# Floating-point numbers

Spring 2021 A. Donev

N - digits

Scientific / exponential format

$\pm 3.1578\ldots \cdot 10^{5}$

$\pm 0.31578 \cdot 10^{6}$

zero

non zero

exponent

$$X = (-1)^{s} \cdot 0.\underbrace{\overbrace{a_1 a_2 a_3 \ldots a_t}^{m}}_{t \text{ digits}} \cdot \beta^{e}$$

$\underbrace{s = 0 \text{ or } 1}_{\text{one bit}}$

$$\rightarrow \quad = (-1)^{s} \cdot m \cdot \beta^{e-t}$$

mantissa

fixed

On computer  floating point number

$$[\, s \mid m \mid e \,]$$

sign  mantissa  exponent

$\beta = 2$

$$134_{10} = 1 \cdot 10^2 + 3 \cdot 10^1 + 4 \cdot 10^0$$

$$100101_2 = 1 \cdot 2^5 + 1 \cdot 2^2 + 1 \cdot 2^0$$

\# of bits = 32, 64, 128 bit

4 bytes  8 bytes  16 byts

Standard  IEEE

- format
- rounding
- exceptions  $1/0$, $\sqrt{-1}$

# IEEE formats

Single precision $= 4$ bytes
$= 32$ bits

$$\underset{\substack{\text{sign} \\ s}}{1} + \underset{\substack{\text{exponent} \\ e}}{\underbrace{8}} + \underset{\substack{\text{mantissa} \\ m}}{23} = 32$$

## Double precision

$= 64$ bits
$= 8$ bytes

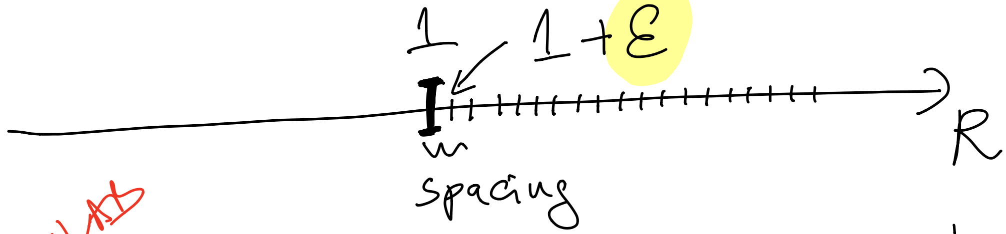
$$1 + 11 + 52 = 64$$

## Rounding errors

$$\hat{x} = fl(x)$$

$x \in \mathbb{R}$

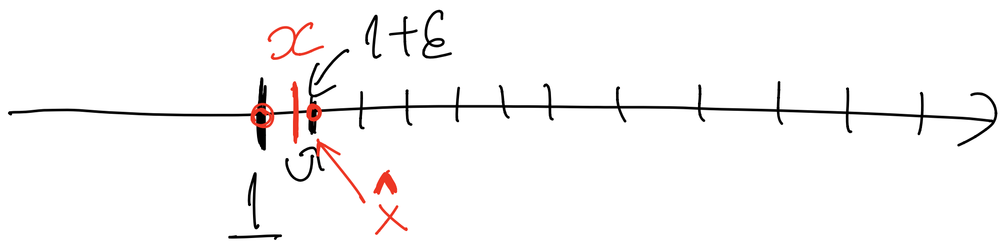
$$\hat{x} \approx x$$




eps in MATLAB

$$\varepsilon = 2^{-\#\text{ of bits in Mantissa}}$$

$$\varepsilon = \begin{cases} 2^{-23} & \text{single precision} \\ 2^{-53} & \text{double precision} \end{cases}$$




If we round to nearest

$$\underbrace{\frac{|x-\hat{x}|}{|x|}}_{\text{relative error}} \le \frac{\varepsilon}{2} = u$$

roundoff unit
machine precision

$$= \begin{cases} 2^{-24} \approx 6\cdot 10^{-8} & SP \\ 2^{-53} \approx 1\cdot 10^{-16} & DP \end{cases}$$

Double precision gives you at most 16 digits of accuracy

There is a largest & smallest number $\sim 10^{300}$

$$fl(x) \odot fl(y) = fl(x \circ y)$$

hardware

$+, -, *, /$

$*$ and $+$ not associative

$$(x+y)+z \neq x+(y+z)$$

$$1/0 = \underline{\underline{\text{inf}}}$$

$$\sqrt{-1} \text{ or } 0/0 = \text{NaN}$$

not a number

Matlab uses double precision by default

To know:

1) Do not compare floating-point numbers

$x = y$ ?

2) $x + y - z$ use parenthesis to (try to) control order

$(x+y) - z$

$x + (y - z)$

3) Built in "standard" functions $\sin$, $\cos$, $\ln$, $\exp$ give you 16 digits.

# Propagation of roundoff error

**Mathematically equivalent expressions are not equivalent in finite-precision arithmetic**

Multiplication: $\underline{x}\, y =$

$$\frac{|(x+\delta x)(y+\delta y) - xy|}{|xy|}$$

$$= \left| \frac{\delta x}{x} + \frac{\delta y}{y} + \overbrace{\cancel{\left(\frac{\delta x}{x}\right)\left(\frac{\delta y}{y}\right)}}^{\text{small}} \right|$$

$$\underbrace{\delta x / x}_{\sim 10^{-16}} \ll 1, \quad \delta y / y \ll 1$$

$$\leq \left| \frac{\delta x}{x} \right| + \left| \frac{\delta y}{y} \right|$$

$$\varepsilon_{xy} \leq \varepsilon_x + \varepsilon_y$$ relative errors

<u>Good</u>

Multiplication / division are "safe" - they don't loose digits

<u>Addition</u>

$$\begin{array}{ll} \underbrace{1.0010} & \text{5 digits} \\ 0.000\underbrace{13678} & \text{5 digits} \\ \hline \underbrace{1.0013}\,678 \leftarrow & \text{lost} \end{array}$$

If round to nearest

1.0014

$13\boxed{678} \leftarrow$ Lost 3 digits

14 000

Catastrophic cancellation

Add numbers of widely different magnitude

Subtraction

$$\begin{array}{r} 1.0012, 321 \\ -\ 1.0011, 020 \\ \hline 0.0001, \underbrace{201}_{\text{lost these digits}} \end{array}$$

Subtraction of numbers that are very close to each other

# Example

Harmonic sum

$$H(N) = \sum_{i=1}^{N} \frac{1}{i}$$

forward or backward summation

$$\lim_{N \to \infty} \left( H(N) - \ln N \right) = \gamma =$$

Euler constant (not e)

Backward is better

(Kahn summation) see Wiki

Example Solve

$$x^2 - 2x + c = 0$$

$$x = 1 - \sqrt{1-c}$$

Assume $|c| << 1$

$\Rightarrow$ loss of digits

If $c = 10^{-9}$

$$1 - c = 1.\underbrace{\overbrace{000000\cdots\cdots}^{16}}_{9 \text{ zeros}}\cdots$$

I loose $16 - 9 = 7$

$$\cancel{1 - \sqrt{1-c}} = \frac{c}{1+\sqrt{1-c}}$$

Good even for small c

If $|c| \ll 1$ $\simeq \frac{c}{2}$

$$1-\sqrt{1-c} \simeq \frac{c}{2} + O(c^2)$$

Taylor series around $c=0$