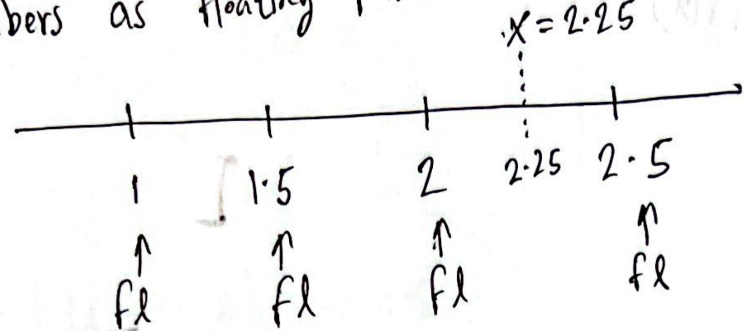
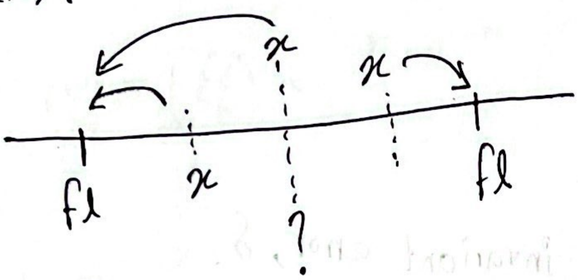
Rounding

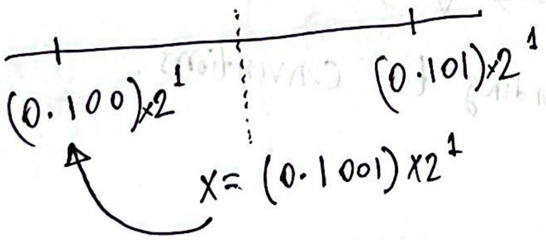
In the number line we can only represent DISCRETE set of numbers as floating points.



~~x is alwa~~ if x is given, x should always be converted to $fl(x)$.



rule → if perfectly in the middle, round it to the nearest even fl.

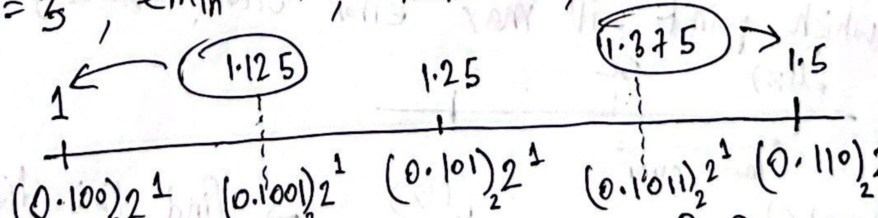


For binary, if number ends in 0 → even
" " 1 → odd

Example

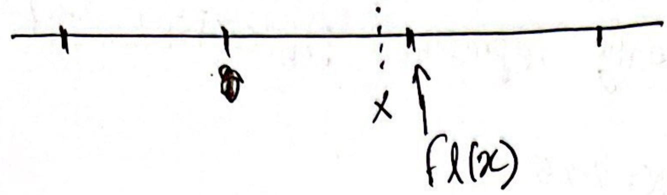
$\beta = 2$, $m = 3$, $e_{min} = -1$, $e_{max} = 2$, convention 1.

$\boxed{\text{when } e = 1}$



If numbers are in the middle of fl, round it off to the ~~pai~~ nearest even.

Rounding Error



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

$\uparrow$
scale invariant error

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

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

$$\boxed{fl(x) = x(1+\delta)}$$

Machine Epsilon

→ Maximum possible scale invariant error, δ.

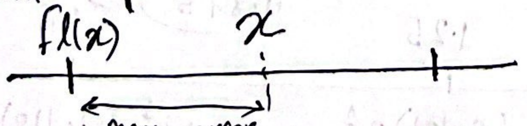
$$\delta = \frac{|fl(x) - x|}{|x|} \quad \begin{matrix} \rightarrow \uparrow \\ \rightarrow \downarrow \end{matrix} \Big\} \text{ results in maximum } \delta.$$

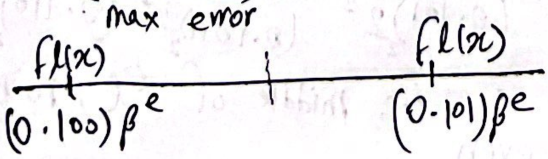
→ δ would change according to conventions.

① Convention 1:

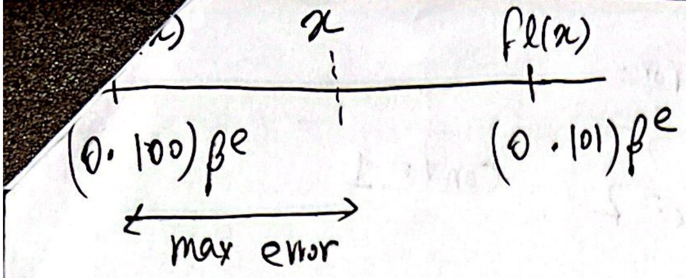
$(0.d_1 d_2 \cdots d_m)_\beta \times \beta^e$

At which point will max^m error occur? exactly in the middle.





