L2.1.

Dowód. Najpierw pokażemy, że każdą liczbę x można zapisać w postaci $sm\mathrm{B}^c$. s to oczywiście znak. Teraz, $x = m\mathrm{B}^c$ (z dokładnością do modułu, ale to nie jest istotne). Skoro tak, to $m = \frac{x}{\mathrm{B}^c}$, innymi słowy przesuwamy binarną reprezentację x o c miejsc w lewo. c możemy dobrać tak, żeby nasze m było z przedziału $[\frac{1}{\mathrm{B}}, 1)$. Oczywiście jest to możliwe, trzeba się tylko trochę zastanowić (przesuwamy przecinek tak, by przed przecinkiem było 0, za przecinkiem będzie 1 i dalej coś, czyli liczba z zakresu $[\frac{1}{\mathrm{B}}, 1)$). Wtedy m mamy już obliczone. Czyli każdego x możemy przedstawić w tej postaci.

Teraz jednoznaczność.
Załóżmy nie wprost, że istnieją dwie różne reprezentacje x w postaci $sm\mathrm{B}^c$. Oczywiście, znak jest stały. Mamy więc, że:

$$x = sm_1\mathrm{B}^{c_1} = sm_2\mathrm{B}^{c_2} \Rightarrow m_1\mathrm{B}^{c_1} = m_2\mathrm{B}^{c_2} \Rightarrow \log m_1 + c_1 = \log m_2 + c_2 \Rightarrow \log \frac{m_1}{m_2} = c_2 - c_1$$

- $c_1 = c_2$
  $$\log \frac{m_1}{m_2} = 0 \Rightarrow \frac{m_1}{m_2} = 1 \Rightarrow m_1 = m_2$$
  Sprzeczność.
- $c_1 > c_2$
  $$\log \frac{m_1}{m_2} = c_2 - c_1 \Rightarrow \log \frac{m_1}{m_2} \leqslant -1 \Rightarrow \frac{m_1}{m_2} \leqslant \frac{1}{\mathrm{B}} \Rightarrow m_1 \leqslant \frac{1}{\mathrm{B}} m_2$$
  Teraz musimy skorzystać z tego, że $m_1, m_2 \in [\frac{1}{\mathrm{B}}, 1)$. Widać już, że mamy sprzeczność, bo żeby m_1 było w dobrym przedziale, to m_2 musiałoby być większe(bądź równe) 1, co jest niemożliwe.
- $c_1 < c_2$
  $$\log \frac{m_1}{m_2} = c_2 - c_1 \Rightarrow \log \frac{m_1}{m_2} \geqslant 1 \Rightarrow \frac{m_1}{m_2} \geqslant \mathrm{B} \Rightarrow m_1 \geqslant \mathrm{B} m_2$$
  Ponownie korzystamy z tego, że $m_1, m_2 \in [\frac{1}{\mathrm{B}}, 1)$. Żeby m_1 było w dobrym przedziale m_2 musiałoby być mniejsze od $\frac{1}{\mathrm{B}}$, co jest niemożliwe.

Notka: Każdy logarytm, o którym mowa wyżej, jest o podstawie B □

L2.2.

(2)

$$f = \pm(0.1xyz)_2 \cdot 2^{\pm c} \qquad x, y, z, c \in \{0,1\}$$

$$f = \pm\left(\frac{1}{2} + x \cdot \frac{1}{4} + y \cdot \frac{1}{8} + z \cdot \frac{1}{16}\right) \cdot 2^{\pm c}$$

$$2^{\pm c} \in \left\{\frac{1}{2}, 1, 2\right\}$$

$$f = \pm\left(\frac{16}{32} + x \cdot \frac{8}{32} + y \cdot \frac{4}{32} + z \cdot \frac{2}{32}\right) \cdot a \qquad a \in \left\{\frac{1}{2}, 1, 2\right\}$$

