Numerikus módszerek 1.

1. előadás: Gépi számábrázolás, Hibaszámítás

Krebsz Anna

ELTE IK

Tartalomjegyzék

Tartalomjegyzék

1. furcsa jelenség

Mennyi $sin(\pi)$ értéke?

`1.224646799147353e-016`

2. furcsa jelenség

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 ...

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

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

Próbáljuk ki mindkét számolási módot!

```
0.011134504483941
```

```
0.016926965158418
```

4. furcsa jelenség

A Matlab-ban

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

Mennyi lesz $a + b$ értéke?
Igaz-e az asszociativatás a Matlab-ban?

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

Próbáljuk ki!

5. furcsa jelenség

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!

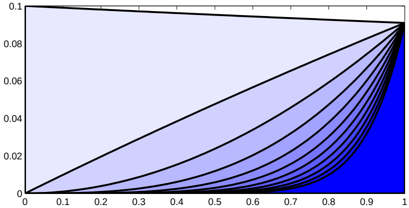
```
0
```

```
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. (Mindig pozitív és nullához tart.)



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

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

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 rekuzió alapján a T_{20}. tagot!

Rendezzük át a rekuziót csökkenően:

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

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 kapott rekuzió alapján a T_{20}. tagot! A két algoritmus közül melyik stabil?

