Unit - IV

Finite Word Length Effect

fixed Point and floating Point number representation - ADC - Quantization - truncation and rounding - quantization noise - input / output quantization - Coefficient quantization error - Product quantization error - Overflow error - limit Cycle oscillations due to Product quantization and Summation - Scaling to Prevent Overflow.

Finite Word Length Effects :-

In general the effect due to finite Precision representation of in a digital System are Commonly referred to as finite word length effects.

The following are Some of the finite word length effects in digital filters.

* Quantization of input data
* Quantization of filter Co-efficients
* Rounding the Product in multiplication.
* overflow in addition
* Limit Cycle oscillation.

The two major methods of representing binary numbers are

* Fixed Point representation
* Floating Point representation.

Fixed Point Representation :-

In this method the digits allotted for integer Part and fraction Part are fixed, and So the Position of binary Point is fixed.

Since the number of digits is fixed it is impossible to represent too large and too Small number.

∴ ~~The~~ The range of numbers that Can be represented in this method is less.

Floating Point Representation :-

In this method, the binary Point Can be Shifted to desired Position so that number of digits in the integer Part and fraction part of a number Can be Varied. This leads to larger range of number Can be represent in floating Point representation

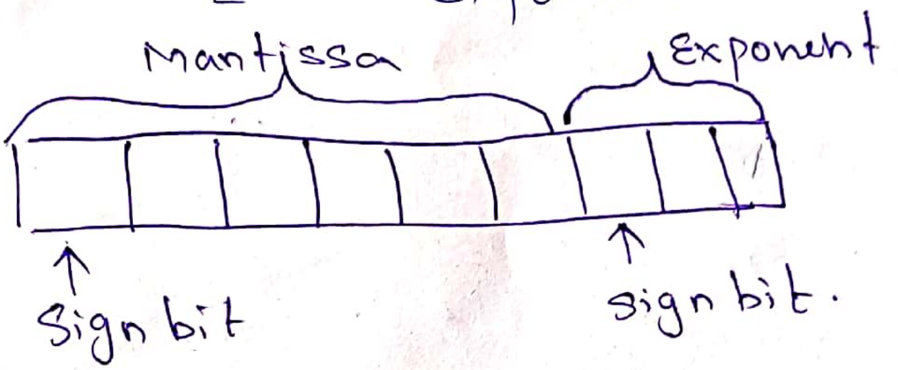
Floating Point number

$$N_f = M \times 2^E$$

where

M is mantissa

E is Exponent.



Fixed Point Representation Formats :-

There are three different formats

1. Sign- magnitude format
2. 1's Complement format
3. 2's Complement format

Binary number in fixed Point representation

Sign-magnitude	1's Complement	2's Complement	Decimal
0000	0000	0000	0
0001	0001	0001	$\frac{1}{8} = 0.125$
0010	0010	0010	$\frac{2}{8} = 0.25$
0011	0011	0011	$\frac{3}{8} = 0.375$
0100	0100	0100	$\frac{4}{8} = 0.5$
0101	0101	0101	$\frac{5}{8} = 0.625$
0110	0110	0110	$\frac{6}{8} = 0.75$
0111	0111	0111	$\frac{7}{8} = 0.875$
1000	1111	~~1001~~	-0
1001	1110	1111	$-\frac{1}{8} = -0.125$
1010	1101	1110	$-\frac{2}{8} = -0.25$
1011	1100	1101	$-3/8 = -0.375$
1100	1011	1100	$-4/8 = -0.5$
1101	1010	1011	$-5/8 = -0.625$
1110	1001	1010	$-6/8 = -0.75$
1111	1000	1001	$-7/8 = -0.875$
		1000	$-8/8 = -1$

Problem 1:

Convert $+0.125_{10}$ and -0.125_{10} to Sign-magnitude format of binary and Verify the result by converting the binary to decimal.

0.125×2

0.250×2

0.5×2

1.00

$0.0 \quad 0 \quad 1$

$+0.125 \xrightarrow[\text{to binary}]{\text{Convert}} 0.001 \xrightarrow[\text{Sign bit}]{\text{Append}} 0.001 = \boxed{0001}$

$-0.125 \xrightarrow[\text{to binary}]{\text{Convert}} -0.001 \xrightarrow[\text{Sign bit}]{\text{Append}} 1.001 = \boxed{1001}$

Problem 2:

Convert 0.25_{10} and -0.25_{10} to 1's Complement format.

0.25×2

0.5×2 Integer part is removed

$1.0 = 0.0 \times 2$

0

$0.0 \quad 1 \quad 0$

$+0.25 \xrightarrow[\text{to binary}]{\text{Convert}} +0.010 \xrightarrow[\text{Sign bit}]{\text{Append}} 0.010 = [0010]$

$-0.25 \xrightarrow[\text{to binary}]{\text{Convert}} -0.010 \xrightarrow[\text{Sign bit}]{\text{Append}} 1.010 \xrightarrow[\text{Complement fraction}]{\text{1's}}$

Part → $1.101 = [1101]$

Problem 3:

Convert +0.5 and −0.5 to 2's complement format

$$
\begin{array}{l}
0.5 \times 2 \\
\hline
1.0 \\
\quad 0.0 \times 2 \\
\quad \overline{\quad 0.0 \times 2} \\
\quad \quad \overline{\quad 0.0}
\end{array}
$$

$0.1\;0\;0$

$+0.5 \xrightarrow[\text{to binary}]{\text{Convert}} +0.100 \xrightarrow[\text{Sign bit}]{\text{append}} 0.100 = [0100]$

$-0.5 \xrightarrow[\text{to binary}]{\text{Convert}} -0.100 \xrightarrow[\text{Sign bit}]{\text{append}} 1.100 \xrightarrow[\text{complement fraction part}]{\text{2's}}$

$\rightarrow 1.100 = [1100]$

Problem 4 :

Convert $+25_{10}$ and -25_{10} to 32-bit IEEE-754 format of floating Point representation.

$$\begin{array}{r|l} 2 & 25 \\ \hline 2 & 12 - 1 \\ \hline 2 & 6 - 0 \\ \hline 2 & 3 - 0 \\ \hline & 1 - 1 \end{array}$$

