Rounding Error

Rounding

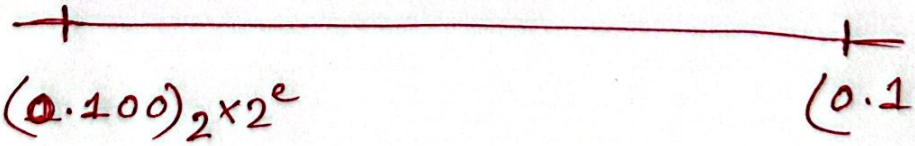
* Computers can read upto certain bit or significant figures that why we need rounding.
* If we want to round any number we need to draw number line.

→ Real number to F is known as rounding. It is denoted by $fl(x)$.

Example: (Prac 1)

Given that $\beta = 2$, $m = 3$. The given number $(0.1000100)_2$ needs to rounded or we need to round this number.

* For rounding any number at first we need to draw a number line to see where this number lines.



* Now we need to find the middle point

The process of finding the midpoint

$(0.\overset{-1}{1}00)_2$

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

$= \frac{1}{2}$

$(0.\overset{-1\ -2\ -3}{1\ 0\ 1})_2$

$= 1 \times 2^{-1} + 1 \times 2^{-3}$

$= \frac{1}{2} + \frac{1}{2^3}$

$= \frac{1}{2} + \frac{1}{8}$

$= \frac{4+1}{8}$

$= 5/8$

$$\frac{\left(\frac{1}{2} + \frac{5}{8}\right)}{2} = \frac{9}{16}$$

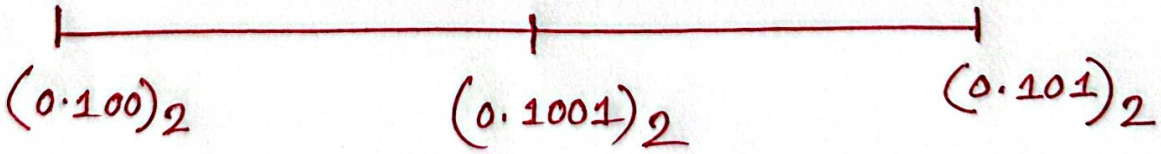
$= \frac{8}{16} + \frac{1}{16}$

$= \frac{1}{2} + \frac{1}{2^4}$

$= 2^{-1} + 2^{-4}$

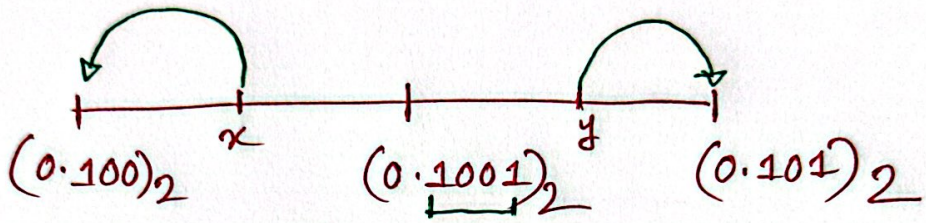
$= (0.1001)_2$

↳ mid point



* How to round ?

Ans:- 1) If a number lies on the left side of the number line then round the number to the left number.



Ex: $(0.\underbrace{100}_{\text{left}}00)_2 = (0.100)_2$

↳ This number lies on the left side of the number line. That's why we rounded it to the left number $(0.100)_2$.

2) If a number lies on the right side of the number line then round the number to the right number.

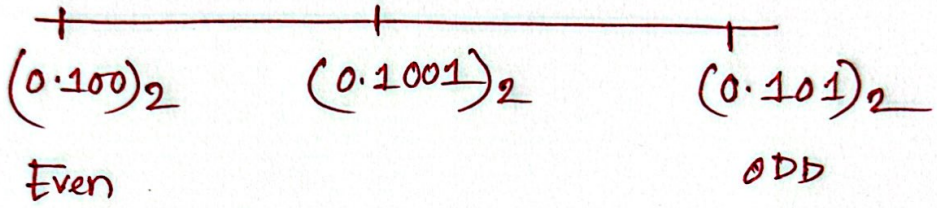
Ex: $(0.\underbrace{1001}_{}01000\underline{1})_2 \approx (0.101)_2$

↳ This number lies on the right side of the number line. That's why we round it to the right number $(0.101)_2$

3) If the number lies exactly at middle or indicate middle point then it will it will get rounded to the even one in that range.

