Numerikus módszerek C

1. előadás: Gépi számábrázolás

Krebsz Anna

ELTE IK

Tartalomjegyzék

Tartalomjegyzék

Mennyi $sin(\pi)$ értéke?

Mennyi $sin(\pi)$ értéke?

`1.224646799147353e-016`

Mennyi $\sum_{k=1}^{+\infty} \frac{1}{k}$ értéke?

Mennyi $\sum_{k=1}^{+\infty} \frac{1}{k}$ értéke?

Mennyi az n-edik részletösszeg, valamely nagy n-re? $\left(\sum_{k=1}^{n} \frac{1}{k}\right)$

Mennyi $\sum_{k=1}^{+\infty} \frac{1}{k}$ értéke?

Mennyi az n-edik részletösszeg, valamely nagy n-re? $\left(\sum_{k=1}^{n} \frac{1}{k}\right)$

Összegezhetünk oda vagy vissza ...

$n = 100000000$-re

`18.997896413852555`

`18.997896413853447`

3. furcsa jelenség

Mennyi $\sqrt{2017} - \sqrt{2016}$ értéke?

Mennyi $\sqrt{2017} - \sqrt{2016}$ értéke?
Más alakban is számolható:

$$\sqrt{2017} - \sqrt{2016} = (\sqrt{2017} - \sqrt{2016}) \cdot \frac{\sqrt{2017} + \sqrt{2016}}{\sqrt{2017} + \sqrt{2016}} =$$
$$= \frac{2017 - 2016}{\sqrt{2017} + \sqrt{2016}} = \frac{1}{\sqrt{2017} + \sqrt{2016}}.$$

Mennyi $\sqrt{2017}-\sqrt{2016}$ értéke?
Más alakban is számolható:

$$\sqrt{2017}-\sqrt{2016}=(\sqrt{2017}-\sqrt{2016})\cdot\frac{\sqrt{2017}+\sqrt{2016}}{\sqrt{2017}+\sqrt{2016}}=$$
$$=\frac{2017-2016}{\sqrt{2017}+\sqrt{2016}}=\frac{1}{\sqrt{2017}+\sqrt{2016}}.$$

Próbáljuk ki mindkét számolási módot!

```
0.011134504483941
0.016926965158418
```

4. furcsa jelenség Matlab-ban

A Matlab-ban

$$a = 1e - 20\,(= 10^{-20}), \quad b = 1.$$

Mennyi lesz $a + b$ értéke?

4. furcsa jelenség Matlab-ban

A Matlab-ban

$$a = 1e - 20\,(= 10^{-20}), \quad b = 1.$$

Mennyi lesz $a + b$ értéke?

$$1$$

4. furcsa jelenség Matlab-ban

A Matlab-ban

$$a = 1e - 20\,(= 10^{-20}), \quad b = 1.$$

Mennyi lesz $a + b$ értéke?

$$1$$

Igaz-e az asszociativatás a Matlab-ban?

$$(a + b) - b, \quad a + (b - b) =?$$

4. furcsa jelenség Matlab-ban

A Matlab-ban

$$a = 1e - 20\,(= 10^{-20}), \quad b = 1.$$

Mennyi lesz $a + b$ értéke?

1

Igaz-e az asszociativatás a Matlab-ban?

$$(a + b) - b, \quad a + (b - b) = ?$$

Próbáljuk ki!

```
1
1.000000000000000e-020
```

5. furcsa jelenség

A Matlab-ban mennyi $\cosh(20) - \sinh(20)$ és $exp(-20)$ értéke?

A Matlab-ban mennyi $\cosh(20) - \sinh(20)$ és $exp(-20)$ értéke?

$$\cosh(20) - \sinh(20) = \frac{\exp(20) + \exp(-20)}{2} - \frac{\exp(20) - \exp(-20)}{2} = \\ = \exp(-20)$$

Próbáljuk ki a kétféle számítási módot!

