Quantization Noise:

For most of the engineering applications, input signal is continuous time signal (or) analog signal. The signal is to be converted into digital by using ADC. (Sampling and quantization). The signal $\hat{x(t)}$ is first sampled at regular intervals $t=nT$ where $n=0,1,2....$ to create a sequence $x(n)$. This is done by sampler. Then numeric equivalent of each sample $x(n)$ is expressed by a finite no. of bits giving $x_q(n)$.

quantisation noise (or) error (or) ADC noise $\Big\}$ $e(n) = x_q(n) - x(n)$

If $(b+1)$ ~~bits~~ bits are used to represent ~~xqxe~~ each sequence (including sign bit), then no. of levels available for quantizing $x(n)$ is 2^{b+1}.

Thus the interval between successive level is said to be quantization step size (q).

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

eg: when $b=3$, $q = 2^{-3} = 0.125$

The common methods of quantization are:

1) Truncation
2) Rounding.

Truncation:

→ process of discarding all bits less significant than LSB that is retained. Suppose if we truncate binary numbers from 8 bits to 4 bits,

$0.00110011 \Longrightarrow 0.0011$
$1.01001001 \Longrightarrow 1.0100$

When it is truncated, its value is approximated by the highest quantization level that is not greater than the signal.

Rounding:

→ process of rounding the number closest to the original number unrounded.

eg: 8 bits $0.11010101 \Longrightarrow$ 4-bits 0.1101 (or) 4-bits $0.1110 \leftarrow$ ~~0.1101~~ $+\,1$ $= 0.1110$

$\Longrightarrow$ have negligible effect on accuracy of computation.

Error due to truncation of fixed point numbers:

If a binary number is truncated, its value is approximated by the nearest level that does not exceed it.

i.e. eg: decimal number $\Rightarrow (0.1289625)_{10}$

binary equivalent $\Rightarrow (0.00100001)_2 = x$

b=4: if x is truncated to 4-bits,

$$x_T = (0.0010)_2 = (0.125)_{10}$$

(x_T → truncated value)

$$q = x_T - x = (-0.00390625)_{10} \Rightarrow > -2^{-4} = (-0.0625)_{10}$$

In this case, error: $x_T - x$ is negative or zero

here, $|x| < 0$ (assumption)

$$\therefore \boxed{0 \geq (x_T - x) > -2^{-b}} \quad (1) \Leftarrow \text{inequality condn.}$$

This condition is satisfied for sign-magnitude, 1's complement and 2's complement if $x > 0$ (+ve, fraction).

We have to find whether the above inequality condition is satisfied or not.

Let us consider first the two's complement representation,

the magnitude of the negative number is given by,

$$x = 1 - \sum_{i=1}^{b} d_i 2^{-i} \longrightarrow (2)$$

If we truncate x to N' bits, then

$$x_T = 1 - \sum_{i=1}^{N} d_i 2^{-i} \longrightarrow (3)$$

The change in magnitude is,

$$x_T - x = \sum_{i=1}^{b} d_i 2^{-i} - \sum_{i=1}^{N} d_i 2^{-i}$$

$$= \sum_{i=N}^{b} d_i 2^{-i} \geq 0 \longrightarrow (4)$$

$x_T - x \geq 0 \Rightarrow$ Change in magnitude is +ve due to truncation, which implies that error is negative and satisfy the inequality,

$$\boxed{0 \geq (x_T - x) > -2^{-b}} \longrightarrow (5)$$

← 2's complement error

For 1's complement representation, the magnitude of negative number with 'b' bits is given by,

$$x = 1 - \sum_{i=1}^{b} d_i 2^{-i} - 2^{-b} \longrightarrow (6)$$

when the number is truncated to 'N' bits, then

$$x_T = 1 - \sum_{i=1}^{N} d_i 2^{-i} - 2^{-N} \longrightarrow (7)$$

The change in magnitude is,

$$x_T - x = \sum_{i=N}^{b} d_i 2^{-i} - (2^{-N} - 2^{-b}) \quad \rightarrow (8)$$