Ex: $(0.1001)_2 = (0.100)_2$

This is the even number.

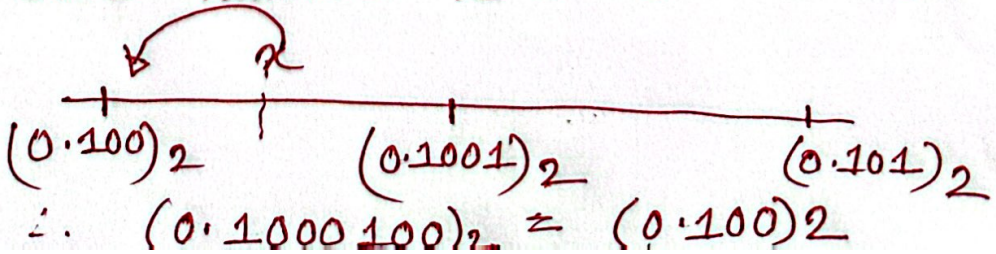


If a number ends in 0, it is even $(0.100)_2$ Here it ends in 0 that's why it is an even number.

If a number ends in 1, it is an odd. (0.101) Here it ends in 1 that's why it is an odd number

Our Example (previous) : Pract 1.

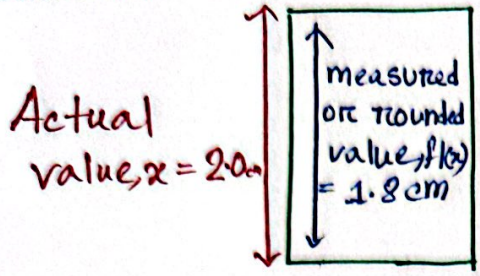
Give number $x = (0.1000100)_2$



$\therefore (0.1000100)_2 \approx (0.100)_2$

Rounding Errors

Case 1:



$$\text{Error} = |fl(x) - x|$$

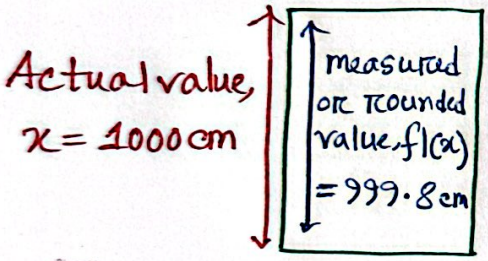
(fl(x): rounded value; x: actual value; |·|: modulus)

$$= |1.8 - 2.0|$$

$$= |-0.2|$$

$$= 0.2$$

Case 2:



$$\text{Error} = |fl(x) - x|$$

$$= 0.2$$

From these two cases it is hard to understand the impact or significance just based on Error.

Though the error is same in both cases, the impact is different, and we understand this through

scale invariant rounding error.

This is denoted by δ. (delta)

Formula

$$\therefore \text{Scale invariant rounding error, } \delta = \frac{|fl(x) - x|}{|x|}$$

$$\delta = \frac{|fl(x) - x|}{|x|}$$

$$\Rightarrow \delta x = fl(x) - x$$

$$\Rightarrow fl(x) = \delta x + x$$

$$= x(1+\delta)$$

However, we deal with **maximum scale invariant rounding error δ_{max}.** This is also known as **Machine Epsilon, ϵ**

