But $\sigma_{01}^2 = \sigma_{02}^2 = \sigma_e^2 \sum_{n=-\infty}^{\infty} h^2(n)$, *which gives us*

$$\sigma_0^2 = 2.\frac{2^{-2b}}{12} \sum_{n=0}^{\infty} r^{2n} \frac{\sin^2(n+1)\theta}{\sin^2\theta}$$

$$= 2.\frac{2^{-2b}}{12} \frac{1}{2\sin^2\theta} \sum_{n=0}^{\infty} r^{2n}\left[1-\cos 2(n+1)\theta\right] \qquad \therefore cos2\theta = 1-2sin^2\theta$$

$$= \frac{2^{-2b}}{6} \frac{1}{2\sin^2\theta}\left[\sum_{n=0}^{\infty} r^{2n} - \sum_{n=0}^{\infty} r^{2n}\cos 2(n+1)\theta\right]$$

$$= \frac{2^{-2b}}{6} \frac{1}{2\sin^2\theta}\left[\frac{1}{1-r^2} - \frac{1}{2}\left(\sum_{n=0}^{\infty} r^{2n} e^{j2(n+1)\theta} + \sum_{n=0}^{\infty} r^{2n} e^{-j2(n+1)\theta}\right)\right]$$

$$= \frac{2^{-2b}}{6} \frac{1}{2\sin^2\theta}\left[\frac{1}{1-r^2} - \frac{1}{2}\left(\frac{e^{j2\theta}}{1-r^2e^{2j\theta}} + \frac{e^{-j2\theta}}{1-r^2e^{-2j\theta}}\right)\right]$$

$$= \frac{2^{-2b}}{6} \frac{1}{2\sin^2\theta}\left[\frac{1}{1-r^2} - \frac{\cos 2\theta - r^2}{1-2r^2\cos 2\theta + r^4}\right]$$

$$= \frac{2^{-2b}}{6} \frac{1}{2\sin^2\theta}\left[\frac{(1+r)^2(1-\cos 2\theta)}{(1-r^2)(1-2r^2\cos 2\theta + r^4)}\right]$$

$$= \frac{2^{-2b}}{6} \frac{(1+r)^2}{(1-r^2)(1-2r^2\cos 2\theta + r^4)}$$

****

Co-efficient quantization error

- We know that the IIR Filter is characterized by the system function

$$H(Z) = \frac{\sum_{k=0}^{N} b_k z^{-k}}{1+\sum_{k=1}^{N} a_k z^{-k}}$$

- After quantizing ,

$$[H(Z)]_q = \frac{\sum_{k=0}^{N} [b_k]_q z^{-k}}{1+\sum_{k=1}^{N} [a_k]_q z^{-k}}$$

Where $[a_k]_q = a_k + \Delta a_k$

$[b_k]_q = b_k + \Delta b_k$

- The quantization of filter coefficients alters the positions of the poles and zeros in z-plane.
  1. If the poles of desired filter lie close to the unit circle, then the quantized filter poles may lie outside the unit circle leading into instability of filter.
  2. Deviation in poles and zeros also lead to deviation in frequency response.

****

Consider a second order IIR filter with $H(z) = \frac{1.0}{(1-0.5z^{-1})(1-0.45z^{-1})}$ **find the effect on quantization o n pole locations of the given system function in direct form and in cascade form. Take b=3bits.[Apr/May-10] [Nov/Dec-11]**

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

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

The roots of the denominator of H(z) are the original poles of H(z). let the original poles of H(z) be p_1 and p_2.

Here p_1=0.5 and p_2=0.45

Direct form I:

$$H(z) = \frac{1.0}{(1 - 0.5z^{-1})(1 - 0.45z^{-1})}$$

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

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

Let us quantize the coefficients by truncation.

$.95_{10}$ —Convert to Binary→ $.1111_2$ —Truncate to 3-bits→ $.111_2$ —Convert to decimal→ $.875_{10}$

$.225_{10}$ —Convert to Binary→ $.0011_2$ —Truncate to 3-bits→ $.001_2$ —Convert to decimal→ $.125_{10}$

Let $\overline{H}(z)$ be the transfer function of the IIR system after quantizing the coefficients.

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

$$let\ \overline{H}(z) = \frac{Y(z)}{X(z)} = \frac{1}{1 - 0.875z^{-1} + 0.125z^{-2}}$$

On cross multiplying the above equation we get,

$$Y(z) - 0.875z^{-1}Y(z) + 0.125z^{-2}Y(z) = X(z)$$

$$Y(z) = X(z) + 0.875z^{-1}Y(z) - 0.125z^{-2}Y(z)$$

Cascade form:

Given that

$$H(z) = \frac{1.0}{(1 - 0.5z^{-1})(1 - 0.45z^{-1})}$$

In cascade realization the system can be realized as cascade of first order sections.

$H(z)=H_1(z)+H_2(z)$

Where,

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

Let us quantize the coefficients of $H_1(z)$ and $H_2(z)$ by truncation.

$.5_{10}$ —Convert to Binary→ $.1000_2$ —Truncate to 3-bits→ $.100_2$ —Convert to decimal→ $.5_{10}$

$.45_{10}$ —Convert to Binary→ $.0111_2$ —Convert to 3-bits→ $.011_2$ —Convert to decimal→ $.375_{10}$

let , $\overline{H}_1(z)$ and $\overline{H}_2(z)$ be the transfer function of the first-order sections after quantizing the coefficients.

$$\overline{H_1(z)} = \frac{1}{1 - 0.5z^{-1}}$$

$$\overline{H_2(z)} = \frac{1}{1 - 0.375z^{-1}}$$

$$let, \overline{H}_1(z) = \frac{Y_1(z)}{X(z)} = \frac{1}{1 - 0.5z^{-1}}$$

$$Y_1(z) - 0.5\,z^{-1}\,Y_1(z) = X(z)$$

$$Y_1(z) = X(z) + 0.5\,z^{-1}\,Y_1(z)$$

$$let, \underline{H}_2(z) = \frac{Y(z)}{Y_1(z)} = \frac{1}{1 - 0.375z^{-1}}$$

on cross multiplying the above equation we get,

$$Y(z) - 0.375\,z^{-1}\,Y(z) = Y_1(z)$$

$$Y(z) = Y_1(z) + 0.375\,z^{-1}Y(z)$$

****

Round off effects and overflow in digital filter:

***Explain in detail about round off effects in digital filters.**

- The presence of one or more quantizer in the realization of a digital filter results in a non-linear device. i.e. recursive digital filter may exhibit undesirable oscillations in its output
- In the finite arithmetic operations, some registers may overflow if the input signal level becomes large.
- These overflow represents non-linear distortion leading to limit cycle oscillations
- There are two types of limit cycle oscillations which includes
  1. Zero input limit cycle oscillations (Low amplitude compared to overflow limit cycle oscillations)
  2. Over flow limit cycle oscillations.

Zero input limit cycle oscillations

- The arithmetic operations produces oscillations even when the input is zero or some non zero constant values. Such oscillations are called zero input limit cycle oscillations.

Overflow limit cycle oscillations

- The limit cycle occurs due to the overflow of adder is known as overflow limit cycle oscillations.

Dead Band:

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

$$|y(n-1)| \le \frac{2^{-b}}{2}$$
$$(1-|a|)$$

Consider a first order filter

$$y(n) = ay(n-1) + x(n); \qquad n > 0$$

After rounding the product

$$y_q(n) = Q[a * y(n-1)] + x(n);$$

The round off error

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

where, e_r → difference between the quantized value and the actual value.

$$Q[ay(n-1)] - ay(n-1)] \le \frac{2^{-b}}{2}$$

The dead band of the filter for the limit cycle oscillations are

$$Q[ay(n-1)] = \begin{cases} y(n-1) & a > 0 \\ -y(n-1) & a < 0 \end{cases}$$

$$|y(n-1)| \quad a|y(n-1)| \le \frac{2^{-b}}{2}$$

$$y(n-1)\left(1-|a|\right) \le \frac{2^{-b}}{2}$$

Dead band of the filter, $|y(n-1)| \le \dfrac{\frac{2^{-b}}{2}}{\left(1-|a|\right)}$

****

Problem: Consider a 1st order FIR system equation $y(n) = x(n) + ay(n-1)$ with

$$x(n) = \begin{cases} 0.875 & , n = 0 \\ 0 & , otherwise \end{cases}$$

Find the limit cycle effect and the dead band. Assume b=4 and a=0.95. (Nov/Dec-12)(Nov/Dec-15) [May/June-2016]

Solution:

Given:

$$x(n) = \begin{cases} 0.875 & , n = 0 \\ 0 & , otherwise \end{cases}$$

$$Dead\ band = \frac{2^{-b}}{2\left(1-|a|\right)} = \frac{2^{-4}}{2\left(1-|0.95|\right)} = 0.625$$

$$y(n) = x(n) + 0.95y(n-1)$$

n	***x(n)***	$y(n-1)$	$ay(n-1)$	$Q[ay(n-1)]$ *(round off to 4-bits)*	$y(n) = x(n) + Q[ay(n-1)]$
0	0.875	0	0	0.0000	y(0)=0.875
1	0	0.875	$0.875 * 0.95 = (0.83125)_{10} = (0.11010)_2$	$= (0.1101)_2 = 0.8125$	y(1)=0.8125
2	0	0.8125	$0.8125 * 0.95 = (0.77187)_{10} = (0.110001)_2$	$= (0.1100)_2 = 0.75$	$y(2) = 0.75$
3	0	0.75	$0.75 * 0.95 = (0.7125)_{10} = (0.1011011)_2$	$= (0.1011)_2 = 0.6875$	$y(3) = 0.6875$
4	0	0.6875	$0.6875 * 0.95 = (0.653125)_{10} = (0.101001)_2$	$= (0.1010)_2 = 0.625$	$y(4) = 0.625$
5	0	0.625	$0.625 * 0.95 = (0.59375)_{10} = (0.10011)_2$	$= (0.1010)_2 = 0.625$	$y(5) = 0.625$
6	0	0.625	$0.625 * 0.95 = (0.59375)_{10} = (0.10011)_2$	$= (0.1010)_2 = 0.625$	$y(6) = 0.625$

Conclusion:

The dead band of the filter is 0.625. When limit cycle oscillations.