A Matlab-ban mennyi $\cosh(20) - \sinh(20)$ és $exp(-20)$ értéke?

$$\cosh(20) - \sinh(20) = \frac{\exp(20) + \exp(-20)}{2} - \frac{\exp(20) - \exp(-20)}{2} = \\ = \exp(-20)$$

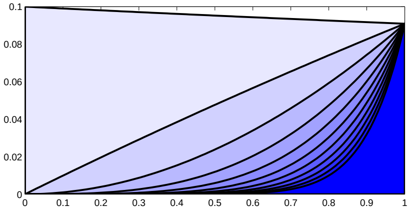
Próbáljuk ki a kétféle számítási módot!

```
0
2.061153622438558e-009
```

Mennyi a

$$T_n := \int_0^1 f_n(x) = \int_0^1 \frac{x^n}{x+10}\,dx$$

határozott integrál értéke? Analitikusan nehéz megadni az értékét. (A geometriai szemléltetésből látszik, hogy mindig pozitív és nullához tart az integrál értéke.)



$$T_n := \int_0^1 \frac{x^n}{x+10}\,dx = \int_0^1 \frac{(x+10-10)x^{n-1}}{x+10}\,dx =$$
$$= \int_0^1 x^{n-1}\,dx - 10\cdot\int_0^1 \frac{x^{n-1}}{x+10}\,dx = \frac{1}{n} - 10\cdot T_{n-1}$$

$$T_n := \int_0^1 \frac{x^n}{x+10}\,dx = \int_0^1 \frac{(x+10-10)x^{n-1}}{x+10}\,dx =$$
$$= \int_0^1 x^{n-1}\,dx - 10\cdot\int_0^1 \frac{x^{n-1}}{x+10}\,dx = \frac{1}{n} - 10\cdot T_{n-1}$$

$$T_0 = \int\limits_0^1 \frac{1}{x+10}\,dx = [\ln(x+10)]_0^1 = \ln(11) - \ln(10) = \ln(1.1)$$

$$T_n := \int_0^1 \frac{x^n}{x+10}\,dx = \int_0^1 \frac{(x+10-10)x^{n-1}}{x+10}\,dx =$$
$$= \int_0^1 x^{n-1}\,dx - 10 \cdot \int_0^1 \frac{x^{n-1}}{x+10}\,dx = \frac{1}{n} - 10 \cdot T_{n-1}$$

$$T_0 = \int\limits_0^1 \frac{1}{x+10}\,dx = [\ln(x+10)]_0^1 = \ln(11) - \ln(10) = \ln(1.1)$$

Tehát a rekuzió:

$$T_0 := \ln(1.1), \quad T_n := \frac{1}{n} - 10 \cdot T_{n-1} \;\; (n = 1, 2 \ldots).$$

Számoljuk a kapott rekurzió alapján a T_{20}. tagot Matlab-bal!

Rendezzük át a rekurziót csökkenően:

$$\begin{aligned} 10\,T_{n-1} &= \frac{1}{n} - T_n \quad \Leftrightarrow \\ T_{n-1} &= \frac{1}{10} \cdot \left(\frac{1}{n} - T_n \right) \end{aligned}$$

Rendezzük át a rekurziót csökkenően:

$$\begin{aligned} 10\,T_{n-1} &= \frac{1}{n} - T_n \quad \Leftrightarrow \\ T_{n-1} &= \frac{1}{10} \cdot \left(\frac{1}{n} - T_n \right) \end{aligned}$$

Indítsuk a rekurziót egy $M >> n$ értékből,

$$T_M := 0, \quad T_{n-1} = \frac{1}{10} \cdot \left(\frac{1}{n} - T_n \right) \; (n = M, \ldots, m+1).$$

Számoljuk a második rekurzió alapján is a T_{20}. tagot! A két algoritmus közül melyik stabil?

Rendezzük át a rekurziót csökkenően:

$$10\,T_{n-1} = \frac{1}{n} - T_n \quad \Leftrightarrow$$
$$T_{n-1} = \frac{1}{10} \cdot \left(\frac{1}{n} - T_n\right)$$

Indítsuk a rekurziót egy $M >> n$ értékből,

$$T_M := 0, \quad T_{n-1} = \frac{1}{10} \cdot \left(\frac{1}{n} - T_n\right) \;\; (n = M, \ldots, m+1).$$

Számoljuk a második rekurzió alapján is a T_{20}. tagot! A két algoritmus közül melyik stabil?

```
7.483468021084803e+003
0.004347035818028
```

Definíció:

A *numerikus algoritmus* aritmetikai és logikai műveletek véges sorozata.

