Analiza numeryczna (L)
14.10.2020 r.

Informacje

- repetytorium
- grupy zdalne, wykład hybrydowy – zapisy
- obowiązkowe deklaracje elektroniczne
- Portal „Przedmioty"

Wykład 2. Arytmetyka zmiennopozycyjna i podstawy teorii błędów – część 1.

- Błąd względny i bezwzględny. Liczba cyfr dokładnych

$x := 1.23456789 \qquad \tilde{x} := 1.2345679$

$y := 10^{50} + 1 \qquad \tilde{y} := 10^{50}$

$|x-\tilde{x}| = 10^{-8} \qquad |y-\tilde{y}| = 1$ ← błąd bezwzględny

$\left|\frac{x-\tilde{x}}{x}\right| \approx 0.8\cdot 10^{-8} \qquad \left|\frac{y-\tilde{y}}{y}\right| = 10^{-50}$ ← $\boxed{\text{błąd względny}}$!

↑ to jest lepsze rozwiązanie

liczba cyfr dokładnych (dziesiętnych) → $acc(v, v') := -\log_{10}\left|1 - \frac{v'}{v}\right|$

v – wartość dokładna

v' – wartość przybliżona

$acc(x, \tilde{x}) = 8.091\ldots$

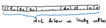
$acc(y, \tilde{y}) = 50$

Maszynowa reprezentacja liczb

Rozpatrujemy model uproszczony $\xrightarrow{\text{rzeczywistość}}$ IEEE 754 ⇓ ćwiczenia

a) liczby całkowite

$l \in \mathbb{Z}$: $l = \pm \sum_{i=0}^{n} e_i 2^i$, $e_i \in \{0,1\}$, $e_n := 1$


d+1 bitów na liczbę całkowitą ze znakiem

$\boxed{n \geq d}$ ⇒ problem! nadmiar

$\boxed{n < d}$ ⇒ wszystko OK

b) liczby rzeczywiste

$x \in \mathbb{R}\setminus\{0\}$

Tw

$x = s\cdot m\cdot 2^c$ ← reprezentacja jest jednoznaczna

$m \in [\frac{1}{2}, 1)$ ← mantysa

$c \in \mathbb{Z}$ ← cecha

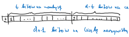
$s \in \{+1, -1\}$ ← znak

reprezentacja mantysy

$$m = \sum_{i=1}^{\infty} e_{-i} 2^{-i}, \quad e_{-i} \in \{0,1\}, \; e_{-1} := 1$$

Problem

Jak reprezentować liczbę rzeczywistą w komputerze biorąc pod uwagę, że prawie zawsze rozwinięcie (do domu: kiedy?) mantysy jest nieskończone



d+1 bitów na liczbę rzeczywistą ze znakiem

Rozważymy dwa sposoby zaokrąglenia mantysy

i) ucięcie (chopping)

$m_t = m_t^c \Rightarrow \underset{\uparrow}{e^*_{-i}} := e_{-i}$ dla $i = 1, 2, \ldots, t$

prawdziwe cyfry mantysy; cyfry reprezentacji mantysy

lepsze rozwiązanie → ii) zaokrąglenie symetryczne (symmetric rounding)

$$m_t = m_t^r := \sum_{i=1}^{t} e_{-i}\cdot 2^{-i} + e_{-(t+1)}\cdot 2^{-t}$$

Zadanie do domu

Jak dokładnie wyglądają cyfry reprezentacji mantysy, tzn. e^*_{-i}, w wypadku zaokrąglenia symetrycznego

Tw

Błąd zaokrąglenia mantysy w wypadku zaokrąglenia symetrycznego nie przekracza $\frac{1}{2}\cdot 2^{-t}$, tzn.

$$|m - m_t^r| \leq \frac{1}{2}\cdot 2^{-t}$$

Dowód → do domu

Reprezentacja maszynowa liczby rzeczywistej $x \in \mathbb{R}\setminus\{0\}$ ($x \approx (s, m, c)$) to

i) ucięcie $\boxed{chop(x) := s\cdot m_t^c\cdot 2^c}$

ii) zaokrąglenie symetryczne $\boxed{rd(x) := s\cdot m_t^r\cdot 2^c}$

Tw

$(x\neq 0)$ $\left|\frac{rd(x)-x}{x}\right| \leq 2^{-t}$, $\left|\frac{chop(x)-x}{x}\right| \leq \boxed{2}\, 2^{-t}$

błąd dwa razy większy

Dowód → ćwiczenia

Przykład

$x := \frac{2}{3} = (0.101010\ldots)_2$

Dla $t = 24$ ⇒ $chop(x) = (0.1010\ldots010)_2$

$rd(x) = (0.1010\ldots011)_2$

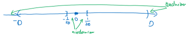
$\left|\frac{x - chop(x)}{x}\right| = 2\cdot 2^{\overset{(-24)?}{-25}}$, $\left|\frac{x - rd(x)}{x}\right| = 2^{\overset{(-24)}{-25}}$?

Jakie liczby, z jakiego zakresu, możemy reprezentować w komputerze?

największa cecha → $c_{max} = -\underset{\uparrow}{c_{min}} = 2^{d-t-1} - 1$

najmniejsza cecha

$D := 2^{c_{max}} \Rightarrow \frac{1}{2D} \leq |x| < D \quad (x \neq 0)$



W rzeczywistości w komputerze możemy reprezentować liczby ze zbioru $X' := (-D, D)$. W pamięci pojawiają się liczby ze zbioru

$X_{fl} := rd(X')$ ← to już jest zbiór dyskretny zależny od wartości t oraz d

zbiór ciągły!