$n \geq 5$ the output remains constant at 0.625 causing

Finite Word length Effects:

- In the design of FIR Filters, The filter coefficients are determined by the system transfer functions. These filter co-efficient are quantized/truncated while implementing DSP System because of finite length registers.
- Only Finite numbers of bits are used to perform arithmetic operations. Typical word length is 16 bits, 24 bits, 32 bits etc.
- This finite word length introduces an error which can affect the performance of the DSP system.
- The main errors are
  1. Input quantization error
  2. Co-efficient quantization error
  3. Overflow & round off error (Product Quantization error)
- The effect of error introduced by a signal process depend upon number of factors including the
  1. Type of arithmetic
  2. Quality of input signal
  3. Type of algorithm implemented

1. Input quantization error

- The conversion of 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 the input signal by a fixed number of digits in A/D conversion process.

2. Co-efficient quantization error

- The filter coefficients are compared to infinite precision. If they are quantized the frequency response of the resulting filter may differ from the desired frequency response.
  i.e poles of the desired filter may change leading to instability.

3. Product Quantization error

- It arises at the output of the multiplier
- When a 'b' bit data is multiplied with another 'b' bit coefficient the product ('2b' bits) should be stored in 'b' bits register. The multiplier Output must be rounded or truncated to 'b' bits. This known as overflow and round off error.

****

Types of number representation:

There are two common forms that are used to represent the numbers in a digital or any other digital hardware.

1. Fixed point representation
2. Floating point representation

*** Explain the various formulas of the fixed point representation of binary numbers.**

1. Fixed point representation

- In the fixed point arithmetic, the position of the binary point is fixed. The bit to the right represents the fractional part of the number and to those to the left represents the integer part.
- For example, the binary number 01.1100 has the value 1.75 in decimal.

  $(0*2^{1)} + (1*2^{0}) + (1*2^{-1}) + (1*2^{-2}) + (0*2^{-3}) = 1.75$

  In general, we can represent the fixed point number 'N' to any desired accuracy by the series

$$N = \sum_{i=n_i}^{n_2} C_i r^i$$

Where, r is called as radix.

- If r=10, the representation is known as decimal representation having numbers from 0 to 9. In this representation the number

$$30.285 = \sum_{i=-3}^{1_2} C_i 10^i$$

$$= (3*10^1) + (0*10^0) + (2*10^{-1}) + (8*10^{-2}) + (5*10^{-3})$$

- If r=2, the representation is known as binary representation with two numbers 0 to 1.
- For example, the binary number

$$110.010 = (1*2^2) + (1*2^1) + (0*2^0) + (0*2^{-1}) + (1*2^{-2}) + (0*2^{-3}) = 6.25$$

Examples:

Convert the decimal number 30.275 to binary form

2	30	
2	15	--0
2	7	--1
2	3	--1
	1	--1

0.55 * 2	➔1.10	➔1
0.10 * 2	➔0.20	➔0
0.20 * 2	➔0.40	➔0
0.40 * 2	➔0.80	➔0
0.80 * 2	➔1.60	➔1
0.60 * 2	➔1.20	➔1
0.20 * 2	➔0.40	➔0

$(30.275)_{10}$ = $(11110.01000110)_2$

0.275 * 2 ➔0.55 ➔0

In fixed point arithmetic =, the negative numbers are represented by 3 forms.

1. Sign-magnitude form
2. One's complement form
3. Two's complement form

1.1 Sign-magnitude form:

- Here an additional bit called sign bit is added as MSB.
  - If this bit is zero → It is a positive number
  - If this bit is one → It is a positive number
- For example
  - 1.75 is represented as 01.110000.
  - -1.75 is represented as 11.110000

1.2 One's complement form:

- Here the positive number is represented same as that in sign magnitude form.
- But the negative number is obtained by complementing all the bits of the positive number
- For eg: the decimal number -0.875 can be represented as
  - $(0.875)_{10}=(0.111000)_2$
  - $(-0.875)_{10}=(1.000111)_2$

0.111000

↓↓↓↓↓↓ (Complement each bit)1.00011

1.3 Two's complement form:

- Here the positive numbers are represented as same in sign magnitude and one's complement form.
- The negative numbers are obtained by complementing all the bits of the positive number and adding one to the least significant bit

$(0.875)_{10}=(0.111000)_2$

↓↓↓↓↓↓ (Complement each bit)

1.000111

\+ 1

1.001000

$(-0.875)_{10}=(1.001000)_2$

Examples:

- **Find the sign magnitude, 1's complement, 2's complement for the given numbers.**
  1. $-\frac{7}{32}$
  2. $-\frac{7}{8}$
  3. $+\frac{7}{8}$

1. $-\frac{7}{32}$

0.21875 * 2 ➔0.43750 ➔0
0.43750 * 2 ➔0.87500 ➔0
0.87500* 2 ➔ 1.750000 ➔1
0.75* 2 ➔1.50 ➔1
0.50* 2 ➔1.00 ➔1

$-\frac{7}{32}=(-0.21875)_{10}=(1.00111)_2$

Sign magnitude form = 1.00111
1's complement form = 1.11000
2's complement form = 1.11001

2. $-\frac{7}{8}$

0.875 * 2 ➔1.75 ➔1
0.750 * 2 ➔1.500 ➔1
0.500 * 2 ➔1.000 ➔1

$-\frac{7}{8}=(-0.875)_{10}=(0.111)_2$

Sign magnitude form = 0.111
1's complement form = 1.000
2's complement form = 1.001

3. $+\frac{7}{8}$

Sign magnitude form = 0.111
1's complement form = 0.111
2's complement form = 0.111

Addition of two fixed point numbers:

- **Add $(0.5)_{10}+(0.125)_{10}$**

$(0.5)_{10}=(0.100)_2$
$(0.125)_{10}=(0.001)_2$

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

- Addition of two fixed point numbers causes an overflow.
  For example

  $(0.100)_2$

  $(0.101)_2$

  $(1.001)_2 = (-0.125)_{10}$ in sign magnitude form

Subtraction of two fixed point numbers:

- Subtraction of two numbers can be easily performed easily by using two's complement representation.
- **Subtract 0.25 from 0.5**

  0.25 * 2 ➔0.50 ➔0 Sign magnitude form = $(0.010)_2$
  0.50 * 2 ➔1.00 ➔1 1's complement form = $(1.101)_2$
  0.00 * 2 ➔0.00 ➔0 2's complement form = $(1.110)_2$

  $(0.5)_{10}$ = $(0.100)_2$
  $-(0.25)_{10}$ = $(1.110)_2$ ➔Two's complement of -0.25
  $(10.010)_2$

  Here the carry is generated after the addition. Neglect the carry bit to get the result in decimal.
  $(0.010)_2 = (0.25)_{10}$
- **Subtract 0.5 from 0.25**

  0.5 * 2 ➔1.00 ➔1 Sign magnitude form = $(0.100)_2$
  0.00 * 2 ➔0.00 ➔0 1's complement form = $(1.011)_2$
  0.00 * 2 ➔0.00 ➔0 2's complement form = $(1.100)_2$

  $(0.25)_{10}$ = $(0.010)_2$
  $-(0.5)_{10}$ = $(1.100)_2$
  $(1.110)_2$

  Here the carry is not generated after the addition. So the result is negative.

Multiplication in fixed point arithmetic:

- Here the sign magnitude components are separated.
- The magnitudes of the numbers are multiplied. Then the sign of the product is determined and applied to the result.
- In the fixed point arithmetic, multiplication of two fractions results in a fraction.
- For multiplications with fractions, overflow can never occur.
- Eg:

  $(0.1001)_2 * (0.0011)_2$ = $(0.00011011)_2$

2. Floating point representation

- Here, a number 'x' is represented by
  $X=M.r^e$
  Where, M → Mantissa which requires a sign bit for representing positive number and negative numbers.
  R → base (or) radix
  e → exponent which require an additional and it may be either positive or negative.
  - For eg, 278 can be represented in floating point representation.