$(25)_{10} = (11001)_2$

$$25_{10} \xrightarrow[\text{to binary}]{\text{Convert}} +11001 \xrightarrow[\text{in 1.M format}]{\text{Represent}} 1.1001_2 \times 2^4.$$

$$\xrightarrow[\text{to E-127}]{\text{Convert Exponent}} 1.1001_2 \times 2^{131-127}$$

The number N in IEEE-754 format is

$$N = (-1)^S \times 1.M \times 2^{E-127}.$$

$$+25_{10} = (-1)^0 \times 1.1001_2 \times 2^{131-127}$$

$$-25_{10} = (-1)^1 \times 1.1001_2 \times 2^{131-127}.$$

$1.M = 1.1001 \xrightarrow[\text{fraction Part to 23-bits}]{\text{Convert}}$

$1.1001 \; 0000 \; 0000 \; 0000 \; 0000 \; 000$

$E = 131_{10} \xrightarrow[\text{to binary}]{\text{Convert}} (10000011)_2$

$$
\begin{array}{r|l}
2 & 131 \\
2 & 65 - 1 \\
2 & 32 - 1 \\
2 & 16 - 0 \\
2 & 8 - 0 \\
2 & 4 - 0 \\
2 & 2 - 0 \\
 & 1 - 0
\end{array}
$$

$$+25_{10} = \underbrace{0}_{\substack{\text{1-bit} \\ \text{sign}}} \; \underbrace{1000\;0011}_{\substack{\text{8-bit} \\ \text{exponent}}} \; \underbrace{1001\;0000\;0000\;0000\;0000\;000}_{\substack{\text{23-bit} \\ \text{mantissa}}}$$

$$-25_{10} = \underbrace{1}_{\substack{\text{1-bit} \\ \text{sign}}} \; \underbrace{1000\;0011}_{\substack{\text{8-bit} \\ \text{exponent}}} \; \underbrace{1001\;0000\;0000\;0000\;0000\;000}_{\text{23-bit mantissa.}}$$

Quantization by Truncation and Rounding :-

In fixed Point (or) floating Point arithmetic the Size of the result of an operation (Sum (or) Product) may be exceeding the Size of binary used in the number System. In Such Cases the low order bits has to be eliminated in order to Store the result. The two methods of eliminating these low order bits are

* Truncation
* Rounding

This Process is also referred to as quantization Via truncation and rounding. The Error occured due to quantization Process is called Quantization Error

Quantization Steps :-

The decimal numbers that are encountered as filter Coefficients, Sum, Product, etc., in DSP applications will usually lie in the range of -1 to 1. when "B" Bit binary is Selected to represent the decimal numbers, then 2^B binary Codes are Possible. each Step of decimal number is also called Quantization Step.

