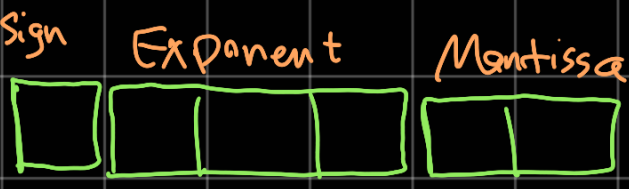
to better understand IEEE 754 standard, here I devise my own 6 bit Floating point number!




$E = (11)_2 = 3$ (half of exponent $(111)_2$)

$$= (-)^{sing} \times a \times 2^{b-E}$$

| | | | |
|---|---|---|---|
| 0 | 0 | 0 | → used to represent non normalized |
| 0 | 0 | 1 | → used to represent normalized |
| 0 | 1 | 0 | |
| 0 | 1 | 1 | |
| 1 | 0 | 0 | |
| 1 | 0 | 1 | |
| 1 | 1 | 0 | |
| 1 | 1 | 1 | → used to represent ± inF, NaN |

★ Normalized and non normalized.

$$(1.a_1a_2)_2 = \left(1 + \frac{1}{2}a_1 + \frac{1}{4}a_2\right) = \text{normalized}$$

$$(0.a_1a_2)_2 = \left(0 + \frac{1}{2}a_1 + \frac{1}{4}a_2\right) = \text{non normalized}$$

exponent 000 signals that a number is non normalized.

## ✯ Possible values for Exponent

| | | | | | |
|---|---|---|---|---|---|
| 0 | 0 | 0 | → 0 | ⇝ | -2 (For non normalized) |
| 0 | 0 | 1 | → 1 | -3 → | -2 |
| 0 | 1 | 0 | → 2 | -3 → | -1 |
| 0 | 1 | 1 | → 3 | -3 → | 0 |
| 1 | 0 | 0 | → 4 | -3 → | 1 |
| 1 | 0 | 1 | → 5 | -3 → | 2 |
| 1 | 1 | 0 | → 6 | -3 → | 3 |
| 1 | 1 | 1 | | | |

## ✯ possible Values For mantissa

Normalized

1.00
1.01
1.10
1.11

⇒

$\begin{pmatrix} 1+0 \\ 1+\frac{1}{4} \\ 1+\frac{2}{4} \\ 1+\frac{3}{4} \end{pmatrix}$