Definíció:

A *numerikus algoritmus* aritmetikai és logikai műveletek véges sorozata.

Definíció:

A numerikus algoritmus *stabil*, ha létezik olyan $C > 0$ konstans, hogy a kétféle B_1, B_2 bemenő adatból kapott K_1, K_2 kimenő adatokra

$$\|K_1 - K_2\| \leq C \cdot \|B_1 - B_2\|.$$

Definíció:

A *numerikus algoritmus* aritmetikai és logikai műveletek véges sorozata.

Definíció:

A numerikus algoritmus *stabil*, ha létezik olyan $C > 0$ konstans, hogy a kétféle B_1, B_2 bemenő adatból kapott K_1, K_2 kimenő adatokra

$$\|K_1 - K_2\| \leq C \cdot \|B_1 - B_2\|.$$

Példa

A Fibonacci sorozat rekurziója instabil. Lásd gyakorlaton.

Tartalomjegyzék

- Gyakorlati és tudományos számításokban sokszor szükségünk van valós számok kezelésére.
- A számítógépeken csak egy véges halmaz elemei közül választhatunk.
- Ráadásul ezek több nagyságrenddel eltérhetnek.

Lebegőpontos számok, normalizált alak: 324 $\rightsquigarrow$ $+0.324 \cdot 10^3$.
Kettes számrendszerben: 101000100 $\rightsquigarrow$ $+0.101000100 \cdot 2^9$.
Általában: $\pm 0.\underbrace{1\,\_\,\_\ldots\_}_{t\ \text{jegy}} \cdot 2^k \qquad (k^- \le k \le k^+)$.

Definíció: Normalizált lebegőpontos szám

Legyen $m = \sum_{i=1}^{t} m_i \cdot 2^{-i}$, ahol $t \in \mathbb{N}$, $m_1 = 1, m_i \in \{0, 1\}$.
Ekkor az $a = \pm m \cdot 2^k$ ($k \in \mathbb{Z}$) alakú számot *normalizált lebegőpontos számnak* nevezzük.
m: a szám *mantisszája*, hossza t
k: a szám *karakterisztikája*, $k^- \le k \le k^+$

Lebegőpontos számok, normalizált alak: 324 $\rightsquigarrow$ $+0.324 \cdot 10^3$.
Kettes számrendszerben: 101000100 $\rightsquigarrow$ $+0.101000100 \cdot 2^9$.
Általában: $\pm 0.\underbrace{1\,\_\,\_\ldots\_}_{t \text{ jegy}} \cdot 2^k \qquad (k^- \leq k \leq k^+)$.

Definíció: Normalizált lebegőpontos szám

Legyen $m = \sum_{i=1}^{t} m_i \cdot 2^{-i}$, ahol $t \in \mathbb{N}$, $m_1 = 1, m_i \in \{0, 1\}$.
Ekkor az $a = \pm m \cdot 2^k$ $(k \in \mathbb{Z})$ alakú számot *normalizált lebegőpontos számnak* nevezzük.
m: a szám *mantisszája*, hossza t
k: a szám *karakterisztikája*, $k^- \leq k \leq k^+$

Lebegőpontos számok, normalizált alak: 324 $\rightsquigarrow$ $+0.324 \cdot 10^3$.
Kettes számrendszerben: 101000100 $\rightsquigarrow$ $+0.101000100 \cdot 2^9$.
Általában: $\pm 0.\underbrace{1\,\_\,\_\ldots\_}_{t \text{ jegy}} \cdot 2^k \qquad (k^- \leq k \leq k^+)$.

Definíció: Normalizált lebegőpontos szám

Legyen $m = \sum_{i=1}^{t} m_i \cdot 2^{-i}$, ahol $t \in \mathbb{N}$, $m_1 = 1, m_i \in \{0, 1\}$.
Ekkor az $a = \pm m \cdot 2^k$ $(k \in \mathbb{Z})$ alakú számot *normalizált lebegőpontos számnak* nevezzük.
m: a szám *mantisszája*, hossza t
k: a szám *karakterisztikája*, $k^- \leq k \leq k^+$

Jelölés: $a = \pm[m_1 \dots m_t | k] = \pm 0.m_1 \dots m_t \cdot 2^k$.