    $278 = \frac{278 \text{ X } 1000}{1000} = 0.278*10_3$

    0.278 → Mantissa (M)
    10 → base (or) radix (r)
    3 → exponents (e)
- Similarly, to represent a binary floating point number
  $X=M.2^e$ in which the fractional part of a number should fall (or) lie in the range of 1/2 to 1.

  $5 = \frac{5 \text{ X } 8}{8}$ $0.625 \text{ X} 2^3$

Mantissa (M) = 0.625
Base (or) radix (r) = 2
Exponent (e) = 3

- Some decimal numbers and their floating point representations are given below:

4.5 → $0.5625 \text{ X } 2^3$ $=0.1001 \text{ X } 2^{011}$
1.5 → $0.75 \text{ X } 2^1$ $=0.1100 \text{ X } 2^{001}$
6.5 → $0.8125 \text{ X } 2^3$ $=0.1100 \text{ X } 2^{011}$
0.625 → $0.625 \text{ X } 2^0$ $=0.1010 \text{ X } 2^{000}$

- 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.
- To represent floating point in multiplication

  Consider $X_1 = M_1 r^{e_1}$

  $X_2 = M_2 r^{e_2}$

  $X_1 X_2 = (M_1 * M_2) r^{(e_1 + e_2)}$

Example

Given $X_1 = 3.5*10^{-12}$, $X_2 = 4.75*10^{6}$. Find the product $X_1 X_2$

$X=(3.5 \text{ X } 4.75)\ 10^{(-12+6)}$

$= (16.625)10^{-6}$ → in decimal

In binary: $(1.5)_{10}$ X $(1.25)_{10}$ = $(2^1 0.75)$ X $(2^1 0.625)$

= 2^{001} X 0.1100 X 2^{001} X 0.1010

= 2^{010} X 0.01111

Addition and subtraction:

- Here the exponent of a smaller number is adjusted until it matches the exponent of a larger number.
- Then, the mantissa are added or subtracted
- The resulting representation is rescaled so that its mantissa lies in the range 0.5 to 1.
- Eg: **Add $(3.0)_{10}$ & $(0.125)_{10}$**

  $(3.0)_{10}$ $=2^{010}$ X 0.1100 = r^{e_1} X M_1

  $(0.125)_{10}$ $=2^{000}$ X 0.0010 = r^{e_2} X M_2

  Now adjust e_2 Such that $e_1=e_2$

  $(0.125)_{10}$ $=2^{010}$ X 0.0000100

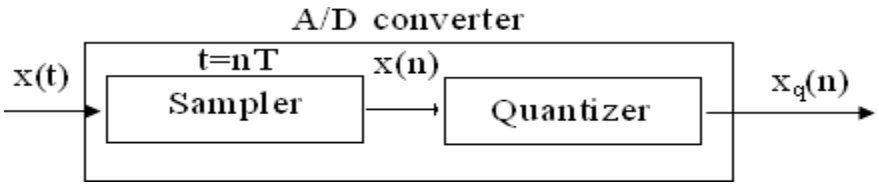
  Addition →2^{010} (0.110000 + 0.0000100) →2^{010} X 0.110010

  Subraction →2^{010} X 1.001101

Compare floating point with fixed point arithmetic.

Sl.No	Fixed point arithmetic	Floating point arithmetic
1	Fast operation	Slow operation
2	Relatively economical	More expensive because of costlier hardware
3	Small dynamic range	Increased Dynamic range
4	Round off errors occurs only for addition	Round off errors can occur with addition and multiplication
5	Overflow occur in addition	Overflow does not arise
6	Used in small computers	Used in large general purpose computers.

****



- The analog signal is converted into digital signal by ADC
- At first, the signal x(t) is sampled at regular intervals t=nT, where n=0,1,2… to create sequence x(n). This is done by a sampler.
- Then the numeric equivalent of each sample x(n) is expressed by a finite number of bits giving the sequence $x_q(n)$
- The difference signal $e(n)= x_q(n)- x(n)$ is called quantization noise (or) A/D conversion noise.
- Let us assume a sinusoidal signal varying between +1 & -1 having a dynamic range 2
- ADC employs (b+1) bits including sign bit. In this case, the number of levels available for quantizing x(n) is 2^{b+1}.
- The interval between the successive levels is

$$q = \frac{2}{2^{b+1}} = 2^{-b}$$

Where q → quantization step size

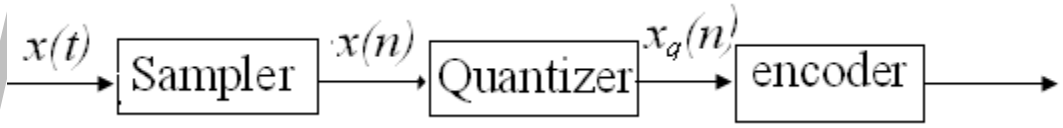
If b=3 bits, then $q=2^{-3}=0.125$

****

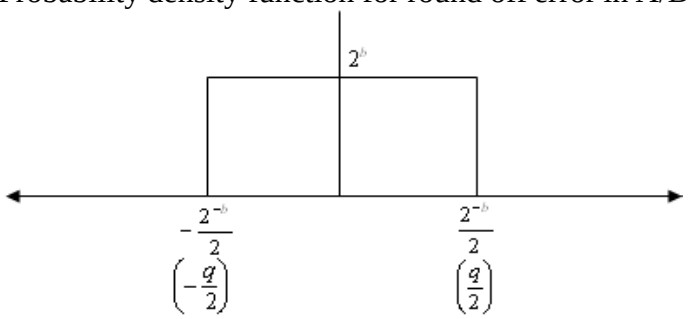
Quantization Noise power:

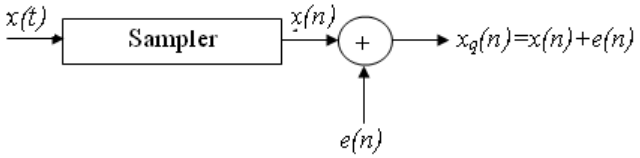
Input Quantization error:

***Derive the equation for quantization noise power (or) Steady state Input Noise Power.**



Probability density function for round off error in A/D conversion is





If rounding is used for quantization, which is bounded by $-\frac{q}{2} \le e(n) \le \frac{q}{2}$, then the error lies between $-\frac{q}{2}$ to $\frac{q}{2}$ with equal probability, where q→ quantization step size.

Properties of analog to digital conversion error, e(n):

1. The error sequence e(n) is a sample sequence of a stationary random process.
2. The error sequence is uncorrelated with x(n) and other signals in the system.
3. The error is a white noise process with uniform amplitude probability distribution over the range of quantization error.

The variance of e(n) is given by

$$\sigma_e^2 = E\left[e^2(n)\right] - E^2\left[e(n)\right] \text{----------} >(1)$$

Where $E\left[e^2(n)\right]$ → Average of $e^2(n)$

$$E\left[e^2(n)\right] = \int_{-\infty}^{\infty} e^2(n)p(e)de \text{ ---------- >(2)}$$

$$p(e) = \frac{1}{q}, -\frac{q}{2} \le e(n) \le \frac{q}{2} \text{ ---------- >(3)}$$

Substituting (3) in (2)

$$E\left[e^2(n)\right] = \int_{-\frac{q}{2}}^{\frac{q}{2}} e^2(n)\frac{1}{q}de$$

$$E\left[e^2(n)\right] = \frac{1}{q}\int_{-\frac{q}{2}}^{\frac{q}{2}} e^2(n)de \text{ ---------- >(4)}$$

$$E\left[e(n)\right] = 0$$

$$E^2\left[e(n)\right] = 0 \text{ ..... >(5)}$$

Substituting (4) and (5) in (1)

$$\sigma_e^2 = \frac{1}{q}\int_{-\frac{q}{2}}^{\frac{q}{2}} e^2(n)de - 0$$

$$= \frac{1}{q}\left[\frac{e^3}{3}\right]_{-\frac{q}{2}}^{\frac{q}{2}}$$

$$= \frac{1}{3q}\left[\left(\frac{q}{2}\right)^3 - \left(-\frac{q}{2}\right)^3\right]$$

$$= \frac{1}{3q}\left[\frac{q^3}{8} - \left(-\frac{q^3}{8}\right)\right]$$

$$= \frac{1}{3q}\left[\frac{q^3}{8} + \frac{q^3}{8}\right]$$

$$= \frac{1}{3q}\left[\frac{2q^3}{8}\right]$$

$$\sigma_e^2 = \frac{q^2}{12} \text{ ----------}>(6)$$

In general, $\frac{1}{2^b} = 2^{-b} = q$ ----------$>(7)$

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

$$\sigma_e^2 = \frac{2^{-2b}}{12} \text{ ----------}>(8)$$

Equation (8) is known as the steady state noise power due to input quantization.

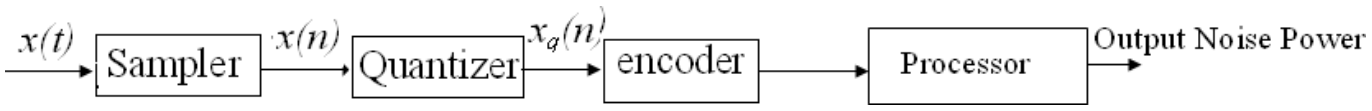
$$q = \frac{R}{2^b} \qquad q = \frac{R}{2^b - 1}$$

→ in two's complement representation.