```
7.483468021084803e+003

0.004347035818028
```

Algoritmus stabilitása

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 rekuziója instabil. Lásd gyakorlaton.

Tartalomjegyzék

Motiváció

- 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 egy modellje

Lebegőpontos számok, normalizált alak: 324 $\quad\rightsquigarrow\quad +0.324 \cdot 10^3$.

Kettes számrendszerben: 101000100 $\quad\rightsquigarrow\quad +0.101000100 \cdot 2^9$.

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 egy modellje

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

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

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

Gyakorlatban még hozzávesszük: $0, \infty, -\infty$, `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):

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

Példa gépi számhalmazra

Példa

$$M(3,-1,2) = 0.1\_\;\_ \cdot 2^k, \qquad (-1 \le k \le 2)$$

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…)

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

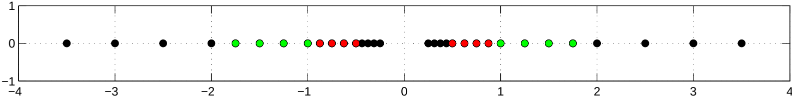
$\varepsilon_1 = [101|1] - 1 = 0.101 \cdot 2^1 - 1 = 0.01 = \frac{1}{4} = 0.25$

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

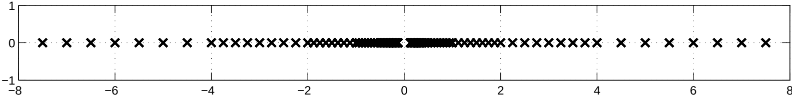
$|M| = 2 \cdot 2^2 \cdot 4 = 32$

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 $fl : \mathbb{R}_M \to M$ 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.

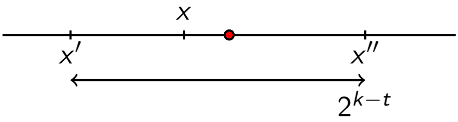
Mennyi a hiba, ha $|x| > M_\infty$?

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:

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

$\square$

Definíció: Hibák jellemzése

Legyen A egy pontos érték, a pedig egy közelítő értéke. Ekkor:

$\Delta a := A - a$	a közelítő érték (pontos) hibája,
$\lvert\Delta a\rvert := \lvert A - a\rvert$	a közelítő érték abszolút hibája,
$\Delta_a \geq \lvert\Delta a\rvert$	az a egy abszolút hibakorlátja,
$\delta a := \frac{\Delta a}{A} \approx \frac{\Delta a}{a}$	az a relatív hibája,
$\delta_a \geq \lvert\delta a\rvert$	az a egy relatív hibakorlátja.

Példa

Vizsgáljuk meg a 3.14 számot mint a π egy közelítő értékét!

Tétel: az alapműveletek hibakorlátai

$$\Delta_{a\pm b} = \Delta_a + \Delta_b \qquad \delta_{a\pm b} = \frac{|a|\cdot\delta_a + |b|\cdot\delta_b}{|a\pm b|}$$

$$\Delta_{a\cdot b} = |b|\cdot\Delta_a + |a|\cdot\Delta_b \qquad \delta_{a\cdot b} = \delta_a + \delta_b$$

$$\Delta_{a/b} = \frac{|b|\cdot\Delta_a + |a|\cdot\Delta_b}{b^2} \qquad \delta_{a/b} = \delta_a + \delta_b$$

Biz.: az összeadást és kivonást azonos előjelű számok között értjük. Az $a \pm b$ hibája

$$\Delta(a\pm b) = (A\pm B) - (a\pm b) = (A-a)\pm(B-b) = \Delta a \pm \Delta b$$

$$|\Delta(a\pm b)| = |\Delta a \pm \Delta b| \le |\Delta a| + |\Delta b| \le \Delta_a + \Delta_b = \Delta_{a\pm b}.$$

Biz.: az összeadás, kivonás hibakorlátai

Nézzük a relatív hibát

$$\frac{\Delta(a \pm b)}{a \pm b} = \frac{\Delta a \pm \Delta b}{a \pm b} = \frac{a \cdot \delta a \pm b \cdot \delta b}{a \pm b}$$

$$\frac{|\Delta(a \pm b)|}{|a \pm b|} = \left|\frac{a \cdot \delta a \pm b \cdot \delta b}{a \pm b}\right| \leq \frac{|a| \cdot |\delta a| + |b| \cdot |\delta b|}{|a \pm b|} \leq$$

$$\leq \frac{|a| \cdot \delta_a + |b| \cdot \delta_b}{|a \pm b|} = \delta_{a \pm b}$$

A szorzás hibája

$$\begin{aligned}\Delta(a \cdot b) &= A \cdot B - a \cdot b = A \cdot B - A \cdot b + A \cdot b - a \cdot b = \\ &= A(B-b) + b(A-a) = A \cdot \Delta b + b \cdot \Delta a = \\ &= (a + \Delta a) \cdot \Delta b + b \cdot \Delta a \approx a \cdot \Delta b + b \cdot \Delta a \\ &(\Delta a \cdot \Delta b \text{ elhanyagolható})\end{aligned}$$

$$|\Delta(a \cdot b)| \leq |a| \cdot |\Delta b| + |b| \cdot |\Delta a| \leq |a| \cdot \Delta_b + |b| \cdot \Delta_a = \Delta_{a \cdot b}$$

