ (-0.5)_{10} = (0.100)_2 \]

\[ (-0.45)_{10} = (0.01110)_2 \]

After truncation we have

\[ (1.011)_2 = (-0.375)_{10} \]

\[ H(z) = \frac{1}{(1-0.5z^{-1})(1-0.875z^{-1})} \]

**Topic IV**

**(Zero-Input) Limit Cycle Oscillations**

When a stable, all-pole digital filter is excited by a finite input sequence, that is constant, output will ideally decay to zero. However, non-linearities and periodic oscillations can occur in the CLP, and such oscillations may recur.

www.studentsfocus.com
Systems are called zero input limit cycle oscillations.

Consider a first order IIR filter with difference equation

\[ y(n) = x(n) + \alpha y(n-1) \rightarrow 0 \]

Let us assume \( \alpha = \frac{1}{2} \) and the data register length is \( 8 \) bits plus a sign bit. If the input is

\[ x(n) = \begin{cases} \frac{0.875}{2^n} & \text{for } n = 0 \\ 0 & \text{otherwise} \end{cases} \rightarrow 2 \]

and rounding applied after the arithmetic operation, the table given below illustrates the limit cycle behaviour.

**Table 1: \( \alpha = \frac{1}{2} \)**

| \( n \) | \( x(n) \) | \( y(n-1) \) | \( \alpha y(n-1) \) | \( Q(x y(n-1)) \) | \( y(n) = x(n) + \alpha y(n-1) \) |
|---------|-------------|--------------|-----------------|----------------|
| 0       | 0.875       | 0.0          | 0.0             | 0.000          | \( -\frac{1}{8} \) |
| 1       | 0           | 0.125        | 0.0             | 0.010          | \( \frac{1}{4} \) |
| 2       | 0           | 0.125        | 0.0             | 0.010          | \( \frac{1}{4} \) |
| 3       | 0           | 0.125        | 0.0             | 0.001          | \( \frac{1}{8} \) |
| 4       | 0           | 0.125        | 0.0             | 0.001          | \( \frac{1}{8} \) |
| 5       | 0           | 0.125        | 0.0             | 0.001          | \( \frac{1}{8} \) |

Fig 1 gives the (zero input) limit cycle oscillation. It is seen from the table.
From the table it is given that for
\( n \geq 3 \), the output remains constant and given
\( \frac{1}{8} \) at steady output causing limit cycle behaviour.

1. From the table it is noted that output oscillates below
\( 0.125 \) at \( 0.125 \).

Commonly it is inferred that, beyond \( n=4 \)
the value of \( [x(n-1)] y(n+1) + \frac{1}{16} \)
= \( 0.0001 \) as \( 0.000100 \) after rounding gives

\[ \begin{array}{c|c|c|c|c|c}
 n & x(n) & y(n-1) & y(n-1) & Q(x(n), y(n-1)) & y(n) - x(n) + Q(x(n), y(n-1)) \\
--- & --- & --- & --- & --- & --- \\
 0 & 0.875 & 0 & 0 & 0 & 7/8 \\
 1 & 0 & 7/8 & 0 & 1.00 & -1/2 \\
 2 & 0 & -1/2 & 1/4 & -1/8 & 0.10 \\
 3 & 0 & 1/4 & -1/8 & 1.00 & -1/8 \\
 4 & 0 & -1/8 & 1/16 & 0.00 & -1/8 \\
 5 & 0 & 1/16 & -1/16 & 1.00 & -1/8 \\
 6 & 0 & -1/16 & 1/16 & 0.00 & -1/8 \\
\end{array} \]
IV.1) **Dead band**

The limit cycle occurs as a result of the quantization effects in multiplications. The amplitude of the output during a limit cycle are confined to a range of values that is called the **dead band** of the filter.