$$x, y, z \in \{0,1\}$$

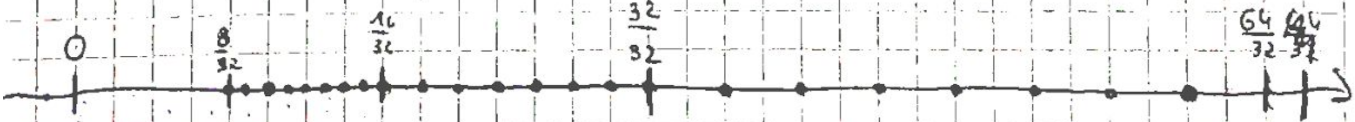
$$\downarrow$$

$$\pm\left\{\frac{16}{32}, \frac{18}{32}, \frac{20}{32}, \frac{22}{32}, \frac{24}{32}, \frac{26}{32}, \frac{28}{32}, \frac{30}{32}\right\}$$

$$\times \left\{\frac{1}{2}, 1, 2\right\}$$

$$\pm\left\{\frac{8}{32}, \frac{9}{32}, \frac{10}{32}, \frac{11}{32}, \frac{12}{32}, \frac{13}{32}, \frac{14}{32}, \frac{15}{32}\right\}$$

$$\pm\left\{\frac{32}{32}, \frac{36}{32}, \frac{40}{32}, \frac{44}{32}, \frac{48}{32}, \frac{52}{32}, \frac{56}{32}, \frac{60}{32}\right\}$$



$$f \in \left\langle -\frac{60}{32}, -\frac{8}{32} \right\rangle \cup \left\langle \frac{8}{32}, \frac{60}{32} \right\rangle$$

$$f \in \left\langle -\frac{60}{32}, \frac{60}{32} \right\rangle$$

Dla ujemnych symetrycznie. Ogólnie jest 48 liczb zapisywalnych w tym systemie.
Wnioski: Dla dodatnich mamy takie 4 przedziały: pierwszy pusty, każdy kolejny 2 razy rzadszy od poprzedniego.

L2.3.

Pokaż, że

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

Dowód.

$$\frac{|rd(x) - x|}{|x|} = \frac{|sm_t 2^c - sm2^c|}{|sm2^c|} = \frac{|(s2^c)(m_t - m)|}{|s2^c m|} = \frac{|m_t - m|}{|m|} \overset{(1)}{\leqslant} \frac{2^{-t}}{2m} \overset{(2)}{\leqslant} 2^{-t}$$

Gdzie:
(1) bo w treści mamy, że $|m_t - m| \leqslant \frac{1}{2}2^{-t}$
(2) bo maksymalizujemy ułamek. Skoro maksymalizujemy ułamek to minimalizujemy licznik. $m \in [\frac{1}{2}, 1)$. Czyli $2m$ może minimalnie wynieść 1. Jeśli ułamek byłby mniejszy, działałoby tym bardziej. □

L2.4.

[Co student PWR o IEEE 754 wiedzieć powinien]
TODO: wypisać tutaj różnice. Na pewno IEEE ma specjalne wartości, jak NaN czy Infinity, ale to nie wszystko...

L2.5.

Załóżmy, że maksymalna wartość $X_{fl} = 2^{64}$ i weźmy $x = y = 2^{60}$.

$$\sqrt{x^2 + y^2} = \sqrt{(2^{60})^2 + (2^{60})^2} = \sqrt{2^{120} + 2^{120}} = \sqrt{2^{120}(1+1)} = 2^{60}\sqrt{2}$$

$2^{60}\sqrt{2} \in X_{fl}$. No ale 2^{120} już się nie mieści. Jak moglibyśmy to naprawić?
Załóżmy, że $x \geqslant y$, jeśli nie to możemy podmienić.

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

Skoro $x \geqslant y$ to pod pierwiastkiem mamy maksymalnie dwójkę, więc nie grozi nam nadmiar (bo $\sqrt{2}max(x, y) \in X_{fl}$). Przerabiamy to na algorytm, co zostawiam jako proste ćwiczenie.