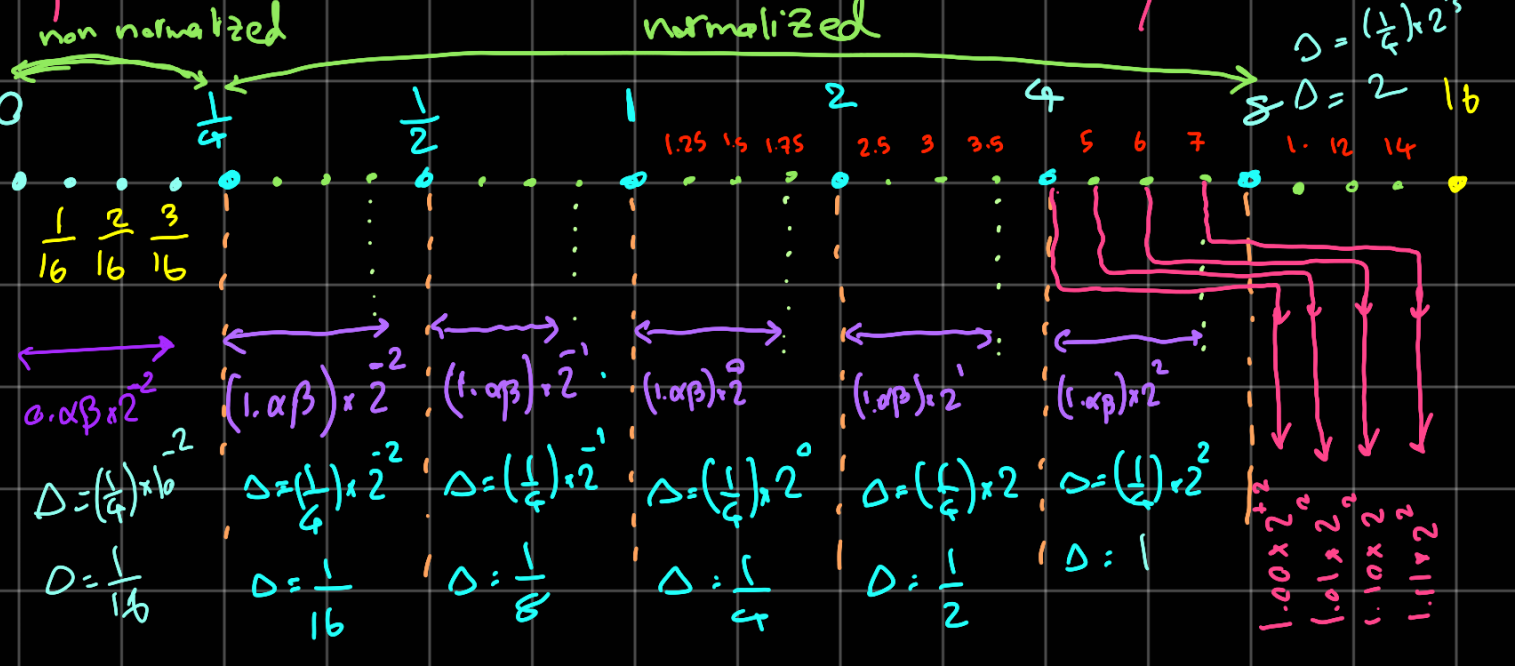
Non Normalized

0.00
0.01
0.10
0.11

⇒

$\begin{pmatrix} 0+0 \\ 0+\frac{1}{4} \\ 0+\frac{2}{4} \\ 0+\frac{3}{4} \end{pmatrix}$

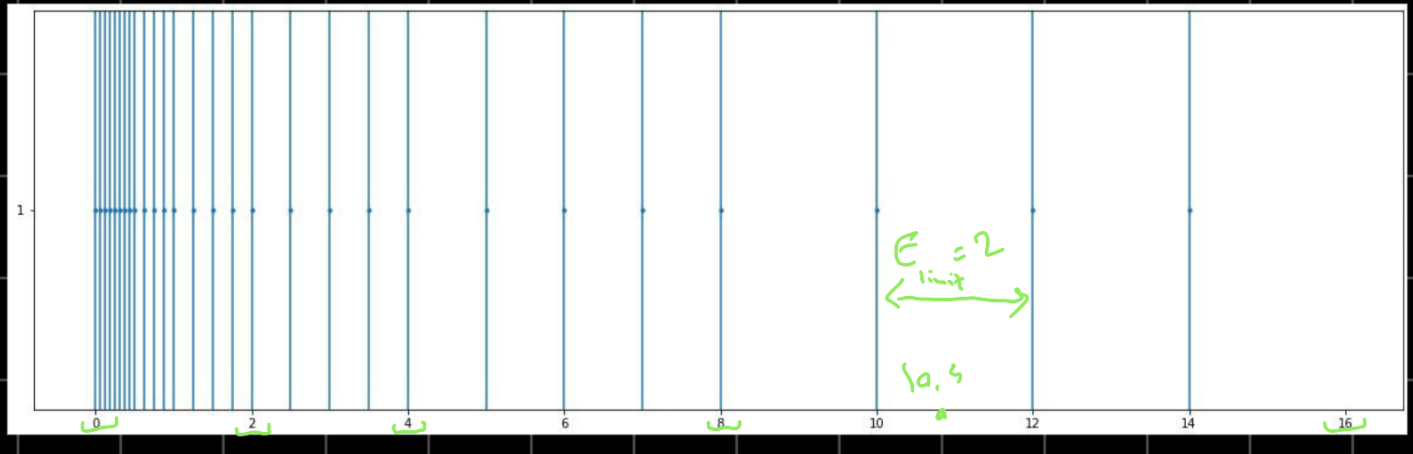
## ✯ possible values to show with my ieee standard



So in fact we only have the following numbers in our representation.

| | | | | |
|---|---|---|---|---|
| 0 | 0.0625 | 0.125 | 0.1875 | → non normalized |
| 0.25 | 0.3125 | 0.375 | 0.4375 | normalized |
| 0.5 | 0.625 | 0.75 | 0.875 | |
| 1 | 1.25 | 1.5 | 1.75 | |
| 2 | 2.5 | 3 | 3.5 | |
| 4 | 5 | 6 | 7 | |
| 8 | 10 | 12 | 14 | |

and here is a plot of their spacing




Suppose that our true number is 10.5. If we want to represent it with my representation

we will have $(1.01)_2 \times 2^3$. the difference (Error) will be bounded by $(0.01)_2 \times 2^3 = 2$ (error will not be bigger than 2)