∴ Quantization Step Size

$$q = \frac{R}{2^B} = \frac{1-(-1)}{2^B} = \frac{2}{2^B}$$

$$= \frac{1}{2^B\,2^{-1}} = \frac{1}{2^{B-1}} = \frac{1}{2^b} = 2^{-b}$$

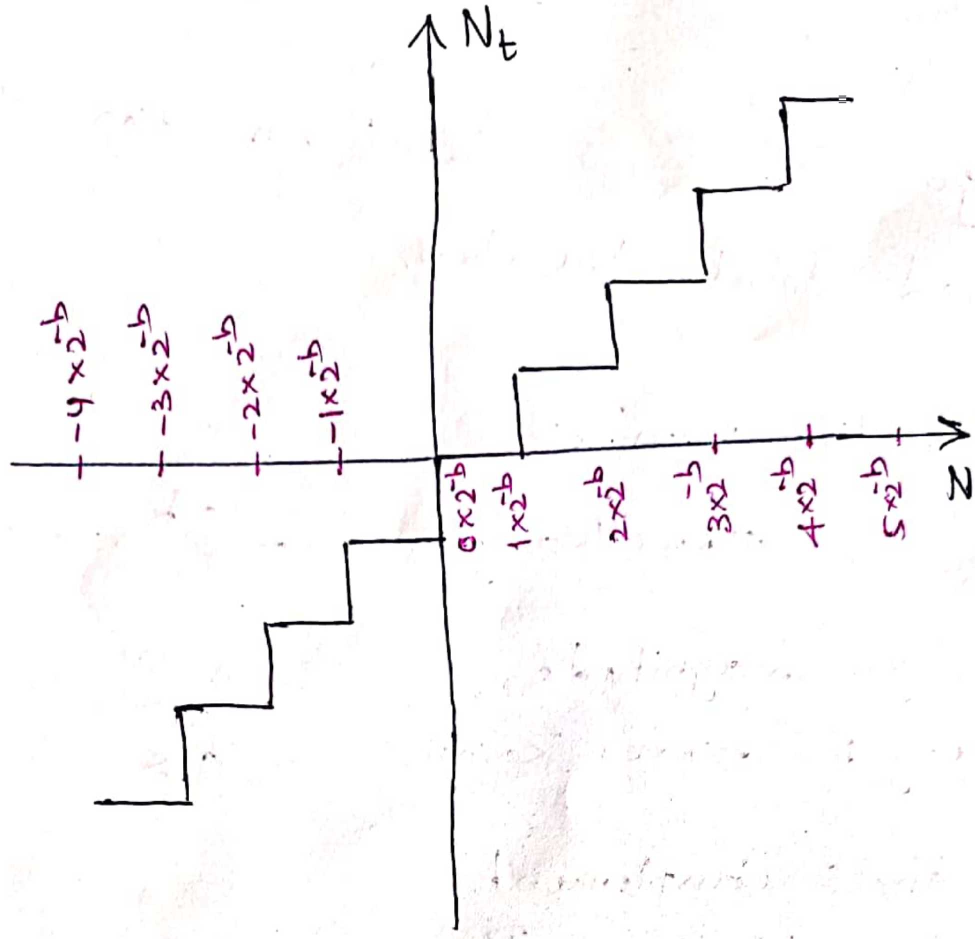
where, R = Range of decimal number

B = Size of binary including sign bit

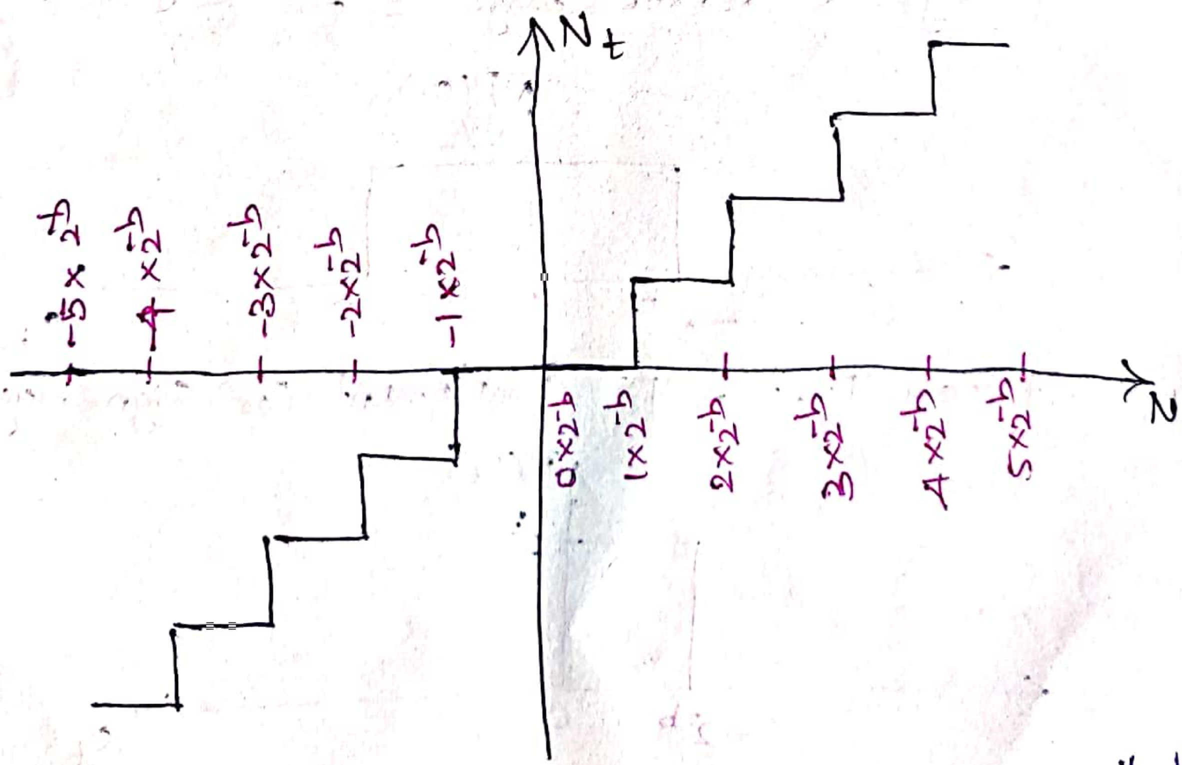
b = B-1 = Size of binary excluding sign bit

Truncation :

The truncation is the process of reducing the size of binary number (or reducing the number of bits in a binary number) by discarding all bits less significant than the least significant bit that is retained. all the less significant bits beyond b^{th} bit are discarded.



Truncation Characteristics of two's Complement Quantizer.



Truncation characteristics of Sign-magnitude (or) One's Complement Quantizer

Range of Errors in Truncation of fixed Point Number.

Positive number

error Range $= -2^{-b} < e < 0$

Sign - magnitude

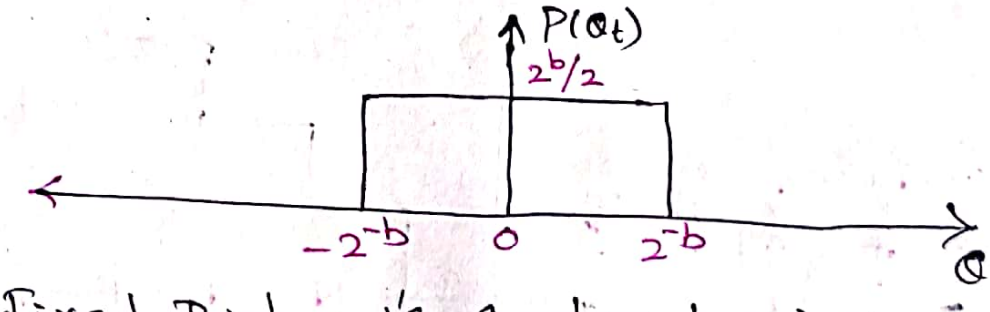
negative error Range $= 0 \leq e < 2^{-b}$

one's Complement

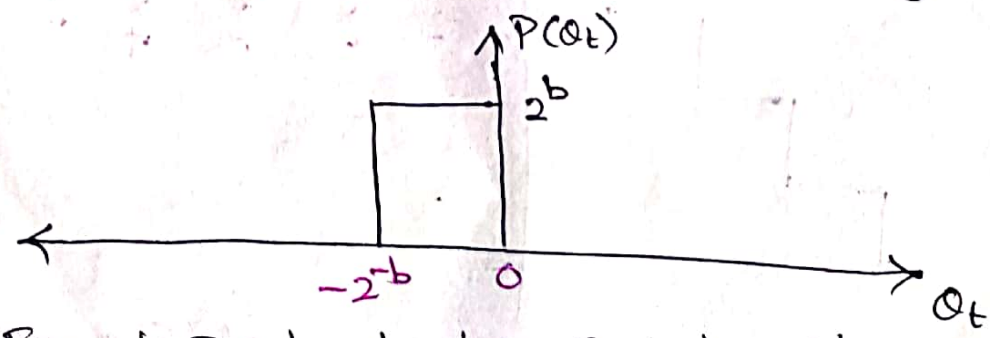
negative error Range $= 0 \leq e < 2^{-b}$