**Example**

Consider a single pole 2nd order system where the difference equation is

\[ y(n) = \alpha y(n-1) + x(n), \quad n \geq 0 \rightarrow (2) \]

After rounding the product term we have

\[ y_q(n) = \left\lfloor \alpha y(n-1) + x(n) \right\rfloor \rightarrow (3) \]

During the limit cycle oscillation,

\[ \alpha \left\lfloor \alpha (y(n-1)) \right\rfloor = y(n+1) \quad \text{for} \quad \alpha > 0 \quad \text{(Refer Table 1)} \]

\[ = -y(n+1) \quad \text{for} \quad \alpha < 0 \quad \text{[By Table 2]} \]

\[ \rightarrow \left| \alpha \left\lfloor \alpha (y(n+1)) \right\rfloor - \alpha y(n+1) \right| \leq \frac{a^2}{2} \rightarrow (4) \]

Sub (5) in (6)

\[ \rightarrow \left| y(n+1) - \alpha y(n+1) \right| \leq \frac{a-b}{2} \]

\[ \rightarrow \left| y(n+1) \right| \leq \frac{a^2}{2} \]
Thus Eqn 1 gives the dead band of the first order filter.

**Topic V**

**Overflow limit cycle oscillations**

**Overflow Error**

Due to rounding in the multiplication, there occurs limit cycle oscillations and there are different types of limit cycle oscillations caused by addition, which make the filter output oscillate between maximum and minimum amplitudes. Such limit cycles is referred as overflow oscillations. The circuit cause this is an adder.

**Example:**

Let us consider two positive nos $n_1$ and $n_2$:

$n_1 = 0.111 \rightarrow 7/8$

$n_2 = 0.110 \rightarrow 6/8$

$n_1 + n_2 = 1.101 \rightarrow -5/8$ in sign magnitude

When two positive numbers are added, the sum is wrongly interpreted as a negative number.
Transfer characteristics of adder

Fig. 1: Transfer characteristics of adder

Here \( n \) is the input given to the adder.

and \( \text{flip} \) is the corresponding 1's and 0's overflow occurs if the total input is out of range.

This problem can be eliminated by modifying the adder characteristics as shown in Fig. 2 called saturation adder transfer characteristics.

Fig. 2: Saturation adder transfer characteristics

Here when an overflow is detected, the sum of adder is set equal to the maximum value.
Topic VI

Signal scaling

Saturation arithmetic eliminates limit cycles due to overflow, but it causes undesirable signal distortion due to the non-linearity of the clipper. In order to limit the amount of non-linear distortion, it is important to scale the input signal and the unit sample response by the input and any internal summing node in the system, such that overflow becomes a rare event.

![Diagram showing signal scaling](image)

Fig. 1: Realization of a second order IIR filter

Let us consider a second order IIR filter as in Fig. 1. A scale factor is introduced by the input, \( x(n) \) and the adder, to prevent overflow at the output adder.
Now the overall input - output transfer function

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

\[ \frac{N(z)}{P(z)} = \frac{b_0}{a_2} \rightarrow (2) \]

From Fig 1,

\[ H'(z) = \frac{w(z)}{x(z)} = \frac{b_0}{a_2} = \frac{b_0}{a_2} \rightarrow (3) \]

Since \( w(n) \) has finite energy in the input sequence, there will not be any overflow.

From eqn (3), we have

\[ w_0(z) = \frac{b_0}{b_2} \frac{x(z)}{p(z)} = \frac{b_0}{b_2} d(z) x(z) \rightarrow (4) \]

where \( d(z) = \frac{1}{p(z)} \)

\[ \Rightarrow w_0(n) = \frac{b_0}{2\pi} \int \tilde{r}(\epsilon) x(e^{i\theta}) d\theta \rightarrow (5) \]

Which gives

\[ w^2(n) = \frac{b_0^2}{4\pi^2} \left| \int \tilde{r}(\epsilon) x(e^{i\theta}) \left( e^{-i\theta} \right) d\theta \right|^2 \rightarrow (6) \]

Using Schwartz inequality
Now the overall input-output transfer function

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

\[ = \frac{N(z)}{D(z)} \]

From Fig 1

\[ H(z) = \frac{W(z)}{K(z)} = \frac{b_0}{1 + a_1 z^{-1} + a_2 z^{-2}} = \frac{b_0}{D(z)} \]

Since \( W(n) \) has finite energy in the input sequence, there will not be any overflow.