→ in sign magnitude (or) one's complement representation.

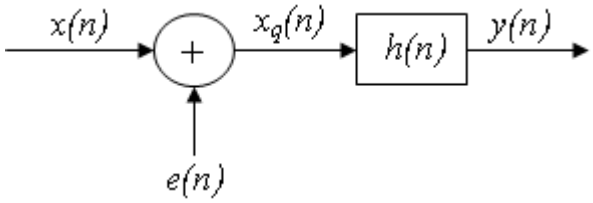
R → Range of analog signal to be quantized.

Steady state Output Noise power:



After quantization, we have noise power σ_e^2 as input noise power. Therefore, Output noise power of system is given by

$$\sigma_{eo}^2 = \sigma_e^2 \left[\sum_{n=0}^{\infty} h^2(n) \right] \quad \text{----------} \quad (9)$$



where h(n) → impulse response of the system.

Let error E(n) be output noise power due to quantization

Error $E(n) = e(n) * h(n)$

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

The variance of error E(n) is called output noise power, σ_e^2.

By using Parseval's theorem,

$$\sigma_{eo}^2 = \sigma_e^2 \sum_{n=0}^{\infty} h^2(n)$$

$$= \sigma_e^2 \frac{1}{2\pi j} \oint H(Z) H(Z^{-1}) \frac{dZ}{z}$$

Where the closed contour integration is evaluated using the method of residue by taking only the poles that lie inside the unit circle.

Z transform of h(n), $H(Z) = \sum_{n=0}^{\infty} h(n) z^{-n}$

Z transform of $h^2(n)$ = $Z[h^2(n)]$ $= \sum_{n=0}^{\infty} h^2(n) z^{-n} = \sum_{n=0}^{\infty} h(n) h(n) z^{-n}$ ---------- (10)

By Inverse Z transform,

$$h(n) = \frac{1}{2\pi j}\oint H(Z)Z^{n-1}dZ \text{ ----------} > (11)$$

Substituting (11) in (10)

$$\sum_{n=0}^{\infty} h^2(n)z^{-n} = \sum_{n=0}^{\infty} \frac{1}{2\pi j}\int H(Z)Z^{n-1}dZ \; h(n)z^{-n}$$

$$= \frac{1}{2\pi j}\oint H(Z)\left[\sum_{n=0}^{\infty} h(n)Z^{-1}\right]dZ$$

$$\sum_{n=0}^{\infty} h^2(n) = \frac{1}{2\pi j}\oint H(Z)\left[\sum_{n=0}^{\infty} h(n)Z^{-1}\right]\frac{dZ}{Z^{-n}}$$

$$= \frac{1}{2\pi j}\oint H(Z)\left[\sum_{n=0}^{\infty} h(n)\left(Z^{-n}\right)^{-1}Z^{-1}dZ\right]$$

$$\sum_{n=0}^{\infty} h^2(n) = \frac{1}{2\pi j}\oint H(Z)H\left(Z^{-1}\right)\frac{dZ}{Z} \text{ ----------} > (12)$$

Substituting (12) in (9)

$$\sigma^2 = \sigma^2\left[\frac{1}{2\pi j}\int H(Z)H\left(Z^{-1}\right)Z^{-1}dZ\right]$$

Problem:

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

$H(z) = \frac{(1-a)z}{z-a}$ *for* $0 < a < 1$. **Find the steady state output noise power due to quantization at the output of the digital filter. [Nov/Dec-2015]**

Solution:

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

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

Substituting $H(z)$ *and* $H(z^{-1})$ *in equation (1), we have*

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

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

$$= \sigma_e^2\left[(z-a)\frac{(1-a)^2 z^{-1}}{(z-a)(z^{-1}-a)} + 0\right]$$

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

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

****

Find the steady state variance of the noise in the output due to quantization of input for the first order filter. [Apr/May'11] [Nov/Dec-2016]

$$y(n) = ay(n-1) + x(n)$$

Solution:

The impulse response for the above filter is given by $h(n) = a^n u(n)$

$$\sigma_\in^2 = \sigma_e^2\sum_{k=0}^{\alpha} h^2(n)$$

$$= \sigma_e^2\sum_{k=0}^{\alpha} a^{2n}$$

$$= \sigma_e^2\left[1 + a^2 + a^4 + \ldots\infty\right]$$

$$= \sigma_e^2\frac{1}{1-a^2}$$

$$= \frac{2^{-2b}}{12}\left[\frac{1}{1-a^2}\right] \qquad (or)$$

Taking Z-transform on both sides we have

$$Y(z) = az^{-1}Y(z) + X(z)$$

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1}{1-az^{-1}} = \frac{z}{z-a}$$

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

We know

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

Substituting $H(z)$ and $H(z^{-1})$ values in the above equation we get

$$\sigma_e^2 = \sigma_\in^2 \frac{1}{2\pi j}\oint_c \frac{z}{z-a}\frac{z^{-1}}{z^{-1}-a}z^{-1}dz$$

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

$$= \sigma_e^2 \begin{bmatrix} \text{residue of } \dfrac{z^{-1}}{(z-a)(z^{-1}-a)} \text{ at } z=a \\ + \text{residue of } \dfrac{z^{-1}}{(z-a)(z^{-1}-a)} \text{ at } z=1/a \end{bmatrix}$$

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

$$= \sigma_e^2 \frac{a^{-1}}{a^{-1}-a} = \sigma_e^2 \frac{1}{1-a^2}$$

****

The output of the A/D converter is applied to a digital filter with the system function

$$H(Z) = \frac{0.45Z}{Z-0.72}$$

Find the output noise power of the digital filter, when the input signal is quantized to 7 bits.

Given:

$$H(Z) = \frac{0.45Z}{Z-0.72}$$

Solution:

$$H(Z)H(Z^{-1})Z^{-1} = \frac{0.45Z}{Z-0.72} \times \frac{0.45Z^{-1}}{Z^{-1}-0.72} \times Z^{-1}$$

$$= \frac{0.45^2 Z^{-1}}{(Z-0.72)\left(\frac{1}{} - 0.72\right)}$$

$$= \frac{0.2025Z^{-1}}{(Z-0.72)\left(\frac{1-0.72Z}{Z}\right)}$$

$$= \frac{0.2025Z^{-1}Z}{(Z-0.72)\left(Z-\frac{1}{0.72}\right)}$$

$$= \frac{-0.28125}{(Z-0.72)(Z-1.3889)}$$

Now the poles of $H(Z)H(Z^{-1})Z^{-1}$ are $p_1=0.72$, $p_2=1.3889$

Output noise power due to input quantization

$$\sigma_{eo}^2 = \sigma_e^2 \left[\frac{1}{2\pi j}\oint H(Z)H(Z^{-1})Z^{-1}dZ\right]$$

$$= \sigma_e^2 \sum_{i=1}^{N} \text{Res}\left[H(Z)H(Z^{-1})Z^{-1}\right]_{z=p_i}$$

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

Where $p_1, p_2, \ldots p_n$ are the poles of $H(Z)H(Z^{-1})\,Z^{-1}$ that lies inside the unit circle in z-plane.

$$\sigma_{eo}^2 = \sigma_e^2 \times (Z - 0.72) \times \frac{-0.28125}{(Z - 0.72)(Z - 1.3889)}\Big|_{Z=0.72}$$

$$= \sigma_e^2 \times \frac{-0.28125}{0.72 - 1.3889}$$

$$= 0.4205\sigma_e^2$$

****

Consider the transfer function $H(z) = H_1(z)H_2(z)$ **where** $H_1(z) = \frac{1}{1 - a_1 z^{-1}}$ *and* $H_2(z) = \frac{1}{1 - a_2 z^{-1}}$

Find the output round off noise power. Assume $\alpha_1 = 0.5$ *and* $\alpha_2 = 0.6$ **and find output round off noise power.**

Solution:

The round off noise model for $H(z) = H_1(z)H_2(z)$ is given by,

From the realization we can find that the noise transfer function seen by noise source $e_1(n)$ is H(z), where,

$$H(z) = \frac{1}{(1 - a_1 z^{-1})(1 - a_2 z^{-1})} \quad \text{----------(1)}$$

Whereas, the noise transfer function seen by $e_2(n)$ *is,*

$$H_2(z) = \frac{1}{(1 - a_2 z^{-1})} \quad \text{----------(2)}$$

The total steady state noise variance can be obtained, we have

$$\sigma_0^2 = \sigma_{01}^2 + \sigma_{02}^2 \quad \text{----------(3)}$$

$$\sigma_{01}^2 = \frac{1}{2\pi j}\oint_c H(z)H(z^{-1})z^{-1}dz$$

$$= \sigma_e^2 \frac{1}{2\pi j}\oint_c \frac{1}{1 - a_1 z^{-1}} \frac{1}{1 - a_2 z^{-1}} \frac{1}{1 - a_1 z} \frac{1}{1 - a_2 z} z^{-1}dz$$

$$= \sigma_e^2 \left[\sum \text{of residue of } H(z)H(z^{-1})z^{-1} \text{ at poles } z = a_1, z = a_2, z = \frac{1}{a_1} \text{ and } z = \frac{1}{a_2}\right]$$