Two's Complement

negative error Range $= -2^{-b} < e \leq 0$



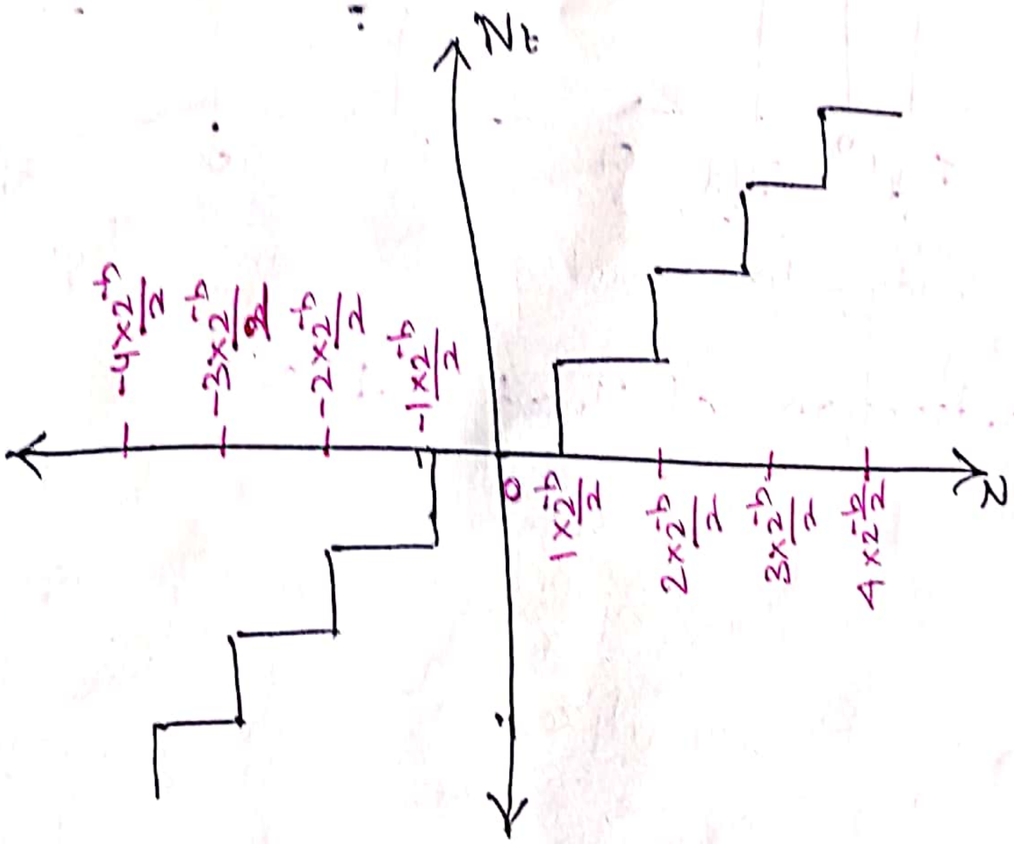
Fixed Point - 1's Complement (or) Sign-magnitude



Fixed Point - two's Complement.

Rounding :

Rounding is the Process of reducing the size of a binary number to finite word size of b-bits such that the rounded b-bit number is closest to the original unquantized number.



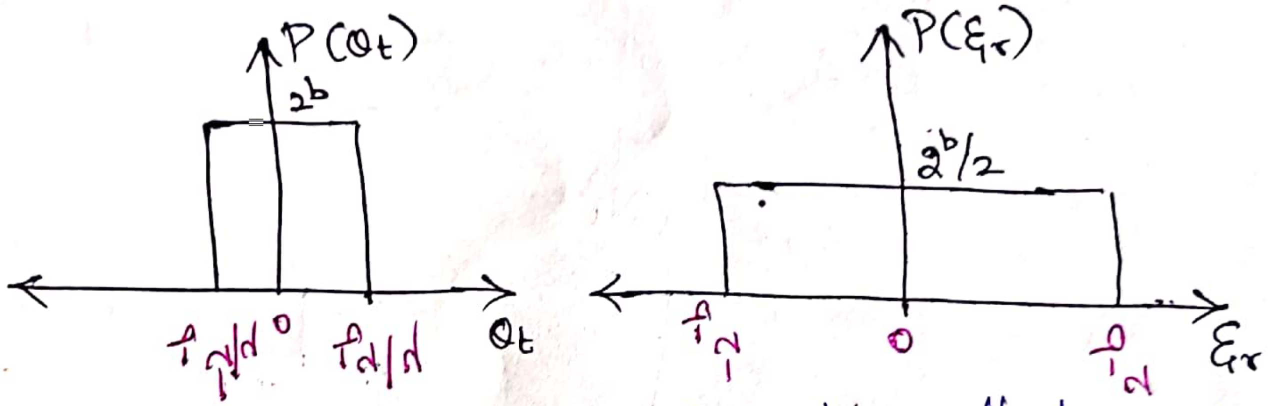
Input - output characteristics of quantizer used for rounding

In fixed Point representation the range of error made by rounding to b bit are

$$-\frac{2^{-b}}{2} \le e_r \le \frac{2^{-b}}{2}$$

In floating Point representation the range of error made by rounding to b-bits is given by

$$-2^{-b} \le \varepsilon_r \le 2^{-b}$$



Rounding - fixed Point   Rounding - floating Point.

Quantization of Input Data :-

For Processing of analog Signal using a digital System the analog Signal has to be digitized by A/D Converter. The A/D Converter Consists of Sampler and quantizer. The Sampler will Sample the Value of analog Signal at uniform intervals to produce a sequence of unquantized values of the Signal.

The quantizer will quantize the analog value and Produce the Corresponding binary Codes. The Process of assigning binary number to quantized analog Value is also called Coding.

Quantization Step Size $q = \frac{R}{2^B} = \frac{R}{2^{b+1}}$

where,

B = Size of binary including Sign bit

$b = B - 1$ = Size of binary excluding Sign bit.

Let $x(n)$ = Unquantized Sample of the Signal

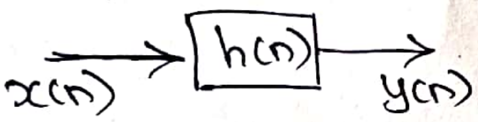
$x_q(n)$ = Quantized Sample of the Signal.

Quantization error, $e(n) = x_q(n) - x(n)$

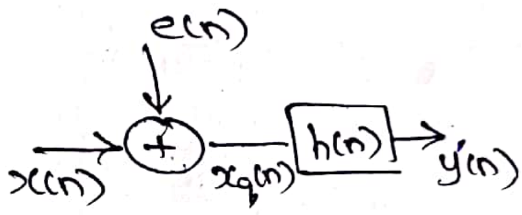
The quantization error for rounding will be in the range of $\frac{-q}{2}$ to $+q/2$

Steady State Output Noise Variance (Power) Due to the Quantization Error Signal.

The quantized input signal of a digital system can be represented as a sum of unquantized signal $x(n)$ and error signal $e(n)$ as shown in figure.



