

Metody numeryczne w fizyce

FZP002934wcl

rok akademicki 2021/22

semestr letni

Wykład 1

Karol Tarnowski

karol.tarnowski@pwr.edu.pl

L-1 p. 220



Plan prezentacji

- Zapis liczb w różnych układach
- Typy danych numerycznych
- Reprezentacja zmiennoprzecinkowa liczb
- Dokładność operacji na liczbach zmiennoprzecinkowych

Na podstawie:

- D. Kincaid, W. Cheney, *Analiza numeryczna*

Arytmetyka komputerowa

Zapis liczb w różnych układach

$$814,72 =$$

$$= 8 \times 10^2 + 1 \times 10^1 + 4 \times 10^0 + 7 \times 10^{-1} + 2 \times 10^{-2}$$

$$(1110,10100)_2 =$$

$$= 1 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 0 \times 2^0 + 1 \times 2^{-1} +$$

$$0 \times 2^{-2} + 1 \times 2^{-3} + 0 \times 2^{-4} + 0 \times 2^{-5} =$$

$$= 8 + 4 + 2 + 0,5 + 0,125 = (14,625)_{10}$$

$$\varphi = 1,618033988 7\dots$$



Arytmetyka komputerowa

Zapis liczb w różnych układach

$$\frac{1}{2} = (0,5)_{10} = (0,1)_2$$

$$\frac{1}{10} = (0,1)_{10} = (0,00011001100\dots)_2$$

$$\frac{1}{3} = (0,33333\dots)_{10} = (0,01010101\dots)_2$$



Typy danych

- całkowite
- zmiennoprzecinkowe



Typy danych

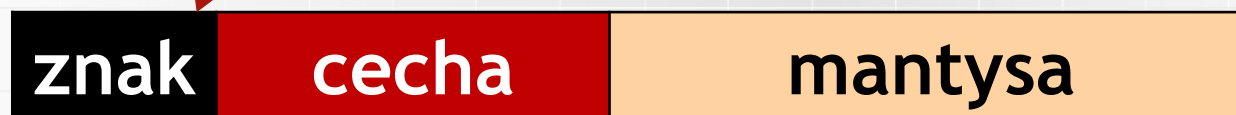
<code>double</code>	podwójna precyzja (64 bity)
<code>single</code>	pojedyncza precyzja (32 bity)
<code>int8</code>	8-bitowa liczba całkowita ze znakiem (signed integer)
<code>int16</code>	16-bitowa liczba całkowita ze znakiem
<code>int32</code>	32-bitowa liczba całkowita ze znakiem
<code>int64</code>	64-bitowa liczba całkowita ze znakiem
<code>uint8</code>	8-bitowa liczba całkowita bez znaku (unsigned integer)
<code>uint16</code>	16-bitowa liczba całkowita bez znaku
<code>uint32</code>	32-bitowa liczba całkowita bez znaku
<code>uint64</code>	64-bitowa liczba całkowita bez znaku

Typy danych

Liczby zmiennoprzecinkowe

$$x = \pm r \times 10^n, \quad r \in [1, 10), \quad n \in \mathbb{C}$$

$$x = \pm q \times 2^n, \quad q \in [1, 2), \quad n \in \mathbb{C}$$



$$x = (-1)^s \cdot (1, q_1 q_2 q_3 \dots) \times 2^{n-b}$$

bias
przesunięcie

Typy danych

Liczby zmiennoprzecinkowe

- rozmiar i zachowanie zależy od implementacji
- standard IEEE 754 określa arytmetykę liczb pojedynczej (32 bity) oraz podwójnej (64 bity) precyzji

Arytmetyka komputerowa

Reprezentacja zmiennoprzecinkowa liczb

Liczby pojedynczej precyzji (przesunięcie 127)

z	c	c	c	c	c	c	c	c	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m			
1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2