If a1 and a2 are less than the poles $z=1/a_1$ and $z=1/a_2$ lies outside of the circle $|z| = 1$. So, the residue of H(z) $H(z^{-1})$ z^{-1}at $z=1/a_1$ and $z=1/a_2$ are zero. Consequently we have,

$$\sigma_{01}^2 = \left[\sum of\ residue\ of\ H(z)H(z^{-1})z^{-1}\ at\ po1es\ z = a_1, z = a_2\right]$$

$$= \left[(z-a_1)\frac{z^{-1}}{(1-a_1z^{-1})(1-a_2z^{-1})(1-a_1z)(1-a_2z)}\right|_{z=a_1} + (z-a_2)\frac{z^{-1}}{(1-a_1z^{-1})(1-a_2z^{-1})(1-a_1z)(1-a_2z)}\Big|_{z=a_2}\Bigg]$$

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

$$\sigma_{01}^2 = \sigma_e^2\left[\frac{a_1}{a_1-a_2}\cdot\frac{1}{1-a_1^2}\cdot\frac{1}{1-a_1a_2} + \frac{a_2}{a_2-a_1}\cdot\frac{1}{1-a_2^2}\cdot\frac{1}{1-a_1a_2}\right] \text{----------(4)}$$

In the same way,

$$\sigma_{02}^2 = \frac{\sigma_e^2}{2\pi j}\oint_c H(z)H(z^{-1})z^{-1}dz$$

$$= \frac{\sigma_e^2}{2\pi j}\oint_c \frac{1}{1-a_2 z}\cdot\frac{1}{1-a_2 z^{-1}}z^{-1}dz$$

$$= \sigma_e^2\left[\frac{z^{-1}}{(1-a_2 z^{-1})(1-a_2 z)}(z-a_2)\right]_{z=a_2}$$

$$= \sigma_e^2\left[\frac{z^{-1}}{(1-a_2 z^{-1})(1-a_2 z)}(z-a_2 z^{-1})\right]_{z=a_2}$$

$$= \sigma_e^2\left[\frac{1}{1-a_2^2}\right] \quad \text{----------(5)}$$

$$\sigma_0^2 = \sigma_e^2\left[\frac{1}{1-a_2^2} + \frac{a_1}{a_1-a_2}\cdot\frac{1}{1-a_2^2}\cdot\frac{1}{1-a_1a_2} + \frac{a_2}{a_2-a_1}\cdot\frac{1}{1-a_2^2}\cdot\frac{1}{1-a_1a_2}\right]$$

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

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

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

The steady state noise power for $a_1 = 0.5, a_2 = 0.6$ is given by

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

$$\left(\frac{2^{-2b}}{12}\right) = 5.4315$$

****

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

Solution:

Given:

$$H(z) = \frac{1}{1-2r\cos\theta z^{-1}+r^2z^{-2}}$$

The quantization noise model is,

we know, $\sigma_0^{\,2} = \sigma_{01}^{\,2} + \sigma_{02}^{\,2}$

Both noise sources see the same transfer function

$$H(z) = \frac{1}{1-2r\cos\theta z^{-1}+r^2z^{-2}}$$

The impulse response of the transfer function is given by

$$h(n) = r^n \frac{\sin(n+1)\theta}{\sin\theta}\,u(n)$$

Now the steady state output noise variance is,

$$\sigma_0^{\,2} = \sigma_{01}^{\,2} + \sigma_{02}^{\,2}$$

But $\sigma_{01}^2 = \sigma_{02}^2 = \sigma_e^2 \sum_{n=-\infty}^{\infty} h^2(n)$, *which gives us*

$$\sigma_0^2 = 2.\frac{2^{-2b}}{12} \sum_{n=0}^{\infty} r^{2n} \frac{\sin^2(n+1)\theta}{\sin^2\theta}$$

$$= 2.\frac{2^{-2b}}{12} \frac{1}{2\sin^2\theta} \sum_{n=0}^{\infty} r^{2n}\left[1-\cos 2(n+1)\theta\right] \qquad \therefore cos2\theta = 1-2sin^2\theta$$

$$= \frac{2^{-2b}}{6} \frac{1}{2\sin^2\theta} \left[\sum_{n=0}^{\infty} r^{2n} - \sum_{n=0}^{\infty} r^{2n}\cos 2(n+1)\theta\right]$$

$$= \frac{2^{-2b}}{6} \frac{1}{2\sin^2\theta} \left[\frac{1}{1-r^2} - \frac{1}{2}\left(\sum_{n=0}^{\infty} r^{2n} e^{j2(n+1)\theta} + \sum_{n=0}^{\infty} r^{2n} e^{-j2(n+1)\theta}\right)\right]$$

$$= \frac{2^{-2b}}{6} \frac{1}{2\sin^2\theta} \left[\frac{1}{1-r^2} - \frac{1}{2}\left(\frac{e^{j2\theta}}{1-r^2e^{2j\theta}} + \frac{e^{-j2\theta}}{1-r^2e^{-2j\theta}}\right)\right]$$

$$= \frac{2^{-2b}}{6} \frac{1}{2\sin^2\theta} \left[\frac{1}{1-r^2} - \frac{\cos 2\theta - r^2}{1-2r^2\cos 2\theta + r^4}\right]$$

$$= \frac{2^{-2b}}{6} \frac{1}{2\sin^2\theta} \left[\frac{(1+r)^2(1-\cos 2\theta)}{(1-r^2)(1-2r^2\cos 2\theta + r^4)}\right]$$

$$= \frac{2^{-2b}}{6} \frac{(1+r)^2}{(1-r^2)(1-2r^2\cos 2\theta + r^4)}$$

****

Co-efficient quantization error

- We know that the IIR Filter is characterized by the system function

M

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

- After quantizing ,

$$[H(Z)]_q = \frac{\sum_{k=0}^{M} [b_k]_q z^{-k}}{1 + \sum_{k=1}^{N} [a_k]_q z^{-k}}$$

Where $[a_k]_q = a_k + \Delta a_k$

$[b_k]_q = b_k + \Delta b_k$

- The quantization of filter coefficients alters the positions of the poles and zeros in z-plane.

1. If the poles of desired filter lie close to the unit circle, then the quantized filter poles may lie outside the unit circle leading into instability of filter.
2. Deviation in poles and zeros also lead to deviation in frequency response.

****

Consider a second order IIR filter with $H(z) = \frac{1.0}{(1-0.5z^{-1})(1-0.45z^{-1})}$ find the effect on quantization on pole locations of the given system function in direct form and in cascade form. Take b=3bits.

[Apr/May-10] [Nov/Dec-11]

Solution:

Given that,

$$H(z) = \frac{1.0}{(1-0.5z^{-1})(1-0.45z^{-1})}$$

****

Overflow Limit cycle oscillations:

***What are called overflow oscillations? How it can be prevented?**

- We know that the limit cycle oscillation is caused by rounding the result of multiplication.
- The limit cycle occurs due to the overflow of adder is known as overflow limit cycle oscillations.\
- Several types of limit cycle oscillations are caused by addition, which makes the filter output oscilate between maximum and minimum amplitudes.
- Let us consider 2 positive numbers n_1 & n_2

  n_1=0.111→7/8

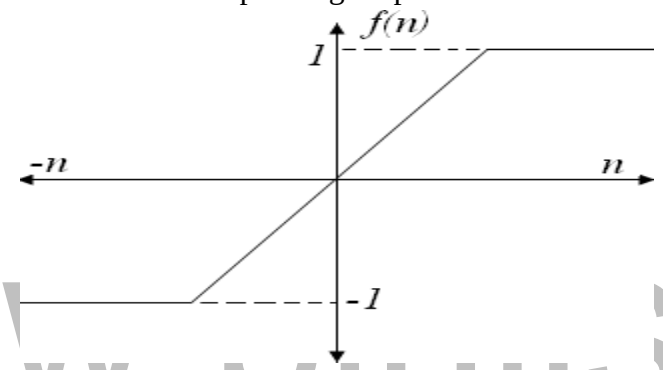
  n_2=0.110→6/8

  n_1 + n2=1.101→-5/8 in sign magnitude form.

  The sum is wrongly interpreted as a negative number.
- The transfer characteristics of an saturation adder is shown in fig below

  where n → The input to the adder

  f(n) → The corresponding output



Saturation adder transfer characteristics

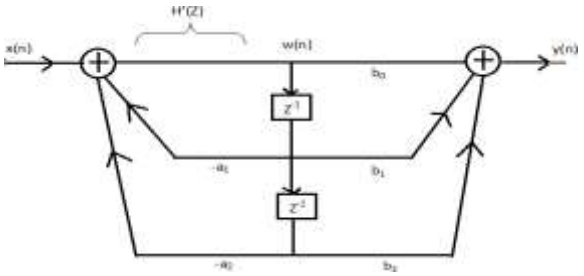
- From the transfer characteristics, we find that when overflow occurs, the sum of adder is set equal to the maximum value.

