

3 MEGOLDÁSOK

3.1 Lebegőpontos számok

1. • Legnagyobb ábrázolható lebegőpontos szám:

$$\begin{aligned} M_\infty &:= a^{k_+} \sum_{i=1}^t \frac{a-1}{a^i} = a^{k_+} \cdot \left(\frac{a-1}{a} + \frac{a-1}{a^2} + \dots + \frac{a-1}{a^t} \right) = \\ &= a^{k_+} \cdot (1 - a^{-t}). \end{aligned}$$

- Legkisebb ábrázolható pozitív szám:

$$\varepsilon_0 := a^{k_-} \cdot \left(\frac{1}{a} + 0 + \dots + 0 \right) = a^{k_- - 1}.$$

- Az 1 jobboldali szomszédja:

$$[+|1|1, 0, \dots, 0, 1] = a^1 \cdot \left(\frac{1}{a} + \frac{1}{a^t} \right) = 1 + a^{1-t} = 1 + \varepsilon_1.$$

- Az 1 baloldali szomszédja:

$$\begin{aligned} [+|0| \quad a-1, \quad a-1, \quad \dots, \quad a-1] &= a^0 \cdot \left(\frac{a-1}{a} + \dots + \frac{a-1}{a^t} \right) = \\ &= 1 - a^{-t}. \end{aligned}$$

2. A pozitív lebegőpontos számok alakja:

$$[+ \quad |k| \quad m_1, m_2, \dots, m_t].$$

Itt k helyére $(k_+ + |k_-| + 1)$ -féle számot írhatunk, m_1 helyére $(a-1)$ -féle, míg m_2, \dots, m_t helyére a -féle értéket.

Tehát a pozitív lebegőpontos számok száma:

$$(k_+ + |k_-| + 1) \cdot (a-1) \cdot a^{t-1}.$$

3. Lebegőpontos alak:

$$\pm a^k \cdot \left(\frac{m_1}{a} + \frac{m_2}{a^2} + \frac{m_3}{a^3} + \frac{m_4}{a^4} \right).$$

- $\frac{3}{16}$ lebegőpontos alakja:

$$\frac{3}{16} = 2^{-2} \cdot \left(\frac{2}{4} + \frac{1}{4} \right) = 2^{-2} \cdot \left(\frac{1}{2} + \frac{1}{4} \right).$$

$$[+ \quad | - 2| \quad 1 \quad 1 \quad 0 \quad 0].$$

- $\frac{11}{4}$ lebegőpontos alakja:

$$\frac{11}{4} = 2^2 \cdot \frac{11}{16} = 2^2 \cdot \left(\frac{1}{2} + \frac{1}{2^3} + \frac{1}{2^4} \right).$$

$$[- \quad | 2| \quad 1 \quad 0 \quad 1 \quad 1].$$

- 3,25 lebegőpontos alakja:

$$3,25 = 2^2 \cdot \frac{13}{16} = 2^2 \cdot \left(\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^4} \right).$$

$$[+ \quad | 2| \quad 1 \quad 1 \quad 0 \quad 1].$$

- $\frac{5}{8}$ lebegőpontos alakja:

$$\frac{5}{8} = 2^0 \cdot \frac{5}{8} = 2^0 \cdot \left(\frac{1}{2} + \frac{1}{2^3} \right).$$

$$[+ \quad | 0| \quad 1 \quad 0 \quad 1 \quad 0].$$

- $\frac{15}{128}$ lebegőpontos alakja:

$$\frac{15}{128} = 2^{-3} \cdot \frac{15}{16} = 2^{-3} \cdot \left(\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} \right).$$

$$[+ \quad | - 3| \quad 1 \quad 1 \quad 1 \quad 1].$$

4. • $\frac{1}{3}$

Megkeressük az $\frac{1}{3}$ -hoz legközelebbi lebegőpontos számot. Azaz megkeressük a nála kisebb lebegőpontos számok közül a legnagyobbat, és a nála nagyobbak közül a legkisebbet. Az így megtalált két lebegőpontos szám közül a keresett számhoz közelebb eső lesz a szabályos kerekítéssel megadott szám. Levágás esetén pedig az előbb megtalált kettő lebegőpontos szám közül a nullához közelebbi lebegőpontos számot ábrázoljuk.

$$\frac{1}{3} = 2^{-1} \cdot \frac{2}{3}.$$

$$\frac{1}{2} + \frac{1}{8} < \frac{2}{3} < \frac{1}{2} + \frac{1}{8} + \frac{1}{16}.$$

$$\text{azaz: } \frac{30}{48} < \frac{32}{48} < \frac{33}{48}.$$

Lebegőpontos alak szabályos kerekítés esetén: $[+|-1| \quad 1 \quad 0 \quad 1 \quad 1]$.