Liczby podwójnej precyzji (przesunięcie 1023)

z	c												m																							
1	2-12												13-64																							

8-bitowa liczba zmiennopozycyjna (przesunięcie 3)

z	c	c	c	m	m	m	m
1	2	3	4	5	6	7	8



Arytmetyka komputerowa

Reprezentacja zmiennoprzecinkowa liczb

0 0 1 1 0 0 0 0

$$+1. 0 0 0 0 \times 2^{3-3} = 1,$$

0 0 1 1 1 0 0 0

$$+1. 1 0 0 0 \times 2^{3-3} = 1,5$$

0 0 1 1 0 0 0 1

$$+1. 0 0 0 1 \times 2^{3-3} = 1,0625 = 1+2^{-4}$$

0 0 1 0 0 0 0 0

$$+1. 0 0 0 0 \times 2^{2-3} = 0,5$$

0 0 0 1 0 0 0 0

$$+1. 0 0 0 0 \times 2^{1-3} = 0,25 = 2^{-2}$$

0 1 1 0 0 0 0 0

$$+1. 0 0 0 0 \times 2^{6-3} = 8,0$$

0 1 1 0 1 1 1 1

$$+1. 1 1 1 1 \times 2^{6-3} = 15,5 = 2^3(2-2^{-4})$$

1 1 0 1 0 0 1 0

$$-1. 0 0 1 0 \times 2^{5-3} = -4,5$$

Arytmetyka komputerowa

Reprezentacja zmiennoprzecinkowa liczb

0 0 0 0 0 0 0 0 = +0,

1 0 0 0 0 0 0 0 = -0,

0 1 1 1 0 0 0 0 = +Inf (infinity)

1 1 1 1 0 0 0 0 = -Inf (infinity)

0 0 0 0 1 0 0 0 +0. 1 0 0 0 $\times 2^{1-3}$ = 0,125

0 0 0 0 0 0 0 1 +0. 0 0 0 1 $\times 2^{1-3}$ = 0,015625 = $2^{-2} \times 2^{-4}$

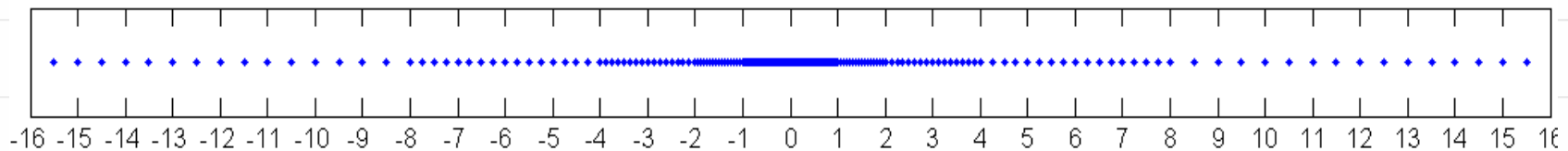
0 1 1 1 1 0 0 0 = NaN (not a number)



Arytmetyka komputerowa

Reprezentacja zmiennoprzecinkowa liczb

Rozmieszczenie liczb zmiennopozycyjnych



- precyzja arytmetyki (epsilon maszynowy ϵ) $0,0625 = 2^{-4}$
- największa liczba zmiennopozycyjna $15,5 = (2-2^{-4})2^3$
- najmniejsza liczba zmiennopozycyjna
 - znormalizowana $0,25 = 2^{-2}$
 - zdenormalizowana $0,015625 = 2^{-6}$

Arytmetyka komputerowa

Reprezentacja zmiennoprzecinkowa liczb

Jak wygląda reprezentacja liczby $4/9$ w rozważanej arytmetyce?

$$4/9 = (0,0111000(111000)\dots)_2$$

po normalizacji

$$4/9 = (1,\underline{11000}111000\dots)_2 \times 2^{-2} = (1,\underline{11000}111000\dots)_2 \times 2^{1-3}$$



$$fl(4/9) = 1,1100 \times 2^{-2} = (1+0,5+0,25) \times 0,25 = 0,4375$$

$$\text{błąd względny} \quad |\delta| \leq \frac{\varepsilon}{2}$$



Arytmetyka komputerowa

Działania na liczbach zmiennoprzecinkowych

Wynikiem operacji matematycznych na liczbach maszynowych zwykle nie jest liczba maszynowa. Przyjmujemy, że po wykonaniu działania mantysa jest normalizowana, a cecha odpowiednio korygowana.

W celu ilustracji rozpatrzmy arytmetykę liczb dziesiętnych z mantysą pięciocyfrową.