$$\boxed{\delta_{max} == \epsilon}$$

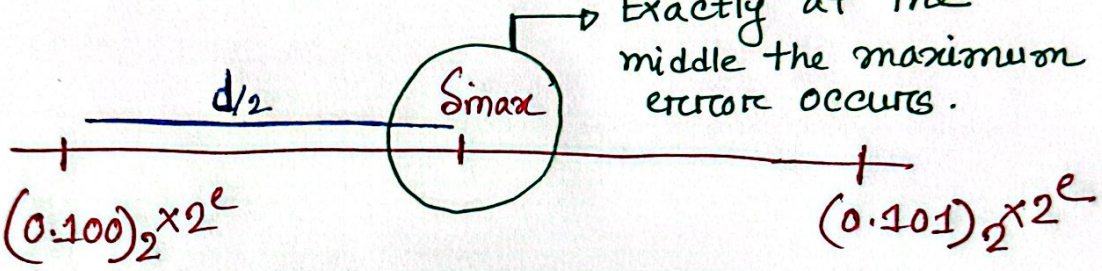
$$\delta_{max} = \epsilon = \frac{|fl(x) - x|_{max} \uparrow}{|x|_{min} \downarrow}$$

Formula of Machine Epsilon (ϵ) for different conventions.

Convention 1

$$(0.d_1 d_2 \cdots d_m)_\beta \times \beta^e$$

Let, $\beta = 2$, $m = 3$



$$d = (0.001)_2 \times 2^e$$

$$= (0.00\overset{-3}{1})_2 \times 2^e$$

$$= 1 \times 2^{-3} \times 2^e$$ [We know, Given that m = 3]

$$= 1 \times \beta^{-m} \times \beta^e$$

$$|fl(x) - x|_{max} = d/2$$

$$= \frac{\beta^{-m} \times \beta^e}{2}$$

$$= \boxed{\frac{1}{2} \beta^{-m} \times \beta^e}$$ (important)

$$|x|_{min} = (0.\underset{min}{1}00)_2 \times 2^e$$

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

$$= \boxed{\beta^{-1} \times \beta^e}$$ (important)

$$\therefore \text{Machine Epsilon}(\epsilon) = \frac{|fl(x) - x|_{max}}{|x|_{min}}$$

$$= \frac{\frac{1}{2} \beta^{-m} \times \beta^{e}}{\beta^{-1} \times \beta^{e}}$$

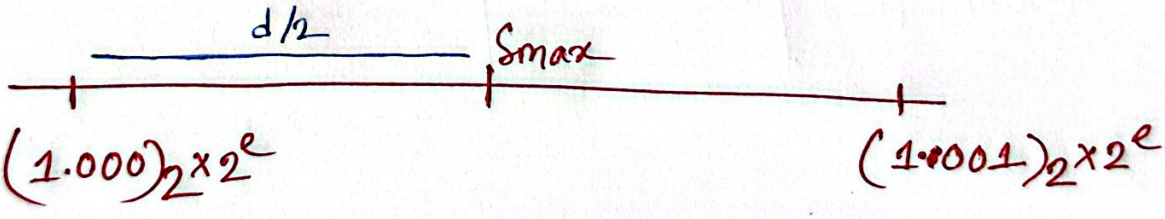
$$= \frac{1}{2} \beta^{1-m} \quad \text{(important)}$$

$$\therefore \text{Machine Epsilon}(\epsilon) = \delta_{max} = \frac{1}{2} \beta^{1-m}$$

Convention 2: (Normalized Form)

$$(1.d_1 d_2 \cdots d_m)_\beta \times \beta^e$$

Let, $\beta = 2$, $m = 3$



$$d = (0.001)_2 \times 2^e$$

$$= 1 \times 2^{-3} \times 2^e$$

$$= 1 \times \beta^{-3} \times \beta^e$$

$$= 1 \times \beta^{-m} \times \beta^e$$

$$|fl(x) - x|_{max} = d/2 = \frac{1}{2} \times \beta^{-m} \times \beta^e \quad \text{(important)}$$

$$|x|_{min} = (\overset{0}{1}.000)_2 \times 2^e$$

$$= 1 \times 2^0 \times 2^e$$

$$= 2^e$$

$$= \boxed{\beta^e} \quad \text{(important)}$$

$$\therefore \text{Machine Epsilon}, \epsilon = \delta_{max} = \frac{\frac{1}{2}\beta^{-m}\beta^e}{\beta^e}$$