From Eqn 3, we have

\[ W(z) = \frac{b_0 K(z)}{D(z)} = \frac{b_0 S(z) K(z)}{D(z)} \]

where \( S(z) = \frac{1}{D(z)} \)

\[ \Rightarrow W(n) = \frac{b_0}{2\pi} \int K(z) K(z) e^{j\theta} d\theta \]

which gives

\[ W^2(n) = \frac{b_0^2}{4\pi^2} \left| \int K(z) K(z) e^{j\theta} d\theta \right|^2 \]

Using Schwartz inequality
\[
\omega^2(n) \leq \delta_0^2 \int_{-\pi}^{\pi} |x(e^{j\omega})|^2 d\omega \leq \frac{1}{2\pi} \int_{-\pi}^{\pi} |x(e^{j\omega})|^2 d\omega
\]

Applying Parseval's theorem, we get

\[
\omega^2(n) \leq \delta_0^2 \sum_{n=0}^{\infty} x^2(n) \int_{-\pi}^{\pi} |x(e^{j\omega})|^2 d\omega
\]

We know \( z = e^{j\omega} \), differentiating with respect to \( \omega \) we have

\[
d\omega = j \omega e^{j\omega} d\omega
\]

which gives

\[
d\omega = \frac{dz}{jz}
\]

Sub Eqn (4) and (6) in Eqn (8)

\[
\Rightarrow \omega^2(n) \leq \delta_0^2 \sum_{n=0}^{\infty} x^2(n) \int_{-\pi}^{\pi} |x(e^{j\omega})|^2 d\omega
\]

We can obtain

\[
\omega^2(n) \leq \sum_{n=0}^{\infty} x^2(n) \text{ when } \int_{-\pi}^{\pi} \frac{1}{2\pi} s(z) s(z^{-1}) |dz| = 1
\]
which gives

\[ S(z) = \frac{1}{\Delta} \int_{C} S(z) S(z^{-1}) z^{-1} dz \]

\[ S(z) = \frac{1}{\Delta} \int_{C} z^{-1} dz \]

\[ S(z) = \frac{1}{\Delta} \int_{C} \frac{z^{-1} dz}{D(z) D(z^{-1})} \]

\[ S(z) = \frac{1}{\Delta} \int_{C} \frac{z^{-1} dz}{D(z) D(z^{-1})} \]

where \( z = \frac{1}{\Delta} \int_{C} \frac{z^{-1} dz}{D(z) D(z^{-1})} \)

Problems on Signal Scaling

1) Given \( H(z) = \frac{0.5 + 0.4z^{-1}}{1 - 0.812z^{-1}} \) is the transfer function of a digital filter, find the scaling factor \( s_0 \) to avoid overflow in the adder 1 of the digital filter shown in Fig. below.

Fig. Realization of transfer function
We have $\delta_o^2 = \frac{1}{T}$ where

$$z = \frac{1}{2 \pi j} \oint_C \frac{z^{-1} \, dz}{(1 - 0.812z^{-1})(1 - 0.812z)}$$

Residue of $\frac{z^{-1}}{(1 - 0.812z^{-1})(1 - 0.812z)}$ due to pole $z = \frac{1}{0.812}$

From the given problem and by condition $x = \delta_o t$

$$x = \frac{1}{2 \pi j} \oint_C \frac{z^{-1} \, dz}{(1 - 0.812z^{-1})(1 - 0.812z)}$$

$$= \frac{(z - 0.812)(z - 0.812)}{1 - 0.812^2} \Bigg|_{z = 0.812}$$

$$= \frac{1}{1 - 0.812^2}$$

$$\delta_o = \left\lfloor \frac{1}{1.1078} \right\rfloor = 0.9501$$
2) For a second order digital filter

\[ H(z) = \frac{1}{1 - 2\cos(\omega_0)z^{-1} + z^{-2}}, \quad |z| > 1.0 \]

Draw the direct form II realization and find the scale factor so to avoid overflow.

Problems for Practice

1) Express the following binary numbers in decimal:

(i) \((101110.1110)_2\)

(ii) \((101110.1111)_2\)

(iii) \((10111.011)_2\)

2) Consider the transfer function

\[ H(z) = \frac{1.0}{1 - 0.94z^{-1} + 0.64z^{-2}} \]

Find the pole location and effect due to rounding to 3 bit (excluding sign bit).

3) Study the limit cycle behaviour of the SMI

\[ y(n) = 0.95y(n-1) + x(n) \]

Determine the deadband of the SMI.
\( y(n) \) limit cycle behaviour

Let us assume 4 bit sign magnitude representation (excluding sign bit).