A relatív hiba

$$\delta(a \cdot b) = \frac{\Delta(a \cdot b)}{a \cdot b} \approx \frac{a \cdot \Delta b + b \cdot \Delta a}{a \cdot b} = \frac{\Delta b}{b} + \frac{\Delta a}{a} = \delta b + \delta a$$

$$|\delta(a \cdot b)| \leq |\delta a| + |\delta b| \leq \delta_a + \delta_b = \delta_{a \cdot b}$$

Az osztás hibája

$$\Delta\left(\frac{a}{b}\right) = \frac{A}{B} - \frac{a}{b} = \frac{A\cdot b - a\cdot B}{Bb} =$$
$$= \frac{A\cdot b - a\cdot b + a\cdot b - a\cdot B}{Bb} = \frac{b\cdot(A-a) - a\cdot(B-b)}{Bb} =$$
$$= \frac{b\cdot\Delta a - a\cdot\Delta b}{(b+\Delta b)\cdot b} \approx \frac{b\cdot\Delta a - a\cdot\Delta b}{b^2}$$

($\Delta b\cdot b$ elhanyagolható)

$$\left|\Delta\left(\tfrac{a}{b}\right)\right| \le |a|\cdot|\Delta b| + |b|\cdot|\Delta a| \le |b|\cdot\Delta_a + |a|\cdot\Delta_b = \Delta_{a/b}$$

A relatív hiba

$$\delta\left(\tfrac{a}{b}\right) = \frac{\Delta(a\cdot b)}{a\cdot b} \approx \frac{a\cdot\Delta b + b\cdot\Delta a}{a\cdot b} = \frac{\Delta b}{b} + \frac{\Delta a}{a} = \delta b + \delta a = \delta\left(\tfrac{a}{b}\right)$$

$$\left|\delta\left(\tfrac{a}{b}\right)\right| \le |\delta a| + |\delta b| \le \delta_a + \delta_b = \delta_{a/b} \qquad \square$$

1. Tétel: a függvényérték hibája

Ha $f \in C^1(k_{\Delta_a}(a))$ és $k_{\Delta_a}(a) = [a - \Delta_a; a + \Delta_a]$, akkor

$$\Delta_{f(a)} = M_1 \cdot \Delta_a,$$

ahol $M_1 = \max\left\{\, |f'(\xi)| \,:\, \xi \in k_{\Delta_a}(a) \,\right\}.$

Biz.: a Lagrange-féle középértéktétel felhasználásával.

$$\Delta f(a) = f(A) - f(a) = f'(\xi) \cdot (A - a) = f'(\xi) \cdot \Delta a,$$

valamely $\xi \in k_{\Delta_a}(a)$ értékre. Vizsgáljuk az abszolút hibát.

Jó felső becslést adva nyerjük az abszolút hibakorlátot:

$$|\Delta f(a)| = \left|f'(\xi)\right| \cdot |\Delta a| \le M_1 \cdot \Delta_a = \Delta_{f(a)},$$

$\square$

2. Tétel: a függvényérték hibája

Ha $f \in C^2(k_{\Delta_a}(a))$ és $k_{\Delta_a}(a) = [a - \Delta_a; a + \Delta_a]$, akkor

$$\Delta_{f(a)} = \left|f'(a)\right| \Delta_a + \frac{M_2}{2} \cdot \Delta_a^2,$$

ahol $M_2 = \max\left\{\, |f''(\xi)| \,:\, \xi \in k_{\Delta_a}(a) \,\right\}$.

Biz.: a Taylor-formula felhasználásával.

$$\Delta f(a) = f(A) - f(a) = f'(a) \cdot (A - a) + \frac{f''(\xi)}{2} \cdot (A - a)^2,$$

valamely $\xi \in k_{\Delta_a}(a)$ értékre. Vizsgáljuk az abszolút hibát.

Jó felső becslést adva nyerjük az abszolút hibakorlátot:

$$|\Delta f(a)| = |f'(a)| \cdot |\Delta a| + \frac{|f''(\xi)|}{2} \cdot |\Delta a|^2 \leq$$
$$\leq |f'(a)| \cdot |\Delta a| + \frac{M_2}{2} \cdot \Delta_a^2 = \Delta_{f(a)},$$

$\square$

Következmény: függvényérték relatív hibája

Ha Δ_a kicsi, akkor $\delta_{f(a)} = \frac{|a||f'(a)|}{|f(a)|} \cdot \delta_a$.

Definíció: Az f függvény a-beli kondíciószáma

A $c(f,a) = \frac{|a||f'(a)|}{|f(a)|}$ mennyiséget az f függvény a-beli kondíciószámának nevezzük.

Biz.: Ha Δ_a kicsi, akkor a 2. tételben szereplő eredményben a Δ_a^2-es tagot elhanyagolhatjuk, így felhasználva, hogy $\Delta_a = |a| \cdot \delta_a$

$$|\delta f(a)| \approx \frac{|f'(a)| \cdot \Delta_a}{|f(a)|} = \frac{|a|\,\delta_a \cdot |f'(a)|}{|f(a)|} = \quad = \frac{|a|\,|f'(a)|}{|f(a)|} \cdot \delta_a.$$

$\square$

Magyarázat a „furcsa” jelenségekre

1. $sin(\pi)$
2. $\sum_{k=1}^{n} \frac{1}{k}$
3. $\sqrt{2017} - \sqrt{2016}$ („kivonási jegyveszteség”)
4. $T_{n+1} = \frac{1}{n} - 10 \cdot T_n$

További furcsaságok

Gépi számokkal végzett műveletek során...

- sérülhet a pozitivitás

$$\exists a, b \in M; a, b > 0 : a + b = a$$

- sérülhet az asszociativitás

$$\exists a, b, c \in M : (a + b) + c \neq a + (b + c)$$

Példák Matlab-ban

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