1. előadás

Lineáris algebra numerikus módszerei

Hiba

A feladatok megoldása során különféle hibaforrásokkal találkozunk:

- **Modellhiba**, amikor a valóságnak egy közelítését használjuk a feladat matematikai alakjának felírásához. (Pl. egy fizikai törvényekkel leírt modellt.)
- **Mérési vagy öröklött hiba**, amikor a modell adatai a pontos értékeknek csak közelítő értékei. Általában a mérés pontosságától függnek.
- **Műveleti (kerekítési-) és input hiba**, amely az adatok számítógépen való ábrázolásából adódnak. A racionális számoknak is csak egy részhalmaza ábrázolható a lebegőpontos aritmetikában. A műveletvégzés során kerekítés, túl- illetve alulcsordulás léphet fel.
- **Képlethiba**, amikor egy végtelen eljárást véges számú lépés után leállítunk, közelítő algoritmusokat alkalmazunk.

Az egész számok

Az egész számokat a számítógépben előjeles vagy előjel nélküli bináris számként képzelhetjük el, így jellemezhetőek a használt bináris jegyek számával. Ez utóbbi nem rögzített, hanem bizonyos mértékben választható. A szokásos az, hogy 2- és 4-byte-os egész számok állnak rendelkezésre, ahol a byte nyolc bitet tartalmaz, azaz nyolc bináris jeggyel rendelkezik (sok gépnél a byte a legkisebb elérhető, címezhető tárolási egység).

Az egész számokkal való aritmetikai műveletek nagyságrenddel gyorsabbak a lebegőpontos számokénál és hibamenteseknek tekinthetők, ezért használatuk döntő mértékben felgyorsíthatja egy adott algoritmus futását a számítógépen.

Az egész számokkal való számítás minden lépését viszont figyelmesen át kell gondolni, mert ilyenkor valójában maradékosztályokban dolgozunk.

A lebegőpontos számok

A számítógépek egy véges számhalmazt ábrázolnak és a számításokat is ezekkel a számokkal végzik. Leggyakrabban a lebegőpontos aritmetikát használják. Nézzük ennek a modelljét:

Definíció

A nemnulla lebegőpontos számok általános alakja:

$$\pm a^k \left(\frac{m_1}{a} + \frac{m_2}{a^2} + \cdots + \frac{m_t}{a^t} \right),$$

ahol $a > 1$ a számábrázolás alapja, $\pm$ az előjel, $t > 1$ a számjegyek száma, $k \in \mathbb{Z}$ a kitevő.

Az m_1 számjegy normalizált, azaz $1 \leq m_1 \leq a - 1$ (ez garantálja a számábrázolás egyértelműségét).
A többi számjegyre: $0 \leq m_i \leq a - 1$ $(i = 2, ..., t)$

A nulla nem normalizált! Ebben az esetben
$k = 0, m_1 = m_2 = \cdots = m_t = 0$, előjele általában $+$.

A számábrázolás alapja lehet $2, 10, 16, \ldots$ (általában a programozási nyelven múlik, hogy melyiket használja)

$t = 8$: egyszeres pontosság, $t = 16$: dupla pontosság.

A lebegőpontos számokat $[\pm, k, m_1, m_2, \ldots, m_t]$ alakban tároljuk (a valóságban ettől eltérhet...), ahol $m := (m_1, m_2, \ldots, m_t)$ a mantissza, míg k a szám karakterisztikája.

A géptől és a pontosságtól függően m tárolására 4,8,16 byte áll rendelkezésre. Ezzel párhuzamosan nő a k értékkészlete. Adott pontosság mellett:

$$L \leq k \leq U,$$

A legnagyobb ábrázolható szám:

$$M^{\infty} = a^{U} \cdot \sum_{i=1}^{t} \frac{a-1}{a^{i}} = a^{U} \cdot \left(\frac{a-1}{a} + \frac{a-1}{a^{2}} + \cdots + \frac{a-1}{a^{t}} \right)$$

$$= a^{U} \left(\frac{a-1}{a} \cdot \frac{1 - \left(\frac{1}{a}\right)^{t}}{1 - \left(\frac{1}{a}\right)} \right) = a^{U} \left(\frac{a-1}{a} \cdot \frac{a^{t}-1}{a^{t}} \cdot \frac{a}{a-1} \right)$$

$$= a^{U} \left(1 - a^{-t}\right)$$

A legkisebb ábrázolható szám: $-M^{\infty}$.

A lebegőpontos számok a $[-M^{\infty}, M^{\infty}]$-beli számok diszkrét (racionális) részhalmazát alkotják és ez a részhalmaz a 0-ra nézve szimmetrikus.

