Floating point Arithmetics

Fixed Point Representation:

Regular numbers (স্বাভাবিক সংখ্যা) that we use.

Example: Decimal → $(10.215)_{10}$, Binary → $(110.11)_2$

Base

Base

Here the point is fixed.

General form:

Point

$$x = \pm \left(\underbrace{d_1\ d_2\ d_3 \cdots d_{k-1}}_{\text{whole part}} \bullet \underbrace{d_k\ d_{k+1} \cdots d_n}_{\text{fraction part}} \right)_\beta$$

where,

d_i = digits

β = Base

But our main focus is on,

Floating Point Representation:

* Floating point number, F is subset of all Real number, $\mathbb{R}$.

$$\therefore F \subset \mathbb{R}$$

It follows a certain format. For example

$$(10.215)_{10} = (0.10215)_{10} \times 10^2$$

General form,

$$F = \pm \left(0 \cdot \underbrace{d_1 d_2 d_3 \cdots d_m}_{\text{fraction/mantissa/significand}} \right)_\beta \times \beta^{\overbrace{e}^{\text{exponent}}}$$

Base

exponent: an integer. has a range defined as, e_{min}, e_{max}.

For example, $(11.101)_2$ this is a fixed point number. To convert it into a floating point number, we have to move the decimal point & adjust the power accordingly.

$$(11.101)_2 \times 2^0$$
$$= (1.1101)_2 \times 2^1$$
$$= (0.11101)_2 \times 2^2$$ ← this is following the general form.

There are 3 conventions of floating point that we will discuss in this course.

Convention 1: (General/Standard form)

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

where $d_1 \neq 0$.

For binary,

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

where $d_1 = 1$

Note,
why $d_1 = 1$ or $d_1 \neq 0$?
In codes, if $d_1 = 0$,
then, $\pm (0.000)$ if
can produce error in
code).
Also, this can lead
unnecessary bit usage as
0.001×2^2.
↳ this 2 bits are
unnecessary.

Convention 2: IEEE Normalized form

For binary,

$$\pm (0.1\, d_1 d_2 d_3 \ldots\ldots d_m)_\beta \times \beta^e$$

Here, d_1 can be either 0 or 1.

Convention 3: IEEE Denormalized form

For binary,

$$\pm (1.d_1 d_2 d_3 \ldots\ldots d_m)_\beta + \beta^e$$

Here, d_1 can be either 0 or 1.

Note, in previous lecture notes & videos, conv 2 and conv 3 were swapped. Follow this lecture note!!!

Examples: ① Given, $\beta = 2$, $m = 3$ and a exponent e. What is the highest value using convention 1, 2, 3 ?

Soln:

conv 1 ; $\pm (0 \cdot d_1\ d_2\ d_3)_2 \times 2^e$ where $d_1 = 1$.

$\therefore$ max = $(0 \cdot 1\ 1\ 1)_2 \times 2^e$

conv 2 : $\pm (0 \cdot 1\ d_1\ d_2\ d_3)_2 \times 2^e$

$\therefore$ max = $(0 \cdot 1\ 1\ 1\ 1)_2 \times 2^e$

conv 3 : $\pm (1 \cdot d_1\ d_2\ d_3)_2 \times 2^e$

$\therefore$ max = $(1 \cdot 1\ 1\ 1)_2 \times 2^e$

② Given, $\beta = 2$, $m = 3$, $e = [-1, 2]$ ($e_{min} = -1$, $e_{max} = 2$), using conv 1 and conv 3 how many combination is possible ? (non negative)

Soln:

Here, $e_{min} = -1$, $e_{max} = 2$ $\therefore e = \{-1, 0, 1, 2\}$

We know,

conv 1 ; $(0 \cdot d_1\ d_2\ d_3)_2 \times 2^e$ where $d_1 = 1$.

For each value of e,

$(0 \cdot 1\ d_2\ d_3)_2 \times 2^e$

d_2	d_3	
0	0	} 4 combinations
0	1	
1	0	
1	1	

in conv 1, $d_1 = 1$. It is fixed. So, we only find out the number of combination possible using $d_2, d_3 \ldots$.

We get 4 combination.

$\therefore$ for 4 values of e $(-1, 0, 1, 2)$ we get, $4+4+4+4 = 16$ combination.

conv 2: $(1 \cdot d_1\ d_2\ d_3)_2 \times 2^e$

$d_1\ d_2\ d_3$: 0 0 0, 0 0 1, ⋮, 1 1 0, 1 1 1 } total $2^3 = 8$ combinations for each e.