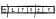
Tragiczne skutki nieprzemyślanego użycia arytmetyki zmiennopozycyjnej.

⇓

na ćwiczeniach

Przykład

Rozważmy arytmetykę dla $d = 5$, $t = 3$:



$? \in \{0, 1\}$

Jak wygląda rozkład tych liczb?

Arytmetyka zmiennopozycyjna

$fl(x\circ y) := rd(x\circ y) = (x\circ y)(1 + \varepsilon_\circ)$ $\qquad x, y \in X_{fl}$, $\circ \in \{+, -, *, /\}$

! ↳ $|\varepsilon_\circ| \leq 2^{-t}$

$$\left|\frac{fl(x\circ y) - (x\circ y)}{(x\circ y)}\right| = |\varepsilon_\circ|$$

Do symulowania zmiennopozycyjnej realizacji algorytmów liczbowych często używa się tzw. twierdzenia o kumulacji błędów.

Tw

Jeśli $|\delta_i| \leq 2^{-t}$ $(i = 1, 2, \ldots, n)$, a

$$\prod_{i=1}^{n}(1+\delta_i) =: 1 + \sigma_n,$$

to

$$\sigma_n = \boxed{\sum_{i=1}^{n}\delta_i} + O(2^{-2t}).$$

← ta część dominuje!

↑ to jest zaniedbywalne

Jeśli $n2^{-t} < 2$, to

$$|\sigma_n| \leq \gamma_n := \frac{n\cdot 2^{-t}}{1 - \frac{1}{2}n2^{-t}} \approx n\cdot 2^{-t}$$

⇓

$|\sigma_n| \lesssim n\cdot 2^{-t}$ (przy realistycznych założeniach jest tak rzeczywiście w praktyce obliczeniowej)

Przykład

Sumowanie liczb x_1, x_2, x_3, x_4, x_5 w naturalnej kolejności w arytmetyce zmiennopozycyjnej.

```
S := x_1
  FOR i := 2 TO 5
    [ S := S + x_i
RETURN (S)
```

$S := \sum_{i=1}^{5} x_i$ ← w „normalnym" świecie

$$fl(S) = ((((x_1 + x_2)(1+\varepsilon_2) + x_3)(1+\varepsilon_3) + x_4)(1+\varepsilon_4) + x_5)(1+\varepsilon_5) =$$

$|\varepsilon_2| \leq 2^{-t}$

$|\varepsilon_3| \leq 2^{-t}$

$|\varepsilon_4| \leq 2^{-t}$

$|\varepsilon_5| \leq 2^{-t}$

$= x_1(1+\varepsilon_2)(1+\varepsilon_3)(1+\varepsilon_4)(1+\varepsilon_5) +$
$x_2(1+\varepsilon_2)(1+\varepsilon_3)(1+\varepsilon_4)(1+\varepsilon_5) +$
$x_3(1+\varepsilon_3)(1+\varepsilon_4)(1+\varepsilon_5) +$
$x_4(1+\varepsilon_4)(1+\varepsilon_5) +$
$x_5(1+\varepsilon_5) =$

$= x_1(1+E_1) + x_2(1+E_2) + x_3(1+E_3) + x_4(1+E_4) + x_5(1+E_5)$

$1 + E_1 := \prod_{i=2}^{5}(1+\varepsilon_i)$

$1 + E_2 := \prod_{i=2}^{5}(1+\varepsilon_i)$

$1 + E_3 := (1+\varepsilon_4)(1+\varepsilon_5)(1+\varepsilon_3)$

$1 + E_4 := (1+\varepsilon_4)(1+\varepsilon_5)$

$1 + E_5 := 1 + \varepsilon_5$

⇓ Tw o kumulacji błędów

$|E_1| \leq \gamma_4 \approx 4\cdot 2^{-t}$

⇓

$|E_1| \lesssim 4\cdot 2^{-t}$

$|E_2| \lesssim 4\cdot 2^{-t}$

$|E_3| \lesssim 3\cdot 2^{-t}$

$|E_2| \lesssim 2\cdot 2^{-t}$

$|E_5| \lesssim 2^{-t}$

Jaki błąd popełniamy:

$$\left|\frac{fl(S) - S}{S}\right| = \left|\frac{\sum_{i=1}^{5} x_i(1+E_i) - \sum_{i=1}^{5} x_i}{\sum_{i=1}^{5} x_i}\right| = \left|\frac{\sum_i x_i\cdot E_i}{\sum_i x_i}\right| \leq$$

$$\leq \frac{\sum_i |x_i||E_i|}{|\sum_i x_i|} \leq \max_i |E_i| \frac{\sum |x_i|}{|\sum x_i|} \lesssim 4\cdot 2^{-t} \underbrace{\boxed{\frac{\sum|x_i|}{|\sum x_i|}}}_{K(x)}\ ?$$

- $x_i > 0$ (lub $x_i < 0$) ⇒ $K(x) = 1$
- Jeśli $\sum|x_i| > 0$, a $|\sum x_i| \approx 0$ ⇒ $K(x)$ może być dowolnie DUŻE!

Pytanie?

W jakiej kolejności „najlepiej" dodawać liczby zmiennopozycyjne?

Odpowiedź

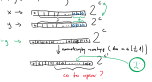
a) Jeśli liczby są dodatnie, to od najmniejszej do największej.

b) i inne strategie ...

Zjawisko utraty cyfr znaczących

Pokażmy na przykładzie w czym jest kolejny problem.

$x > y > 0$, $x \approx y$ ← bliskie sobie liczby



co to wpisać?

To jest tzw. utrata cyfr znaczących

Czy można zaradzić temu problemowi? TAK! ⇒ Przykłady