LTI System with Unquantized input



LTI System with quantized input.

$$y'(n) = x_q(n) * h(n)$$

$$= [x(n) + e(n)] * h(n)$$

$$= [x(n) * h(n)] + [e(n) * h(n)]$$

Let, $y'(n) = y(n) + \varepsilon(n)$

where $y(n) = x(n) * h(n)$ — o/p due to I/p signal $x(n)$

$\varepsilon(n) = e(n) * h(n)$ — o/p due to error signal $e(n)$

The Variance of the Signal $\xi(n)$ is Called output noise Power (or) Steady State output noise Power (or variance) due to the quantization error signal. Using auto correlation function and the diffinition for variance of a discrete time signal,

Steady State output noise power due to input quantization errors $\Bigg\}$ $\sigma_{eoi}^2 = \sigma_e^2 \sum_{n=0}^{\infty} h^2(n)$

The above equation is variance of error signal σ_e^2 using Parseval's theorem.

$$\sigma_{eoi}^2 = \sigma_e^2 \sum_{n=0}^{\infty} h^2(n) = \sigma_e^2 \frac{1}{2\pi j} \oint_c H(z) H(z^{-1}) z^{-1} dz$$

where $\oint_c$ denote integration around unit Circle $|z| = 1$, in the anti clockwise direction.

$$\sigma_{eoi}^2 = \sigma_e^2 \frac{1}{2\pi j} \oint_c H(z) H(z^{-1}) z^{-1} dz$$

$$= \sigma_e^2 \sum_{i=1}^{N} Res\left[H(z)H(z^{-1})z^{-1}\right]\Bigg|_{z=P_i}$$

$$= \sigma_e^2 \sum_{i=1}^{N} \left[(z-P_i)\, H(z)H(z^{-1})z^{-1}\right]\Bigg|_{z=P_i}$$

where $P_1, P_2 - - - - P_N$ are Poles of $H(z)$ $H(z^{-1})z^{-1}$.

Problem 5 :-

An LTI System is Characterized by the difference equation, $y(n) = 0.68\,y(n-1) + 0.15\,x(n)$. The input Signal $x(n)$ has a range of $-5V$ to $+5V$, represented by 8-bits. Find the quantization Step size, Variance of the error Signal and variance of the quantion noise at the output.

<u>Give that</u>

Range $R = -5$ to $+5 = 5-(-5) = 10$

Size of binary, $B = 8$ bits (including sign bit)

Quantization Step size, $q = \frac{R}{2^8} = \frac{10}{2^8} = 0.0390625$

Variance of error signal

$$\sigma_e^2 = \frac{q^2}{12} = \frac{0.0390625^2}{12} = 1.2716 \times 10^{-4}$$

The difference equation governing the LTI System is

$$y(n) = 0.68\, y(n-1) + 0.15\, x(n)$$

On taking z-transform of above equation we get

$$Y(z) = 0.68\, z^{-1}\, Y(z) + 0.15\, X(z)$$

$$Y(z) - 0.68\, z^{-1}\, Y(z) = 0.15\, X(z)$$

$$Y(z)[1 - 0.68\, z^{-1}] = 0.15\, X(z)$$

$$\frac{Y(z)}{X(z)} = \frac{0.15}{1-0.68\, z^{-1}} = H(z)$$

$$H(z)H(z^{-1})z^{-1} = \frac{0.15}{1-0.68\, z^{-1}} \times \frac{0.15}{1-0.68\, z} \times z^{-1}$$

$$= \frac{0.0225\, z^{-1}}{\left[1 - \frac{0.68}{z}\right](-0.68)\left[z - \frac{1}{0.68}\right]}$$

$$= \frac{-0.0331 z^{-1}}{\left(\frac{z-0.68}{z}\right)(z-1.4706)}$$

$$H(z)H(z^{-1})z^{-1} = \frac{-0.0331}{(z-0.68)(z-1.4706)}$$

Now, Poles of $H(z)H(z^{-1})z^{-1}$ are

$$P_1 = 0.68, \quad P_2 = 1.4706$$

Here $P_1 = 0.68$ is the only Pole that lies inside the unit circle in z-plance

Variance of the I/p Quantization noise at the output

$$\sigma_{eoi}^2 = \sigma_e^2 \frac{1}{2\pi j} \oint_c H(z)H(z^{-1})z^{-1}dz$$

$$= \sigma_e^2 \sum_{i=1}^{N} Res\left[H(z)H(z^{-1})z^{-1}\right]\Big|_{z=P_i}$$

$$= \sigma_e^2 \sum_{i=1}^{N} \left[(z-P_i)H(z)H(z^{-1})z^{-1}\right]\Big|_{z=P_i}$$

$$\sigma_{eoi}^2 = \sigma_e^2 \times \cancel{(z-0.68)} \times \frac{-0.0331}{\cancel{(z-0.68)}(z-1.4706)}\Bigg|_{z=0.68}$$

$$= \sigma_e^2 \times \frac{-0.0331}{0.68-1.4706}$$

$$= 0.0419\,\sigma_e^2$$

$$= 0.0419 \times 1.2716 \times 10^{-4}$$

$$\boxed{\sigma_{eoi}^2 = 5.328 \times 10^{-6}}$$

Quantization of Filter Coefficients :-

In the realization of FIR and IIR filters in hardware (or) in Software, the accuracy with which filter Coefficients Can be Specified is limited by the word length of the register used to Store the Co-efficients are quantized to the word Size of the register used to store them either by truncation (or) by rounding.

The location (or the value) of Poles and zeros of the digital filters directly depends on the value of filter Co-efficients.

will modify the value of Poles and Zeros, and so the location of the Poles and Zeros will be Shifted from the desired location. This will Create deviations in the frequency response of the System. Hence we obtain a filter having a frequency response that is different from the frequency response of the filter with unquantized Co-effients..

Problem 6:-

For Second-order IIR filter, H(z)

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

Study the effect of Shift in pole location with 3-bit Co-efficient representation in direct and cascade form.