∴ For 4 values of e $(-1, 0, 1, 2)$, we get total $4 * 8 = 32$ values.

③ Given, $\beta = 2$, $m = 3$, $e = [-1, 2]$ (e_{min}, e_{max}), using conv 1 & conv 2 what is the smallest possible non negative value & highest value?

Soln:

$$e = \{-1, 0, 1, 2\}$$

We know,

conv 1: $(0 \cdot d_1\ d_2\ d_3)_2 \times 2^e$ where $d_1 = 1$.

∴ smallest possible non negative value $= (0 \cdot 1\ 0\ 0)_2 \times 2^{-1}$

∴ highest value $= (0 \cdot 1\ 1\ 1)_2 \times 2^2$

in conv 1, $d_1 = 1$ and $2^{e_{min}}$ makes the smallest value. and $2^{e_{max}}$ makes the largest value

We know,

conv 2: $(0 \cdot 1\ d_1\ d_2\ d_3)_2 \times 2^e$

∴ smallest possible non negative value $= (0 \cdot 1\ 0\ 0\ 0)_2 \times 2^{-1}$

∴ highest value $= (0 \cdot 1\ 1\ 1\ 1)_2 \times 2^2$

④ Given, $\beta = 2$, $m = 3$, $e = [-1, 2]$ (e_{min}, e_{max}), Find the lowest value in conv 1.

Soln:

conv 1: $\pm(0 \cdot d_1\ d_2\ d_3)_2 \times 2^e$ where $d_1 = 1$.

lowest value $= -(0 \cdot 1\ 1\ 1)_2 \times 2^2$

Sometimes the questions may ask for the decimal representation. In that case, you have to convert the number into decimal.

Recap: how to convert bases of a number

Number Conversion:

Type 1: Decimal to other bases

Type 2: Other Base to Decimal

Type 1

Decimal to other Base:

$(43.3125)_{10}$ ← Base

Whole number | Fractions

* Whole number → Repeated Division by Base R.
* Fraction number → Repeated Multiplication by Base R.

Example: $(43.3125)_{10} = (?)_2$

2 | 43	→ 1	MSB
2 | 21	→ 1	
2 | 10	→ 1	
2 | 5	→ 0	
2 | 2	→ 1	
2 | 1	→ 0	
0	→ 1	LSB

We stop at 0.

$\therefore (43)_{10} = (101011)_2$

	Carry
.3125 × 2 = 0.625	0 (LSB)
.625 × 2 = 1.25	1
.25 × 2 = 0.5	0
.5 × 2 = 1.0	1 (MSB)

We stop when it reach 0 or the limit will be given

$\therefore (.3125)_{10} = (.0101)_2$

LSB = Least Significant Bit
MSB = Most Significant Bit

$\therefore (43.3125)_{10} = (101011.0101)_2$

* $(34.215)_{10} = (?)_5$

5	34	
5	6	→ 4
5	1	→ 1
	0	→ 1

$(34)_{10} = (114)_5$

.215 × 5	= 1.075	1
.075 × 5	= 0.375	0
.375 × 5	= 1.875	1
.875 × 5	= 4.375	4
.375 × 5	= 1.875	1
.875 × 5	= 4.375	4

$(.215)_{10} = (101414\ldots)_5$

$\therefore (34.215)_{10} = (114.101414\ldots)_5$

Type 2

Any Base to Decimal.

$(1101.101)_2 = (?)_{10}$

Positions: ..... 3 2 1 0 −1 −2 −3 .....

1 1 0 1 . 1 0 1

$= 1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 + 1 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3}$

2^3: Base (Position)

$= (13.625)_{10}$

* $(572.6)_8 = (?)_{10}$

2 1 0 −1

5 7 2 . 6

$\Rightarrow 5 \times 8^2 + 7 \times 8^1 + 2 \times 8^0 + 6 \times 8^{-1}$

$= (378.75)_{10}$

⑤ Given, $\beta = 2$, $m = 4$, $e = [-1, 3]$ (e_{min}, e_{max}), Find the highest & lowest value in conv 1 in decimal.

