Lista 2

Zadanie 1

Niech $B \in \{2, 3, 4...\}$. Pokażmy, że każda niezerowa liczba rzeczywista x ma jednoznaczne przedstawienie w postaci $x = smB^c$, gdzie $s = sgnx,\ c \in \mathbb{Z},\ m \in [\frac{1}{B}, 1)$.

Załóżmy nie wprost, że istnieją dwie różne reprezentacje takiej liczby x i oznaczmy je jako:

$$x = s_1 m_1 B^{c_1} = s_2 m_2 B^{c_2},$$

gdzie $m_1, m_2 \in [\frac{1}{B}, 1)$ oraz $c_1, c_2 \in \mathbb{Z}$.

Przy czym $s = s_1 = s_2$. Mamy zatem:

$$sm_1 B^{c_1} = sm_2 B^{c_2}$$

$$m_1 B^{c_1} = m_2 B^{c_2}$$

$$\frac{m_1}{m_2} = \frac{B^{c_2}}{B^{c_1}}$$

$$\frac{m_1}{m_2} = B^{c_2 - c_1}$$

Rozpatrzmy przypadki:

$1^\circ\ c_1 = c_2$
Wtedy:

$$\frac{m_1}{m_2} = 1 \implies m_1 = m_2$$

Sprzeczność z założeniem o dwóch różnych reprezentacjach.

$2^\circ\ c_1 > c_2 \implies c_2 - c_1 \leq -1$

Czyli $B^{c_2 - c_1} \leq \frac{1}{B}$.

$$\frac{m_1}{m_2} = B^{c_2 - c_1} \leq \frac{1}{B}$$

$$m_1 \leq m_2 \frac{1}{B}$$

Z założenia $m_2 < 1$, więc otrzymujemy:

$$m_1 < \frac{1}{B}$$

A ponieważ $m_1 \in [\frac{1}{B}, 1)$ to otrzymujemy sprzeczność.

3° $c_1 < c_2 \implies c_2 - c_1 \geq 1$

Czyli $B^{c_2 - c_1} \geq B$.

$$\frac{m_1}{m_2} = B^{c_2 - c_1} \geq B$$

$$m_1 \geq Bm_2$$

Ponieważ $m_2 \geq \frac{1}{B}$ otrzymujemy:

$$m_1 \geq 1$$

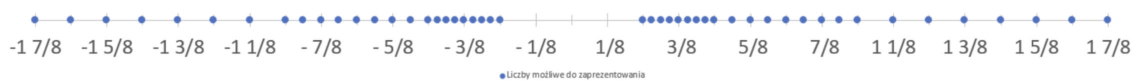
Sprzeczność, ponieważ $m_1 \in [\frac{1}{B}, 1)$.

Zadanie 2

$m_{(2)}$	$m_{(10)}2^0$	$m_{(10)}2^{-1}$	$m_{(10)}2^1$
± 0.1111	$\pm\frac{15}{16}$	$\pm\frac{15}{32}$	$\pm\frac{15}{8}$
± 0.1110	$\pm\frac{14}{16}$	$\pm\frac{14}{32}$	$\pm\frac{14}{8}$
± 0.1101	$\pm\frac{13}{16}$	$\pm\frac{13}{32}$	$\pm\frac{13}{8}$
± 0.1100	$\pm\frac{12}{16}$	$\pm\frac{12}{32}$	$\pm\frac{12}{8}$
± 0.1011	$\pm\frac{11}{16}$	$\pm\frac{11}{32}$	$\pm\frac{11}{8}$
± 0.1010	$\pm\frac{10}{16}$	$\pm\frac{10}{32}$	$\pm\frac{10}{8}$
± 0.1001	$\pm\frac{9}{16}$	$\pm\frac{9}{32}$	$\pm\frac{9}{8}$
± 0.1000	$\pm\frac{8}{16}$	$\pm\frac{8}{32}$	$\pm\frac{8}{8}$

Najmniejszy przedział zawierający wszystkie te liczby to $[-\frac{15}{8}, \frac{15}{8}]$

Rozkład liczb na osi w tym przedziale:



Zauważmy, że im większe (co do modułu) liczby, tym bardziej ‘rozrzucone’ są na osi OX, czyli tracimy możliwość precyzyjnego zaprezentowania liczby. A dla liczb co modułu mniejszych od $\frac{1}{4}$ tracimy możliwość reprezentacji (w przyjętym modelu).