Solution:-

$$H(z) = \frac{1}{(1-0.5z^{-1})(1-0.45z^{-1})} = \frac{z^2}{(z-0.5)(z-0.45)}$$

The Poles of H(z) $P_1 = 0.5$, $P_2 = 0.45$

Case(i): Direct form Realization

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

$$= \frac{1}{1-0.5z^{-1}-0.45z^{-1}+0.225z^{-2}}$$

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

$$0.95_{10} \xrightarrow[\text{binary}]{\text{Convert to}} 0.1111_2 \xrightarrow[\text{to 3-bits}]{\text{Truncate}}$$

$$0.\cancel{0}11 \xrightarrow[\text{Decimal}]{\text{Convert to}} 0.875_{10}$$

$$0.225 \xrightarrow{\text{binary}} 0.0011 \xrightarrow[\text{to 3-bits}]{\text{Truncate}}$$

$$0.001 \xrightarrow[\text{Decimal}]{\text{Convert to}} 0.125_{10}$$

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

$$= \frac{z^2}{z^2-0.875z+0.125}$$

$$\hat{P}_1 = 0.695 \qquad \hat{P}_2 = 0.18$$

we can observe that the poles of $\hat{H}(z)$ deviate very much from the original poles.

Case(iii) :- Cascade realization

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

$$H_1(z) = \frac{1}{1-0.5z^{-1}} \qquad H_2(z) = \frac{1}{1-0.45z^{-1}}$$

$$0.5 \xrightarrow{\text{binary}} 0.1000 \xrightarrow[\text{to 3-bits}]{\text{Truncate}} 0.100$$

$$\xrightarrow{\text{Decimal}} 0.5_{10}$$

$$0.45 \xrightarrow{\text{binary}} 0.011 \xrightarrow[\text{to 3-bits}]{\text{Truncate}} 0.011$$

$$\xrightarrow{\text{Decimal}} 0.375$$

$$\hat{H}(z) = \frac{1}{(1-0.5z^{-1})(1-0.375z^{-1})}$$

$$\hat{P}_1 = 0.5 \qquad \hat{P}_2 = 0.375$$

on comparing the poles of the cascade system very close to original pole.

Product Quantization Error :-

In realization Structure of IIR System, multipliers are used to multiply the signal by Constants. The output of the multipliers i.e., the products are quantized to finite word length in order to store them in register.

Two b-bits numbers results in a product of length 2b-bits. The error due to the quantization of the output of the multiplier is referred to as Product quantization error.

Limit Cycle in Recursive Systems.

Zero Input Limit Cycles :-

In recursive systems, when the input is zero (or) some non zero constant value, the non linearities due to finite precision arithmetic operations may cause periodic oscillations in the

the output. During Periodic oscillations, the output $y(n)$ of a System will oscillate between a finite Positive and negative Value for increasing n (or) the output will became Constant for increasing n. Such oscillations are called limit Cycle.

for a first order System described by the equation, $y(n) = a\,y(n-1) + x(n)$, the dead band is given by

$$\text{Dead band} = \pm \frac{2^{-B}}{1-|a|}$$

B = No. of binary bits including sign bit.

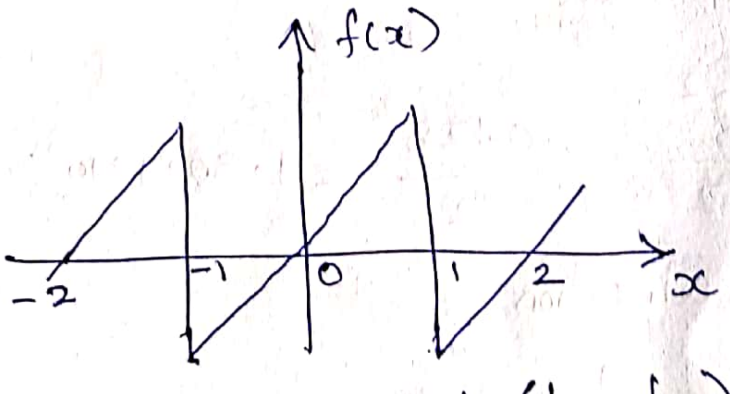
for a second order system described by the equation $y(n) = a_1 y(n-1) + a_2 y(n-2) + x(n)$ the dead band is given by

$$\text{Dead band} = \frac{\pm 2^{-B}}{1-|a_2|}$$

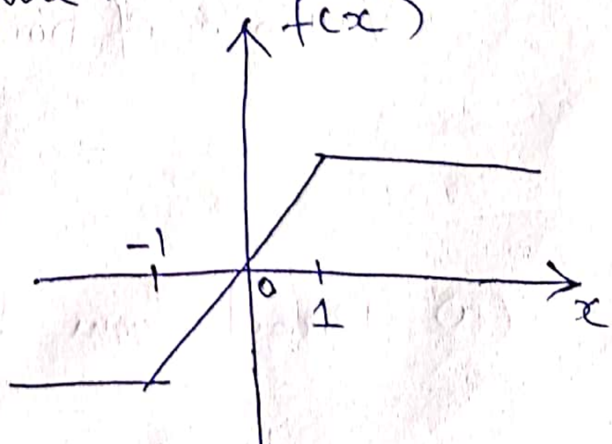
Over flow Limit Cycle :

In fixed point addition of two binary numbers the overflow occurs when the sum exceeds the finite word length of the register used to store the sum. The overflow in addition may lead to oscillations in the output which is ~~self~~ referred to as overflow limit cycles.

If the dynamic range is (-1, 1) The overflow is explained by considering 4-bit binary fraction number in 2's complement representation. The input-output (transfer) characteristics of 2's complement adder is shown in figure.



Input-output (transfer) characteristics of 2's complement adder.



Characteristics of Saturation adder

Problem 7 :-

In the IIR System given below the Products are rounded to 4 - bit (including sign bit)

$$H(z) = \frac{1}{(1-0.15z^{-1})(1-0.43z^{-1})}$$

Find the output rounded off noise Power in (a) Direct form realization (b) Cascade realization.

