

Obliczenia naukowe

Wykład nr 1

Paweł Zieliński

Katedra Podstaw Informatyki,
Wydział Informatyki i Telekomunikacji,
Politechnika Wrocławska



Literatura

Literatura podstawowa

- [1] D. Kincaid, W. Cheney, *Analiza numeryczna*, WNT, 2005.
- [2] M. T. Heath, *Scientific Computing, An Introductory Survey*, The McGraw-Hill Companies, 1997.
- [3] A. Bjorck, G. Dahlquist, *Metody numeryczne*, PWN, 1987.
- [4] M. Dryja, J. i M. Jankowscy, *Przegląd metod i algorytmów numerycznych*, cz. 2, WNT, 1988.
- [5] Z. Fortuna, B. Macukow, J. Wąsowski, *Metody numeryczne*, WNT 1998.
- [6] J. i M. Jankowscy, *Przegląd metod i algorytmów numerycznych*, cz. 1, WNT, 1988.
- [7] J. Stoer, R. Bulirsch, *Wstęp do analizy numerycznej*, PWN, 1987.

Literatura uzupełniająca

- [8] Kielbasiński, H. Schwetlick, *Numeryczna algebra liniowa*, WNT, 1993.

Znaczna część wykładu została przygotowana na podstawie książki [1].



Numeryczna reprezentacja liczb

Każdą liczbę rzeczywistą $x \neq 0$ można zapisać w postaci *znormalizowanej*:

$$x = \pm m \cdot \beta^c = \pm(0.a_1, a_2, \dots)_\beta \cdot \beta^c,$$

gdzie β jest bazą rozwinięcia, np. $\beta = 2, 8, 10, 16$, liczba $m \in [1/\beta, 1)$ nazywana jest *mantysą*, c jest liczbą całkowitą zwaną *cechą*, $a_i \in \{0, 1, \dots, \beta - 1\}$, $a_1 \neq 0$.

Przykład

$$\beta = 10$$

$$732.5051 = 0.7325051 \cdot 10^3$$

$$-0.005612 = -0.5612 \cdot 10^{-2}$$

$$\beta = 2$$

$$(1001.11101)_2 = (0.100111101)_2 \cdot 2^4$$

$$(0.0010111)_2 = (0.10111)_2 \cdot 2^{-2}$$



Numeryczna reprezentacja liczb...

Punktem wyjścia dla reprezentacji liczb w arytmetyce numerycznej jest zapis

$$x = \pm m \cdot \beta^c = \pm (0.a_1, a_2, \dots)_\beta \cdot \beta^c.$$

Zakłada się: stałą liczbę cyfr mantysy t

$$rd(x) = \pm m_t \cdot \beta^c = \pm (0.a_1, a_2, \dots, a_t)_\beta \cdot \beta^c,$$

ustalony zakres cechy c , $c \in [c_{\min}, c_{\max}]$.

Tę arytmetykę nazywamy *arytmetyką zmiennopozycyjną* lub w skrócie *fl*.
Reprezentację x w arytmetyce *fl* oznaczamy przez $rd(x)$ albo $\tilde{x}$.



Numeryczna reprezentacja liczb...

Wykonanie jakichkolwiek obliczeń w arytmetyce *fl* wymaga zastąpienia liczb rzeczywistych x , pojawiających się jako dane początkowe oraz jako wyniki pośrednie, możliwie bliskimi im liczbami $rd(x)$ w arytmetyce *fl* - liczbami maszynowymi.

Wyróżniamy dwa sposoby zastąpienia: **zaokrąglenie** i **obcięcie**
Liczbę $rd(x)$ dla x wyznaczamy następująco:

krok 1: przedstaw x w postaci $x = \pm m\beta^c$, $m \in [1/\beta, 1)$,

krok 2: zaokrąglij mantysę m do najbliższej liczby t -cyfrowej (w przypadku obcięcia rozwinięcie mantysy obcinamy po t cyfrach), otrzymując $rd(m)$,

krok 3: podstaw $rd(x) := \pm rd(m)\beta^c = \pm m_t\beta^c$.