Zadanie 3

Niech
$x = sm2^c$, gdzie $s = sgnx,\ c \in \mathbb{Z},\ m \in [\frac{1}{2}, 1)$,
$rd(x) = sm_t^r 2^c$, gdzie $m_t^r \in [\frac{1}{2}, 1)$ oraz $|m - m_t^r| \leq \frac{1}{2}2^{-t}$

Pokazać, że

$$\frac{|rd(x) - x|}{|x|} \leq 2^{-t}$$

Dowód:

$$\frac{|rd(x) - x|}{|x|} = \left|\frac{sm_t^r 2^c - sm2^c}{sm2^c}\right| = \left|\frac{m_t^r - m}{m}\right| =$$

$$\frac{|m - m_t^r|}{m} \leq \frac{1}{2}2^{-t}\frac{1}{m} \overset{(1)}{\leq} 2^{-t}$$

(1) $m \in [\frac{1}{2}, 1) \implies \frac{1}{m} \in (1, 2]$

Zadanie 4

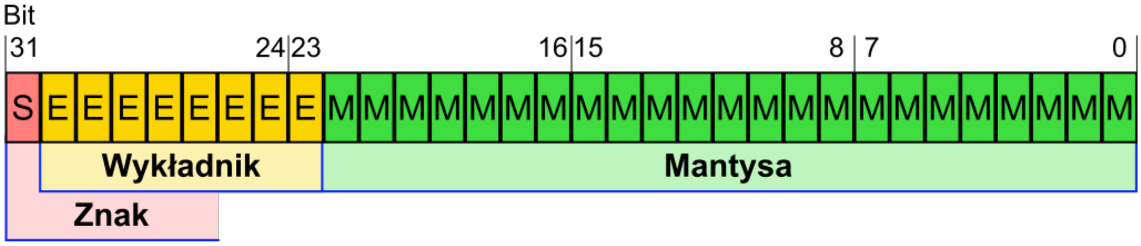
25 lutego 1991r. bateria American Patriot Missile nie zdołała wyśledzić oraz przechwycić nadlatującego irackiego pocisku. Jak się okazuje, śmierć 28 żołnieży i zranienie kolejnych 100 osób spowodował błąd w obliczeniach.

Czas w zegarze komputera liczony był w dziesiątkach sekund (od rozpoczęcia pracy) i był reprezentowany liczbami całkowitymi. Do uzyskania czasu w sekundach ta liczba mnożona była przez $\frac{1}{10}$ (operacja na liczbach zmiennoprzecinkowych). Ponieważ $\frac{1}{10}$ ma nieskończone rozwinięcie w systemie binarnym, a rejestry w tym komputerze były 24-bitowe to każda taka operacja obarczona była małym błędem.

Do odpowiedniego przewidzenia, gdzie rakieta powinna się znajdować użyto funkcji korzystającej z jej prędkości oraz czasu, w którym ostatni raz została wykryta przez radar. Tamtego dnia system pracował około 100 godzin, co przekłada się na 0.34s błędu. Biorąc pod uwagę, że owy typ rakiety leci z prędkością 1676 $\frac{m}{s}$ to wykroczył poza monitorowany obszar.

Zadanie 5

Kodowanie liczby zmiennoprzecinkowej pojedynczej precyzji:



Mamy format numeryczny: $(-1)^s 2^E M$, gdzie $M \in [1, 2)$, $s \in \{0, 1\}$.

Bit znaku kodowany jest jako 0 (dodatnia) lub 1 (ujemna).

$E = wykładnik - BIAS$
$BIAS = 2^{k-1} - 1$, gdzie k to ilość bitów przeznaczonych na wykładnik
Mantysa w słowie maszynowym to M zapisane bez wiodącej jedynki, czyli Mantysa $\in [0, 1)$

W przypadku $wykładnik = 1 - BIAS$ mamy liczby zdenormalizowane (-0, $+0$ oraz liczby bardzo blisko zera).

Dla $wykładnik = BIAS$ możemy otrzymać $-\infty, +\infty$ (dla Mantysy równej 0) oraz NaN (gdy Matysa nie jest równa 0) - powstający np. z pierwiastkowania liczby ujemnej.