So the relative error will be:

worst error

$$\left(\frac{a-\tilde{a}}{a}\right) \quad |a-\tilde{a}| \leq (0.01)_2 \times 2^3 \qquad \frac{(0.01)_2 \times 2^3}{(1.00)_2 \times 2^3} = (0.01)_2$$

to Calculate the worst error I set a to be smallest possible

Important

No matter what valid number (in valid range) we want to represent with my Ieee standard, the relative error will always be less than $(0.01)_2$ which is $\frac{1}{4}$

Relative error will not be larger than $1/4$

# Example

True number 13

nearest number in our set $\Rightarrow$ 12

relative error $= \frac{13-12}{12} = \frac{1}{12} < \frac{1}{4}$ ✓

True number 5.5

nearest number $\longrightarrow$ 5

relative error: $\frac{0.5}{5} = \frac{1}{10} < \frac{1}{4}$

# example

$$F_{(x_0+\Delta h)} = F(x_0) + \frac{dF}{dx}\Delta h + \frac{1}{2}\frac{d^2F}{dx^2}\Delta h^2 + \frac{1}{6}\frac{d^3F}{dx^3}\Delta h^3 + \dots$$

(estimate) $\tilde{F}(x_0+\Delta h) = F(x_0) + \frac{dF}{dx}\Delta h$

relative error $= \frac{\tilde{F}(x_0+\Delta h) - F(x_0+\Delta h)}{\tilde{F}(x_0+\Delta h)} = \frac{\frac{1}{2}\frac{d^2F}{dx^2}\Delta h^2 + \dots}{F(x+\Delta h)}$

we should set the value of $\Delta h$ in a way that the relative error is less than the accuray of our floating point number ($10^{-7}$ in case of 32 bit, $10^{-15}$ for 64 bit and $\frac{1}{64}$ for my own standard!)

note that if we use reduced scales in our computation/simulation, then the $F(x+\Delta x) \simeq F(x)$ will not be so big or so small. will be around 1.

---

## ☆ Deriving smallest and largest numbers 8

**My own 6 bit standard (Positives)**

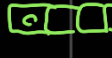
Smallest non normalized

000 01

$= (\frac{1}{2}^2) \times (\frac{1}{2}^2) = \frac{1}{2}^4 = 0.0625$

**IEEE 32 bit standard (positive)**

Smallest non normalized



ex 00 ex 0 00001

$= (\frac{1}{2}^{126}) \times (\frac{1}{2}^{23}) = \frac{1}{2}^{149}$

## Largest non normalized number

000 11

$= (2^{-2}) \times (\frac{3}{4}) = \frac{3}{16} = .1875$

0x00 0x7FFFFF

$(2^{-126}) \times (\frac{1}{2^{23}} + \frac{1}{2^{22}} + \cdots + \frac{1}{2})$

$= (2^{-126}) \times (1 - 2^{-23})$

## Smallest normal value

(1-3)

001 00

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

(1-127)

0x01 0x000000

$2^{-126} \times 1 = 2^{-126}$

## largest normal value

(6-3) $(1.11)_2$

110 11

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

$= \frac{7}{4} \times 8 = 14$

(254-127)

0xFE 0x7FFFFF

$2^{127} \times (2 - 2^{-23})$

## one

(3-3) $(1.00)_2$

011 00

$(2^0) \times (1) = 1$

127-127

0x7F 0x000000

$(2^0) \times (1) = 1$

## Smallest number bigger than one

(3-3) $(1.01)_2$

011 01

$(2^0) \times (1 + \frac{1}{4}) = 1.25$

0x7F 0x000001

$(2^0) \times (1 + 2^{-23}) = 1 + 2^{-23}$

# Largest value less than one

$(2_{-3})$ $(1.11)_2$

010 11

$(2^{-1})\left(1+\frac{3}{4}\right) = \frac{7}{8}$

$(126_{-127})$

0x7E 0x7FFFFF

$(2^{-1}) * (2 - 2^{-23}) = 1 - 2^{-24}$