A 0-hoz legközelebbi pozitív lebegőpontos számot ε_0-val jelöljük.

$$\varepsilon_0 = a^L \left(\frac{1}{a} + 0 + 0 + \cdots + 0 \right) = a^{L-1}.$$

Így a 0-n kívül a $(-\varepsilon_0, \varepsilon_0)$ intervallumban nincs más lebegőpontos szám (lehetnek nem normalizált számok, de azokkal nem foglalkozunk).

Az ε_0-hoz legközelebbi pozitív lebegőpontos szám:

$$a^L\left(\frac{1}{a}+0+0+\cdots+\frac{1}{a^t}\right)=\varepsilon_0+a^{L-t}=\varepsilon_0(1+a^{1-t}).$$

Az 1 mindig lebegőpontos szám: $1=[+,1,1,0,0,\ldots 0]$.

Az 1 után a $[+,1,1,0,0,\ldots 0,1]$ lebegőpontos szám következik, ez $1+a^{1-t}=1+\varepsilon_1$, ahol $\varepsilon_1=a^{1-t}$.

Definíció

Ezt az ε_1-et a gép relatív pontosságának, vagy gépi epszilonnak nevezzük.

Az $\varepsilon_0,\varepsilon_1$ számok abszolút és relatív hibakorlátot jelentenek az inputnál és a négy alapműveletnél.

Legyen adott

$$0 < x = [+, k, m] = a^k \left(\frac{m_1}{a} + \frac{m_2}{a^2} + \cdots + \frac{m_t}{a^t} \right) < M^\infty$$

Az x-hez legközelebb eső, x-nél nagyobb lebegőpontos szám: $x + a^{k-t}$, ugyanis

$$\overline{x} = x + a^k \left(0 + 0 + 0 + \cdots + \frac{1}{a^t} \right) = x + a^{k-t},$$

tehát $\delta_x = \overline{x} - x = a^{k-t}$. Mivel k karakterisztikájú számok közül a legkisebb lehetséges érték $a^k \cdot \frac{1}{a}$, ezért (mivel $a^{k-1} \leq x$)

$$\delta_x = \overline{x} - x = a^{k-t} = a^{k-1+1-t} = a^{k-1} \cdot a^{1-t} = a^{k-1} \cdot \varepsilon_1 \leq x \cdot \varepsilon_1,$$

Input hibája

Legyen $x \in \mathbb{R}$, $|x| \leq M^\infty$ és legyen $\mathrm{fl}(x)$ az x-hez hozzárendelt lebegőpontos szám (ez lehet kerekítéssel vagy levágással).

Kerekítés esetén:

$$\mathrm{fl}(x) = \begin{cases} 0, & \text{ha } |x| < \varepsilon_0 \\ \text{az } x\text{-hez legközelebbi lebegőpontos szám}, & \text{ha } \varepsilon_0 \leq |x| \leq M^\infty \end{cases}$$

Ekkor kerekítés esetén

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

Levágás esetén $\frac{1}{2}\varepsilon_1|x|$ helyett $\varepsilon_1|x|$ áll (ez pontatlanabb, de könnyebb levágni, mint kerekíteni).

Alapműveletek hibája

Legyen a $\Diamond$ az alapműveletek bármelyike $(+, -, \cdot, /)$. Ekkor kerekítés esetén

$$|\,\mathrm{fl}(x \Diamond y) - x \Diamond y| \leq \begin{cases} \varepsilon_0, & \text{ha } |x \Diamond y| < \varepsilon_0 \\ 1/2 \cdot \varepsilon_1 \cdot |x \Diamond y|, & \text{ha } |x \Diamond y| \geq \varepsilon_0 \end{cases}$$

vagy az ε_0-lal kapcsolatos eseteket elhagyva:

$$|\,\mathrm{fl}(x \Diamond y) - x \Diamond y| \leq \varepsilon_1 |x \Diamond y| \begin{cases} 1, & \text{levágás esetén} \\ 1/2, & \text{kerekítés esetén} \end{cases}$$

Levágás esetén:

$$-\varepsilon_1 |x \Diamond y| \leq \mathrm{fl}(x \Diamond y) - x \Diamond y \leq \varepsilon_1 |x \Diamond y|$$

Alapműveletek hibája

ebből adódik,hogy

$$\mathrm{fl}(x \Diamond y) - x \Diamond y = \varepsilon_1 \cdot |x \Diamond y| \cdot s \qquad \text{ahol } -1 \leq s \leq 1.$$

Ekkor viszont

$$\mathrm{fl}(x \Diamond y) = x \Diamond y + \varepsilon_1 \cdot |x \Diamond y| \cdot s = x \Diamond y(1 + \mathrm{sgn}(x \Diamond y) \cdot \varepsilon_1 \cdot s)$$