Solution :-

Let assume that the range of Product is -1 to $+1$

$$R = 1-(-1) = 2$$

Quantization Step size $\delta = \frac{R}{2^B} = \frac{2}{2^4} = 0.125$

Variance of noise Source

$$\sigma_e^2 = \frac{\delta^2}{12} = \frac{0.125^2}{12} = 1.3021 \times 10^{-3}.$$

(a) Direct form Realization

$$H(z) = \frac{1}{(1-0.15z^{-1})(1-0.43z^{-1})}$$

$$H(z) = \frac{1}{1 - 0.43z^{-1} - 0.15z^{-1} + 0.0645z^{-2}}$$

$$H(z) = \frac{1}{1 - 0.58z^{-1} + 0.0645z^{-2}}$$

Noise signal $e_1(n) = H(z) = \frac{1}{(1-0.15z^{-1})(1-0.43z^{-1})}$

Noise signal $e_2(n) = H(z) = \frac{1}{(1-0.15z^{-1})(1-0.43z^{-1})}$

$$\sigma^2_{e1\,o/p} = \sigma_e^2 \frac{1}{2\pi j} \oint T_1(z)\, T_1(z^{-1})\, z^{-1} dz$$

$$= \sigma_e^2 \sum_{i=1}^{N} \text{Res}\left[T_1(z) T_1(z^{-1}) z^{-1}\right]\Big|_{z=P_i}$$

$$= \sigma_e^2 \sum_{i=1}^{N} \left[(z-P_i)\, T_1(z)\, T_1(z^{-1})\, z^{-1}\right]\Big|_{z=P_i}$$

$$T_1(z) = \frac{1}{(1-0.15z^{-1})(1-0.43z^{-1})} = \frac{1}{\left(\frac{z-0.15}{z}\right)\left(\frac{z-0.43}{z}\right)}$$

$$T_1(z) = \frac{z^2}{(z-0.15)(z-0.43)}$$

$$T_1(z^{-1}) = \frac{1}{\left(1-0.15(z^{-1})^{-1}\right)\left(1-0.43(z^{-1})^{-1}\right)} = \frac{1}{(1-0.15z)(1-0.43z)}$$

$$\sigma_{e1}^2\, o/p = 1.3021 \times 10^{-3} \left[\cancel{(z-0.15)} \frac{z^{\cancel{2}\,1}}{\cancel{(z-0.15)}(z-0.43)} \cdot \frac{1}{(1-0.15z)(1-0.43z)} \times \cancel{z^{-1}} \Big|_{z=0.15} \right.$$

$$\left. + \cancel{(z-0.43)} \frac{z^{\cancel{2}\,1}}{(z-0.15)\cancel{(z-0.43)}} \cdot \frac{1}{(1-0.15z)(1-0.43z)} \times \cancel{z^{-1}} \Big|_{z=0.43} \right]$$

$$\sigma_{e1}^2\, o/p = 1.3021 \times 10^{-3} \left[\frac{0.15}{\underbrace{(0.15-0.43)}_{(-0.28)} \underbrace{(1-0.15\times0.15)}_{0.9775} \underbrace{(1-0.43\times0.15)}_{0.9355}} \right.$$

$$\left. + \frac{0.43}{\underbrace{(0.43-0.15)}_{0.28} \underbrace{(1-0.15\times0.43)}_{0.9355} \underbrace{(1-0.43\times0.43)}_{0.8151}} \right]$$

$$= 1.3021 \times 10^{-3} [-0.58583 + 2.014]$$

$$\sigma_{e1}^2\, o/p = 1.8596 \times 10^{-3}$$

$$\sigma_{e1}^2\, o/p = \sigma_{e2}^2\, o/p \quad \text{in direct form}$$

$$\therefore \sigma_e^2\, o/p = \sigma_{e1}^2\, o/p + \sigma_{e2}^2\, o/p \qquad 3$$

$$= 1.8596 \times 10^{-3} + 1.8596 \times 10^{-3}$$

$$\boxed{\sigma_e^2\, o/p = 3.7192 \times 10^{-3}}$$

output noise Power in direct form.

(b) Cascade Realization.

$$H(z) = \frac{1}{(1-0.15z^{-1})(1-0.43z^{-1})}$$

Case (i)

$$H(z) = H_1(z)\, H_2(z)$$

$$H_1(z) = \frac{1}{1-0.15z^{-1}} \qquad H_2(z) = \frac{1}{1-0.43z^{-1}}$$

$$e_1(n) = T_1(z) = H(z) = \frac{1}{(1-0.15z^{-1})(1-0.43z^{-1})}$$

$$e_2(n) = T_2(z) = H_2(z) = \frac{1}{1-0.43z^{-1}} = \frac{z}{z-0.43}$$

$$T_2(z^{-1}) = \frac{1}{1-0.43(z^{-1})^{-1}} = \frac{1}{1-0.43z}$$

$$\sigma_{e\,tot\,o/p}^2 = \sigma_{e\,o/p}^2 + \sigma_{e2\,o/p}^2 \quad \text{(Case 1)}$$
(Case 1)

Direct form noise power

$$\sigma_{e\,o/p}^2 = 3.7192 \times 10^{-3}$$

$$\sigma_{e2\,o/p}^2 = \sigma_e^2 \sum_{i=1}^{N} \text{Res}\left[T_2(z) \cdot T_2(z^{-1})z^{-1}\right]_{z=P_i}$$

$$= 1.3021 \times 10^{-3} \left[\cancel{(z-0.43)} \frac{\cancel{z}}{\cancel{(z-0.43)}} \cdot \frac{1}{1-0.43z} \cdot \cancel{z^{-1}} \right]_{z=0.43}$$

$$\sigma_{e2\,o/p}^2 = 1.3021 \times 10^{-3} \times \frac{1}{1-0.43 \times 0.43} = 1.597 \times 10^{-3}$$

$$\sigma_{e\,tot\;o/p\;(case\ 1)}^2 = 3.7192\times10^{-3} + 1.597\times10^{-3}$$

$$\sigma_{e\,tot\;o/p\;(case\ 1)}^2 = 5.3166\times10^{-3}$$

Case (ii)

$$H(z) = H_2(z)\,H_1(z)$$

$$e_1(n) = T_1(z) = \frac{1}{(1-0.43z^{-1})(1-0.015z^{-1})}$$

$$e_2(n)\quad T_2(z) = \frac{1}{(1-0.15z^{-1})} = \frac{z}{z-0.015}$$

$$T_2(z^{-1}) = \frac{1}{1-0.15(z^{-1})^{-1}} = \frac{1}{1-0.15z}$$

$$\sigma_{e\,tot\;o/p\;(case\ 2)}^2 = \sigma_e^2\ o/p + \sigma_{e3}^2\ o/p\ (case 2)$$

$$\sigma_e^2\ o/p = 3.7192\times10^{-3}$$ ← direct form o/p noise.