Numeryczna reprezentacja liczb...

Prawidłowe **zaokrąglenie** m , odpowiednio, dla baz $\beta = 10$ i $\beta = 2$

$$rd(m) := \begin{cases} 0.a_1, \dots, a_t & \text{jeśli } a_{t+1} \leq 4, \\ 0.a_1, \dots, a_t + 10^{-t} & \text{jeśli } 5 \leq a_{t+1} \leq 9, \end{cases}$$

$$rd(m) := \begin{cases} 0.a_1, \dots, a_t & \text{jeśli } a_{t+1} = 0, \\ (0.a_1, \dots, a_t)_2 + 2^{-t} & \text{jeśli } a_{t+1} = 1. \end{cases}$$



Numeryczna reprezentacja liczb...

Zakłada się, że zaokrąglenie jest domyślnym sposobem otrzymania przybliżenia $rd(x)$ liczby x .

Weźmy dwie sąsiadujące niezerowe liczby (liczby maszynowe)

$$x^- = m_t^- \beta^c \text{ i } x^+ = m_t^+ \beta^c$$

$$x^+ = (m_t^- + \beta^{-t}) \beta^c, \quad x^- = m_t^- \beta^c,$$

wówczas

$$|x^+ - x^-| = \beta^{-t} \beta^c = \beta^{c-t}.$$

Oszacujmy więc błąd względny dla zaokrąglenia.

$$|\delta| = \frac{|rd(x) - x|}{|x|} \leq \frac{1}{2} \frac{|x^+ - x^-|}{|x|} = \frac{1}{2} \frac{\beta^{c-t}}{|m_x \beta^c|} \leq \frac{1}{2} \frac{\beta^{-t}}{\beta^{-1}} = \frac{1}{2} \beta^{1-t},$$

gdzie x^- , x i x^+ są sąsiadującymi liczbami. Dla obciążenia $|\delta| \leq \beta^{1-t}$.



Numeryczna reprezentacja liczb...

$$|\delta| = \frac{|rd(x) - x|}{|x|}$$

lub inaczej

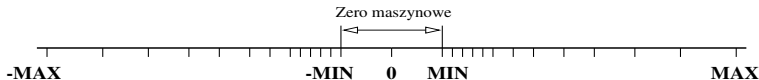
$$rd(x) = x(1 + \delta), \quad |\delta| \leq \epsilon, \quad \epsilon = 0.5\beta^{1-t}.$$

Liczbę ϵ nazywamy *precyzją arytmetyki*.



Numeryczna reprezentacja liczb...

Liczby w arytmetyce *fl* (liczby maszynowe) tworzą siatkę symetryczną niejednorodną (zob. poniżej).



Założenia, że $m_t \in [1/\beta, 1)$ wynika, że arytmetyce *fl* możemy reprezentować liczby x spełniające zależność:

$$\text{MIN} = \frac{1}{\beta} \beta^{c_{\min}} \leq |x| \leq (1 - \beta^{-t}) \beta^{c_{\max}} = \text{MAX}.$$

Jeśli $|x| > \text{MAX}$, to mówimy o *nadmiarze* (mogą być przerywane obliczenia).

Jeśli $|x| < \text{MIN}$, to $rd(x) = 0$ i mówimy o *niedomiarze*. Błąd tej reprezentacji jest równy 100%.

Numeryczna reprezentacja liczb...

Przykład

Rozważmy arytmetykę dwójkową ($\beta = 2$), w której 1 bit przeznaczono na zapis znaku s liczby x , 6 bitów przeznaczono na zapis cechy c (wraz z bitem znaku) i 16 bitów przeznaczono na zapis mantysy ($t = 16$), wówczas $c \in [-31, 32]$, $MIN = \frac{1}{2}2^{-31} \approx 4.66 \cdot 10^{-10}$, $MAX = (1 - 2^{-16})2^{32} \approx 4.29 \cdot 10^9$.



