QUANTIZATION DUE TO TRUNCATION AND ROUNDIND,QUANTIZATION NOISE

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
  $0.11001_2 \rightarrow (0.7815)_{10}$
- This 5 bit number is truncated to a 4 bit number
  $0.1100_2 \rightarrow (0.75)_{10}$
  i.e. 5 bit number → 0.11001 has 'l' bits
  4 bit number → 0.1100 has 'b' bits
- Truncation error, $e_t = 0.1100 - 0.11001$
  $= -0.00001 \rightarrow (-0.03125)_{10}$
- 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

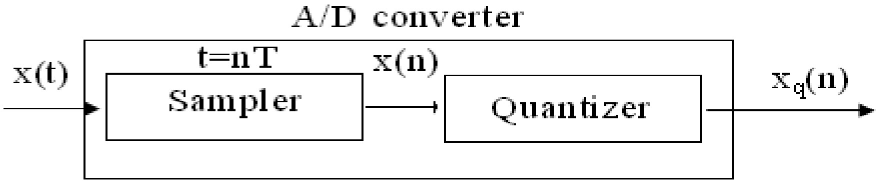
$$-\frac{2^{-b}-2^{-l}}{2} \le e_r \le \frac{2^{-b}-2^{-l}}{2}$$

****

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 $q=2^{-3}=0.125$

****