Teoretyczny model reprezentacji liczb maszynowych z wykładu:

$\pm$	1	m_2	m_3	...	m_b	$\pm$	c_1	c_2	...	c_{d-b-1}

$x = sm2^c$, gdzie $s \in \{-1, 1\}, m \in [\frac{1}{2}, 1), c \in \mathbb{Z}$

Mamy $d + 1$ bitów na liczbę rzeczywistą ze znakiem, z czego kolejno:

- 1 bit na znak s liczby,
- b bitów na mantysę m,
- $d - b$ bitów na cechę c (z czego jeden na jej znak).

Główne różnice między reprezentacjami:

- w uproszczonej reprezentacji nie można zapisać 0, nieskończoności, NaN, ani liczb bardzo blisko 0,
- korzystając ze standardu IEEE 754 możemy zapisać więcej liczb mając do dyspozycji tą samą ilość bitów,
- przedział możliwych do zaprezentowania liczb jest symetryczny (dla wersji z wykładu).

Zadanie 6

Algorytm do obliczania wartości $d := \sqrt{x^2 + y^2}$:

```
u := x*x;
u := u + y*y;
d := sqrt(u)
```

Pokażmy, że przy zaproponowanym algorytmie może wystąpić zjawisko nadmiaru.

Niech $X_{fl} = [-16, 16]$ oraz $x = y = 8$, stosując algorytm:

```
u := 8*8       // (1)
u := 64 + 8*8 // (2)
d := sqrt(128)
```

Jak widzimy, w miejscach (1) oraz (2) algorytmu wystąpiło zjawisko nadmiaru, mimo że $d = \sqrt{128} = 8\sqrt{2} \approx 11,3 \in X_{fl}$.

Niech $M = max(|x|, |y|) \neq 0$
Wtedy:

$$\sqrt{x^2 + y^2} = M\sqrt{\frac{x^2}{M^2} + \frac{y^2}{M^2}} = M\sqrt{(\frac{x}{M})^2 + (\frac{y}{M})^2} \leq \sqrt{2}max(|x|, |y|) \in X_{fl}$$

Algorytm unikający zjawiska nadmiaru:

```
u := min(x,y)/max(x,y);       \\ u <= 1
u := 1 + u*u;
u := abs(max(x,y)) + sqrt(u);
```

Długość euklidesowa wektora (w $\mathbb{R}^n$):
$||x|| = \sqrt{x \cdot x} = (\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}}$

Niech $M = max(|x_1|, |x_2|, ...|x_n|)$, oraz wiedząc, że $max(|x_1|, |x_2|, ..., |x_n|)\sqrt{n} \in X_{fl}$

$$\sqrt{x_1^2 + x_2^2 + ... + x_n^2} = M\sqrt{\frac{x_1^2}{M^2} + \frac{x_2^2}{M^2} + ... + \frac{x_n^2}{M^2}} =$$

$$M\sqrt{(\frac{x_1}{M})^2 + (\frac{x_2}{M})^2 + ... + (\frac{x_n}{M})^2} \leq M\sqrt{n} = max(|x_1|, |x_2|, ..., |x_n|)\sqrt{n}$$

Algorytm:

```
m := max(x[1], x[2], ..., x[n]);
s := 0;
For i = 1, 2, ... n
    t := x[i]/m;
    s := s + t*t;
s := sqrt(s);
d := abs(m) * s;
```

Zadanie 8

$$f(x) = 4040\frac{\sqrt{x^{11}+1}-1}{x^{11}}$$

Z poprzedniej listy wiemy, iż już $f(0.001)$ daje niewiarygodny wynik. Dzieje się tak ze względu na zjawisko utraty cyfr znaczących.

Dla małych x, $\sqrt{x^{11}+1}-1$ interpretowany jest jako 0, dlatego też dla całego wyrażenia otrzymujemy 0.

Wzór możemy przekształcić:

$$f(x) = 4040\frac{\sqrt{x^{11}+1}-1}{x^{11}} = 4040\frac{(\sqrt{x^{11}+1}-1)(\sqrt{x^{11}+1}+1)}{x^{11}(\sqrt{x^{11}+1}+1)} =$$

$$\frac{4040}{\sqrt{x^{11}+1}+1}$$

Tym samym pozbywając się zjawiska utraty cyfr znaczących dla małych x.

tags: `anl`