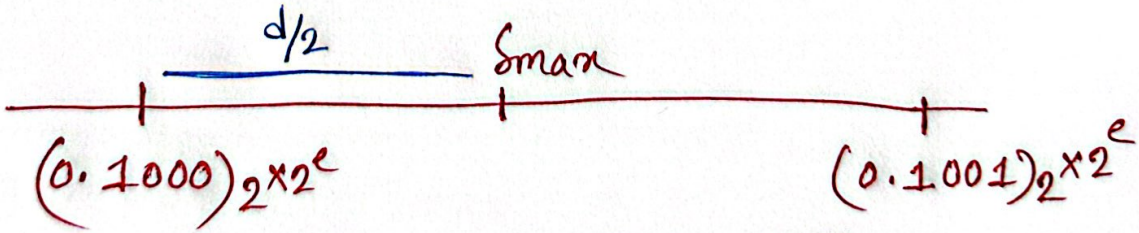
$$= \boxed{\frac{1}{2}\beta^{-m}} \quad \text{(important)}$$

$$\boxed{\therefore \text{Machine Epsilon } (\epsilon) = \delta_{max} = \frac{1}{2}\beta^{-m}}$$

Convention 3: (Denormalized form :)

$$(0.1d_1d_2 - - d_m)_\beta \times \beta^e$$

Let, $\beta = 2$, $m = 3$



$$d = (0.\overset{-1}{0}\overset{-2}{0}\overset{-3}{0}\overset{-4}{1})_2 \times 2^e$$

$$= 1 \times 2^{-4} \times 2^e$$

$$= 1 \times 2^{-3-1} \times 2^e$$

$$= 1 \times \beta^{-m-1} \times \beta^e$$

$$|fl(x) - x|_{max} = \frac{d}{2} = \frac{1}{2} \beta^{-m-1} \times \beta^e$$

$$= \boxed{\frac{1}{2} \beta^{-(m+1)} \times \beta^e} \quad \text{(important)}$$

$$|x|_{min} = (0.1000)_2 \times 2^e$$

$$= (0.\overset{-1}{1}000)_2 \times 2^e$$

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

$$= \boxed{\beta^{-1} \times \beta^e} \quad \text{(important)}$$

$$\text{Machine Epsilon, } \epsilon = \delta_{max} = \frac{|fl(x) - x|_{max}}{|x|_{min}}$$

$$= \frac{\frac{1}{2} \beta^{-m-1} \beta^e}{\beta^{-1} \beta^e}$$

$$= \boxed{\frac{1}{2} \beta^{-m}} \quad \text{(important)}$$

$$\therefore \text{Machine Epsilon, } \epsilon = \delta_{max} = \frac{1}{2} \beta^{-m}$$

Note: Machine Epsilon, ϵ or δ_{max} doesn't affect with the value of exponent, e.

Example: (Practice Problem)

Given that, $\beta = 2$, $m = 3$, $x = 5/8$, $y = 7/8$

Find. $fl(x * y)$.

$$\Rightarrow \quad x * y = 5/8 \times 7/8$$

$$= \frac{35}{64}$$

Now convert $\frac{35}{64}$ into binary.