Numeryczna reprezentacja liczb...

Przykład

Jaka jest reprezentacja liczby 9.13 w arytmetyce fl, $\beta = 2$, z mantysą o długości $t = 6$? Dla tej arytmetyki $\epsilon \approx 1.56 \cdot 10^{-2}$

$$x = 9.13 = (1001.0010\dots)_2 \xrightarrow{\text{normalizacja}} (0.10010010\dots)_2 \cdot 2^4 \xrightarrow{\text{zaokrąglenie}}$$

$$rd(x) = (0.100101)_2 \cdot 2^4. |\delta| = |rd(x) - x|/|x| \approx 1.32 \cdot 10^{-2}.$$



Kilka słów o standardzie IEEE 754

Liczbę rzeczywistą $x \neq 0$, zgodnie ze standard IEEE 754, zapisujemy w postaci

$$x = \pm m \cdot 2^c = \pm(1.f)_2 \cdot 2^c = \pm(1.a_1, a_2, \dots)_2 \cdot 2^c,$$

gdzie $m \in [1, 2)$, f jest częścią ułamkową mantysy, $(f)_2 < 1$, cecha jest liczbą całkowitą, $a_i \in \{0, 1\}$.

Niech $t - 1$ będzie liczbą bitów przeznaczoną na zapisanie f , a d będzie liczbą bitów przeznaczoną na zapisanie cechy c .

Zakres cechy jest następujący

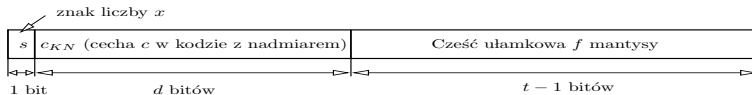
$$-2^{d-1} + 1 \leq c \leq 2^{d-1}.$$



Kilka słów o standardzie IEEE 754

Stosowany jest kod z nadmiarem w zapisie c

$$c_{KN} = c + \text{bias} = c + 2^{d-1} - 1 \text{ Stąd } 0 \leq c_{KN} \leq 2^d - 1,$$
$$c = c_{KN} - \text{bias}.$$



Wartości brzegowe c_{KN} tj. 0 i $2^d - 1$ są zarezerwowane dla wyjątkowych przypadków ± 0 i $\pm \infty$. Więc

$$1 \leq c_{KN} \leq 2^d - 2 \text{ ponieważ } c = c_{KN} - \text{bias} = c_{KN} - (2^{d-1} - 1).$$

Ostatecznie rzeczywisty zakres cechy jest następujący

$$c_{\min} = -2^{d-1} + 2 \leq c \leq 2^{d-1} - 1 = c_{\max}.$$



Zakres, zero maszynowe i precyzja IEEE 754

$x = \pm m \cdot 2^c$, gdzie mantysa $m \in [1, 2)$.

$$\text{MIN}_{\text{sub}} = 2^{-(t-1)} 2^{c_{\min}} < \text{MIN}_{\text{nor}} = 2^{c_{\min}} \leq |rd(x)| \leq (2 - 2^{-(t-1)}) 2^{c_{\max}} = \text{MAX}$$

Niech $x^- = m_{t-1}^- 2^c$ i $x^+ = m_{t-1}^+ 2^c$ będą sąsiadującymi liczbami maszynowymi

$$x^+ = (m_{t-1}^- + 2^{-(t-1)}) 2^c, \quad x^- = m_{t-1}^- 2^c,$$

$$|x^+ - x^-| = 2^{-(t-1)} 2^c = 2^{c-(t-1)}.$$

$$|\delta| = \frac{|rd(x) - x|}{|x|} \leq \frac{1}{2} \frac{|x^+ - x^-|}{|x|} = \frac{1}{2} \frac{2^{c-(t-1)}}{|m_x 2^c|} \leq \frac{1}{2} \frac{2^{-(t-1)}}{1} = 2^{-t} = \epsilon.$$

format	d	t	MIN_{sub}	MIN_{nor}	MAX	$\epsilon = 2^{-t}$
single	8	24	$1.4 \cdot 10^{-45}$	$1.2 \cdot 10^{-38}$	$3.4 \cdot 10^{38}$	$5.96 \cdot 10^{-8}$
double	11	53	$4.9 \cdot 10^{-324}$	$2.2 \cdot 10^{-308}$	$1.8 \cdot 10^{308}$	$1.11 \cdot 10^{-16}$