****

Signal Scaling:

***Explain how reduction of round-off errors is achieved in digital filters. [Nov/Dec-2016]**

- Saturation arithmetic eliminates limit cycles due to overflow, but it causes undeniable 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 input signal and unit sample response between input and any internal summing node in the system to avoid overflow.



Realization of a second order IIR Filter

- Let us consider a second order IIR filter as shown in the above figure. Here a scale factor S_0 is introduced between the input x(n) and the adder 1 to prevent overflow at the output of adder 1.
- Now the overall input-output transfer function is

Now the transfer function

$$H(z) = S_0 \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}$$

$$= S_0 \frac{N(z)}{D(z)}$$

From figure

$$H'(z) = \frac{W(z)}{X(z)} = \frac{S_0}{1 + a_1 z^{-1} + a_2 z^{-2}} = \frac{S_0}{D(z)}$$

$$W(z) = \frac{S_0 X(z)}{D(z)} = S_0 \, S(z)\, x(Z)$$

Where $S(z) = \frac{1}{D(z)}$

wehave

$$w(n) = \frac{S_0}{2\pi} \int S(e^{j\theta}) X(e^{j\theta}) (e^{jn\theta}) \, d\theta$$

$$w(n)^2 = \frac{S_0^2}{2\pi^2} \left| \int S(e^{j\theta}) X(e^{j\theta}) (e^{jn\theta}) \, d\theta \right|^2$$

Using Schwartz inequality

$$w(n)^2 \le S_0^2 \left[\int_{2\pi} S(e^{j\theta})^2 \, d\theta \right] \left[\int_{2\pi} X(e^{j\theta})^2 \, d\theta \right]$$

Applying parsevals theorem

$$w(n)^2 \le S_0^2 \sum_{n=0}^{\infty} x^2(n) \frac{1}{2\pi} \int_{2\pi} \left| S(e^{j\theta}) \right|^2 d\theta$$

if $z = e^{j\theta}$ *then* $dz = je^{j\theta} d\theta$

which gives

$$d\theta = \frac{dz}{jz}$$

By substituting all values

$$w(n)^2 \le S_0^2 \sum_{n=0}^{\infty} x^2(n)$$

$$\sum_{n=0} x^2(n) \frac{1}{2\pi j} \int_c |S(z)|^2 z^{-1} dz$$

$$w(n)^2 \leq S_0^2 \sum_{n=0}^{\infty} x^2(n) \frac{1}{2\pi j} \int_c S(z)S(z^{-1}) z^{-1} dz$$

$$w^2(n) \leq \sum_{n=0}^{\infty} x^2(n) \quad when$$

$$S_0^2 \frac{1}{2\pi j} \int_c S(z)S(z^{-1}) dz = 1$$

Which gives us,

$$S_0^2 = \frac{1}{\frac{1}{2\pi j} \int_c S(z)S(z^{-1}) z^{-1} dz}$$

$$= \frac{1}{\frac{1}{2\pi j} \int_c \frac{z^{-1} dz}{D(z)D(z^{-1})}}$$

$$S_0^2 = \frac{1}{I}$$

Where I=

$$\frac{1}{2\pi j}\int_c \frac{z^{-1}dz}{D(z)D(z^{-1})}$$

Note:

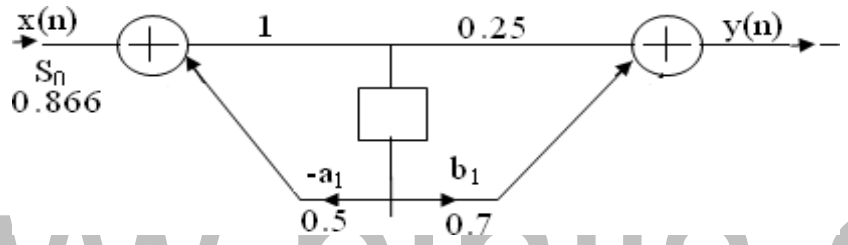
- Because of the process of scaling, the overflow is eliminated. Here so is the scaling factor for the first stage.
- Scaling factor for the second stage = S_{01} and it is given by $S_{01}^2 = \frac{1}{S_0^2 I_2}$

Where $I_2 = \frac{1}{2\pi j}\oint_c \frac{H_1(Z)H_1(Z^{-1})Z^{-1}dZ}{D_2(Z)D_2(Z)}$

****

For the given transfer function, $H(Z) = \frac{0.25 + 0.7Z^{-1}}{1 - 0.5Z^{-1}}$, find scaling factor so as to avoid ******

overflow in the adder '1' of the filter.



Given:

$D(Z) = 1-0.5Z^{-1}$

$D(Z^{-1}) = 1-0.5\,Z$

Solution:

$$I = \frac{1}{2\pi j}\oint_c \frac{1}{D(Z)D(Z^{-1})}\frac{dZ}{Z}$$

$$= \frac{1}{2\pi j}\oint_c \frac{1}{(1-0.5Z^{-1})(1-0.5Z)}\frac{dZ}{Z}$$

$$= \frac{1}{2\pi j}\oint_c \frac{Z}{(Z-0.5)}\frac{1}{(1-0.5Z)}\frac{dZ}{Z}$$

Residue of $\left.\frac{Z}{(Z-0.5)}\frac{1}{(1-0.5Z)}\right|_{Z=0.5} + 0$

I=1.3333

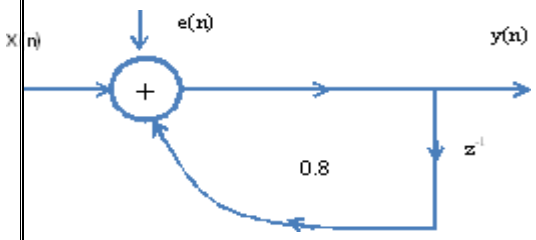
$$S_0 = \frac{1}{\sqrt{I}}$$

$$S_0 = \frac{1}{\sqrt{1.333}}$$

$= 0.866$

****

Consider the recursive filter shown in fig. The input x(n) has a range of values of ±100V, represented by 8 bits. Compute the variance of output due to A/D conversion process. (6)



Solution:

Given the range is ±100V

The difference equation of the system is given by $y(n) = 0.8y(n-1) + x(n)$, whose impulse response h(n) can be obtained as

$h(n) = (0.8)^n u(n)$

$$quantization\ step\ size = \frac{range of\ the\ signal}{No.of\ quantization\ levels}$$

$$= \frac{200}{2^8}$$

$$= 0.78125$$

Variance of the error signal

$$\sigma_e^2 = \frac{q^2}{12}$$

$$= \frac{(0.78125)^2}{12}$$

$$\sigma_e^2 = 0.05086$$

Variance of output

$$\sigma_{\not e}^2 = \sigma_e^2 \sum_{n=0}^{\alpha} h^2(n)$$

$$= (0.05086)\sum_{n=0}^{\alpha}(0.8)^{2n}$$

$$= \frac{0.05086}{1-(0.8)^2} = 0.14128$$

****

The input to the system y(n)=0.999y(n-1)+x(n) is applied to an ADC. What is the power produced by the quantization noise at the output of the filter if the input is quantized to a) 8 bits b)16 bits. May-07

Solution:

y(n)=0.999y(n-1)+x(n)

Taking z-transform on both sides

$Y(z)=0.999z^{-1}Y(z)+X(z)$

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1}{1-0.999z^{-1}}$$

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

$$0.999)(z^{-1} - 0.999)z$$

$$= \frac{-0.001}{(z-0.999)(-0.999)(z-\frac{1}{0.999})}$$

$$= \frac{-0.001}{(z-0.999)(z-0.001)}$$

$$\left.\begin{array}{l}\text{output noise power due}\\ \text{to input quantization}\end{array}\right\}\sigma^2_{eoi} = \sigma^2_e \frac{1}{2\pi j}\int_c H(z)H(z^{-1})z^{-1}dz$$

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

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

Where $p_1, p_2, \ldots\ldots p_N$ are poles of $H(z)H(z^{-1})z^{-1}$, that lies inside the unit circle in z-plane.

$$\sigma_{eoi^2} = \sigma_{e^2}(z-0.999)\left(\frac{0.001}{(z-0.999)(z-0.001)}\right)\Big|_{z=0.999}$$

$$= \sigma_{e^2} 500.25$$

a) b+1=8 bits(Assuming including sign bit)

$$\sigma^2_\varepsilon = \frac{2^{2(7)}}{12}(500.25) = 2.544 \times 10^{-3}$$

b) b+1=16 bits

$$\sigma^2_\varepsilon = \frac{2^{2(15)}}{12}(500.25) = 3.882 \times 10^{-8}$$

****

Find the effect of coefficient quantization on pole locations of the given second order IIR system, when it is realized in direct form I and in cascade form. Assume a word length of 4 bits through truncation.

$$H(z) = \frac{1}{1-0.9z^{-1}+0.2z^{-2}}$$

Solution:

Direct form I

Let b=4 bits including a sign bit

$(0.9)_{10} = (0.111001\ldots)_2$

Integer part

$\frac{0.9\times 2}{1.8} \rightarrow 1 \downarrow$

$\frac{0.8\times 2}{1.6} \rightarrow 1$

$\frac{0.6\times 2}{1.2} \rightarrow 1$

$\frac{0.2\times 2}{0.4} \rightarrow 0$

$\frac{0.4\times 2}{0.8} \rightarrow 0$

$\frac{0.8\times 2}{1.6} \rightarrow 1$

After truncation we get

$(0.111)_2 = (0.875)_{10}$

$(0.2)_{10} = (0.00110\ldots)_2$

$(0.2)_{10} = \dfrac{0.2 \times 2}{0.4}$

$\mapsto \quad 0 \quad \downarrow$

$\dfrac{0.4 \times 2}{0.8}$

$\mapsto \quad 0$

$\dfrac{0.8 \times 2}{1.6}$

$\mapsto \quad 1$

$\dfrac{0.6 \times 2}{1.2}$

$\mapsto \quad 1$

$\dfrac{0.2 \times 2}{0.4}$

$\mapsto \quad 0$

After truncation we get

$(0.001)_2=(0.125)_{10}$

The system function after coefficient quantization is

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

Now the pole locations are given by

$z_1=0.695$

$z_2=0.178$

If we compare the Poles of H(z) and $\overline{H(z)}$ we can observe that the poles of $\overline{H(z)}$ deviate very much from the original poles.