Soln:

conv 1 ; $\pm (0.d_1 d_2 d_3 d_4)_2 \times 2^e$ where $d_1 = 1$.

highest value $= (0.1111)_2 \times 2^3$ (positions $-1, -2, -3, -4$)

$= (1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4}) \times 2^3$

$= \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}\right) \times 2^3$

$= \frac{15}{2}$

$= 7.5$

lowest value $= -7.5$ (Ans)

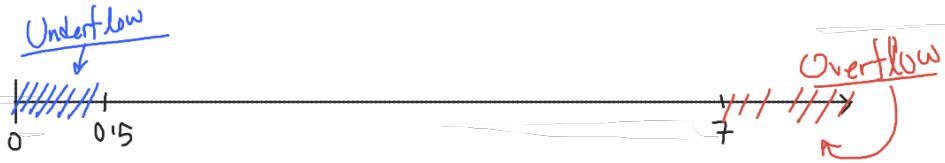
Think, how we directly found the lowest value.

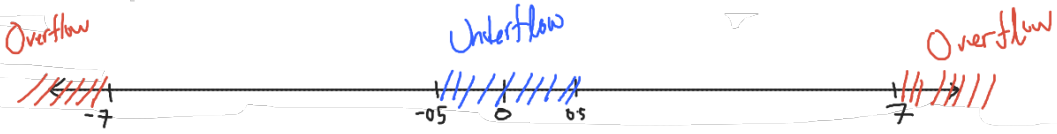
#Now, if we put the highest and lowest value for

$\beta = 2$, $m = 3$ $e = [0, 3]$

highest v $= (0.111)_2 \times 2^3 = (2^{-1} + 2^{-2} + 2^{-3}) \times 2^3 = 7$

lowest (non-ve) v $= (0.100)_2 \times 2^0 = 2^{-1} \times 2^0 = 0.5$





Floating point Representation Limits:

1. Range are finite and limited. Causes Overflow & Underflow
2. Possible numbers are not equally spaced.

Now, what does this mean :

Possible numbers are not equally spaced.

Let, $\beta = 2$, $m = 3$, $e = [-1, 2]$

Possible numbers are,

$(0.100)_2 \times 2^{-1} = 1/4$
$(0.101)_2 \times 2^{-1} = 5/16$
$(0.110)_2 \times 2^{-1} = 3/8$
$(0.111)_2 \times 2^{-1} = 7/16$

$(0.100)_2 \times 2^{0} = 1/2$
$(0.101)_2 \times 2^{0} = 5/8$
$(0.110)_2 \times 2^{0} = 3/4$
$(0.111)_2 \times 2^{0} = 7/8$

$(0.100)_2 \times 2^{1} = 1$
$(0.101)_2 \times 2^{1} = 5/4$
$(0.110)_2 \times 2^{1} = 3/2$
$(0.111)_2 \times 2^{1} = 7/4$

$(0.100)_2 \times 2^{2} = 2$
$(0.101)_2 \times 2^{2} = 5/2$
$(0.110)_2 \times 2^{2} = 3$
$(0.111)_2 \times 2^{2} = 7/2$

If we put this values in a number line,

Observation:

① If the value of e stays the same, then the number are equally spaced.

② the lower the value of e, the narrower is the gap. So, more precise the value gets.

IEEE Standard (1985) For double Precision (64 bits)

To ensure every system are following a standard.

sign bit — exponent — Fraction / Mantissa

0 = +ve
1 = -ve

∴ Sign 1 bit → provides + -

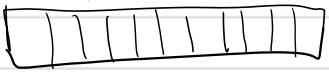
Exponent 11 bits → provides Range

fraction 52 bits → Provides precision

total 64 bit

Starting with De normalized form: (Conv 3),

$(1.d_1 d_2 \ldots\ldots d_{52}) \times 2^e$ and $e = 11$ bits $= 2^{11}$ combination $= 2048$



$\therefore e_{min} = (00000000000)_2 = 0$

$\therefore e_{max} = (11111111111)_2 = 2047$

$\therefore e = [0, 2047]$

$\therefore$ largest number $= (1.11\ldots1) \times 2^{2047}$ = too large value

$\therefore$ Smallest " $= (1.0\ldots0) \times 2^0 = 1$ = Not too small.
(non negative)