$$< 0$$

∴ the magnitude decreases with truncation which implies that error is positive and satisfy the inequality,

$$\boxed{0 \leq (x_T - x) < 2^{-b}} \quad \rightarrow (9)$$

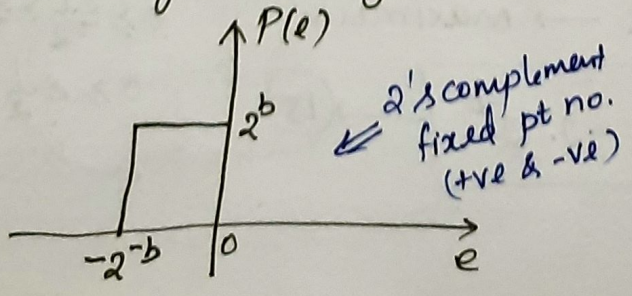
⇐ 1's complement & sign-magnitude error

The above condition holds for sign-magnitude representation also.

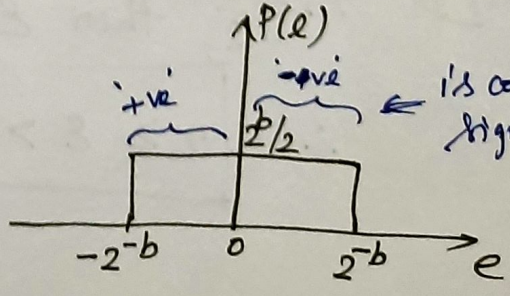
Range of errors in truncation of fixed pt nos:

number & its representation	error range
positive number (all 3 rep.)	$0 \geq e > -2^{-b}$
Sign-magnitude negative no.	$0 \leq e < 2^{-b}$
1's complement " "	$0 \leq e \leq 2^{-b}$
2's complement " "	$0 \geq e > -2^{-b}$

probability density function P(e) for truncation of fixed pt nos.



⇐ 2's complement fixed pt no. (+ve & -ve)



⇐ 1's complement & sign-mag. fixed pt nos

Errors due to truncation of floating point numbers:

In floating point systems, the truncation effect is visible only in the mantissa.

Let the mantissa is truncated to 'N'-bits,

If $x = M \cdot 2^{exp}$, then

$$x_T = 2^{exp} M_T$$

Error, $e = x_T - x = 2^{exp} \cdot M_T - 2^{exp} M = 2^{exp}(M_T - M)$ → (10)

From equation number (5), with 2's complement representation of mantissa, we have

$$0 \geq M_T - M > -2^{-b} \quad \rightarrow (11)$$

$$0 \geq e > -2^{-b} \cdot 2^{exp} \quad \rightarrow (12)$$

$\left[\because e = 2^{exp}(M_T - M)\right.$

e → error

Let us define relative error $\varepsilon = \frac{x_T - x}{x} = \frac{e}{x}$

$$\Rightarrow \boxed{e = \varepsilon x}$$

$\therefore$ (12) $\Rightarrow$ $0 \geq \varepsilon x > -2^{-b} \cdot 2^{exp}$ → (13)

$$0 \geq \varepsilon \cdot 2^{exp} \cdot M > -2^{-b} \cdot 2^{exp} \quad \left[\because x = e^{exp} \cdot M\right.$$

$$0 \geq \varepsilon \cdot M > -2^{-b} \quad \rightarrow (14)$$

If $M = \frac{1}{2}$, then $\varepsilon \rightarrow$ maximum

$$\therefore \boxed{0 \geq \varepsilon > -2 \cdot 2^{-b}} \quad \rightarrow (15)$$

