3. előadás

Lineáris algebra numerikus módszerei

Definíció

A $\|.\|_M : \mathbb{R}^{n\times n} \to \mathbb{R}$ mátrixnormát a $\|.\|_V : \mathbb{R}^n \to \mathbb{R}$ vektornorma által indukált mátrixnormának nevezzük, ha

$$\|A\|_M = \max\{\|Ax\|_V : \|x\|_V = 1\}.$$

Az indukált mátrixnorma geometriai jelentése: az egységnormájú x vektorok megnyújtásának (Ax) maximális mértéke. Könnyen igazolható, hogy $\|A\|_1$ az $\|x\|_1$, $\|A\|_2$ az $\|x\|_2$, $\|A\|_\infty$ pedig az $\|x\|_\infty$ vektornorma által indukált mátrixnorma.

Példa

Felhasználva az indukált mátrixnorma definícióját, igazoljuk, hogy $a, b \in R^n$ esetén $\left\|ab^T\right\|_2 = \|a\|_2 \|b\|_2$.

Megoldás

Az értelmezés szerint

$$\left\|ab^T\right\|_2 = \max_{\|x\|_2=1} \left\|ab^T x\right\|_2 = \|a\|_2 \max_{\|x\|_2=1} \left|b^T x\right|$$

Tehát a $|\sum_{i=1}^n b_i x_i| \to \max$, $\sum_{i=1}^n x_i^2 = 1$ feltételes szélsőérték feladatot kell megoldanunk ($b \neq 0$). Analitikus eszközökkel könnyen előállítható a megoldás: $x = \pm b/\|b\|_2$. Eredményünket az egyenlőséglánc jobboldalába helyettesítve megkapjuk a példa állítását.

Tétel

Indukált mátrixnormában $\|AB\|_M \leq \|A\|_M \|B\|_M$ $(\forall A, B \in \mathbb{R}^{n\times n})$.

Bizonyítás

Először igazoljuk, hogy indukált mátrixnormában

$$\|Ax\|_V \leq \|A\|_M \|x\|_V \quad \left(A \in \mathbb{R}^{n\times n},\ x \in \mathbb{R}^n\right).$$

(A továbbiakban az M és V normaindexeket elhagyjuk.)

Bizonyítás

Ha $x \neq 0$, az indukált mátrixnorma definíciója alapján

$$\|Ax\| = \left\| A \|x\| \frac{x}{\|x\|} \right\| = \|x\| \left\| A \frac{x}{\|x\|} \right\| \leq \|x\| \|A\|,$$

ahonnan

$$\|ABx\| \leq \|A\| \|Bx\| \leq \|A\| \|B\| \|x\|$$

Azt az 1 normájú x-et választva melynél az $\|ABx\|$ maximális, éppen az állításunk adódik.

Jelölje $A(i)$ azt az $(n-1)\times(n-1)$-es mátrixot, amelyet az $A \in \mathbb{R}^{n\times n}$ az i-edik sora és *első* oszlopa elhagyásával kapunk:

$$A(i) = \begin{bmatrix} a_{12} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{i-1,2} & \cdots & a_{i-1,n} \\ a_{i+1,2} & \cdots & a_{i+1,n} \\ \vdots & \ddots & \vdots \\ a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

Definíció

A

$$\det(A) = a_{11}, \qquad ha\ n = 1$$

$$\det(A) = a_{11}a_{22} - a_{12}a_{21}, \qquad ha\ n = 2$$

$$\det(A) = \sum_{i=1}^{n} (-1)^{i+1} \cdot a_{i1} \cdot \det(A(i)), \qquad ha\ n \geq 3$$

előírásokkal számított valós számot a négyzetes, $A \in \mathbb{R}^{n \times n}$ mátrix determinánsának nevezzük.

Definíció

Az $X \in \mathbb{R}^{n \times n}$ mátrixot a négyzetes, $A \in \mathbb{R}^{n \times n}$ mátrix inverzének nevezzük, ha

$$AX = XA = I,$$

ahol I az egységmátrix.

Ha az inverz mátrix létezik, akkor egyértelmű. Az inverz mátrix jelölése $A^{-1} = X$.

Tétel

Az inverz mátrixra fennállnak az alábbi tulajdonságok:

$$\left(A^{-1}\right)^{-1} = A, \quad (AB)^{-1} = B^{-1}A^{-1}, \quad \left(A^T\right)^{-1} = \left(A^{-1}\right)^T := A^{-T}.$$

Azokat a mátrixokat, melyeknek létezik inverze, nemszinguláris mátrixoknak nevezzük.

Tétel

Az $A \in \mathbb{R}^{n\times n}$ mátrixnak akkor és csak akkor van inverze, ha $\det(A) \neq 0$.

Ha $\det(A) = 0$, akkor a mátrixot szinguláris nak nevezzük.