Cascade form

$$H(z) = \frac{1}{1 - 0.5z^{-1}(1 - 0.4z^{-1})}$$

$(0.5)_{10} = (0.1000)_2$

After truncation we get

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

After truncation we get

$(0.011)_2=(0.375)_{10}$

$$(0.4)_{10} = \frac{0.4\times 2}{0.8} \mapsto 0 \downarrow$$

$$\frac{0.8\times 2}{1.6} \mapsto 0$$

$$\frac{0.6\times 2}{1.2} \mapsto 1$$

$$\frac{0.2\times 2}{0.4} \mapsto 1$$

$$\frac{0.4\times 2}{0.8} \mapsto 0$$

$(0.4)_{10}=(0.01100\ldots)_2$

The system function after coefficient quantization is

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

The pole locations are given by
z_1=0.5

z_2
=0.375
on comparing the poles of the cascade system with original poles we can say that one of the poles is same and other pole is very close to original pole.
****

A LTI system is characterized by the difference equation y(n)=0.68y(n-1)+0.5x(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 quantization noise at the output.

Solution:

Given

Range R=-5V to +5V = 5-(-5) =10
Size of binary, B= 8 bits (including sign bit)
Quantization step size,

$$q = \frac{R}{2^8} = \frac{10}{2^8} = 0.0390625$$

$$\text{var}\,iance of\ error\ signal, \sigma_e^2 = \frac{q^2}{12} = \frac{0.0390625^2}{12} = 1.27116*10^{-4}$$

The difference equation governing the LTI system is
Y (n) =0.68y (n-1) +0.15x (n)
On taking z transform of above equation we get

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

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

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

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

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

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

$$H(z)H(z^{-1})z^{-1} = \frac{\left(1 - \frac{0.68}{z}\right)0.225z^{-1}}{(-0.68)\left(z - \frac{1}{0.68}\right)}$$

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

Now, poles of H (z) H (z^{-1}) z^{-1} are p_1=0.68, p_2=1.4706
Here, p_1=0.68 is the only pole that lies inside the unit circle in z-plane
Variance of the input quantization noise at the output.

$$\sigma^2_{eoi} = \sigma^2_e \frac{1}{2\pi j}\int_c H(z)H(z^{-1})z^{-1}dz$$

$$\sigma^2_{eoi} = \sigma^2_e \sum_{i=1}^{N}\left[\text{Res}\, H(z)H(z^{-1})z^{-1}\right]_{z=pi}$$

$$\sigma^2_{eoi} = \sigma^2_e \sum_{i=1}^{N}\left[(z-p_i)H(z)H(z^{-1})z^{-1}\right]_{z=pi}$$

$$\sigma^2_{eoi} = \sigma^2_e (z-0.68)* \frac{-0.0331}{(z-0.68)(z-1.4706)}_{z=0.68}$$

$$\sigma^2_{eoi} = \sigma^2_e * \frac{-0.0331}{(0.68-1.4706)} = 0.0419\sigma^2_e$$

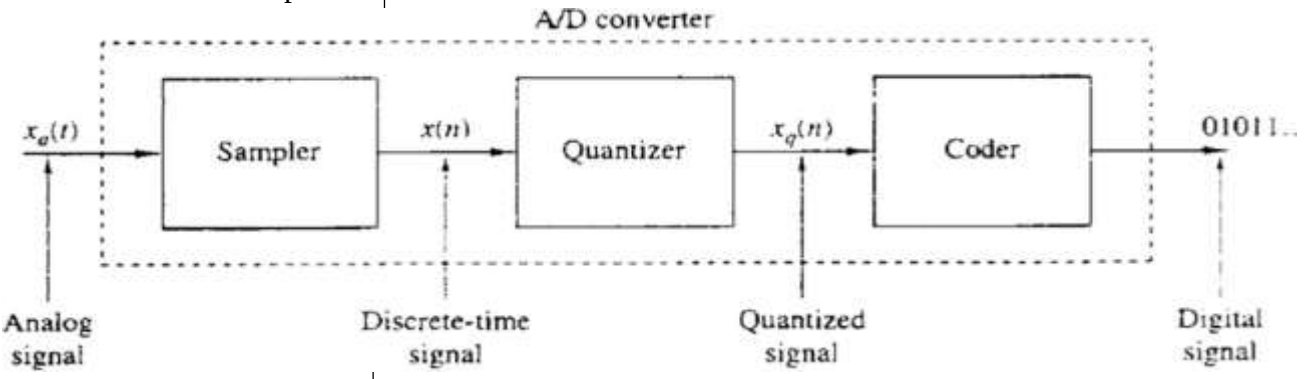
$$\sigma^2_{eoi} = 0.0419*1.2716*10^{-4}$$

$$\sigma^2_{eoi} = 5.328*10^{-6}$$

Analog to digital conversion:

10. Explain the ADC and DAC in detail.

A/D conversion has three process,



Basic parts of an analog-to digital (A/D) converter

1. Sampling

- Sampling is the conversion of a continuous- tome signal into a discrete-time signal obtained by taking the samples of continuous-time signal at discrete instants.
- Thus if $x_a(t)$ is the input to the sampler, the output is $x_a(nT) \equiv x(n)$, where T is called the sampling interval.

2. Quantisation

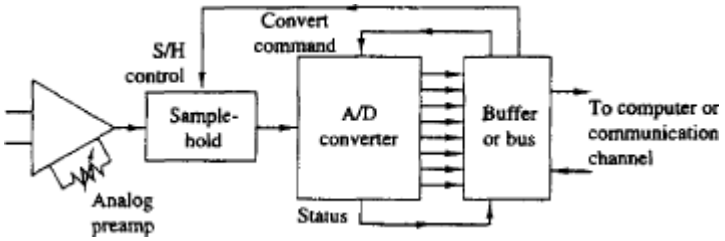
- The process of converting a discrete-time continuous amplitude signal into digital signal is called quantization.
- The value of each signal sample is represented by a value selected from a finite set of possible values.
- The difference between the unquantised sample x(n) and the quantized output $x_q(n)$ is called the quantization error or quantization noise.

$$e_q(n)= x_q(n)-x(n)$$

- To eliminate the excess bits either discard them by the process of truncation or discard them by rounding the resulting number by the process of rounding.
- The values allowed in the digital signals are called the quantization levels
- The distance Δ between two successive quantization levels is called the quantization step size or resolution.
- The quality of the output of the A/D converter is measured by the signal-to-quantization noise ratio.

3. Coding

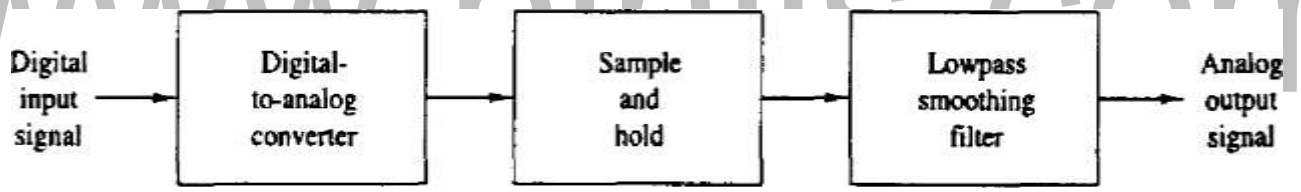
- In the coding process, each discrete value $x_q(n)$ is represented by a b-bit binary sequence.



Block diagram of basic elements of an A/D Converter

Digital to analog conversion:

- To convert a digital signal into an analog signal, digital to analog converters are used.



Basic operations in converting a digital signal into an analog signal

- The D/A converter accepts, at its input, electrical signals that corresponds to a binary word, and produces an output voltage or current that is proportional to the value of the binary word.
- The task of D/A converter is to interpolate between samples.
- The sampling theorem specifies the optimum interpolation for a band limited signal.
- The simplest D/A converter is the zero order hold which holds constant value of sample until the next one is received.
- Additional improvement can be obtained by using linear interpolation to connect successive samples with straight line segment.
- Better interpolation can be achieved y using more sophisticated higher order interpolation techniques.
- Suboptimum interpolation techniques result in passing frequencies above the folding frequency. Such frequency components are undesirable and are removed by passing the output of the interpolator through a proper analog filter which is called as post filter or smoothing filter.
- Thus D/A conversion usually involve a suboptimum interpolator followed by a post filter.

