$2^0 \quad 2^0 + \varepsilon_m \quad 2^0 + 2\varepsilon_m$

Single $\varepsilon_m \approx 10^{-8}$

Double $\varepsilon_m \approx 10^{-16}$

**Def:** - Let $x$ denote a float representable number then $fl(x)$ is the associated float.

- $$\frac{|x - fl(x)|}{|x|} \leq \varepsilon_m$$

---

## Floating pt arithmetic

Q: How many real # are there?

Uncountably infinite.

Q: Can the machine represent all of them?

No.

The largest number that you can represent on

The largest number that you can represent on the machine is $1.79 \times 10^{308}$. This number is called the overflow (OFL). The smallest # that you can represent on the machine is $2.23 \times 10^{-308}$. This is called the underflow (UFL).

On the machine, numbers are represented by truncated binary series of the form

$$\alpha = (a_0 2^0 + a_1 2^{-1} + a_2 2^{-2} + \cdots + a_p 2^{-(p-1)}) 2^e$$

Where $p$ is the precision, & $e$ is the exponential. (common factor in base 2)

Def: The representation of a real # on the machine is called is floating pt representation.

For $x \in \mathbb{R}$ that is represented w/ a float, $fl(x)$ denotes it floating pt representation.

Types of precision

| | p | Machine Epsilion $\varepsilon_m$ |
|---|---|---|
| single | 24 | $\varepsilon_m = 2^{-23} \approx 10^{-8}$ |
| double | 53 | $\varepsilon_m = 2^{-52} \approx 10^{-16}$ |

## Warm-up

Let $\varepsilon_m = 10^{-2}$

(i) set $x = 1/3$. What is $\tilde{x} = fl(x)$?

What is $\frac{|x-\tilde{x}|}{|x|}$?

(ii) set $x = 1234$. What is $\tilde{x} = fl(x)$?

what is $\frac{|x-\tilde{x}|}{|x|}$?

Soln: (i) $\tilde{x} = fl(x) = 0.33$

absolute error $|x - \tilde{x}| = 0.0033333$

relative error $\frac{|x-\tilde{x}|}{|x|} = 10^{-2}$

(ii) $\tilde{x} = 1200$

absolute error $|x - \tilde{x}| = 1234 - 1200 = 34 > 10^{-2}$

relative error $\frac{|x-\tilde{x}|}{|x|} = \frac{34}{1234} \sim 10^{-2}$

$0.027$

$2.7 \times 10^{-2}$

Def: The most accurately we can approximate a real # that is representable w/ a float is machine

# that is representable w/ a float is machine epsilon $\varepsilon_m$.

i.e. let $x \in \mathbb{R}$ s.t. representable by a float, let $\tilde{x} = fl(x)$

then $\frac{|x - \tilde{x}|}{|x|} \leq \varepsilon_m$

The biggest float $\sim 10^{308}$. This # is called the overflow #. (OFL). The smallest float is

$\sim 10^{-308}$. This is called the underflow #. (UFL)

The ordering of everything you can represent on the computer is

$$0 < UFL < \varepsilon_m < OFL$$

## Exceptional values

| Computer | name | How to make it |
|---|---|---|
| Inf | Inifinity | $\pm ?/0$ or go over OFL |
| NaN | Not a # | $0/0$, $0 \cdot Inf$, $Inf/Inf$ |