Legyen $\varepsilon_\Diamond := \mathrm{sgn}(x \Diamond y) \cdot \varepsilon_1 \cdot s \leq \varepsilon_1$. Ekkor

$$\mathrm{fl}(x \Diamond y) = x \Diamond y(1 + \varepsilon_\Diamond),$$

ahol

$$|\varepsilon_\Diamond| \leq \varepsilon_1 \cdot \begin{cases} 1, & \text{levágás esetén} \\ 1/2, & \text{kerekítés esetén} \end{cases}$$

Alapműveletek hibája

Ez az összefüggés a 0 körüli hézagban nem érvényes! Továbbá akkor sem, ha a művelet eredménye $> M^{\infty}$ (azaz túlcsordulás esetén)

Ha a művelet eredménye $\neq 0$, de eleme a $(-\varepsilon_0, \varepsilon_0)$ intervallumnak, akkor alulcsordulást kapunk (általában 0-nak veszi a gép hibajelzés nélkül!)

Definíció

Legyen A pontos érték, a pedig annak valamilyen közelítése. A $\Delta a = A - a$ mennyiséget az a közelítés hibájának nevezzük, a $|\Delta a| = |A - a|$ számot pedig az abszolút hibájának. Azt a δa értéket pedig, amelyre fennáll, hogy $|A - a| = |\Delta a| \leq \delta a$, az a abszolút hibakorlátjának mondjuk.

A definíció értelmében használjuk az $A = a \pm \delta a$ hivatkozást is, ami annyit jelent, hogy $A \in [a - \delta a, a + \delta a]$. Nyilván annál jobb a közelítés, más szóval annál élesebb a becslés (és erre törekedni kell), minél kisebb a δa.
A közelítés jóságát ezért az abszolút hiba és az abszolút hibának a pontos érték egységére eső része – a relatív hiba – együtt jellemzi.

Definíció

Az A szám valamely a közelítő értékének relatív hibája a $\frac{\delta a}{|A|}$ mennyiség.

Minthogy az A pontos érték általában nem ismeretes, ezért a $\frac{\delta a}{|A|}$ helyett a $\frac{\delta a}{|a|}$ közelítést használjuk. Az így elkövetett hiba mértéke:

$$\left|\frac{\delta a}{|A|} - \frac{\delta a}{|a|}\right| = \delta a \frac{||a| - |A||}{|a|\,|A|} \leq \delta a \frac{|a - A|}{|a|\,|A|} \leq \frac{(\delta a)^2}{|a|\,|A|}.$$

Szokás a relatív hiba helyett annak százalékos érték ét megadni, azaz $\frac{\delta a}{|A|} \Leftrightarrow 100\frac{\delta a}{|A|}$

Jelölések

A következő jelöléseket és elnevezéseket használjuk: x, y pontos értékek, a és b a közelítő értékeik, δa és δb hibakorlátokkal, azaz $|x - a| = |\Delta a| \leq \delta a$ és $|y - b| = |\Delta b| \leq \delta b$.

Jelölje $\Diamond$ a $+$, $-$, $\cdot$, $/$ műveletek bármelyikét. Az $a \Diamond b$ művelet eredményét az $x \Diamond y$ elméleti eredmény közelítésének tekintjük és a

$$|\Delta (a \Diamond b)| \leq \delta (a \Diamond b),$$

illetve a

$$\frac{|\Delta (a \Diamond b)|}{|(x \Diamond y)|} \leq \frac{\delta (a \Diamond b)}{|(x \Diamond y)|} \approx \frac{\delta (a \Diamond b)}{|(a \Diamond b)|}$$

becsléseket keressük, ahol $\Delta (a \Diamond b)$ a művelet hibáját, $\delta(a \Diamond b)$ pedig abszolút hibakorlátját jelöli. Az additív műveletek (összeadás, kivonás) hibaszámítás szempontjából egymás között hasonlóságot mutatnak, ezért egyetlen tételben adjuk meg a megfelelő hibakorlátokat.

Tétel

Az additív műveletek abszolút hibakorlátjai a következők:

$$\begin{aligned} \delta(a+b) &\leq \delta a + \delta b, \\ \delta(a-b) &\leq \delta a + \delta b. \end{aligned}$$

Bizonyítás

$$\begin{aligned} |\Delta(a \pm b)| &= |(x \pm y) - (a \pm b)| \\ &= |[(a+\Delta a) \pm (b+\Delta b)] - (a \pm b)| \\ &= |\Delta a \pm \Delta b| \leq |\Delta a| + |\Delta b| \leq \delta a + \delta b, \end{aligned}$$

amiből a fenti állításunk következik.

