Zadanie 1. Jednoznaczne przedstawienie każdej niezerowej liczby rzeczywistej x postaci $x = s \cdot m \cdot B^c$, gdzie $s = \operatorname{sgn} x$, $m \in [\frac{1}{B}, 1)$, $c \in \mathbb{Z}$, $B \in \{2, 3, 4, \ldots\}$. w postaci znormalizowanej.

Każdą liczbę można przedstawić jako smB^c:

(1) Skoro s to znak, to możemy go pominąć, wtedy $x = mB^c$ (z dokładnością do $|x|$).

(2) Dzięki temu mamy $x = mB^c \Rightarrow m = \frac{x}{B^c}$, więc dla liczb w systemie B-liczbowym (dwójkowy, trójkowy, czwórkowy etc.) możemy przesunąć je o c miejsc w lewo, a wartość c dobierzemy tak, aby $m \in [\frac{1}{B}, 1)$ - w tym celu przecinek przesuwamy tak, aby przed nim było 0, a za nim liczba w systemie liczbowym dwójkowym przy B.

Stąd każda liczba rzeczywista x ma przedstawienie w tym systemie.

Jednoznaczność reprezentacji x:

(1) Załóżmy, że istnieją dwie różne reprezentacje $sm B^c$ jednej liczby. Znak jest stały, więc go pomijamy:

$$x = sm_1B^{c_1} = sm_2B^{c_2} \Rightarrow m_1B^{c_1} = m_2B^{c_2} \Rightarrow \frac{m_1}{m_2} = B^{c_2 - c_1}$$

(2) Rozpatrzmy przypadki (dopuszczamy do sprzeczności):

$c_1 = c_2 \rightarrow$ wtedy $B^0 = 1 = \frac{m_1}{m_2} \Rightarrow m_1 = m_2$

$c_1 > c_2 \rightarrow$ wtedy $B^{c_2 - c_1} \leq B^{-1} \Rightarrow \frac{m_1}{m_2} \leq \frac{1}{B} \Rightarrow m_1 \leq \frac{m_2}{B}$, jednak $m_1, m_2 \in [\frac{1}{B}, 1)$, więc gdyby m_1 było w dobrym zakresie, to musiałoby zachodzić $m_2 \geq 1$.

$c_1 < c_2 \rightarrow$ wtedy $B^{c_2 - c_1} \leq B^1 \Rightarrow \frac{m_1}{m_2} \geq B \Rightarrow m_1 \geq m_2 B$, $m_1, m_2 \in [\frac{1}{B}, 1)$, więc gdyby m_1 było w dobrym przedziale, to m_2 musiałoby być $m_2 \leq \frac{1}{B}$.

■

Zadanie 2. Wszystkie liczby zmiennopozycyjne postaci:

$$x = \pm (0.1\, e_{-2} e_{-3} e_{-4})_2 \cdot 2^{\pm c}; \quad e_i, c \in \{0,1\}$$

2^{-1}	2^0	2^1
0.01000	0.1000	1.0000
0.01001	0.1001	1.0010
0.01010	0.1010	1.0100
0.01011	0.1011	1.0110
0.01100	0.1100	1.1000
0.01101	0.1101	1.1010
0.01110	0.1110	1.1100
0.01111	0.1111	1.1110

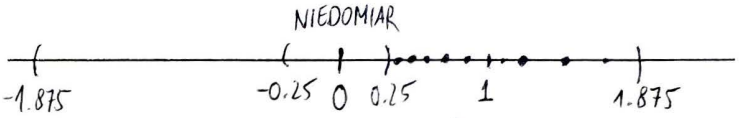
24 liczby dodatnie i tyle samo liczb ujemnych, więc łącznie mamy 48 takich liczb.

Najmniejsza / największa wartość co do modułu:

$\min(|x|) = 0.01000_2 = 0.25$

$\max(|x|) = 1.1110_2 = 1.875$

Stąd mamy, że $x \in [-1.875, 1.875]$, a niedomiar dla $x \in [-0.25, 0.25]$.



wiele liczb niedaleko $\min(|x|)$, im dalej $\max(|x|)$, tym mniej liczb

Zadanie 3. Wykazać, że $\dfrac{|rd(x) - x|}{|x|} \leq 2^{-t}$ dla $x = sm2^c$.

Informacje: $s = \operatorname{sgn} x$, $c \in \mathbb{Z}$, $m \in \left[\frac{1}{2}, 1\right)$, $rd(x) = sm_t^r 2^c$, $m_t^r \in \left[\frac{1}{2}, 1\right)$,

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

Dowód:

$$\frac{|rd(x) - x|}{|x|} = \frac{|sm_t^r 2^c - sm2^c|}{|sm2^c|} = \frac{|(s2^c)(m_t^r - m)|}{|s2^c \cdot m|} = \frac{|m_t^r - m|}{|m|} \leq$$

$$\leq \frac{\frac{1}{2} \cdot 2^{-t}}{m} = \frac{2^{-t}}{2m} \leq 2^{-t} \quad \blacksquare$$

małe sygnalizujemy ułamek, a skoro $m \in \left[\frac{1}{2}, 1\right)$, to $2m \geq 1$, a dla każdego m większego od $\frac{1}{2}$ ułamek maleje

Zadanie 4. IEEE 754 - standard reprezentacji binarnej i operacji na liczbach zmiennoprzecinkowych

32 bity (64)

bit znaku
1 - ujemna
0 - dodatnia

(11) 8 bitów cechy, tzn. wykładnik Z, zakres wykładników $[-127, 128]$, min i max wykładnika ma znaczenie specjalne

(52) 23 bity mantysy z pominięciem wiodącego, niezerowego bitu (czyli dla 1.1100... mantysa będzie miała postać 1100...)