Definíció

A $cond(A) = \|A\| \left\|A^{-1}\right\|$ mennyiséget az $A \in \mathbb{R}^{n\times n}$ mátrix kondíciószámának nevezzük.

Külön foglalkozunk az egy- és a többváltozós esetekkel.

Egyváltozós eset

Legyen $f : \mathbb{R} \to \mathbb{R}$ legalább kétszer folyonosan differenciálható függvény, $x = a \pm \delta a$. Az $f(x)$ helyett $f(a)$-t számoljuk. Az

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(\xi)}{2}(x-a)^2 \quad (\xi \in (a-\delta a, a+\delta a))$$

másodrendű Taylor-formulából kapjuk, hogy

$$|f(x) - f(a)| = \left| f'(a)(x-a) + \frac{f''(\xi)}{2}(x-a)^2 \right| \leq \left|f'(a)\right| \delta a + M(\delta a)^2,$$

ahol $M \geq \frac{1}{2}|f''(x)|$ $(x \in [a-\delta a, a+\delta a])$. A másodrendű $M(\delta a)^2$ tagot elhanyagolva kapjuk, hogy a függvénybehelyettesítés abszolút hibája

$$\delta\left(f\left(a\right)\right) \approx \left|f'(a)\right| \delta a.$$

Többváltozós eset

Legyen $f : \mathbb{R}^n \to \mathbb{R}$ legalább kétszer folytonosan differenciálható függvény és $x, a \in \mathbb{R}^n, \Delta a = x - a$, valamint $x_i = a_i \pm \delta a_i$, ahol $x_i, a_i, \delta a_i \in \mathbb{R}$. A többváltozós Taylor-formulából az egyváltozós esethez hasonlóan a másodrendű tagot elhanyagolva (megjegyezzük, hogy nem mindig lehet) kapjuk:

$$f(x) \approx f(a) + \nabla f(a)^T (x - a),$$

ahol $\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right]^T$. Ebből pedig adódik a

$$\delta(f(a)) \approx \sum_{i=1}^{n} \left| \frac{\partial f(a)}{\partial x_i} \right| \delta a_i$$

becslés.

Definíció (Függvények relatív hibája)

$$\frac{\delta\left(f\left(a\right)\right)}{|f(x)|} \approx \frac{\delta\left(f\left(a\right)\right)}{|f(a)|} \approx \frac{|f'(a)|\,\delta a}{|f(a)|}.$$

Egy függvény kiszámítása rendszerint egy algoritmussal történik, ezért érdekes megvizsgálni, hogy az input adat relatív hibáját az algoritmus hányszorosra nagyítja fel, amit a

$$\frac{|f(a+\Delta a)-f(a)|}{|f(a)|} : \frac{|\Delta a|}{|a|}$$

mennyiség fejez ki.

Egyszerű átalakításokkal adódik, hogy

$$\frac{|f(a+\Delta a)-f(a)|}{|f(a)|} : \frac{|\Delta a|}{|a|} \approx \frac{|f'(a)|\,|\Delta a|}{|f(a)|} \cdot \frac{|a|}{|\Delta a|} = \frac{|f'(a)|\,|a|}{|f(a)|}.$$

Definíció (Függvények kondíciószáma)

A

$$c(f,a) = \frac{|f'(a)|\,|a|}{|f(a)|}$$

mennyiséget az $f : \mathbb{R} \to \mathbb{R}$ függvény a pontbeli kondíciószámának nevezzük.

Definíció

Egy függvényt numerikusan instabilnak, vagy rosszul kondicionáltnak nevezünk, ha nagy a kondíciószáma. A függvény stabil, vagy jól kondicionált, ha a kondíciószám kicsi.

Példa

Vizsgáljuk az $f(x) = \log x$ függvényt. Ennek kondíciószáma $c(f, x) = c(x) = 1/|\log x|$, amely $x \approx 1$ esetén nagy. Tehát az $x \approx 1$ értékekre a relatív direkt hiba nagy lesz.

Példa

Az $f(x) = 1 + \sqrt{x-1}$ és $x > 1$. Ekkor

$$c(f,x) = \frac{|x|}{2\left(\sqrt{x-1} + (x-1)\right)},$$

ami tetszőlegesen nagy lehet, ha x elég közel van 1-hez. Ezért a példa függvénye numerikusan instabil. Ha bevezetjük az új $x = 1 + t$ változót, akkor kapjuk, hogy $g(t) = f(1+t) = 1 + \sqrt{t}$. Ennek a függvénynek a $t > 0$ helyen vett kondíciószáma

$$c(g,t) = \frac{\sqrt{t}}{2 + 2\sqrt{t}}.$$

Ha $t \approx 0$, azaz $x \approx 1$, akkor a kondíciószám kicsi marad. Tehát stabilizáltuk a számítást egy egyszerű átalakítással.