Jelölés: $M = M(t, k^-, k^+)$ a gépi számok halmaza, adott $k^-, k^+ \in \mathbb{Z}$ és $t \in \mathbb{N}$ esetén. (Általában $k^- < 0$ és $k^+ > 0$.)

Definíció: Gépi számok halmaza

$$M(t, k^-, k^+) =$$
$$= \left\{ a = \pm 2^k \cdot \sum_{i=1}^{t} m_i \cdot 2^{-i} \;:\; \begin{array}{l} k^- \le k \le k^+, \\ m_i \in \{0, 1\}, m_1 = 1 \end{array} \right\} \bigcup \{0\}$$

Gyakorlatban még hozzávesszük: $\infty, -\infty$, NaN,...

Jelölés: $a = \pm[m_1 \dots m_t | k] = \pm 0.m_1 \dots m_t \cdot 2^k$.

Jelölés: $M = M(t, k^-, k^+)$ a gépi számok halmaza, adott $k^-, k^+ \in \mathbb{Z}$ és $t \in \mathbb{N}$ esetén. (Általában $k^- < 0$ és $k^+ > 0$.)

Definíció: Gépi számok halmaza

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

$$= \left\{ a = \pm 2^k \cdot \sum_{i=1}^{t} m_i \cdot 2^{-i} \; : \; \begin{array}{l} k^- \le k \le k^+, \\ m_i \in \{0, 1\}, m_1 = 1 \end{array} \right\} \bigcup \{0\}$$

Gyakorlatban még hozzávesszük: $\infty, -\infty$, NaN,...

Jelölés: $a = \pm[m_1 \ldots m_t | k] = \pm 0.m_1 \ldots m_t \cdot 2^k$.

Jelölés: $M = M(t, k^-, k^+)$ a gépi számok halmaza, adott $k^-, k^+ \in \mathbb{Z}$ és $t \in \mathbb{N}$ esetén. (Általában $k^- < 0$ és $k^+ > 0$.)

Definíció: Gépi számok halmaza

$$M(t, k^-, k^+) =$$
$$= \left\{ a = \pm 2^k \cdot \sum_{i=1}^{t} m_i \cdot 2^{-i} \ : \ \begin{array}{l} k^- \le k \le k^+, \\ m_i \in \{0, 1\}, m_1 = 1 \end{array} \right\} \bigcup \{0\}$$

Gyakorlatban még hozzávesszük: ∞, $-\infty$, `NaN`,...

Gépi számok tulajdonságai, nevezetes értékei

1. $\frac{1}{2} \leq m < 1$
2. M szimmetrikus a 0-ra.
3. M legkisebb pozitív eleme:

$$\varepsilon_0 = [100\ldots0|k^-] = \frac{1}{2} \cdot 2^{k^-} = 2^{k^- - 1}$$

4. M-ben az 1 után következő gépi szám és 1 különbsége:

$$\varepsilon_1 = [100\ldots01|1] - [100\ldots00|1] = 2^{-t} \cdot 2^1 = 2^{1-t}$$

5. M legnagyobb eleme:

$$M_\infty = [111\ldots11|k^+] = 1.00\ldots00 \cdot 2^{k^+} - 0.00\ldots01 \cdot 2^{k^+} =$$
$$= (1 - 2^{-t}) \cdot 2^{k^+}$$

6. M elemeinek száma (számossága):

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

Gépi számok tulajdonságai, nevezetes értékei

1. $\frac{1}{2} \leq m < 1$
2. M szimmetrikus a 0-ra.
3. M legkisebb pozitív eleme:

$$\varepsilon_0 = [100\ldots0|k^-] = \frac{1}{2} \cdot 2^{k^-} = 2^{k^- - 1}$$

4. M-ben az 1 után következő gépi szám és 1 különbsége:

$$\varepsilon_1 = [100\ldots01|1] - [100\ldots00|1] = 2^{-t} \cdot 2^1 = 2^{1-t}$$

5. M legnagyobb eleme:

$$M_\infty = [111\ldots11|k^+] = 1.00\ldots00 \cdot 2^{k^+} - 0.00\ldots01 \cdot 2^{k^+} = \\ = (1 - 2^{-t}) \cdot 2^{k^+}$$