Niech $x = 9,7541 \times 10^2$, $y = 2,7849 \times 10^4$, wtedy

$$x + y = 2,882441000 \times 10^4, \text{ fl}(x + y) = 2,8824 \times 10^4, \delta = 1,43 \times 10^{-5}$$

$$x - y = -2,687359000 \times 10^4, \text{ fl}(x - y) = -2,6874 \times 10^4, \delta = 1,53 \times 10^{-5}$$

$$x \times y = 2,716419309 \times 10^7, \text{ fl}(x \times y) = 2,7164 \times 10^7, \delta = 7,1 \times 10^{-6}$$

$$x / y = 3,502495601 \times 10^{-2}, \text{ fl}(x / y) = 3,5025 \times 10^{-2}, \delta = 1,3 \times 10^{-6}$$



Arytmetyka komputerowa

Działania na liczbach zmiennoprzecinkowych

Przykładem sytuacji, w której mogą pojawić się duże błędy względne jest odejmowanie bliskich sobie liczb

$$x = 8,147869223178015,$$

$$\text{fl}(x) = 8,14787,$$

$$y = 8,147235863931790,$$

$$\text{fl}(y) = 8,14724,$$

$$x - y = 0,000633359246225,$$

$$\text{fl}(x) - \text{fl}(y) = 0,00063,$$

$$\text{fl}(\text{fl}(x) - \text{fl}(y)) = 6,3000 \times 10^{-4}$$

$$\left| \frac{(x - y) - \text{fl}[\text{fl}(x) - \text{fl}(y)]}{x - y} \right| = \left| \frac{0,000633359246225 - 0,00063}{0,000633359246225} \right| \approx 0,0053$$



Arytmetyka komputerowa

Działania na liczbach zmiennoprzecinkowych

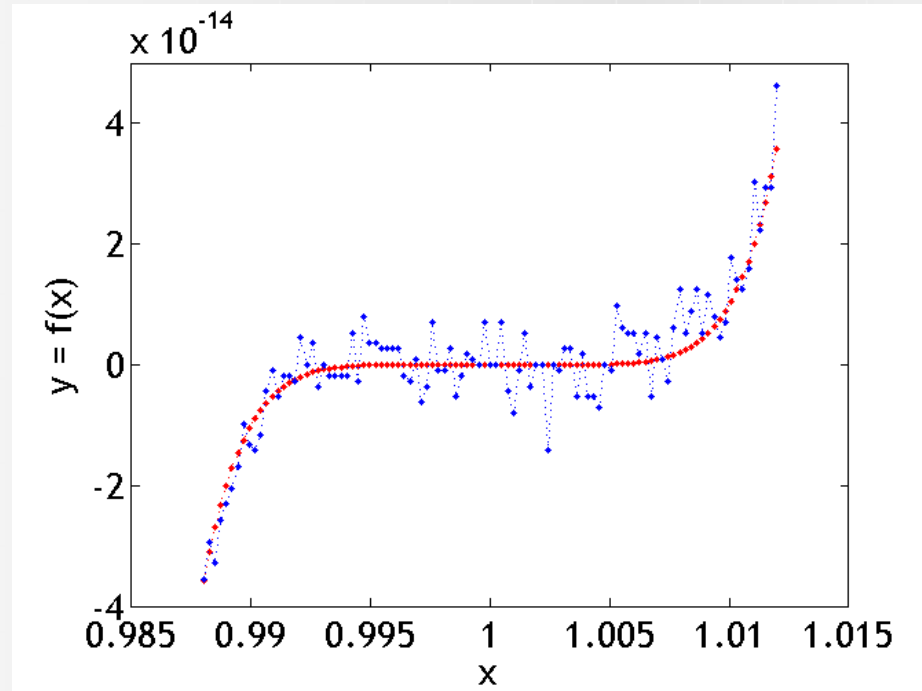
$$y \leftarrow \sqrt{x^2 + 1} - 1$$

$$\sqrt{x^2 + 1} - 1 = \left(\sqrt{x^2 + 1} - 1 \right) \frac{\sqrt{x^2 + 1} + 1}{\sqrt{x^2 + 1} + 1} = \frac{x^2 + 1 - 1}{\sqrt{x^2 + 1} + 1} = \frac{x^2}{\sqrt{x^2 + 1} + 1}$$

$$y \leftarrow \frac{x^2}{\sqrt{x^2 + 1} + 1}$$

Arytmetyka komputerowa

Działania na liczbach zmiennoprzecinkowych



$$f(x) = (x-1)^7 = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1$$



Podsumowanie

- Zapis liczb w różnych układach
- Typy danych numerycznych
- Reprezentacja zmiennoprzecinkowa liczb
- Dokładność operacji na liczbach zmiennoprzecinkowych