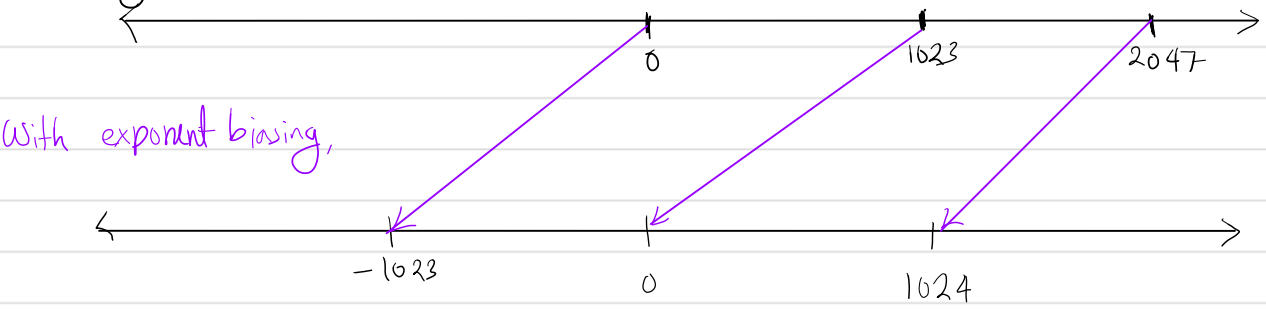
Problem: We can't represent 0.1, 0.11, 0.0001

We have unnecessarily big number & not too small number.

To solve this, we use Exponent Biasing.

Exponent biasing means scaling the middle value of $[0, 2047]$; which is 1023 in such a way that it becomes 0.

Previously,



With exponent biasing,

So, we do this by subtracting 1023 from e...

$$\therefore (1 \cdot d_1 d_2 d_3 \ldots d_{52}) \times 2^{e-1023}$$

this value is called bias.

this subtraction is called exponent biasing.

We will focus in this form.

In case of Normalized form,

$$F = \pm (0 \cdot 1\, d_1 d_2 \ldots d_{52}) \times 2^{e-1022}$$

because,

$$(1 \cdot d_1 d_2 \cdots) \times 2^{1023}$$
$$(0 \cdot 1 d_1 d_2 \cdots) \times 2^{1022}$$

$e \in [0, 2047]$

Previously, $e \in [0, 2047]$

$(e - 1023) \in [-1023, 1024]$

So, largest possible number,

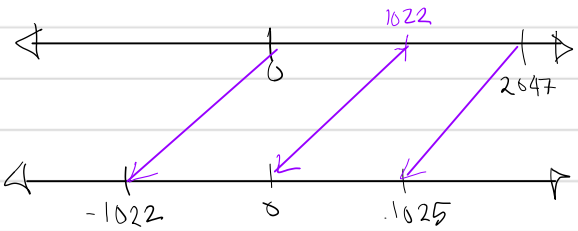
$$(0.111 \ldots 11) \times 2^{1025}$$

smallest possible number,

$$(0.100 \ldots 0) \times 2^{-1022}$$

= a very tiny value close to 0.

in normalized form, exponent is shifted like this,



* But what about absolute 0 and ∞ ?

These are special case, We have reserved exponents for them.

$$0 = 2^{e_{min}} = 2^{-1022}$$
$$\infty = 2^{e_{max}} = 2^{1025}$$

Finally,

the highest possible value (except ∞) $= (0.111\cdots1)_2 \times 2^{1024}$

$\approx 1.798 \times 10^{308}$

the lowest possible non negative value (except 0) $= (0.10\cdots0)_2 \times 2^{-1021}$

$\approx 2.225 \times 10^{-308}$

#Note: You don't have to memorize these decimal values. Just see, how we calculate the value.

Example: What is the lowest value possible (non negative) value if bias is 512 in IEEE Normalized form 64 bit?

$$F = \pm(0.1\, d_1\, d_2 \cdots d_{52}) \times 2^{e - bias}$$
$$= 0.1 \times 2^{0-512}$$
$$= 0.1 \times 2^{-512}$$

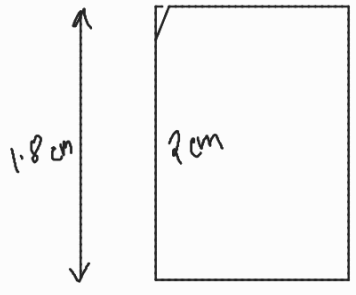
(Ans)