Kilka słów o standardzie IEEE 754

Przykład

Wyznacz reprezentację $x = 2/3$ w formacie *single* ($t - 1 = 23$, $d = 8$).

$$\frac{2}{3} = (0.10101010\dots)_2 \approx (1.0101010\dots 11)_2 \cdot 2^{-1}.$$

$$c = -1, c_{KN} = c + \text{bias} = c + 127. c_{KN} = 126 = (01111110)_2.$$

$$\frac{2}{3} \approx [0 \mid 01111110 \mid 0101010101010101010101011].$$

W języku Julia: `bitstring(Float32(2/3))`



Kilka słów o standardzie IEEE 754 - zero, nieskończoność

W nawiasach podane są wielkości w formacie **single**.

- **Zera** ($c_{KN} = 0, f = 0$):

+0 ([0 | 00000000 | 000000000000000000000000]),

-0 ([1 | 00000000 | 000000000000000000000000]).

-0 lub +0 jest w szczególności skutkiem niedomiaru
($|x| < 2^{-23} \cdot 2^{-126} = \text{MIN}_{sub}$)

- **Nieskończoności** ($c_{KN} = 255, f = 0$):

$+\infty$ ([0 | 11111111 | 000000000000000000000000]),

$-\infty$ ([1 | 11111111 | 000000000000000000000000]).

$+\infty$ jeśli wystąpił nadmiar, $x + \infty = +\infty$, $x \cdot \infty = +\infty$, $\infty/x = +\infty$
dla $x > 0$.



Kilka słów o standardzie IEEE 754 - not a number, liczby denormalizowane

W nawiasach podane są wielkości w formacie **single**.

- **NaN** “not a number” ($c_{KN} = 255, f \neq 0$):
($[0 \mid 11111111 \mid 1000000000000000000000000000]$)
NaN jeśli $0/0, \infty - \infty, NaN + x$.
- **Liczby denormalizowane** $x \in (-MIN_{nor}, -MIN_{sub}] \cup [MIN_{sub}, MIN_{nor})$
($c_{KN} = 0, f \neq 0$):
 $x = 2^{-23} \cdot 2^{-126} ([0 \mid 00000000 \mid 0000000000000000000000000001])$,
 $x = -2^{-23} \cdot 2^{-126} ([1 \mid 00000000 \mid 0000000000000000000000000001])$.



Kilka słów o standardzie IEEE 754 - zaokrąglenie

Rozważmy arytmetykę **single**.

$$x = (1 + 2^{-24}) \cdot 2^0$$

Sąsiednie liczby maszynowe $x^- = 1$ i $x^+ = (1 + 2^{-23}) \cdot 2^0$, $x \in (x^-, x^+)$.
 $|x^+ - x^-| = 2^{-23}$, $|x^+ - x| = 2^{-24}$, $|x - x^-| = 2^{-24}$.

$$x^- = 1 \cdot 2^0 : ([0 \mid 01111111 \mid 000000000000000000000000 \mid])$$

$$x = (1 + 2^{-24}) \cdot 2^0 : ([0 \mid 01111111 \mid 000000000000000000000000 \mid \mathbf{1}00\dots])$$

$$x^+ = (1 + 2^{-23}) \cdot 2^0 : ([0 \mid 01111111 \mid 000000000000000000000000 \mathbf{1} \mid])$$

$$\mathbf{rd}(x) = x^- = 1$$



Działania arytmetyczne w arytmetyce fl

Zakłada się, że argumenty działań arytmetycznych są równe swoim reprezentacjom w arytmetyce fl oraz, że przy obliczaniu wyniku może wystąpić niedomiar lub nadmiar.