$$\sigma_{e3}^2\ o/p = \sigma_e^2 \sum_{i=1}^{N} \mathrm{Res}\left[T_2(z)\,T_2(z^{-1})\,z^{-1}\right]\Big|_{z=p_i}$$

$$= 1.3021\times10^{-3}\left[\cancel{(z-0.015)}\,\frac{\cancel{z}}{\cancel{(z-0.015)}}\,\frac{1}{1-0.015z}\,\cancel{z^{-1}}\Big|_{z=0.15}\right]$$

$$= 1.3021\times10^{-3}\times1.023$$

$$\sigma_{e3}^2\ o/p = 1.332\times10^{-3}$$

$$\sigma_e^2 \text{ tot o/p} \atop (\text{case 2}) = 3.7192\times10^{-3} + 1.0332\times10^{-3}$$

$$\sigma_e^2 \text{ tot o/p} \atop (\text{Case 2}) = 5.0512 \times 10^{-3}$$

Problem 8:

Explain the characteristics of a limit cycle oscillation with respect to the system described by the equation

$$y(n) = 0.82\, y(n-1) + x(n)$$

4 bit rounding with input $x(n) = 0.875$ for $n = 0$ and $x(n) = 0$ for $n \neq 0$. Determine the dead band of the filter.

Solution:

Given $B = 4$ $x(0) = 0.875$
$a = 0.82$

$$y(n) = 0.82\, y(n-1) + x(n)$$

$n = 0$

$$y(0) = 0.82\, \cancel{y(-1)}^{0} + \cancel{x(0)}^{0.875}$$

$$y(0) = 0.875$$

0.875×2
$\boxed{1}.75 \times 2$
$\boxed{1}.5 \times 2$
$\boxed{1}.0 \times 2$
0.0

0.1110

0.111: $1\times2^{-3} = 0.125$, $1\times2^{-2} = 0.25$, $1\times2^{-1} = 0.5$ $= 0.875$

$\hat{y}(0) = 0.875$

$n = 1$

$$y(1) = 0.82 \times \overset{0.875}{\cancel{\hat{y}(1-1)}} + \overset{0}{\cancel{x(1)}}$$

$y(1) = 0.7175$

$$\begin{array}{l} 0.7175 \times 2 \\ \hline 1.435 \times 2 \\ \hline 0.87 \times 2 \\ \hline 1.74 \times 2 \\ \hline 1.48 \end{array}$$

0.1011
1

0.110

$= 0.110$

$0 \times 2^{-3} = 0$
$1 \times 2^{-2} = 0.25$
$1 \times 2^{-1} = 0.5$

0.75

$\hat{y}(1) = 0.75$

$n = 2$

$$y(2) = 0.82\ \overset{0.75}{\cancel{\hat{y}(2-1)}} + \overset{0}{\cancel{x(2)}}$$

$y(2) = 0.615$

$$\begin{array}{l} 0.615 \times 2 \\ \hline 1.23 \times 2 \\ \hline 0.46 \times 2 \\ \hline 0.92 \times 2 \\ \hline 1.84 \end{array}$$

0.1001
1

0.101

$= 0.101$

$1 \times 2^{-3} = 0.125$
$0 \times 2^{-2} = 0$
$1 \times 2^{-1} = 0.5$

0.625

$\hat{y}(2) = 0.625$

$n = 3$

$$y(3) = 0.82\ \overset{0.625}{\cancel{\hat{y}(3-1)}} + \overset{0}{\cancel{x(3)}}$$

$= 0.5125$

$= 0.100$ $\quad = 0$, $= 0$, $1 \times 2^{-1} = 0.5$

$\hat{y}(3) = 0.5$

0.5125×2
1.025×2
0.05×2
0.1×2
0.2

0.1000
0
0.100

$n = 4$

$y(4) = 0.82\, \hat{y}(\cancel{4-1}) + \cancel{x(4)}$ $\quad$ (0.5, 0)

$= 0.41$

$= 0.011$

$1 \times 2^{-3} = 0.125$
$1 \times 2^{-2} = 0.25$
$1 \times 2^{-1} = 0$
0.375

0.41×2
0.82×2
1.64×2
1.28×2
0.56

$0.011\cancel{1}$
0
0.011

$\hat{y}(4) = 0.375$

$n = 5$

$y(5) = 0.82\, \hat{y}(\cancel{5-1}) + \cancel{x(5)}$ $\quad$ (0.375, 0)

$= 0.3075$

$= 0.010$

$0 \times 2^{-3} = 0$
$1 \times 2^{-2} = 0.25$
$0 \times 2^{-1} = 0$
0.25

0.3075×2
0.615×2
1.23×2
0.46×2
0.92

0.0010
0
0.010

$\hat{y}(5) = 0.25$

$n = 6$

$y(6) = 0.82\, \hat{y}(\cancel{6-1}) + \cancel{x(6)}$ $\quad$ (0.25, 0)

$y(6) = 0.205$

$y(6) = 0.010$

$0 \times 2^{-3} = 0$

$1 \times 2^{-2} = 0.25$

$0 \times 2^{-1} = 0$

0.25

0.205×2

0.41×2

0.82×2

1.64×2

1.28

0.0001

0.010

$\hat{y}(6) = 0.25$.

$$\text{Dead band} = \frac{\pm 2^{-B}}{1 - |a|}.$$

$$= \frac{\pm 2^{-4}}{1 - 0.82}$$

$$= \pm 0.347222 \times 0.82$$

$$= 0.284722$$

$$= 0.010$$

$0 \times 2^{-3} = 0$

$1 \times 2^{-2} = 0.25$

$0 \times 2^{-1} = 0$

0.25

0.284722×2

0.569×2

1.388×2

0.277×2

0.555

0.010

0.010

Dead band $= \pm 0.25$

The characteristics of Limit Cycle oscillation for $y(n) = 0.82\,y(n-1) + x(n)$ is

n	$x(n)$	$y(n)$	$\hat{y}(n)$
0	0.875	0.875	0.875
1	0	0.7175	0.75
2	0	0.615	0.625
3	0	0.5125	0.5
4	0	0.41	0.375
5	0	0.3075	0.25
6	0	0.205	0.25
7	0	0.205	0.25

Dead band