Let the input be

\[ x(n) = \begin{cases} 0.875 & \text{for } n = 0 \\ 0 & \text{otherwise} \end{cases} \]

Before adding to \( x(n) \), product 0.95y/ln must be rounded to 4 bits.

\[ y(n) = x(n) + Q \left\lfloor 0.95 \cdot y(n-1) \right\rfloor \]

where \( Q \left\lfloor \cdot \right\rfloor \) stands for quantization.

For \( n = 0 \),

\[ y(0) = Q \left\lfloor 0.95 \cdot y(-1) \right\rfloor = 0.875 \]

\[ y(-1) = 0 \]

For \( n = 1 \),

\[ y(1) = Q \left\lfloor 0.95 \cdot y(0) \right\rfloor + x(1) \]

\[ = Q \left\lfloor 0.95 \cdot 0.875 \right\rfloor + 0.875 \]

\[ = Q \left\lfloor 0.83125 \right\rfloor \]

\[ y(1) = 0.83125 = (1101010)_2 \]
After rounding we get
\[ Q(0.8125) = (0.1101)_{2} = (0.8125)_{10} \]
\[ \therefore y(1) = 0.8125 \]

for \( n = 2 \)
\[ y(2) = Q \left[ 0.95 (y(1)) + x(2) \right] \]
\[ = Q \left[ (0.95) (0.8125) \right] \]
\[ = Q \left[ (0.771875) \right] \]
\[ = Q \left( (0.771875) \right)_{10} = (0.110001)_{2} \]

After rounding we get
\[ Q(0.771875) = (0.1100)_{2} = (0.75)_{10} \]
\[ \therefore y(2) = 0.75 \]

for \( n = 3 \)
\[ y(3) = Q \left[ 0.7125 \right] \]
\[ (0.7125) = (0.101101\ldots)_{2} \]

After rounding
\[ Q(0.7125) = (0.1011)_{2} = (0.6875)_{10} \]
\[ y(3) = 0.6875 \]

for \( n = 4 \)
\[ y(4) = Q \left[ 0.653125 \right] = (0.101001\ldots)_{2} \]
\[ Q(0.653125) = (0.1010)_{2} = (0.625)_{10} \]
\[ y(4) = 0.625 \]
for \( n = 5 \)

\[ y(5) = 0.95 \begin{pmatrix} 0.625 \end{pmatrix} + x(4) \]

\[ y(5) = 0.95 \begin{pmatrix} 0.59375 \end{pmatrix} = \begin{pmatrix} 0.10011 \end{pmatrix} \]

\[ \Rightarrow Q(0.59375) = \begin{pmatrix} 0.1001 \end{pmatrix}^2 = \begin{pmatrix} 0.625 \end{pmatrix} \]

\[ \Rightarrow y(5) = 0.625 \]

for \( n = 6 \)

\[ y(6) = Q(0.59375) = \begin{pmatrix} 0.1001 \end{pmatrix}^2 = \begin{pmatrix} 0.625 \end{pmatrix} \]

\[ y(6) = 0.625 \]

\[ \Rightarrow n \geq 5, \text{ the output remains constant at } 0.625 \]

Causing limit cycle behaviour.

a) The deadband

\[ = \frac{1}{2} a - b \]

For \( b = 1 \)

\[ \text{Deadband} = \frac{1}{2} a - b = \frac{0.625}{1 - 0.95} \]

b) Study the limit cycle behaviour of the

(i) \[ y(n) = 0.7y(n-1) + x(n) \]

(ii) \[ y(n) = 0.65y(n-1) + 0.52y(n-1) + x(n) \]
5) Consider the SLM

\[ H(z) = \frac{1 - \frac{1}{2} z^{-1}}{(1 - \frac{1}{2} z^{-1})(1 + \frac{1}{2} z^{-1})} \]

a) Draw the direct form II and cascade form realization of the SLM.

b) Suppose that we implement the biquad with fixed point sign and magnitude fractional arithmetic using (b+1) bits (one bit used for sign). Each resulting product is rounded off to b bits. Determine the variance of round off noise created by the multipliers at the output of each one of one realization in parts.