Co do długości euklidesowej, to jeśli wiemy, że zapisuje się ją wzorem

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

To już możemy łatwo znaleźć analogię do tego co robiliśmy przed chwilą. Wyciągamy największego x przed pierwiastek. Jeśli $\sqrt{n} \cdot max(|x_1|, |x_2|, /ldots, |x_n|) \in X_{fl}$

L2.6.

2.1 Utrata cyfr znaczących

Utrata cyfr znaczących występuje, gdy odejmujemy od siebie dwie bliskie sobie liczby zmiennoprzecinkowe. Wtedy na początku otrzymamy dużo zer, a dopiero na dalekim miejscu cyfry znaczące. Jednak ponieważ mantysa jest znormalizowana, to liczba zostanie przesunięta w lewo do pierwszej cyfry znaczącej. Zauważmy jednak, że przesuwając liczbę w lewo w „ogonie" dopiszemy liczby, których nie znamy.

Przykład: Miejsce w pamięci na 7 cyfr, $m_1 = 1,23467890123...$, $m_2 = 1,23456789012...$.

$$m_1 - m_2 = \underbrace{1,234678}_{\substack{\text{tyle mamy}\\ \text{w pamięci}}}901... - \underbrace{1,234567}_{\substack{\text{tyle mamy}\\ \text{w pamięci}}}89012... = 0,00011111...$$

Po przesunięciu otrzymamy: $111\underbrace{????}_{\substack{\text{te liczby}\\ \text{były poza}\\ \text{pamięcią}}}$

Najczęściej w takim wypadku te nieznane wartości wypełniamy zerami. W tym przypadku popełniamy niewielki błąd, bo tylko o 1111, ale co byłoby, gdyby miało tam się znajdować 9999?

L2.7.

TODO: Podobno trzeba udowodnić indukcyjnie, że $x_k = 2^k sin(\frac{\pi}{2^{k-1}})$

L2.8.

a) $x^3 + \sqrt{x^6 + 2017}$,

Utrata cyfr znaczących przy dużych ujemnych x-ach.

a) $\left(x^3 + \sqrt{x^6+2017}\right) \dfrac{x^3 - \sqrt{x^6+2017}}{x^3 - \sqrt{x^6+2017}} =$

$= \dfrac{x^6 - (x^6+2017)}{x^3 - \sqrt{x^6+2017}} = \dfrac{-2017}{x^3 - \sqrt{x^6+2017}}$, więc użyj tego wzoru gdy $x<0$,

a skorzystaj z oryginalnego wpw

b) $1 - x^5 - e^{(-x)^5}$,

Utrata cyfr znaczących, gdy $1 \approx x^5$ lub $1 \approx e^{(-x)^5}$

b) $$1 - x^5 - e^{(-x)^5} \overset{\circledast}{=} -\sum_{n=2}^{\infty} \frac{(-x)^{5n}}{n!}$$

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!}$$

$$e^{(-x)^5} = 1 - x^5 + \sum_{n=2}^{\infty} \frac{(-x)^{5n}}{n!} \quad \circledast$$

c) $\log_3 x - 5$,

Utrata cyfr znaczących, gdy $log_3 x \approx 5$

c) $$\log_3 x - 5 \overset{①}{=} \log_3 x - \log_3 243 \overset{②}{=} \log_3\left(\frac{x}{243}\right)$$

$$5 = \log_3 a \iff 3^5 = a = 243 \quad ①$$

$$\log_a b - \log_a c = \log_a\left(\frac{b}{c}\right) \quad ②$$

d) $\sin(x/3) - x/3 + x^3/162 - x^5/29160$,

Utrata cyfr znaczących w wielu miejscach, np. gdy $sin(x/3) \approx x/3$

d) $$\sin(x/3) - x/3 + x^3/162 - x^5/29160 \overset{\circledast}{=} \sum_{n=3}^{\infty} \frac{(-1)^n}{(2n+1)!} (x/3)^{2n+1}$$

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots$$

$$\sin(x/3) = x/3 - \frac{x^3}{162} + \frac{x^5}{29160} + \sum_{n=3}^{\infty} \frac{(-1)^n}{(2n+1)!} (x/3)^{2n+1} \quad \circledast$$