Niech $x = rd(x)$ i $y = rd(y)$ będą liczbami w arytmetyce fl , wówczas przez

$$fl(x * y) \text{ (lub przez } x \circledast y \text{)}$$

oznaczamy wynik działania $x * y$, gdzie $*$ $\in \{+, -, /, \cdot\}$, w arytmetyce fl .



Działania arytmetyczne w arytmetyce fl...

Przykład

Niech dana będzie arytmetyka $\beta = 10$, $t = 5$, $x = rd(x) = 0.31426 \cdot 10^3$, $y = rd(y) = 0.92577 \cdot 10^5$. Używamy akumulatora o podwójnej długości do pośrednich obliczeń

$$x + y \stackrel{\text{po wyrównaniu cech}}{=} 0.9289126000 \cdot 10^5 \xrightarrow{\text{zaokrągl.}}$$

$$fl(x + y) = 0.92891 \cdot 10^5,$$

$$x - y \stackrel{\text{po wyrównaniu cech}}{=} -0.9226274000 \cdot 10^5 \xrightarrow{\text{zaokrągl.}}$$

$$fl(x - y) = -0.92263 \cdot 10^5,$$

$$x \cdot y = 0.2909334802 \cdot 10^8 \xrightarrow{\text{zaokrągl.}} fl(x \cdot y) = 0.29093 \cdot 10^8,$$

$$x/y \approx 0.3394579647 \cdot 10^{-2} \xrightarrow{\text{zaokrągl.}} fl(x/y) = 0.33946 \cdot 10^{-2},$$

Błędy względne działań wynoszą: $|\delta_+| \approx 2.8 \cdot 10^{-6}$, $|\delta_-| \approx 2.8 \cdot 10^{-6}$, $|\delta_\cdot| \approx 8.5 \cdot 10^{-6}$,
 $|\delta_/| \approx 6 \cdot 10^{-6}$.

Precyzja arytmetyki $\epsilon = 5 \cdot 10^{-5}$.

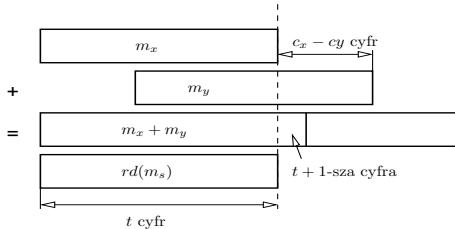


Działania arytmetyczne w arytmetyce fl...

Przykład

Niech $x = rd(x) = m_x \beta^{c_x}$, $y = rd(y) = m_y \beta^{c_y}$, założmy dla uproszczenia $x \geq y > 0$. Ich dokładna suma jest równa

$$x + y = m_x \beta^{c_x} + m_y \beta^{c_y} = (m_x + m_y \beta^{-(c_x - c_y)}) \beta^{c_x} = m_s \beta^{c_x}.$$



Z powyższej rysunku wynika, że dla $x \gg y$, $fl(x + y) = x$.

Działania arytmetyczne w arytmetyce *fl*...

Żądanie

Arytmetyka *fl* powinna zapewnić równość

$$fl(x * y) = rd(x * y), \text{ dla } * \in \{+, -, /, \cdot\}, \text{ i } x * y \in [-MAX, MAX].$$

Jeśli spełnione jest żądanie, to zachodzi równość

$$fl(x * y) = (x * y)(1 + \delta), \text{ gdzie } |\delta| \leq \epsilon, \quad \epsilon = 0.5\beta^{1-t}.$$

Przykład

Obliczmy $x(y + z)$ w arytmetyce *fl*, dla $x = rd(x)$, $y = rd(y)$, $z = rd(z)$.

$$\begin{aligned} fl(x(y + z)) &= fl(xfl(y + z)) = [xfl(y + z)](1 + \delta_1) \\ &= [x(y + z)(1 + \delta_2)](1 + \delta_1) \quad |\delta_i| \leq \epsilon, \quad i = 1, 2 \\ &= x(y + z)(1 + \delta_2 + \delta_1 + \delta_2\delta_1) \approx x(y + z)(1 + \delta_2 + \delta_1) \\ &= x(y + z)(1 + \delta_3), \quad |\delta_3| \leq 2\epsilon. \end{aligned}$$



Działania arytmetyczne w arytmetyce fl...