Lebegőpontos alak levágás esetén: $[+|-1| \quad 1 \quad 0 \quad 1 \quad 0]$.

• $\frac{1}{27}$

Az adott számábrázolási jellemzők mellett a legkisebb pozitív ábrázolható szám: $\varepsilon_0 = \frac{1}{16}$. Mivel $\frac{1}{27} < \frac{1}{16}$, így a gép - függetlenül attól, hogy szabályos kerekítéssel vagy levágással kerekít - az $\frac{1}{27}$ -hez a 0-t rendeli. Alulcsordulás lép fel.

• e

$$e \approx 2,718.$$

2,718-hoz legközelebbi lebegőpontos számok a $2,5 < e < 2,75$.

Szabályos kerekítésnél a 2,75-t kell felírni lebegőpontosan :

$$2,75 = 2 + 0,75 = 2 + \frac{1}{2} + \frac{1}{4} = 2 \cdot \left(1 + \frac{1}{4} + \frac{1}{8}\right) = 2^2 \cdot \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16}\right).$$

Lebegőpontos alak szabályos kerekítés esetén: $[+|2| \quad 1 \quad 0 \quad 1 \quad 1]$.

Levágás esetén a 2,5-t kell felírni lebegőpontosan:

$$2,5 = 2 + 0,5 = 2 + \frac{1}{2} = 2^1 \cdot \left(1 + \frac{1}{4}\right) = 2^2 \cdot \left(\frac{1}{2} + \frac{1}{8}\right).$$

Lebegőpontos alak levágás esetén: $[+|2| \quad 1 \quad 0 \quad 1 \quad 0]$.

5. A 2. feladat alapján összesen 48 db szám lesz: $((2 + (-3) + 1) \cdot 1 \cdot 2^3) = 48$.

$k = 0$ esetén az alábbi számok ábrázolhatóak: $\frac{8}{16}, \quad \frac{9}{16}, \quad \frac{10}{16}, \quad \dots, \quad \frac{15}{16}$.

Ezekből a többi karakterisztikához tartozó számokat úgy kapjuk, hogy a kettő alkalmas hatványával szorozzuk az előző számokat.

6. Két szomszédos, k karakterisztikájú szám távolsága:

$$a^k \cdot \frac{1}{a^t} = a^{k-t}.$$

Ha $k=t$, akkor ez a távolság 1.

$k = t$ karakterisztika mellett a legnagyobb lebegőpontos szám: $a^t \cdot (1 - a^{-t}) = a^t - 1$. Ennek jobboldali szomszédja (a legkisebb $t + 1$ karakterisztikájú szám): $a^{t+1} \cdot \frac{1}{a} = a^t$. $t + 1$ karakterisztika esetén két szomszédos lebegőpontos szám távolsága a , így az $a^t + 1$ nem lebegőpontos.

7. (a) $x \neq y$ és $fl(x - y) = 0$

A legkisebb ábrázolható szám: $\varepsilon_0 := a^{k-1}$.

A fenti adatoknak megfelelően a legkisebb ábrázolható szám: $\varepsilon_0 := 2^{-5} = \frac{1}{32}$.

Olyan lebegőpontos számokat kell keresnünk, melyek távolsága kisebb, mint $\frac{1}{32}$.

Lehetséges értékek: $x = \frac{9}{128}; y = \frac{8}{128}$.

Vagyis: $fl(x - y) = \frac{1}{128}$,

$$\frac{1}{128} < \frac{1}{32}.$$

Mivel a $fl(x - y)$ kisebb, mint ε_0 , ezért $fl(x - y) = 0$.