6. M elemeinek száma (számossága):

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

Gépi számok tulajdonságai, nevezetes értékei

1. $\frac{1}{2} \leq m < 1$
2. M szimmetrikus a 0-ra.
3. M legkisebb pozitív eleme:

$$\varepsilon_0 = [100\ldots0|k^-] = \frac{1}{2} \cdot 2^{k^-} = 2^{k^- - 1}$$

4. M-ben az 1 után következő gépi szám és 1 különbsége:

$$\varepsilon_1 = [100\ldots01|1] - [100\ldots00|1] = 2^{-t} \cdot 2^1 = 2^{1-t}$$

5. M legnagyobb eleme:

$$M_\infty = [111\ldots11|k^+] = 1.00\ldots00 \cdot 2^{k^+} - 0.00\ldots01 \cdot 2^{k^+} =$$
$$= (1 - 2^{-t}) \cdot 2^{k^+}$$

6. M elemeinek száma (számossága):

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

Gépi számok tulajdonságai, nevezetes értékei

1. $\frac{1}{2} \le m < 1$
2. M szimmetrikus a 0-ra.
3. M legkisebb pozitív eleme:
$$\varepsilon_0 = [100\ldots0|k^-] = \frac{1}{2} \cdot 2^{k^-} = 2^{k^- - 1}$$
4. M-ben az 1 után következő gépi szám és 1 különbsége:
$$\varepsilon_1 = [100\ldots01|1] - [100\ldots00|1] = 2^{-t} \cdot 2^1 = 2^{1-t}$$
5. M legnagyobb eleme:
$$M_\infty = [111\ldots11|k^+] = 1.00\ldots00 \cdot 2^{k^+} - 0.00\ldots01 \cdot 2^{k^+} = \\ = (1 - 2^{-t}) \cdot 2^{k^+}$$
6. M elemeinek száma (számossága):
$$|M| = 2 \cdot 2^{t-1} \cdot (k^+ - k^- + 1) + 1$$

Gépi számok tulajdonságai, nevezetes értékei

1. $\frac{1}{2} \leq m < 1$
2. M szimmetrikus a 0-ra.
3. M legkisebb pozitív eleme:

$$\varepsilon_0 = [100\ldots0|k^-] = \frac{1}{2} \cdot 2^{k^-} = 2^{k^- - 1}$$

4. M-ben az 1 után következő gépi szám és 1 különbsége:

$$\varepsilon_1 = [100\ldots01|1] - [100\ldots00|1] = 2^{-t} \cdot 2^1 = 2^{1-t}$$

5. M legnagyobb eleme:

$$M_\infty = [111\ldots11|k^+] = 1.00\ldots00 \cdot 2^{k^+} - 0.00\ldots01 \cdot 2^{k^+} =$$
$$= (1 - 2^{-t}) \cdot 2^{k^+}$$

6. M elemeinek száma (számossága):

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

Gépi számok tulajdonságai, nevezetes értékei

1. $\frac{1}{2} \leq m < 1$
2. M szimmetrikus a 0-ra.
3. M legkisebb pozitív eleme:

$$\varepsilon_0 = [100\dots0|k^-] = \frac{1}{2} \cdot 2^{k^-} = 2^{k^- -1}$$

4. M-ben az 1 után következő gépi szám és 1 különbsége:

$$\varepsilon_1 = [100\dots01|1] - [100\dots00|1] = 2^{-t} \cdot 2^1 = 2^{1-t}$$

5. M legnagyobb eleme:

$$M_\infty = [111\dots11|k^+] = 1.00\dots00 \cdot 2^{k^+} - 0.00\dots01 \cdot 2^{k^+} = \\ = (1-2^{-t}) \cdot 2^{k^+}$$

6. M elemeinek száma (számossága):

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

Példa

$M(3,-1,2)$ gépi számainak alakja: $\pm 0.1\_\ \_ \cdot 2^k, \quad (-1 \leq k \leq 2)$

Elemei $k = 0$ esetén: $0.100, 0.101, 0.110, 0.111$, azaz $\frac{1}{2}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}$.