Megjegyzés

Mivel mindkét művelet esetén ugyanazt az eredményt kaptuk, valójában az előjelükre semmilyen kikötést nem kellett tenni. Az eredmény akárhány, tetszőleges előjelű tagra kiterjeszthető. Tekintsük a

$$\sum_{i=1}^{n} x_i \approx \sum_{i=1}^{n} a_i, \qquad (x_i = a_i \pm \delta a_i, \quad i = 1, 2, \ldots, n)$$

összegzést. Könnyen belátható, hogy $\delta\left(\sum_{i=1}^{n} a_i\right) = \sum_{i=1}^{n} \delta a_i$. Természetesen ez az esetek nagy részében jelentősen túlbecsli a tényleges abszolút hibát, hiszen azt tételezi fel, hogy az egyes tagok hibáinak előjele a legkedvezőtlenebbül alakul. Valószínűségszámítási eszközökkel élesebb becslés is adható, jó megbízhatósággal.

Tétel

A multiplikatív műveletek abszolút hibakorlátjai a következők:

$$\begin{aligned} \delta(ab) &\approx |a|\,\delta b + |b|\,\delta a, \\ \delta(a/b) &\approx \frac{|a|\,\delta b + |b|\,\delta a}{|b|^2}. \end{aligned}$$

Bizonyítás

A szorzat abszolút hibakorlátjára kapjuk, hogy

$$\begin{aligned} |\Delta(ab)| &= |xy - ab| = |(a+\Delta a)(b+\Delta b) - ab| \\ &= |a\Delta b + b\Delta a + \Delta a\Delta b| \le |a|\,\delta b + |b|\,\delta a + |\Delta a|\,|\Delta b| \\ &\approx |a|\,\delta b + |b|\,\delta a. \end{aligned}$$

Ha $|a| \gg |\Delta a|$ és $|b| \gg |\Delta b|$, akkor a $|\Delta a|\,|\Delta b|$ másodrendű hibatagot elhanyagolhatjuk és azzal éppen az állításunkat kapjuk.

Bizonyítás

Az osztás esetén természetesen feltesszük, hogy a nevező nem zérus és azt kapjuk, hogy

$$\left|\frac{x}{y}-\frac{a}{b}\right| = \left|\frac{a+\Delta a}{b+\Delta b}-\frac{a}{b}\right| = \left|\frac{-a\Delta b+b\Delta a}{b(b+\Delta b)}\right|$$

$$\leq \frac{|a|\,|\Delta b|+|b|\,|\Delta a|}{b^2\left|1+\frac{\Delta b}{b}\right|} \leq \frac{|a|\,\delta b+|b|\,\delta a}{b^2\left|1+\frac{\Delta b}{b}\right|}$$

$$\approx \frac{|a|\,\delta b+|b|\,\delta a}{b^2}.$$

Itt pedig hasonló meggondolással a $\frac{\Delta b}{b}$ tagot hanyagolhatjuk el az 1 mellett, amivel állításunk kiadódik.

Tétel

Az aritmetikai műveletek relatív hibakorlátjai a következők (feltéve, hogy nevező sehol sem lehet zérus, és az additív műveleteknél az operandusok előjele megegyező):

$$\frac{\delta(a+b)}{|a+b|} = \max\left\{\frac{\delta a}{|a|}, \frac{\delta b}{|b|}\right\},$$

$$\frac{\delta(a-b)}{|a-b|} = \frac{\delta a + \delta b}{|a-b|},$$

$$\frac{\delta(ab)}{|ab|} \approx \frac{\delta a}{|a|} + \frac{\delta b}{|b|},$$

$$\frac{\delta\left(\frac{a}{b}\right)}{\left|\frac{a}{b}\right|} \approx \frac{\delta a}{|a|} + \frac{\delta b}{|b|}.$$

Bizonyítás

Csak az összeadás relatív hibáját bizonyítani.

$$\begin{aligned}\frac{\delta(a+b)}{|a+b|} &= \frac{\delta a+\delta b}{|a+b|} = \frac{\left(\frac{|a|\delta a}{|a|}+\frac{|b|\delta b}{|b|}\right)}{|a+b|} \leq \\ &\leq \max\left\{\frac{\delta a}{|a|},\frac{\delta b}{|b|}\right\}\frac{|a|+|b|}{|a+b|} = \max\left\{\frac{\delta a}{|a|},\frac{\delta b}{|b|}\right\}.\end{aligned}$$

Az utolsó egyenlőség az a és b azonos előjeléből következik.
A kivonásra adott összefüggés megegyezik a definícióval. A szorzás és osztás relatív hibája behelyettesítés után azonnal adódik.