Rounding

→ We need to do rounding, because computers can read upto certain bits/significant figures

→ To round a number, we need to draw a number line.

The mapping from $\mathbb{R}$ (real no.) to F is called rounding, and it is denoted by $fl(x)$.

Example: A computer has $\beta = 2$, $m = 3$.

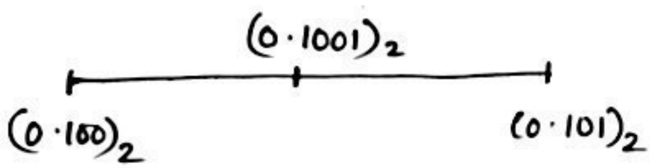
Now, suppose a number is given, $x = (0.\underbrace{100}_{m=3}0100)_2$

We need to round this number.

① Draw number line.

→ We need to see where this number lies.

- Find the middle point



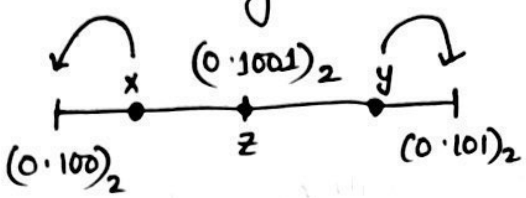
How did we find?

$(0.100)_2 = \frac{1}{2}$

$(0.101)_2 = \frac{5}{8}$

$$\left(\frac{1}{2}+\frac{5}{8}\right)/2 = \frac{9}{16}$$
$$= \frac{8}{16} + \frac{1}{16}$$
$$= \frac{1}{2} + \frac{1}{2^4}$$
$$= 2^{-1} + 2^{-4}$$
$$= (0.1001)_2$$

Rules for rounding.



- if number lies on left side of the number line, then round the number to left number.

  Ex. $(0.\underbrace{100}_{\text{left}}0100)_2 = (0.100)_2$

  middle point is $(0.1001)_2$, so all other no. $(0.1000\ldots)$

- if number lies on right side of the number line, then round the number to right number.

  Ex. $(0.\underbrace{100}_{\text{right}}101001)_2 = (0.101)_2$

- if number is exactly at middle, then it will get rounded to the nearest number.

  Ex. $(0.1001)_2$ [exactly at middle]

  Now, how to know the even numbers in binary?

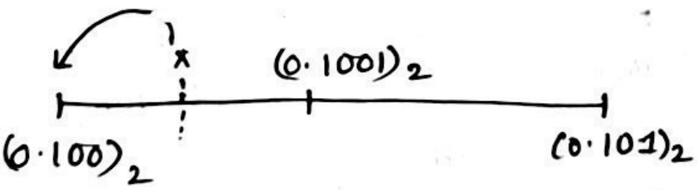
  → if it ends in 0, it is even

  → if it ends in 1, it is odd.

So,

$(0.1001)_2 = (0.100)_2$ (even number)

Now, in our example, $x = (0.1000100)_2$



so, $(0.1000100)_2 = (0.100)_2$ [rounded]

" x (real number) is converted/rounded to $fl(x)$ [rounded no.] "

Rounding Error

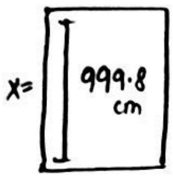
actual
$x = 2.0\text{cm}$

measured x
$= 1.8\text{cm}$

$\text{Error} = (2 - 1.8)$
$= 0.2\text{cm}.$

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

$fl(x) \rightarrow$ rounded value



actual = 1000cm
measured = 999.8cm
error = 0.2 cm. $= |999.8 - 1000|$ ← $fl(x) - x$ ← modulus.
$= 0.2\text{cm}$

It is difficult to understand the impact using Error only, hence, we find scale invariant rounding error, denoted by:

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

$$\delta \cdot x = fl(x) - x \quad [\text{cross multiplication}]$$

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

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

We deal with maximum scaled invariant error called

$\boxed{\text{Machine Epsilon}, \epsilon}$ $\boxed{\delta_{max} == \epsilon}$

Maximum scale invariant error will be:

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

- $\delta \propto |fl(x) - x|$
- $\delta \propto \frac{1}{|x|}$

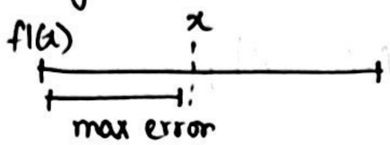
→ δ will change according to conventions.

Convention 1

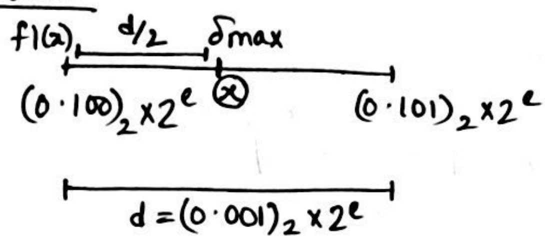
$(0.d_1d_2\ldots d_m)_\beta \times \beta^e$

At which point max. error occurs?

→ exactly at the middle.