Dokładność ~ 7-8 miejsc dziesiętnych, zakres od około $\pm 1.18 \cdot 10^{-38}$ do około $\pm 3.4 \cdot 10^{38}$.

Szczególne przypadki:

wykładnik jest równy -127 w kodzie z nadmiarem lub zero w kodzie binarnym
- $+0$ - wszystkie bity są zerami
- -0 - bit znaku ustawiony, reszta jest zerami
- liczby małe (nienormalne) - mantysa różna od zera
- $\pm \infty$ - ustawiane wszystkie bity wykładnika (128 w kodzie z nadmiarem lub 255 w kodzie binarnym), mantysa równa zero, może się pojawić jako wynik dzielenia przez 0.
- NaN - Not a Number, wykładnik jak wyżej, może się pojawić jako wynik pierwiastkowania liczby ujemnej

Zadanie 5. x, y - liczby maszynowe, pokazać, że algorytm obliczający $d := \sqrt{x^2+y^2}$ postaci: $u := x^*x$, $u := u + y^*y$, $d := \text{sqrt}(u)$ może powodować zjawisko nadmiaru.

Niech $X_{fl} \in 2^{32}$, weźmy $x = y = 2^{30}$. Wtedy $\sqrt{x^2+y^2} = \sqrt{(2^{30})^2 + (2^{30})^2} = \sqrt{2^{60}+2^{60}} = \sqrt{2^{60}(1+1)} = 2^{30}\sqrt{2} \in X_{fl}$, jednak $2^{60} \notin X_{fl}$. Aby tam zapobiec, przekształcamy wzór (dla $x \geq y$, w razie potrzeby swapujemy):

$$\sqrt{x^2+y^2} = \sqrt{x^2\left(1+\frac{y^2}{x^2}\right)} = |x| \cdot \underbrace{\sqrt{1+\left(\frac{y}{x}\right)^2}}$$

skoro $x \geq y$, to $\sqrt{1+\left(\frac{y}{x}\right)^2} \leq 2$, więc $\sqrt{2} \cdot \max(|x|,|y|) \in X_{fl}$.

Długość euklidesowa: $\|x_n\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$

Więc optymalizujemy w następujący sposób (zakładając $x_i \geq x_{i+1}$):

$$\|x_n\| = \sqrt{x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2} =$$

$$= \sqrt{x_1^2 \left(1 + \frac{x_2^2}{x_1^2} + \frac{x_3^2}{x_1^2} + \ldots + \frac{x_n^2}{x_1^2}\right)}$$

$$= |x_1| \cdot \sqrt{1 + \underset{\wedge\!\!\wedge}{\underset{1}{\frac{x_2^2}{x_1^2}}} + \underset{\wedge\!\!\wedge}{\underset{1}{\frac{x_3^2}{x_1^2}}} + \ldots + \underset{\wedge\!\!\wedge}{\underset{1}{\frac{x_n^2}{x_1^2}}}}$$

$$= |x_1| \cdot \sqrt{n} \equiv \max(x_1, x_2, \ldots, x_n) \cdot \sqrt{n}$$

Zadanie 6. Poprawienie $f(x) = 4038 \cdot \frac{\sqrt{x^{11}+1} - 1}{x^{11}}$, aby mieścić się w przedziale double'a.

$$f(x) = 4038 \frac{\sqrt{x^{11}+1} - 1}{x^{11}} = 4038 \frac{\left(\sqrt{x^{11}+1} - 1\right)\left(\sqrt{x^{11}+1} + 1\right)}{x^{11}\left(\sqrt{x^{11}+1} + 1\right)} =$$

$$= 4038 \cdot \frac{x^{11} + 1 - 1}{x^{11}\left(\sqrt{x^{11}+1} + 1\right)} = 4038 \cdot \frac{1}{\sqrt{x^{11}+1} + 1}$$

■

Zadanie 7. $x_{k+1} = 2^k \sqrt{2\left(1 - \sqrt{1 - \left(\frac{x_k}{2^k}\right)^2}\right)}$ $\quad k = 1, 2, \ldots; \; x_1 = 2$

Ciąg x_k jest zbieżny do π, dla $k = 30$ zaczynamy tracić cyfry znaczące. Przekształcamy wzór, aby tam zapobiec:

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

Wtedy $x_{k+1} = 2^k \sqrt{2A} = \sqrt{2 \cdot \frac{x_k^2}{1 + \sqrt{1 - \left(\frac{x_k}{2^k}\right)^2}}}$

Zadanie 8. Dla jakich x stracimy cyfry znaczące?

(a) $f(x) = x^7 + \sqrt{x^{14} + 2019}$ $\rightarrow$ źle, gdy $|x^7| \approx \sqrt{x^{14} + 2019}$, $x < 0$, $|x| \gg 2019$

$$f(x) = \frac{\left(x^7 + \sqrt{x^{14} + 2019}\right)\left(x^7 - \sqrt{x^{14} + 2019}\right)}{x^7 - \sqrt{x^{14} + 2019}} = \frac{-2019}{x^7 - \sqrt{x^{14} + 2019}}$$

(b) $f(x) = x^{-7}\left(\underbrace{\sin x - x}_{\text{to jest problematyczne, bo dla małych } x \text{ (do ok. } 3^\circ) \; \sin x \approx x} + \frac{x^3}{6} - \frac{x^5}{120}\right)$

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} \cdot x^{2n-1} = x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{7!} + \ldots =$$

$$(?) = -\frac{1}{7!} + \frac{x^2}{9!} - \frac{x^4}{11!} + \ldots$$