

Numerikus módszerek 1.

1. előadás: Gépi számábrázolás, Hibaszámítás

Krebsz Anna

ELTE IK

- ① „Furcsa” jelenségek. . .
- ② Gépi számok: a lebegőpontos számok egy modellje
- ③ A hibaszámítás elemei

- 1 „Furcsa” jelenségek. . .
- 2 Gépi számok: a lebegőpontos számok egy modellje
- 3 A hibaszámítás elemei

Mennyi $\sin(\pi)$ értéke?

1.224646799147353e-016

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

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

Összegezzük oda vagy vissza ...

$n = 100000000$ -re

18.997896413852555

18.997896413853447

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

Más alakban is számolható:

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

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?

1

Igaz-e az asszociativitás a Matlab-ban?

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

Próbáljuk ki!

1

1.0000000000000000e-020

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

$$\begin{aligned}\cosh(20) - \sinh(20) &= \frac{\exp(20) + \exp(-20)}{2} - \frac{\exp(20) - \exp(-20)}{2} = \\ &= \exp(-20)\end{aligned}$$

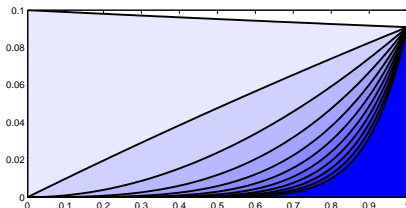
Próbáljuk ki a kétféle számítási módot!

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.)



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

$$T_0 = \int_0^1 \frac{1}{x+10} dx = [\ln(x+10)]_0^1 = \ln(11) - \ln(10) = \ln(1.1)$$

Tehát a rekuzió:

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

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

Rendezzük át a rekurziót csökkenően:

$$10 T_{n-1} = \frac{1}{n} - T_n \Leftrightarrow$$

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

Indítsuk a rekurziót egy $M \gg n$ értékből,

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

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 *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.

- ① „Furcsa” jelenségek. . .
- ② Gépi számok: a lebegőpontos számok egy modellje
- ③ A hibaszámítás elemei

- 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: $324 \rightsquigarrow +0.324 \cdot 10^3$.

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

Általában: $\pm 0. \underbrace{1 \text{ --- } \dots \text{ --- } 1}_{t \text{ jegy}} \cdot 2^k \quad (k^- \leq k \leq k^+)$.

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

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

Gyakorlatban még hozzávesszük: $\infty, -\infty, \text{NaN}, \dots$

Gépi számok tulajdonságai, nevezetes értékei

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

$$\varepsilon_0 = [100 \dots 0 | k^-] = \frac{1}{2} \cdot 2^{k^-} = 2^{k^- - 1}$$

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

$$\varepsilon_1 = [100 \dots 01 | 1] - [100 \dots 00 | 1] = 2^{-t} \cdot 2^1 = 2^{1-t}$$

- 5 M legnagyobb eleme:

$$\begin{aligned} M_\infty &= [111 \dots 11 | k^+] = 1.00 \dots 00 \cdot 2^{k^+} - 0.00 \dots 01 \cdot 2^{k^+} = \\ &= (1 - 2^{-t}) \cdot 2^{k^+} \end{aligned}$$

- 6 M elemeinek száma (számossága):

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

Példa

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

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. . .)

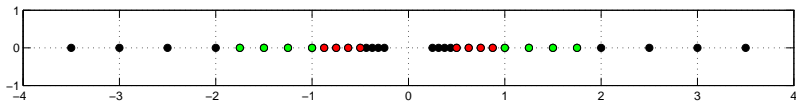
$$\varepsilon_0 = [100|-1] = 0.100 \cdot 2^{-1} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0.25$$

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

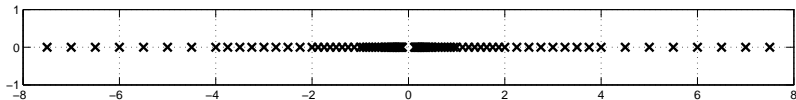
$$M_\infty = [111|2] = 0.111 \cdot 2^2 = \frac{7}{8} \cdot 4 = \frac{7}{2} = 3.5$$

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

$$M(3, -1, 2)$$



$$M(4, -2, 3)$$



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 $fl: \mathbb{R}_M \rightarrow M$ 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 számításakor, 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:

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

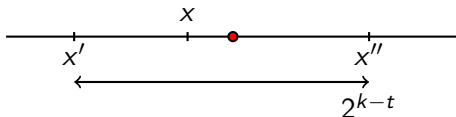
Elegendő csak pozitív x -ekkel foglalkoznunk a 0-ra való szimmetria miatt. Keressük meg azt a két szomszédos gépi számot:

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



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:

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



- 1 „Furcsa” jelenségek. . .
- 2 Gépi számok: a lebegőpontos számok egy modellje
- 3 A hibaszámítás elemei**

Definíció: Hibák jellemzése

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

$\Delta a := A - a$ a közelítő érték (pontos) hibája,

$|\Delta a| := |A - a|$ a közelítő érték abszolút hibája,

$\Delta_a \geq |\Delta a|$ az a egy abszolút hibakorlátja,

$\delta a := \frac{\Delta a}{A} \approx \frac{\Delta a}{a}$ az a relatív hibája,

$\delta_a \geq |\delta a|$ az a egy relatív hibakorlátja.

Példa

Vizsgáljuk meg a 3.14 számot mint a π két tizedesjegyre kerekített értékét!

Tétel: az alpműveletek hibakorlátai

$$\Delta_{a \pm b} = \Delta_a + \Delta_b$$

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

$$\Delta_{a \cdot b} = |b| \cdot \Delta_a + |a| \cdot \Delta_b$$

$$\delta_{a \cdot b} = \delta_a + \delta_b$$

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

$$\delta_{a/b} = \delta_a + \delta_b$$

Megjegyzés: a kapott korlátok két esetben lehetnek nagyságrendileg nagyobbak, mint a kiindulási értékek hibái:

- ① $\delta_{a \pm b}$ esetén, amikor közeli számokat vonunk ki egymásból.
- ② $\Delta_{a/b}$ esetén, amikor kicsi számmal osztunk.

Ezeket az eseteket az algoritmusok implementálásakor el kell kerülni.

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

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

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

Nézzük a relatív hibát

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

$$\begin{aligned} \frac{|\Delta(a \pm b)|}{|a \pm b|} &= \frac{|a \cdot \delta a \pm b \cdot \delta b|}{|a \pm b|} \leq \frac{|a| \cdot |\delta a| + |b| \cdot |\delta b|}{|a \pm b|} \leq \\ &\leq \frac{|a| \cdot \delta_a + |b| \cdot \delta_b}{|a \pm b|} = \delta_{a \pm b} \end{aligned}$$

A szorzás hibája

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

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

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

$$\begin{aligned}\Delta\left(\frac{a}{b}\right) &= \frac{A}{B} - \frac{a}{b} = \frac{A \cdot b - a \cdot B}{Bb} = \\&= \frac{A \cdot b - a \cdot b + a \cdot b - a \cdot B}{Bb} = \frac{b \cdot (A - a) - a \cdot (B - b)}{Bb} = \\&= \frac{b \cdot \Delta a - a \cdot \Delta b}{(b + \Delta b) \cdot b} \approx \frac{b \cdot \Delta a - a \cdot \Delta b}{b^2} \\&(\Delta b \cdot b \text{ elhanyagolható})\end{aligned}$$

$$\left| \Delta\left(\frac{a}{b}\right) \right| \leq \frac{|b| \cdot |\Delta a| + |a| \cdot |\Delta b|}{b^2} \leq \frac{|b| \cdot \Delta a + |a| \cdot \Delta b}{b^2} = \Delta_{a/b}$$

Az osztás relatív hibája

$$\begin{aligned}\delta\left(\frac{a}{b}\right) &= \frac{\Delta\left(\frac{a}{b}\right)}{\frac{a}{b}} \approx \frac{b \cdot \Delta a - a \cdot \Delta b}{b^2} \cdot \frac{b}{a} = \\ &= \frac{b \cdot \Delta a - a \cdot \Delta b}{b \cdot a} = \frac{\Delta a}{a} - \frac{\Delta b}{b} = \\ &= \delta a - \delta b = \delta\left(\frac{a}{b}\right)\end{aligned}$$

$$|\delta\left(\frac{a}{b}\right)| \leq |\delta a| + |\delta b| \leq \delta_a + \delta_b = \delta_{a/b}$$



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

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:

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



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

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

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

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:

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



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 $\Delta_a = |a| \cdot \delta_a$

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

