1. Lebegő pontos számok és tulajdonságaik, Horner algoritmus

A) Ismertesse a lebegőpontos számábrázolás modelljét, és definiálja a gépi számokat. Nevezze meg és számítsa ki a számhalmaz nevezetes mennyiségeit (elemszám, M_∞, ε_0). Szemléltesse a halmaz elemeit számegyenesen. Adjon meg két példát a véges számábrázolásból fakadó furcsaságokra.

Lebegőpontos számábrázolás modellje

Definíció. Legyen $t \in \mathbb{N}$, $k \in \mathbb{Z}$ és

$$m = \sum_{i=1}^{t} m_i \cdot 2^{-i},$$

ahol $m_1 := 1$ és $\forall i = 2, \dots, t : m_i \in \{0, 1\}$. Ekkor az

$$a = \pm m \cdot 2^k$$

számot *normalizált lebegőpontos gépi számnak* nevezzük, ahol t a *mantissza hossza*, k a *karakterisztika*, m pedig a *mantissza*.

Gépi számok definiciója

Definíció. Az alábbi halmazt *gépi számhalmaznak* nevezzük:

$$M = M(t, k^-, k^+),$$

ahol t a mantissza hossza, továbbá $k^- \le k \le k^+$ a karakterisztikák határai. Halmazos jelöléssel:

$$M = \left\{ a \,\middle|\, a = \pm m \cdot 2^k \text{ normalizált lebegőpontos szám és } k^- \le k \le k^+ \right\} \cup \left\{ 0 \right\}.$$

A nullát hozzá kellett vennünk a halmazhoz, hisz mivel $m_1 = 1$ minden gépi számnál, így a nullát nem állítja elő egyik sem.

A legkisebb ábrázolható pozitív szám

$$\varepsilon_0 = +[10\dots0 \,|\, k^-] = \frac{1}{2} \cdot 2^{k^-}$$

A legnagyobb ábrázolható pozitív szám

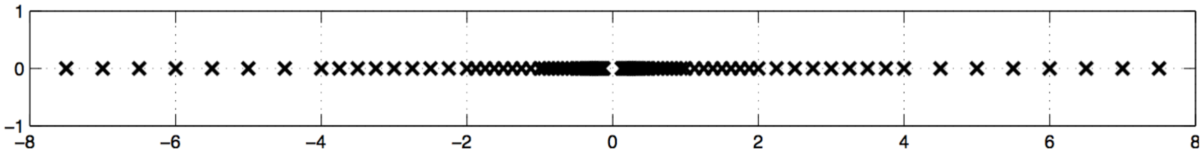
$$M_\infty = +[11\dots1 \,|\, k^+] = \left(\frac{1}{2^1} + \frac{1}{2^2} + \dots + \frac{1}{2^t} \right) \cdot 2^{k^+} = \left(1 - 2^{-t} \right) \cdot 2^{k^+}$$

M elemeinek száma

$$|M| = 2 \cdot 2^{t-1} \cdot (k^+ - k^- + 1) + 1$$

- 1/2 < m < 1
- M szimmetrikus a 0-ra

pl.: M(4, -2, 3)



pl.:
sin(pi) = 0, de gépi számábrázolással 1.22

B) Az input függvény fogalma, tétel az ábrázolt szám hibájáról, ε_1 mennyiség bevezetése és értelmezése.

Input függvény fogalma

Definíció. Az $fl : \mathbb{R}^x \to M$ függvény az *input függvény*, ahol

$$fl(x) = \begin{cases} 0 & \text{ha } |x| < \varepsilon_0 \\ \text{az } x\text{-hez legközelebbi gépi szám} & \text{ha } \varepsilon_0 \leq |x| \leq M_\infty. \end{cases}$$

Ábrázol szám hibájának tétele

Tétel. (Input hiba). Minden $x \in \mathbb{R}^x$ esetén

$$|x - fl(x)| \leq \begin{cases} \varepsilon_0 & \text{ha } |x| < \varepsilon_0 \\ \frac{1}{2}|x|\varepsilon_1 & \text{ha } \varepsilon_0 \leq |x|, \end{cases} \qquad \text{ahol } \varepsilon_1 = 2^{1-t}.$$

ε_1 mennyiség:

Következmény. Ha $\varepsilon_0 \leq |x| \leq M_\infty$, akkor

$$\frac{|x - fl(x)|}{|x|} \leq \frac{1}{2}\varepsilon_1 = 2^{-t},$$

az ábrázolás relatív hibakorlátja. A hiba tehát lényegében ε_1-től, azaz t-től függ.

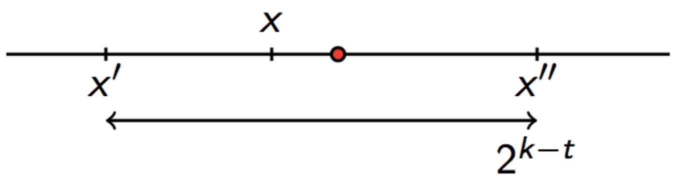
Mennyi a hiba, ha $|x| > M_\infty$?

Bizonyítás:

1. Ha $|x| < \varepsilon_0$, akkor $fl(x) = 0$, így $|x - fl(x)| = |x| < \varepsilon_0$.
2. Ha $|x| \geq \varepsilon_0$ és $x \in M$, akkor $fl(x) = x$, így $|x - fl(x)| = 0$.
3. A meggondolandó eset, amikor $|x| \geq \varepsilon_0$ és $x \notin M$.

Elegendő csak pozitív x-ekkel foglalkoznunk a 0-ra való szimmetria miatt. Keressük meg azt a két szomszédos gépi számot:
$x' < x < x''$ és $x', x'' \in M$, amelyek közrefogják x-et.
Legyen $x' = [1\_\ \ldots\ \_|k]$ alakú. Mennyi x' és x'' távolsága?

Ha x-ben az utolsó helyiértékhez 1-et adunk, akkor x''-t kapjuk.
Tehát $x'' - x' = 2^{-t} \cdot 2^k = 2^{k-t}$.



Ha x az intervallum első felében van, akkor $fl(x) = x'$, ha a második felében, akkor $fl(x) = x''$. Ezért x és $fl(x)$ eltérése legfeljebb az intervallum fele, azaz $\frac{1}{2} \cdot 2^k \cdot 2^{-t}$. Vagyis

$$|x - fl(x)| \leq \frac{1}{2} \cdot 2^k \cdot 2^{-t}.$$

Viszont x abszolút értékére, fenti alakját figyelembe véve
$0.1 \cdot 2^k = \frac{1}{2} \cdot 2^k \leq |x|$ is teljesül, ezért a becslést így folytathatjuk:

$$|x - fl(x)| \leq |x| \cdot 2^{-t} = \frac{1}{2} \cdot |x| \cdot \underbrace{2^{1-t}}_{\varepsilon_1} = \frac{1}{2} \cdot |x| \cdot \varepsilon_1.$$

$\square$