$0 \geq \varepsilon \cdot \frac{1}{2} > -2^{-b}$

$0 \geq \varepsilon \cdot \frac{1}{2} \times 2 > -2 \times 2^{-b}$

If $M = \frac{-1}{2}$, ε will be

$$0 \leq \varepsilon < 2 \cdot 2^{-b} \quad \text{(16)}$$

$0 \geq \varepsilon \times \frac{-1}{2} > -2^{-b}$

$0 \geq \varepsilon \times \frac{-1}{2} \times 2 > -2 \times 2^{-b}$

In 1's complement & sign-magnitude representation, the error for truncation of positive values of mantissa is

$$0 \geq M_T - M > -2^{-b} \text{ (or)}$$

$$0 \geq e \geq -2^{-b} \times 2^{exp} \quad \text{with } \varepsilon = \frac{e}{x}$$

$$0 \geq \varepsilon x > -2^{-b} \cdot 2^{exp}$$

$$\Rightarrow e = \varepsilon x = \varepsilon \cdot M \cdot 2^{exp}$$

when $M = \frac{1}{2}$, $0 \geq \varepsilon \cdot M > -2^{-b}$

we get maximum range of relative error for positive M as

$$0 \geq \varepsilon > -2 \cdot 2^{-b} \quad \text{(17)}$$

For negative mantissa, error value is

$$0 \leq M_T - M < 2^{-b} \text{ (or)}$$

$$0 \leq e < 2^{exp} \cdot 2^{-b} \quad \text{(18)}$$

with $M = -1/2$, the maximum range of the relative error for negative M is

$$0 \geq \varepsilon > -2 \cdot 2^{-b} \quad \text{(19)}$$

Range of errors in truncation of floating point nos:

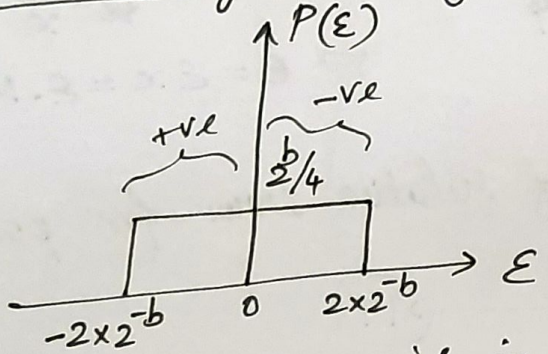
Type of representation for mantissa	Range of error
• 2's complement positive mantissa	$0 \geq \varepsilon > -2 \times 2^{-b}$

- 2's complement -ve mantissa } $0 \le \varepsilon < 2 \times 2^{-b}$

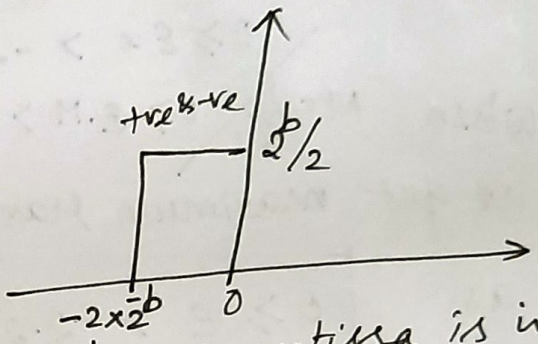
1's complement +ve & -ve mantissa } $0 \le \varepsilon < -2 \times 2^{-b}$

Sign-magnitude +ve & -ve mantissa } $0 \le \varepsilon < -2 \times 2^{-b}$

- Probability density functions $P(\varepsilon)$:



when mantissa is in 2's complement form



when mantissa is in 1's complement (or) sign-mag. form.

