Representation of Numbers:

Integer Numbers:

N bits: 2^N integers

2^{N-1}

$$I = (-1)^s (\alpha_n 2^n + \alpha_{n-1} 2^{n-1} + \ldots + \alpha_0 2^0) \Leftarrow$$

$s = \begin{cases} 0 \to + \\ 1 \to - \end{cases}$ $n = N-2$ $\alpha_i = \begin{cases} 0 \\ 1 \end{cases}$

single precision: 32 bits

$\Rightarrow 2^{31} \sim 2 \times 10^9$

a) the answer respects the range

b) division is interpreted asthe integer part

Real Numbers

$$r = \underbrace{(-1)^s}_{1\text{bit}} \underbrace{(m_1 2^{-1} + m_2 2^{-2} + \ldots + m_{23} 2^{-23})}_{23 \text{ bits}} 2^{\overset{\Rightarrow}{\alpha - 127}} \leftarrow \text{"bias"}$$

8 bits → $\alpha \in [0, 255]$

$\Downarrow$ $2^{-23} \sim 10^{-7} \nwarrow$

$\Downarrow$ $2^{-127} \to 2^{128}$

real numbers: $10^{-44} \lesssim r \lesssim 10^{+38}$

double precision: 10^{-16}

Machine Precision: (ε)

for 32-bits: $\varepsilon = 10^{-7}$ for 64 bits: $\varepsilon = 10^{-16}$

smallest #:

32 bits $\# > 2^{128}$ OVERFlOW

$\# < 2^{-127}$ UNDERFLOW

Roundoff Error:

$$1 + \underbrace{\varepsilon + \varepsilon + \varepsilon \ldots}_{10^{-7} \text{ times}} = 2 \text{ but } 1$$

$$\propto \sqrt{N}\, \varepsilon_m \qquad \propto N$$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \qquad b^2 >> 4ac$$

Truncation Error:

$$f(x) = \sum_{n=0}^{\infty} a_n x^n \quad \Rightarrow \quad f_N(x) = \sum_{n=0}^{N} a_n x^n$$

derivative

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

as $h \to 0$ $\quad x+h \sim x$ in compute