Derivation



[m = 3]

$$|fl(x) - x|_{max} = \frac{1}{2}(0.\overset{-1\,-2\,-3}{001})_2 \times 2^e$$

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

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

$$|x|_{min} = (0.1\underline{0}\underline{0})_2 \times 2^e$$
min

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

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

$\therefore$ Machine Epsilon $(\epsilon) = \dfrac{|fl(x) - x|}{|x|} = \dfrac{\frac{1}{2} \times \beta^{-m} \times \beta^e}{\beta^{-1} \times \beta^e}$

$$= \boxed{\frac{1}{2}\beta^{1-m}} \text{ [Fixed]}$$

Questions

→ calc. ϵ for conv. 1 $= \frac{1}{2}\beta^{1-m}$

→ min. value of x for conv. 1 $= \beta^{-1}\beta^e$

Convention 2 (Normalised form)

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

Similar derivation concept :

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

$$|x|_{min} = \beta^e$$

$$\text{Machine Epsilon} = \boxed{\frac{1}{2}\beta^{-m}}$$

Convention 3 (Denormalised form)

$$(0 \cdot 1 d_1 d_2 \cdots d_m)_\beta \times \beta^e$$

Similar derivation concept:

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

For denormalised form, we need 1 extra bit

Hence,

|fl(x) - x| = 1/2 B^ -(m+1) B^e

= 1/2 B^ -m-1 B^e

$$|x|_{min} = \beta^{-1}\beta^e$$

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

Point to note :

- There is no exponent (e) in ϵ.
- So value of machine epsilon (ϵ) won't be affected due to exponent.
- $\delta \le \epsilon_M$

Example

$\beta = 2$, $m = 3$, $x = \frac{5}{8}$, $y = \frac{7}{8}$

Calculate the value of $\underbrace{fl(x * y)}_{\text{round the output.}}$

① Convert to binary

$$x = \frac{5}{8} = \frac{4}{8} + \frac{1}{8} = \frac{1}{2} + \frac{1}{8} = \frac{1}{2^1} + \frac{1}{2^3}$$

$= 2^{-1} + 2^{-3}$ (there will be 1 in -1 and -3 position)

$= (0.101)_2$

$$y = \frac{7}{8} = \frac{6}{8} + \frac{1}{8} = \frac{4}{8} + \frac{2}{8} + \frac{1}{8} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = 2^{-1} + 2^{-2} + 2^{-3} = (0.111)_2$$

(ii) Actual $(x * y) = (0.101)_2 * (0.111)_2$

$$= \frac{5}{8} \times \frac{7}{8} = \frac{35}{64}$$

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

$$= \frac{32}{64} + \frac{3}{64}$$

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

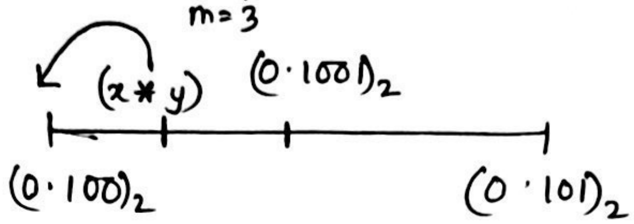
$$= \frac{1}{2} + \frac{1}{2^5} + \frac{1}{2^6}$$

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

$$= (0.100011)_2$$

We need to round ~~to~~ the value since $m = 3$

$$(0.\underbrace{100}_{m=3}011)_2$$



$0.100\overset{\swarrow \text{ left side}}{0}11 \Rightarrow$ round to left side

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

rounded-value: $fl(x) * fl(y) = (0.100)_2$

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

$$\therefore \underset{\text{actual}}{\frac{35}{64}} \neq \underset{\text{rounded.}}{\frac{1}{2}} \quad [\text{rounding error}]$$

Loss of significance

if, $x \neq fl(x) \qquad y \neq fl(y)$

then $fl(x) = x(1+\delta_1) \qquad fl(y) = y(1+\delta_2)$

now, if we want to calc. $x \pm y$

$$\begin{aligned} x \pm y &= fl(x) \pm fl(y) \\ &= x(1+\delta_1) \pm y(1+\delta_2) \\ &= (x \pm y) \pm x\delta_1 \pm y\delta_2 \\ &= (x \pm y)\Big(1 + \underbrace{\frac{x\delta_1 \pm y\delta_2}{x \pm y}}_{\text{scale invariant error}}\Big) \end{aligned}$$

if we want to calc. $x-y$

For scale invariant error, we have

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

Now, if x and y are closer values ($x \approx y$), scale invariant error becomes significantly high.

If $x == y$, then s.i.e becomes infinite (∞)

→ This phenomenon is called $\boxed{\text{loss of significance}}$

→ subtract two close numbers

→ denominator will be very small / approximately zero.

→ scale invariant error will be very high.

Example

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

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

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

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

equal

let's assume, computer can calc. upto 4 sf.

$\sqrt{783} = 27.98$

$x_1 = 28 + 27.98 = 55.98$

$x_2 = 28 - 27.98 = 0.02000$

$\Big\}$ 4 sf.

not equal

(loss of significance) [28 and 27.98 → very close values]

$0.01786 \neq 0.02000$ → Loss of significance

→ denominator very small

→ rounding error / scale invariant error very high

Solution

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

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

$\alpha, \beta \rightarrow$ roots

$$\alpha\beta = 1$$

$\alpha = 28 + 27.98 = 55.98$ [prev. x_1]

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

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

$= 0.01786$ (same as original x_2)

☐ target $\Rightarrow$ avoid subtraction.

Example: Average of 5.01 and 5.02

Actual $= (5.01 + 5.02)/2 = 5.015$

let's say, computer can read to 3 sf.

so, $fl\left(\frac{5.01+5.02}{2}\right)$

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

$= fl\left(\frac{10.0}{2}\right)$

$= 5$

$5.015 \neq 5$

$\boxed{\text{rounding error}}$