Computer Arithmetic

We often want to work with the real number system which consists of all integers, rational and irrational numbers

ex $\{-2, \sqrt{3}, e, \pi\} \subset \mathbb{R}$

In a computer, we have finite storage for numbers

$\Longrightarrow$ not all real numbers can be represented exactly

Clearly, non-repeating / non-terminating decimals cannot be represented, but there are lots of others as well.

This can cause problems with arithmetic.

We typically use base 10 decimal system

eg 427.325

$= 4\times10^{2} + 2\times10^{1} + 7\times10^{0} + 3\times10^{-1} + 2\times10^{-2} + 5\times10^{-3}$

Computers often use the binary (base 2) system

$$(1001.11101)_2 = 1\times 2^3 + 0\times 2^2 + 0\times 2^1 + 1\times 2^0 + 1\times 2^{-1} + 1\times 2^{-2} + 1\times 2^{-3} + 0\times 2^{-4} + 1\times 2^{-5}$$

Verify (not to be handed in)

$$(1001.11101)_2 = (9.90625)_{10}$$

Note that the

$$(\ )_{10} \Longleftrightarrow (\ )_2$$

process can lead to errors!

<u>Ex.</u> What is $\left(\frac{1}{10}\right)_{10}$ in $(\ )_2$?

Assume $\frac{1}{10} = (.a_1\, a_2\, a_3 \ldots)_2$

Multiply by 2 :

$$\frac{2}{10} = (a_1 . a_2\, a_3 \ldots)_2$$

Take integer part of both sides

$$0 = a_1$$

Continue $\frac{4}{10} = (a_2 . a_3 a_4 \ldots)_2$

$\Rightarrow \quad a_2 = 0$

$$\frac{8}{10} = (a_3 . a_4 a_5 \ldots)_2$$

$\Rightarrow \quad a_3 = 0$

$$\frac{16}{10} = (a_4 . a_5 a_6 \ldots)_2$$

$\Rightarrow \quad a_4 = 1$ (taking integer part)

Subtract 1

$$\frac{6}{10} = (. a_5 a_6 a_7 \ldots)_2$$

$$12/10 = (a_5 . a_6 a_7 \ldots)_2$$

$\Rightarrow \quad a_5 = 1$

Subtract 1

$$\frac{2}{10} = (. a_6 a_7 a_8 \ldots)_2$$

repeats

$$\frac{2}{10} = (. a_6 a_7 a_8 \ldots)_2$$

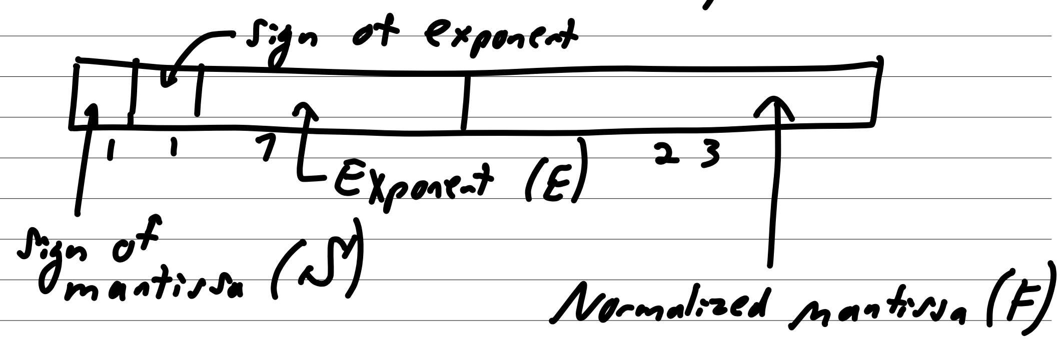
We have

$$\left(\frac{1}{10}\right)_{10} = (0.001\ 1001\ 1001 \ldots)_2$$

Since computers have a finite storage the number on the right cannot be stored exactly.

The decimal has to be truncated somehow ...

Hypothetical storage scheme (32 bit)



Normalization

Can write all real numbers in normalized scientific notation

eg $\quad 732.5051 = 0.7325051 \times 10^{3}$

$\quad -0.005612 = -0.5612 \times 10^{-2}$

if $x \in \mathbb{R}$ then $x = \pm r \times 10^{n}$

$(x \neq 0)$

where $\frac{1}{10} \leq r < 1$ (if $r < \frac{1}{10}$ it is not normalized; shift some more).

In binary

$$x = \pm q \times 2^{m}$$

where $\frac{1}{2} \leq q < 1 \quad (x \neq 0)$

q : mantissa

m : integer exponent

written in base 2