Reprezentacja liczb

Sunday, 28 January 2024 13:29

Floating point

$$\pm 2^{w} \cdot (1+M)$$

w - wykładnik

M - mantysa (0,1)

Wykładnik

Zapis biased na n bitach: $w + 2^{n-1}$

dla 8 bitów (float):

$w \in [-126, 127]$ $w + 127$ w pamięci

Specjalne wartości

0 - wykładnik i mantysa to same 0 (± 0)

∞ - wykładnik to same 1, mantysa same 0 ($\pm \infty$)

NaN - wykładnik same 1, niezerowa mantysa

Denormalized numbers

Wykładnik same 0, dowolna mantysa: $0.xx\ldots x \cdot 2^{-126}$

Błędy

bezwzględny $|X - X^*|$ $3 \pm \varepsilon$

względny $\left|\frac{X - X^*}{X}\right|$ $3 \pm 1\%$

Catastrophic cancellation - odejmowanie bliskich liczb

- alternatywny sposób obliczania wyniku
- sprzęganie

$X^2 = 2 \Rightarrow X^* = 1.4$

forward error: $|\sqrt{2} - 1.4|$

backward error: $|1.4^2 - 2|$

Uwarunkowanie problemu

małe zaburzenie wejścia → małe zaburzenie wyniku

źle uwarunkowany problem: pierwiastki wielomianu

Stabilność jest cechą algorytmu