$$\frac{32}{64} + \frac{2}{64} + \frac{1}{64}$$

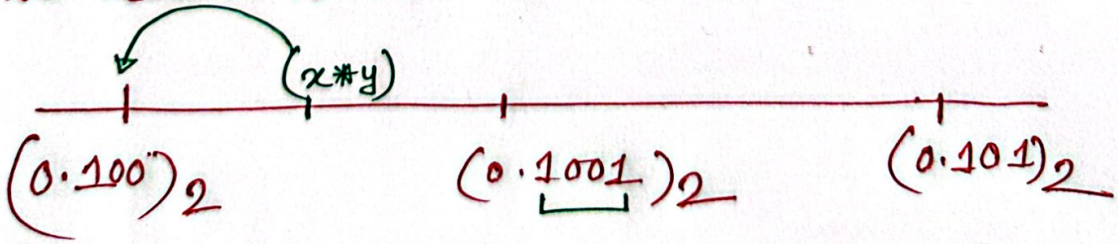
$$= \frac{1}{2} + \frac{1}{32} + \frac{1}{64}$$

$$= 2^{-1} + \frac{1}{2^5} + \frac{1}{2^6}$$

$$= 2^{-1} + 2^{-5} + 2^{-6}$$

$\therefore$ The binary value $= (0.\underbrace{1000}_{\text{left side}}11)_2$ ($m=3$)

We need to round the value because $m = 3$ [given]



So, $(0.100011)_2 = (0.100)_2$

$= \frac{1}{2}$ P.T.O

$$\text{Actual} = \frac{35}{64}$$

$$\text{Rounded} = \frac{1}{2}$$

So rounding error occurs because

$$\frac{35}{64} \neq \frac{1}{2}$$

Loss of Significance

We know,

$$fl(x) \neq x$$

$$fl(y) \neq y$$

Because,

$$fl(x) = x(1+\delta_1)$$

$$fl(y) = y(1+\delta_2)$$

Now if we want to calculate $x \pm y$

$$x \pm y = fl(x) \pm fl(y)$$

$$= x(1+\delta_1) + y(1+\delta_2)$$

$$= (x \pm y)\left(1 + \underbrace{\frac{x\delta_1 \pm y\delta_2}{x \pm y}}\right)$$

Scale invariant Error.

Now if we want to calculate $x - y$

For scale invariant Error, we will get

$$\frac{x\delta_1 - y\delta_2}{x-y}$$

Note: Now, if x and y are more closer values ($x \approx y$), scale invariat error will be significantly high.

Again, if $x = y$ then denominator will be 0 and scale invariant rounding error will be infinite (∞).

This is called loss of significance.

This can occur if 1) we subtract two closer number, 2) Denominator will be very small or 0.

3) Scale invariant error will be very high.

Example:

$$x^2 - 56x + 1 = 0$$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$x_1 = 28 + \sqrt{783} = 55.98$

$x_2 = 28 - \sqrt{783} = 0.01786$

Lets assume computers can calculate upto 4 significant figures.

$\sqrt{783} = 27.98$

$x_1 = 28 + 27.98 = 55.98$

$x_2 = 28 - 27.98 = 0.02000$

equal

not equal

[Loss of significance]

Because 28 and 27.98 are very close to eachother.

$0.01786 \neq 0.02000$ [Loss of significance]

→ denominator very small. That's why rounding errors or scale invarient error very high.

(*) Solution of the previous problem that occurs

$$x^2 - 56x + 1$$

$$x^2 - (\alpha+\beta)x + \alpha\beta$$

$\alpha, \beta \Rightarrow$ roots

$$\alpha\beta = 1$$

$$\alpha = x_1 = 28 + 27.98 = 55.98$$ [From the previous problem

$$\beta = x_2 = \frac{1}{\alpha}$$

$$= \frac{1}{55.98}$$

$$= 0.01786$$ (same as original x_2)

∴ Here target is to avoid subtraction.

Example: Find the average of 5.01 & 5.02.

$$\text{Actual} = \frac{(5.01 + 5.02)}{2} = 5.015$$

Let assume computer can read upto 3 significant figure.

So, $fl\left(\frac{5.01+5.02}{2}\right)$ $\quad 5.015 \neq 5$

$= fl\left(\frac{10.03}{2}\right)$ ∴ rounding occurs

$= fl\left(\frac{10.0}{2}\right)$ → take 3 significant figures.

$= 5$