Error due to rounding of fixed point numbers:

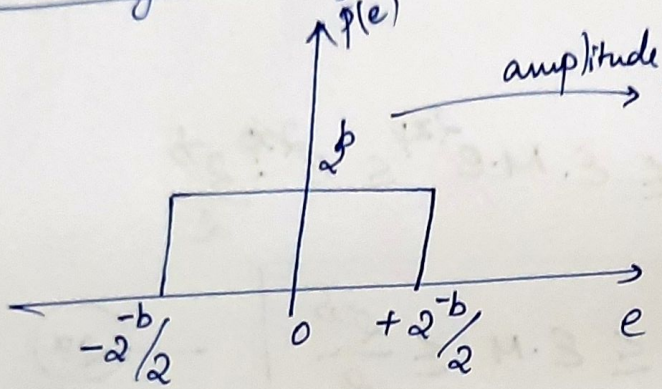
In fixed point arithmetic, error due to rounding a number to 'b'-bits produces an error $'\varepsilon' = x_R - x$ which satisfies inequality,

$$\frac{-2^{-b}}{2} \le x_R - x \le \frac{2^{-b}}{2} \quad \text{---} \quad (20)$$

This is because with rounding, the error lies ~~but~~ half way between 2 levels, ie it can be approx-imated

to either nearest higher level or by the nearest lower level. For fixed point number, the above inequality condition is satisfied for all the 3 types of representation. (Sign-mag, 1C, 2C & x)

Probability density function P(e):



amplitude $= \frac{1}{\frac{2^{-b}}{2} - \left(\frac{-2^{-b}}{2}\right)} = \frac{2}{2 \cdot 2^{-b}}$

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

Error due to rounding of floating point nos:

In floating point arithmetic, only mantissa is affected by rounding. (i.e. quantization)

If $x = M \cdot 2^{exp}$ & $x_R = M_R \cdot 2^{exp}$

then $e = x_R - x = (M_R - M) \cdot 2^{exp}$ ⟶ (21)

But for rounding,

$$\frac{-2^{-b}}{2} \leq (M_R - M) \leq \frac{2^{-b}}{2} \longrightarrow (22)$$

From (21), $(M_R - M) = \frac{e}{2^{exp}} = \frac{x_R - x}{2^{exp}}$ ⟶ (23).

$$-2^{exp}\frac{2^{-b}}{2} \le (x_p - x) \le 2^{exp}\cdot\frac{2^{-b}}{2} \longrightarrow (24)$$

let relative error $\varepsilon = \frac{x_p - x}{x} = \frac{e}{x}$ $\longrightarrow (25)$

$$e = \varepsilon x$$

$$-2^{exp}\cdot\frac{2^{-b}}{2} \le \varepsilon x \le \frac{2^{-b}}{2}\,2^{exp} \longrightarrow (26)$$

wkt, $x = M \cdot 2^{exp}$

$$-2^{exp}\cdot\frac{2^{-b}}{2} \le \varepsilon \cdot M \cdot 2^{exp} \le 2^{exp}\cdot\frac{2^{-b}}{2}$$

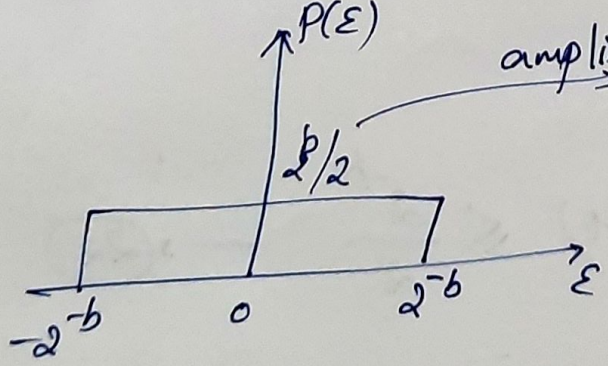
$$\boxed{\frac{-2^{-b}}{2} \le \varepsilon \cdot M \le \frac{2^{-b}}{2}} \longrightarrow (27)$$

wkt, The mantissa range $M \Rightarrow \frac{1}{2} \le M < 1$

If $M = 1/2$, maximum range of relative error is obtained

$$\boxed{-2^{-b} \le \varepsilon < 2^{-b}}$$

Probability density function $P(\varepsilon)$:



$$\text{amplitude} = \frac{1}{2^{-b} - (-2^{-b})} = \frac{1}{2\cdot 2^{-b}} = \frac{2^b}{2}$$

for all 3 types of representation