# Computer Arithmetic & Machine Precision Lecture Notes

BM 531
Numerical Methods and C/C++ Programming

Ahmet Ademoglu, *PhD*
Bogazici University
Institute of Biomedical Engineering

# Float representation : 4 bytes

$x_f = (-1)^s \times mantissa \times 2^{exp-bias}$

$(0.5)_f$ : 0 0111 1111 1000 0000 0000 0000 0000 000

The bias : 0111 1111

$(\text{Maximum Number})_f$ :

0 1111 1111 1111 1111 1111 1111 1111 111 :

$2^{128} = 3.4 \times 10^{38}$

$(\text{Minimum Number})_f$ :

0 0000 0000 1000 0000 0000 0000 0000 000 :

$2^{-128} = 2.9 \times 10^{-39}$

# Machine Precision

$7 + 1.0 \times 10^{-7}$

$(7)_f$ : 0 1000 0010 1110 0000 0000 0000 0000 000

$(10^{-7})_f$ : 0 0110 0000 1101 0110 1011 1111 1001 010

Shift right to align the exponents before adding

$(10^{-7})_f$ :

0 1000 0010 0000 0000 0000 0000 0000 000 (0001101...)

$7 + 1.0 \times 10^{-7} = 7$

If $24^{th}$ bit is 1, round up makes an error of $2^{-23} \approx 10^{-7}$

Float precision is no more reliable after 7 decimal digits

# Relative Error

$x = (-39.9)_{10}$

$: \quad 1 \quad 1000\ 0101 \quad 1001 \quad 1111 \quad 1001 \quad 1001 \quad 1001 \quad 100 \qquad \overline{1100}$

$x_f =$

$: \quad 1 \quad 1000 \quad 0101 \quad 1001 \quad 1111 \quad 1001 \quad 1001 \quad 1001 \quad 101$

rounded-up

$x_f = (-39.90000152587890625)_{10}$

Relative Error$= \epsilon_r = |x_f - x|/|x|$

Maximum Relative Error occurs at $x = 1$ with $\epsilon_r^{max} = 2^{-23}$

# Double precision representation : 8 bytes

52 bits : mantissa
11 bits : exponent
Double Precision Round up error : $2^{-52} \approx 10^{-16}$
Magnitude Range of Double Precision :
$2.225074 \times 10^{-308} : 1.799693 \times 10^{308}$

# Numerical Evaluation

$\log 2 \approx ?$  
$\log 3 \approx ?$  
$e = \lim\limits_{n \to \infty} \left(1 + \frac{1}{n}\right)^n \approx ?$  
$\ln 2 \approx ?$  
Series Expansion  
$(a+b)^n = a^n + na^{n-1}/1! + n(n-1)a^{n-1}b^2/2! + \ldots$  
$(1+x)^n = 1 + nx/1! + n(n-1)x^2/2! + \ldots$  
Taylor Expansion Theorem  
$f(x - x_0) = f(x_0) + f'(x_0)(x - x_0)/1! + f''(x_0)(x - x_0)^2/2! + \ldots$

# Numerical Derivative

The 1st derivative

$f'(x) =_{\lim \Delta x \to 0} \frac{f(x+\Delta x)-f(x)}{\Delta x}$

$(sin(x))' = ?$

Chain Rule

$d(f(g(x)))/dx = d(f(x))/dx \ \ d(g(x))/dx$

$(sin^2(x))' = ?$

The second Derivative

$f''(x) =_{\lim \Delta x \to 0} \frac{f'(x+\Delta x)-f'(x)}{\Delta x}$

$(x^2)'' = ?$

Differential equation

$f''(x) + \alpha f'(x) = x$