Przykład

*Dla danych $x_i > 0$, $i = 0, \dots, n$, obliczyć $S = \sum_{i=0}^n x_i$
w arytmetyce fl za pomocą algorytmu*

```
 $S_0 := x_0;$   
for  $i := 1$  to  $n$  do  
     $S_i := S_{i-1} + x_i$   
end for
```

Twierdzenie 1

Jeśli $|\delta_i| \leq \epsilon$, $i = 1, \dots, n$, $\prod_{i=1}^n (1 + \delta_i) = 1 + E_n$ oraz $n\epsilon < 2$, to

$$|E_n| < \frac{n\epsilon}{1 - 0.5n\epsilon} \approx n\epsilon.$$



Działania arytmetyczne w arytmetyce *fl*...

$$\begin{aligned} fl(S) &= fl(\dots fl(fl(fl(x_0 + x_1) + x_2) + x_3) + \dots + x_n) \\ &= (\dots (((x_0 + x_1)(1 + \delta_1) + x_2)(1 + \delta_2) + x_3)(1 + \delta_3) + \\ &\quad + \dots + x_n)(1 + \delta_n) \quad |\delta_i| \leq \epsilon, \quad i = 1, \dots, n \\ &= x_0 \prod_{i=1}^n (1 + \delta_i) + x_1 \prod_{i=1}^n (1 + \delta_i) + x_2 \prod_{i=2}^n (1 + \delta_i) + x_3 \prod_{i=3}^n (1 + \delta_i) + \dots + x_n (1 + \delta_n) \end{aligned}$$

postawiając $\prod_{i=k}^n (1 + \delta_i) = 1 + E_k$ otrzymujemy

$$\begin{aligned} fl(S) &= x_0 + x_0 E_1 + x_1 + x_1 E_1 + x_2 + x_2 E_2 + x_3 + x_3 E_3 + \dots + x_n + x_n E_n \\ &= S + x_0 E_1 + x_1 E_1 + x_2 E_2 + x_3 E_3 + \dots + x_n E_n \end{aligned}$$

$$\begin{aligned} |fl(S) - S| &= |x_0 E_1 + x_1 E_1 + x_2 E_2 + x_3 E_3 + \dots + x_n E_n| \quad (\text{zauważ, że } x_i > 0) \\ &\leq (x_0 + x_1 + x_2 + x_3 + \dots + x_n) \max_{1 \leq k \leq n} |E_k| = S |E_1|, \end{aligned}$$

gdzie $|E_1| < n\epsilon$ (twierdzenie 1).



Działania arytmetyczne w arytmetyce fl...

Redukcja cyfr znaczących

Przykład

Niech $x = 0.3721478693$ i $y = 0.3720230572$ (zauważ, że $x \approx y$).

Różnica dokładna $x - y = 0.0001248121$.

Policzmy $x - y$ w arytmetyce fl z $t = 5$.

$$rd(x) = 0.37215, \quad rd(y) = 0.37202, \quad fl(rd(x) - rd(y)) = 0.00013.$$

Zauważ, że w tym przypadku $fl(rd(x) - rd(y)) = rd(x) - rd(y)$.

$$|\delta| = \frac{|x - y - [rd(x) - rd(y)]|}{|x - y|} = \frac{|0.0001248121 - 0.00013|}{|0.0001248121|} \approx 4.16 \cdot 10^{-2}.$$

Precyzja arytmetyki $\epsilon = 5 \cdot 10^{-5}$.



Działania arytmetyczne w arytmetyce fl...