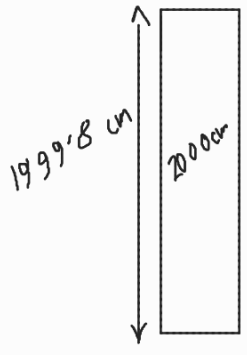
Rounding Error

$x =$ ~~~ $\quad fl(x) =$

$$\text{Error} = \left|(\text{Actual value} - \text{Rounded value})\right|$$

$$= \left| x - fl(x) \right|$$





$(2 - 1.8) = 0.2$ ✓ $\qquad (2000 - 1999.8) = 0.2$ ✓✓

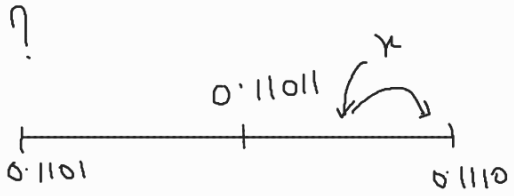
$$\text{Error} = \frac{|\text{Actual} - \text{Rounded}|}{\text{Actual}}$$

$$\frac{0.2}{2} = \underline{0.1} \qquad \frac{0.2}{2000} = \underline{0.0001}$$

Scale Invariant Rounding Error,

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

Let, $x = (0.11011101)_2$ $\quad m = 4$, $\quad \delta = ?$



$$fl(x) = (0.1110)_2$$

$$\delta = \frac{|(0.11011101)_2 - (0.1110)_2|}{|(0.11011101)_2|} =$$

Machine Epsilon (ϵ)

Maximum Scale Invarient Rounding Error;

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

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

Conv 1: $\epsilon = \frac{1}{2} \beta^{1-m}$

Conv 2: $\epsilon = \frac{1}{2} \beta^{-m}$

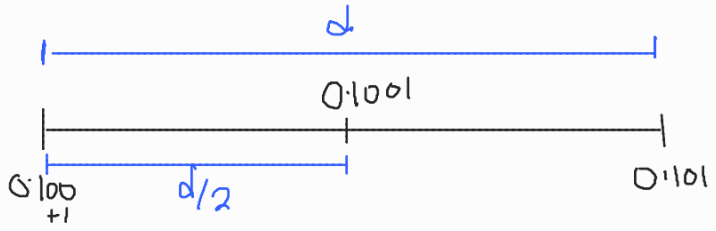
Conv 3: $\epsilon = \frac{1}{2} \beta^{-m}$

Conv 1:

$m = 3$

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

$$(0.\overset{=1}{d_1}d_2d_3)_\beta \times \beta^e$$



$$d = (0.001)_2 \times 2^e \qquad d/2 = \frac{1}{2}\beta^{-m} \times \beta^e$$

$$= 2^{-3} \times 2^e \qquad \therefore |x - fl(x)|_{max} = \frac{1}{2}\beta^{-m} \times \beta^e$$

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

$$|x|_{min} \quad (0.\overset{=1}{d_1}d_2d_3)_2 \times 2^e$$

$$= (0.100)_2 \times 2^e$$

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

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

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

$$= \frac{1}{2}\beta^{-m+1} \qquad \boxed{= \frac{1}{2}\beta^{1-m}}$$

Conv 3:

$m = 3$

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

$$(1.d_1d_2d_3) \times 2^e$$

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

$$|x|_{min} = (1.000)_2 \times 2^{e}$$
$$= 2^{0} \times 2^{e}$$
$$= 2^{e} = \beta^{e}$$

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

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

Convention 2: $\boxed{m=3}$

$$(0.1\, d_1 d_2 d_3)_2 \times 2^{e}$$

$$d = (0.0001)_2 \times 2^{e} \quad (0, -1, -2, -3, -4)$$
$$= 2^{-4} \times 2^{e}$$
$$= \beta^{-m-1} \times \beta^{e}$$

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

$$|x|_{min} = (0.1000)_2 \times 2^{e} \quad (-1)$$
$$= 2^{-1} \times 2^{e}$$
$$= \beta^{-1} \times \beta^{e}$$

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

Example: $\beta = 2$, $\boxed{m = 3}$ calculate rounding error for $(x * y)$.

$$x = \frac{5}{8} = \frac{1}{8} + \frac{4}{8} = \frac{1}{2^3} + \frac{1}{2^1} = 2^{-1} + 2^{-3} = (0.\overset{-1}{1}0\overset{-3}{1})_2$$