****

Example

Given $X_1 = 3.5*10^{-12}$, $X_2 = 4.75*10^{6}$. Find the product X_1X_2

$X=(3.5 \times 4.75)\ 10^{(-12+6)}$

$= (16.625)10^{-6}$ → in decimal

In binary: $(1.5)_{10} \times (1.25)_{10}$ = $(2^1 0.75) \times (2^1 0.625)$

= $2^{001} \times 0.1100 \times 2^{001} \times 0.1010$

= $2^{010} \times 0.01111$

Addition and subtraction:

- Here the exponent of a smaller number is adjusted until it matches the exponent of a larger number.
- Then, the mantissa are added or subtracted
- The resulting representation is rescaled so that its mantissa lies in the range 0.5 to 1.
- Eg: **Add $(3.0)_{10}$ & $(0.125)_{10}$**

  $(3.0)_{10}$ $=2^{010} \times 0.1100 = r^{e_1} \times M_1$

  $(0.125)_{10}$ $=2^{000} \times 0.0010 = r^{e_2} \times M_2$

  Now adjust e_2 Such that $e_1=e_2$

  $(0.125)_{10}$ $=2^{010} \times 0.0000100$

  Addition → 2^{010} (0.110000 + 0.0000100) → $2^{010} \times 0.110010$

  Subraction → $2^{010} \times 1.001101$

Compare floating point with fixed point arithmetic.

Sl.No	Fixed point arithmetic	Floating point arithmetic
1	Fast operation	Slow operation
2	Relatively economical	More expensive because of costlier hardware
3	Small dynamic range	Increased Dynamic range
4	Round off errors occurs only for addition	Round off errors can occur with addition and multiplication
5	Overflow occur in addition	Overflow does not arise
6	Used in small computers	Used in large general purpose computers.

****

Quantization:

***Discuss the various methods of quantization.**

***Derive the expression for rounding and truncation errors**

*** Discuss in detail about Quantization error that occurs due to finite word length of registers.**

The common methods of quantization are

1. Truncation
2. Rounding

1. Truncation

- The abrupt termination of given number having a large string of bits (or)
- Truncation is a process of discarding all bits less significant than the LSB that is retained.
- Suppose if we truncate the following binary number from 8 bits to 4 bits, we obtain
  - 0.00110011 to 0.0011
    (8 bits) (4 bits)
  - 1.01001001 to 1.0100
    (8 bits) (4 bits)
- When we truncate the number, the signal value is approximated by the highest quantization level that is not greater than the signal.

2. Rounding (or) Round off

- 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 unquantised number.

Error Due to truncation and rounding:

- While storing (or) computation on a number we face registers length problems. Hence given number is quantized to truncation (or) round off.
  i.e. Number of bits in the original number is reduced register length.

Truncation error in sign magnitude form:

- Consider a 5 bit number which has value of
  0.11001_2 → $(0.7815)_{10}$
- This 5 bit number is truncated to a 4 bit number
  0.1100_2 → $(0.75)_{10}$
  i.e. 5 bit number →0.11001 has 'l' bits
  4 bit number →0.1100 has 'b' bits
- Truncation error, e_t = 0.1100 – 0.11001
  = -0.00001 → $(-0.03125)_{10}$
- Here original length is 'l' bits. (l=5). The truncated length is 'b' bits.
- The truncation error, e_t = $2^{-b}-2^{-l}$
  = $-(2^{-l}-2^{-b})$
  e_t = $-(2^{-5}-2^{-4})$ = -2^{-1}
- The truncation error for a positive number is
  $-\left(2^{-b}-2^{-l}\right) \le e_t \le 0$ → Non causal
- The truncation error for a negative number is
  $0 \le e_t \le \left(2^{-b}-2^{-l}\right)$ → Causal

Truncation error in two's complement:

- For a positive number, the truncation results in a smaller number and hence remains same as in the case of sign magnitude form.
- For a negative number, the truncation produces negative error in two's complement

$$-\left(2^{-b}-2^{-l}\right) \le e_t \le \left(2^{-b}-2^{-l}\right)$$

Round off error (Error due to rounding):

- Let us consider a number with original length as '5' bits and round off length as '4' bits.

$$0.11001 \xrightarrow{\text{Round off to}} 0.1101$$

- Now error due to rounding $e_r = \dfrac{2^{-b}-2^{-l}}{2}$

  Where b→Number of bits to the right of binary point after rounding
  L→Number of bits to the right of binary point before rounding
- Rounding off error for positive Number:

$$-\frac{2^{-b}-2^{-l}}{2} \le e_r \le 0$$

- Rounding off error for negative Number:

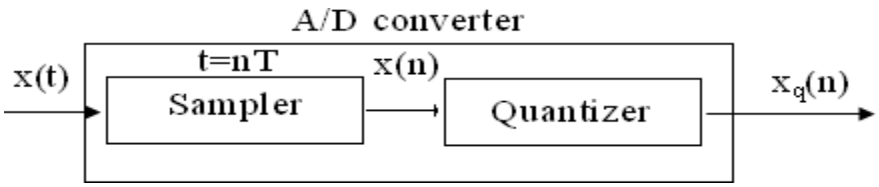
$$0 \le e_r \le \frac{2^{-b}-2^{-l}}{2}$$

- For two's complement

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

****

Quantization Noise:
***Derive the expression for signal to quantization noise ratio**
***What is called Quantization Noise? Derive the expression for quantization noise power.**



- The analog signal is converted into digital signal by ADC
- At first, the signal x(t) is sampled at regular intervals t=nT, where n=0,1,2… to create sequence x(n). This is done by a sampler.
- Then the numeric equivalent of each sample x(n) is expressed by a finite number of bits giving the sequence $x_q(n)$
- The difference signal $e(n)= x_q(n)- x(n)$ is called quantization noise (or) A/D conversion noise.
- Let us assume a sinusoidal signal varying between +1 & -1 having a dynamic range 2
- ADC employs (b+1) bits including sign bit. In this case, the number of levels available for quantizing x(n) is 2^{b+1}.
- The interval between the successive levels is

$$q = \frac{2}{2^{b+1}} = 2^{-b}$$

Where q → quantization step size

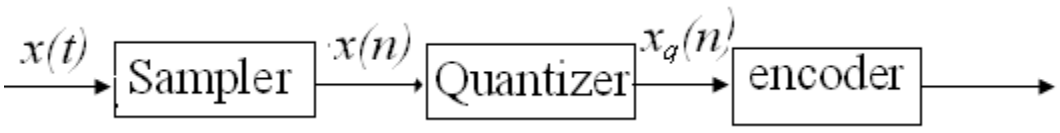
If b=3 bits, then $q=2^{-3}=0.125$

****

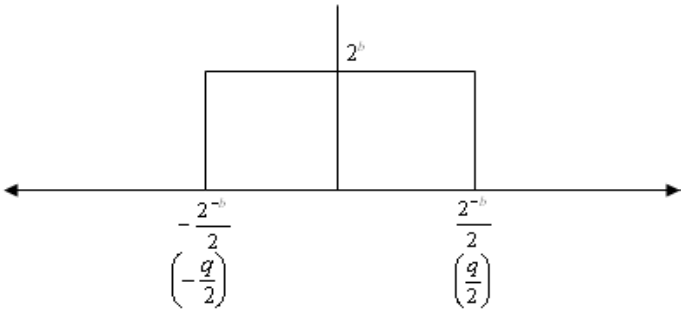
Quantization Noise power:

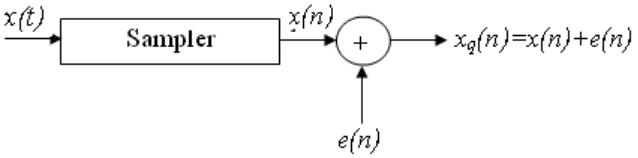
Input Quantization error:

***Derive the equation for quantization noise power (or) Steady state Input Noise Power.**



Probability density function for round off error in A/D conversion is





If rounding is used for quantization, which is bounded by $-\frac{q}{2} \le e(n) \le \frac{q}{2}$, then the error lies between $-\frac{q}{2}$ to $\frac{q}{2}$ with equal probability, where q→ quantization step size.

Properties of analog to digital conversion error, e(n):

1. The error sequence e(n) is a sample sequence of a stationary random process.
2. The error sequence is uncorrelated with x(n) and other signals in the system.
3. The error is a white noise process with uniform amplitude probability distribution over the range of quantization error.

The variance of e(n) is given by

$$\sigma_e^2 = E\left[e^2(n)\right] - E^2\left[e(n)\right] \text{----------} > (1)$$

Where $E\left[e^2(n)\right]$→Average of $e^2(n)$

$E\left[e(n)\right]$→Mean value of e(n).

For rounding, e(n) lies between $-\frac{q}{2}$ and $\frac{q}{2}$ with equal probability