(b) $fl(x + y) = x$

Lehetséges számok: $x = 2; y = \frac{1}{32}$, ugyanis az adott jellemzők mellett a 2 jobboldali szomszédja $2 + \frac{1}{4}$, így a $2 + \frac{1}{32}$ -et a gép 2-re kerekíti, vagyis: $fl(x + y) = 2$.

(c) $x + y \in [-M_\infty, M_\infty]$, de $x + y$ nem lebegőpontos szám!

Lehetséges számok: $x = 2; y = \frac{1}{16}$, mert ezek összege nem lebegőpontos szám.

8. $\det(A) = 1 - s$.

A legrosszabb esetet kell megkeresni. Ebben a helyzetben ez az, amikor az $1 - s$ a legkisebb, azaz az 1 baloldali szomszédját kell megkeresni.

Az 1 baloldali szomszédja: $1 - a^{-t}$.

Tehát $\frac{1}{\det A}$ akkor a legnagyobb, ha $s = 1 - a^{-t}$, vagyis $1 - s = a^{-t}$.

$$\frac{1}{\det A} = \frac{1}{1 - s} = \frac{1}{a^{-t}} = a^t = a^{t+1} \cdot \frac{1}{a} < a^{k+} \cdot \left(\frac{a-1}{a} + \dots + \frac{a-1}{a^t} \right) = M_\infty$$

tehát nincs túlcsordulás.

P1. A feladat Matlab programkódja:

```
e = 1;
while(1 + e) > 1
e = e/2;
end
2 * e
```

A feladat megoldása: $\varepsilon_1 = 2,220446049250313e - 016$.
 ε_1 értékét a Matlab beépített konstansként is ismeri (eps).

P2. A fenti algoritmus MATLAB-ban megírva:

```
x = 100;
for i = 1 : 60
x = sqrt(x);
end
for i = 1 : 60
x = x^2;
end
```

Azt fogja végül kiírni, hogy $x = 1$. Ennek oka, hogy $a > 1$ esetén $\sqrt[n]{a}$ tart 1-hez, ha $n \rightarrow \infty$. Ha a gyökvonások eredményeit kiíratjuk, akkor látható, hogy $i = 55$ esetén a gyökvonás eredményét a gép pontosan 1-nek látja.
Néhány részeredmény kiírva:

i	x
5	1,154781984689458
10	1,004507364254463
20	1,000004391842173
30	1,0000000004288899
40	1,0000000000004188
50	1,0000000000000004

P3.

$$\left(\frac{1}{10x^2} + 1\right)x^2 - x^2 = 0, 1$$

Az azonosság Matlab-ban megírva a következőképpen néz ki:

```
k = 0;
for x = 1 : 100
if((((1/x^2)/10 + 1) * x^2 - x^2) == 0, 1)
k = k + 1;
end
end
```

A fenti MATLAB programot lefuttatva azt tapasztaljuk, hogy k értéke 0. Azaz a MATLAB szerint az egyenlőség egyetlen x -re sem teljesül. Ennek oka, hogy a 0,1 nem lebegőpontos szám.

P4. A feladat Matlab-ban megírva:

```
x = 1;
for i = 1 : 60
    (exp(x) - 1)/x
    x = x/2;
end
```

A ciklust 53-ig futtatva a hányados értékére 1-et ad eredményül, de 54-ig futtatva a hányados értéke már 0. Ennek oka, hogy $\lim_{x \rightarrow 0} e^x = 1$. Így egy idő múlva a gép e^x értékét 1-nek látja, ekkor a hányados értéke is 0.

P5. A feladat Matlab-ban megírva:

```
x = 1/3;
for i = 1 : 40
    x = 4 * x - 1;
end
```

Ha a ciklust 40-szer futtatjuk le az eredmény: 22369621. Ennek oka, hogy az $\frac{1}{3}$ nem lebegőpontos szám (lásd 4.feladat). Így a kerekítési hibák a ciklus minden lépésében halmozódnak. Néhány részeredmény 15 tizedesjegyre kerekítve:

i	eredmény
2	0,333333333333333
3	0,333333333333332
10	0,333333333313931
20	0,333312988281250
26	0,250000000000000
27	0
28	-1
30	-21
35	-21845
40	-22369621