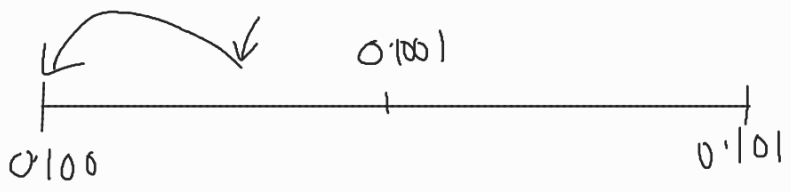
$$y = \frac{7}{8} = \frac{4}{8} + \frac{2}{8} + \frac{1}{8} = \frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3} = 2^{-1} + 2^{-2} + 2^{-3} = (0.111)_2$$

$$x * y = \frac{5}{8} * \frac{7}{8} = \boxed{\frac{35}{64}}$$

$$x * y = \frac{32}{64} + \frac{2}{64} + \frac{1}{64} = \frac{1}{2} + \frac{1}{32} + \frac{1}{64} = \frac{1}{2^1} + \frac{1}{2^5} + \frac{1}{2^6}$$

$$= 2^{-1} + 2^{-5} + 2^{-6} = (0.1\,0\,0\,0\,1\,1)_2$$

$fl(x*y)$



$$fl(x*y) = (0.\overset{-1}{1}00)_2 = 2^{-1} = \boxed{\frac{1}{2}}$$

$$\text{rounding error, } \delta = \frac{|(x*y) - fl(x*y)|}{|x*y|}$$

$$= \frac{\left|\frac{35}{64} - \frac{1}{2}\right|}{\left|\frac{1}{2}\right|} = 0.09375$$

Ans

Loss of significance:

Significant figure:

5 s.f

$19.215|105 = 19.215$

$0.00\boxed{31256}712 = 0.00031256$

$= 6.31256 \times 10^{-2}$

Ex:

$x^2 - 56x + 1 = 0$, $x_1 = ?$ $x_2 = ?$

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

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

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

$$\sqrt{783} = 27.98$$

4 s.f

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

$$x_2 = 28 - 27.98 = 0.02$$

Solution

$$x^2 + (\alpha + \beta)x + \alpha\beta = 0$$

$$\alpha = x_1$$

$$\beta = \frac{1}{x_1} = x_2$$

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

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

In the system

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

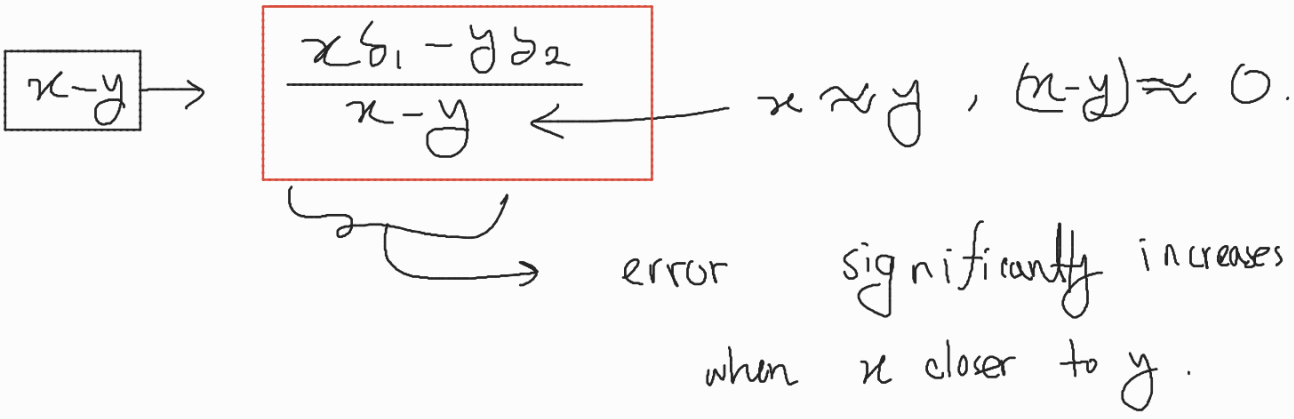
$$\Rightarrow x + \delta_x \pm y + \delta_2 y$$

$$\Rightarrow x(1 + \delta_1) \pm y(1 + \delta_2)$$

$$\Rightarrow (x \pm y) \pm x\delta_1 \pm y\delta_2$$

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

Scale invarient error


$x-y$
$\frac{x\delta_1 - y\delta_2}{x-y}$
$x \approx y$, $(x-y) \approx 0$.
error significantly increases
when x closer to y.