→ find this distance then divide by 2.



$$(0.101)\beta^e - (0.100)\beta^e$$
$$= (0.001)\beta^e$$
$$= \beta^{-3}\cdot\beta^e$$
$$= \beta^{-m}\beta^e$$

$$\therefore \text{max error} = fl(x) - x$$
$$= \frac{1}{2}\beta^{-m}\beta^e$$

$$\delta = \frac{|fl(x) - x| \rightarrow \text{max}}{|x| \longrightarrow \text{min} \rightarrow (0.100)\beta^e \rightarrow \beta^{-1}\beta^e}$$

$$\therefore \text{machine epsilon } (\varepsilon_M) = \frac{\frac{1}{2}\beta^{-m}\beta^e}{\beta^{-1}\beta^e} = \boxed{\frac{1}{2}\beta^{1-m}}$$

② **Normalized Form**

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

③ **Denormalized form**

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

same.

point to notice $\boxed{\delta \le \varepsilon_M}$

FP Arithmetic with Rounding Error:

$\beta = 2 \quad m = 3 \quad e_{min} = -1 \quad e_{max} = 2 \quad conv = 1$

$x = \frac{5}{8}$ $\qquad$ $y = \frac{7}{8}$

$= (0.101)_2\, 2^0$ $\qquad$ $= (0.111)_2\, 2^0$

both are already FPs, bcz
it matches with above specification.

$\therefore fl(x) = (0.101)_2 \times 2^0$ $\qquad$ $fl(y) = (0.111)_2\, 2^0$

$\boxed{\text{Find } x * y}$

$x * y = fl(x) \cdot fl(y)$

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

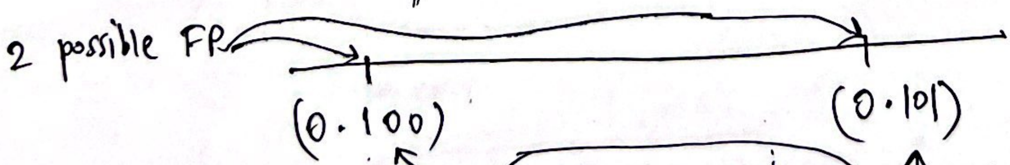
$= \frac{35}{64}$

$= (\underbrace{0.100}\,|\,0\,11)_2 \times 2^0$

→ need to take upto d_3
according to specification.

→ should we take 0.100
or 0.101 ?

If $(m+1)$ digit $= 1$, round it to next number
" " " $= 0$, " " " prev ".



$x * y = 0.100|\underline{0}11$

if $x * y = 0.100|111$

if $x * y$ perfectly in middle, round it to nearest even.

0.1001

$x * y \xrightarrow{\text{mapped to}} fl(xy) = (0.100)_2 \, 2^0$

$= \frac{1}{2}$

$= \frac{32}{64}$

originally $x * y = \frac{35}{64}$. But for toy computer $fl(xy) = \frac{32}{64}$

Bcz of Rounding error

Note:

If initially $fl(x) \neq$, $fl(y) \neq y$

approx x to $fl(x)$

" y to $fl(y)$

Then do arithmetic like $fl(x) + fl(y)$

$= \dots$

$= \dots$

$= (\quad)$

then ~~approach~~ approx again $\boxed{fl(x+y)}$

Loss of Significance

previously $x = fl(x)$, $y = fl(y)$

what if $x \neq fl(x)$, $y \neq fl(y)$?

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

Scanned with CamScanner

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

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

If we want to calculate $\boxed{x-y}$:

For scale invarient error, we have

$$\frac{x\delta_1 - y\delta_2}{x-y} \rightarrow \text{if } x \approx y, \text{ value} \approx 0, \text{ error would } \cancel{\text{increa}} \text{ increase.}$$

This is called Loss of significance.

How to avoid Los:

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

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

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

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

equal

Lets say, my toy computer can only calc upto 4 sf.

$$\sqrt{783} = 27.98$$

$$\therefore x_1 = 28 + 27.98 = 55.98$$

$$x_2 = \boxed{28} - \boxed{27.98} = 0.02000$$

not equal

close numbers.

~~Example 2~~

$$\cancel{f(x) = e^x - \cos(x) - x}$$

$$\cancel{x \in [-5 \times 10}$$

Work Around:

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

$$x^2 - \underbrace{(\alpha + \beta)}_{\alpha,\beta \text{ are roots.}} x + \alpha\beta$$

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

$$\alpha\beta = 1$$

Find α using $28 + 27.98$

$\therefore \quad \alpha = 55.98$

$$\alpha\beta = 1$$

$$55.98\,\beta = 1$$

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

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