Valamint $k = -1$ esetén ezek fele, $k = 1$ esetén ezek kétszerese, $k = 2$ esetén ezek négyszerese. (Továbbá negatív előjellel...)

$\varepsilon_0 = [100|-1] = 0.100 \cdot 2^{-1} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0.25$

$\varepsilon_1 = [101|1] - 1 = 0.101 \cdot 2^1 - 1 = \frac{1}{8} \cdot 2 = \frac{1}{4} = 0.25$

$M_\infty = [111|2] = 0.111 \cdot 2^2 = \frac{7}{8} \cdot 4 = \frac{7}{2} = 3.5$

$|M| = 2 \cdot 2^2 \cdot 4 + 1 = 33$

Példa gépi számhalmazra

Példa

$M(3,-1,2)$ gépi számainak alakja: $\pm 0.1\_\ \_ \cdot 2^k, \quad (-1 \leq k \leq 2)$

Elemei $k = 0$ esetén: $0.100, 0.101, 0.110, 0.111$, azaz $\frac{1}{2}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}$.

Valamint $k = -1$ esetén ezek fele, $k = 1$ esetén ezek kétszerese, $k = 2$ esetén ezek négyszerese. (Továbbá negatív előjellel...)

$\varepsilon_0 = [100|-1] = 0.100 \cdot 2^{-1} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0.25$

$\varepsilon_1 = [101|1] - 1 = 0.101 \cdot 2^1 - 1 = \frac{1}{8} \cdot 2 = \frac{1}{4} = 0.25$

$M_\infty = [111|2] = 0.111 \cdot 2^2 = \frac{7}{8} \cdot 4 = \frac{7}{2} = 3.5$

$|M| = 2 \cdot 2^2 \cdot 4 + 1 = 33$

Példa gépi számhalmazra

Példa

$M(3,-1,2)$ gépi számainak alakja: $\pm 0.1\_\,\_ \cdot 2^k, \quad (-1 \le k \le 2)$

Elemei $k = 0$ esetén: $0.100, 0.101, 0.110, 0.111,$ azaz $\frac{1}{2}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}$.

Valamint $k = -1$ esetén ezek fele, $k = 1$ esetén ezek kétszerese, $k = 2$ esetén ezek négyszerese. (Továbbá negatív előjellel...)

$$\varepsilon_0 = [100|-1] = 0.100 \cdot 2^{-1} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0.25$$

$$\varepsilon_1 = [101|1] - 1 = 0.101 \cdot 2^1 - 1 = \frac{1}{8} \cdot 2 = \frac{1}{4} = 0.25$$

$$M_\infty = [111|2] = 0.111 \cdot 2^2 = \frac{7}{8} \cdot 4 = \frac{7}{2} = 3.5$$

$$|M| = 2 \cdot 2^2 \cdot 4 + 1 = 33$$

Példa

$M(3,-1,2)$ gépi számainak alakja: $\pm 0.1\_\ \_ \cdot 2^k, \quad (-1 \leq k \leq 2)$

Elemei $k = 0$ esetén: $0.100, 0.101, 0.110, 0.111,$ azaz $\frac{1}{2}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}$.

Valamint $k = -1$ esetén ezek fele, $k = 1$ esetén ezek kétszerese, $k = 2$ esetén ezek négyszerese. (Továbbá negatív előjellel...)

$\varepsilon_0 = [100|-1] = 0.100 \cdot 2^{-1} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0.25$

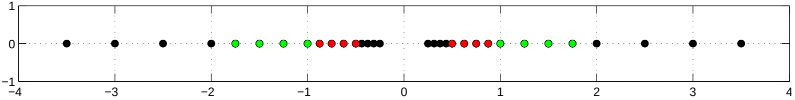
$\varepsilon_1 = [101|1] - 1 = 0.101 \cdot 2^1 - 1 = \frac{1}{8} \cdot 2 = \frac{1}{4} = 0.25$