Przeanalizujemy zjawisko redukcji cyfr znaczących. $rd(x) = x(1 + \delta_x)$ i $rd(y) = y(1 + \delta_y)$.

$$\frac{|x - y - [rd(x) - rd(y)]|}{|x - y|} = \frac{|\delta_x x - \delta_y y|}{|x - y|} \leq \epsilon \frac{|x| + |y|}{|x - y|}$$

Z powyższego oszacowania wynika, że niedokładność danych może przenieść się na wynik z wielkim mnożnikiem $(|x| + |y|)/|x - y|$ nawet gdy samo działanie nie wprowadza błędu.



Działania arytmetyczne w arytmetyce fl...

Przykład

Gdy liczymy wartość wyrażenia $\sqrt{x^2 + 1} - 1$ kiedy $|x|$ **jest mała**. W tym przypadku następuje redukcja cyfr znaczących. Przekształcamy $\sqrt{x^2 + 1} - 1$ następująco

$$y = (\sqrt{x^2 + 1} - 1) \frac{\sqrt{x^2 + 1} + 1}{\sqrt{x^2 + 1} + 1} = \frac{x^2}{\sqrt{x^2 + 1} + 1}.$$

Pytanie. Ile cyfr znaczących tracimy przy odejmowaniu $x - y$, kiedy $x \approx y$?

Twierdzenie 2

Jeśli x i y są dodatnimi liczbami w dwójkowej arytmetyce fl takimi, że $x > y$ i

$$2^{-q} \leq 1 - \frac{y}{x} \leq 2^{-p},$$

wtedy tracimy co najmniej p i co najwyżej q bitów przy odejmowaniu



Dowód Twierdzenia 1

$$\begin{aligned}|E_n| &= |\prod_{i=1}^n (1 + \delta_i) - 1| \leq (1 + \epsilon)^n - 1 \\&= 1 + \binom{n}{1}\epsilon + \binom{n}{2}\epsilon^2 + \binom{n}{3}\epsilon^3 + \dots + \binom{n}{n}\epsilon^n - 1 \\&= n\epsilon + \frac{n\epsilon(n-1)\epsilon}{2!} + \frac{n\epsilon(n-1)\epsilon(n-2)\epsilon}{3!} + \dots + \frac{n\epsilon(n-1)\epsilon \dots 1\epsilon}{n!} \\&= n\epsilon \left(1 + \frac{(n-1)\epsilon}{2!} + \frac{(n-1)\epsilon(n-2)\epsilon}{3!} + \dots + \frac{(n-1)\epsilon \dots 1\epsilon}{n!} \right) \\q &= \frac{n\epsilon}{2} < 1 \\|E_n| &< n\epsilon(1 + q + q^2 + \dots + q^{n-1}) = n\epsilon \frac{1 - q^n}{1 - q} < \frac{n\epsilon}{1 - 0.5n\epsilon} \approx n\epsilon. \quad \square\end{aligned}$$



Dowód Twierdzenia 2

Niech x i y będą dwiema liczbami w dwójkowej arytmetyce *fl*, tj. $x = m_x 2^{c_x}$, $y = m_y 2^{c_y}$, $m_x, m_y \in [1/2, 1)$.

Ponieważ $x > y$, to aby obliczyć różnicę $r = x - y$ musimy wyrównać cechy

$$r = m_x 2^{c_x} - m_y 2^{c_y} = (m_x - m_y 2^{c_y - c_x}) 2^{c_x} = m_r 2^{c_x}$$

Mantysa m_r spełnia

$$m_r = (m_x - m_y 2^{c_y - c_x}) = m_x \left(1 - \frac{m_y 2^{c_y}}{m_x 2^{c_x}} \right) = m_x \left(1 - \frac{y}{x} \right) < 2^{-p}.$$

Aby znormalizować reprezentację $x - y$ arytmetyce *fl*, musimy przesunąć mantysę m_r o co najmniej p bitów w lewo. Wtedy wprowadzane są zera (na prawo), które nie niosą żadnej informacji. Co oznacza redukcję p znaczących bitów.

Rozumowanie dla drugiej części jest podobne.