$M_\infty = [111|2] = 0.111 \cdot 2^2 = \frac{7}{8} \cdot 4 = \frac{7}{2} = 3.5$

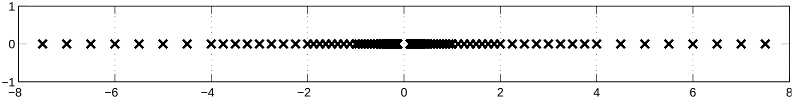
$|M| = 2 \cdot 2^2 \cdot 4 + 1 = 33$

Példa gépi számhalmazra

$M(3, -1, 2)$



$M(4, -2, 3)$

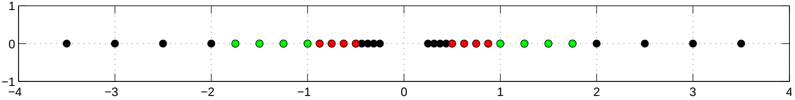


float $\sim M(23, -128, 127)$, double $\sim M(52, -1024, 1023)$

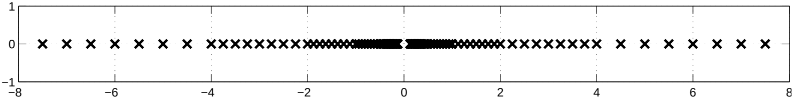
bitek, nevezetes értékek?

Példa gépi számhalmazra

$M(3, -1, 2)$



$M(4, -2, 3)$



float $\sim M(23, -128, 127)$, double $\sim M(52, -1024, 1023)$

bitek, nevezetes értékek?

Hogyan feleltetünk meg egy $\mathbb{R}$-beli számnak egy gépi számot? Jelöljük $\mathbb{R}_M$-mel az ábrázolható számok tartományát, azaz $\mathbb{R}_M := \{x \in \mathbb{R} : \ |x| \leq M_\infty\}$.

Definíció: Input függvény

Az $fl \colon \mathbb{R}_M \to M$ függvényt *input függvénynek* nevezzük, ha

$$fl(x) = \begin{cases} 0 & \text{ha } |x| < \varepsilon_0, \\ \tilde{x} & \text{ha } \varepsilon_0 \leq |x| \leq M_\infty, \end{cases}$$

ahol $\tilde{x}$ az x-hez legközelebbi gépi szám (a kerekítés szabályai szerint).

Tehát már az is egyfajta hibát okoz számításkor, hogy valós számokat számítógépre viszünk. . . de mekkorát?

Tétel: Input hiba

Minden $x \in \mathbb{R}_M$ esetén

$$|x - fl(x)| \leq \begin{cases} \varepsilon_0 & \text{ha } |x| < \varepsilon_0, \\ \frac{1}{2}|x| \cdot \varepsilon_1 & \text{ha } \varepsilon_0 \leq |x| \leq M_\infty, \end{cases}$$

Következmény: Input hiba

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

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

A hiba tehát lényegében ε_1-től, azaz t-től függ.

Tehát már az is egyfajta hibát okoz számításkor, hogy valós számokat számítógépre viszünk... de mekkorát?

Tétel: Input hiba

Minden $x \in \mathbb{R}_M$ esetén

$$|x - fl(x)| \leq \begin{cases} \varepsilon_0 & \text{ha } |x| < \varepsilon_0, \\ \frac{1}{2}|x| \cdot \varepsilon_1 & \text{ha } \varepsilon_0 \leq |x| \leq M_\infty, \end{cases}$$

Következmény: Input hiba

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

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

A hiba tehát lényegében ε_1-től, azaz t-től függ.

Valós számok ábrázolása

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 $x'' - x' = 2^{-t} \cdot 2^k = 2^{k-t}$.

Valós számok ábrázolása

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 $x'' - x' = 2^{-t} \cdot 2^k = 2^{k-t}$.

Valós számok ábrázolása

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 $x' = [1\_ \dots \_|k]$ alakú. Mennyi x' és x'' távolsága?

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

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 $x'' - x' = 2^{-t} \cdot 2^k = 2^{k-t}$.

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 $x'' - x' = 2^{-t} \cdot 2^k = 2^{k-t}$.

Valós számok ábrázolása

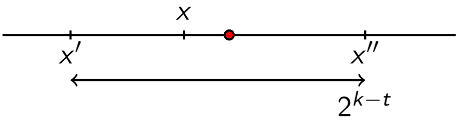
Mennyi a hiba, ha $|x| > M_\infty$?

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:

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

$\square$

Példák Matlab-ban

1. Az említett „furcsa” jelenségek kipróbálása. . .