A kondíciószámot értelmezhetjük az $F = [f_1, \dots f_n]^T : \mathbb{R}^m \to \mathbb{R}^n$ többváltozós (ún. vektor-vektor) függvényre is. A levezetést mellőzve adódik, hogy

$$c(F, a) = \frac{\|a\| \, \|F'(a)\|}{\|F(a)\|},$$

ahol $F'(x) = \left[\frac{\partial f_i}{\partial x_j}\right]_{i,j=1}^{n,m}$, az ún. Jacobi-mátrix.

Megjegyzés

A kondíciószám normafüggő.

A mátrixok kondíciószáma bevezetésének motivációja:
Legyen $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix, $x, y \in \mathbb{R}^n$ és tekintsük az $F : \mathbb{R}^n \to \mathbb{R}^n$ függvényt, ahol

$$F(x) = y = A^{-1}x$$

(azaz y az $Ay = x$ lineáris egyenletrendszer megoldása a jobboldali vektor függvényében). Egyszerű számolással megmutatható, hogy ezen függvény kondíciószáma $c(F, x) \leq \|A\| \left\|A^{-1}\right\|$, és ez a becslés pontos.
Így $\mathrm{cond}(A) = \|A\| \left\|A^{-1}\right\|$ kifejezés az A mátrix kondíciószáma.

A függvényértékek számítása során – mint már említettük – hiba következhet be.

Jelölje x és y a pontos értékeket és legyen pontosan $y = f(x)$, a ténylegesen számított behelyettesítési érték pedig $\hat{y}$. Az eltérést, azaz a $\Delta y = \hat{y} - y$ értéket **direkt hibának** nevezzük. Amennyiben az $\hat{y}$-ra valamely $\hat{x}$ értékkel pontosan fennáll, hogy $\hat{y} = f(\hat{x}) = f(x + \Delta x)$, akkor a Δx értéket **inverz hibának** mondjuk.

Az $x \to x + \Delta x = \hat{x}$ és az $y \to y + \Delta y = \hat{y}$ megváltozást (vagy megváltoztatást) perturbációnak is szoktuk említeni. Az inverz hiba elemzését és becslését **inverz hibaanalízisnek** nevezzük. Ha több inverz hiba is létezik, akkor a (valamilyen normában) legkisebb inverz hiba meghatározása az érdekes. (Gondoljunk például arra, hogy ha $x, \hat{y} \in \mathbb{R}^n$ és $A \in \mathbb{R}^{n \times n}$, akkor többféle $\Delta A \in \mathbb{R}^{n \times n}$ is szolgáltathatja ugyanazt az $\hat{y} = (A + \Delta A)x$ eredményt.)

A direkt és az inverz hiba kapcsolatának vizsgálatához tegyük fel, hogy f kétszer folytonosan differenciálható. Ekkor tehát felírható a következő Taylor-polinom:

$$\hat{y} = f(x + \Delta x) = f(x) + f'(x)\,\Delta x + \frac{f''(x + \vartheta \Delta x)}{2!}(\Delta x)^2,$$

ahol $\vartheta \in (0, 1)$ Így a számított megoldás abszolút hibája

$$\hat{y} - y = f(x + \Delta x) - f(x) = f'(x)\,\Delta x + \frac{f''(x + \vartheta \Delta x)}{2!}(\Delta x)^2.$$

A relatív hiba pedig

$$\frac{\hat{y} - y}{y} = \left(\frac{x f'(x)}{f(x)}\right)\frac{\Delta x}{x} + O\left((\Delta x)^2\right).$$

Innen kapjuk, hogy

$$\frac{\delta(\hat{y})}{|y|} \leq c(f,x)\,\frac{\delta(x)}{|x|}$$

közelítő egyenlőtlenséget, amely szóban kifejezve a következő:

relatív direkt hiba $\leq$ **kondíciószámszám** $\times$ **relatív inverz hiba**

Az egyenlőtlenség azt mutatja, hogy egy rosszul kondicionált probléma számított megoldásának nagy lehet a (relatív) direkt hibája. Egy $y = f(x)$ értéket számító algoritmust **direkt stabilnak** nevezünk, ha a direkt hiba kicsi és **inverz stabilnak** nevezzük, ha bármely x értékre olyan $\hat{y}$ számított értéket ad, amelyre a Δx inverz hiba kicsi. A kicsi jelző környezetfüggő. Egy direkt stabil módszer nem feltétlenül inverz stabil. Ha az inverz hiba és a kondíciószám kicsi, akkor az algoritmus direkt stabil.

Megjegyzés

A gyakorlatban természetesen a számítás végeredményének a hibája, a direkt hiba a fontos. Az inverz hibaanalízis jelentősége abban áll, hogy sokszor az inverz hibát tudjuk becsülni. Az alkalmazott számítógép számábrázolási pontossága rendszerint ismert, gyakran annak (később tárgyalt) mérőszámát vagy az azzal arányos mennyiséget tekinthetjük inverz hibának. Az arányossági tényező megállapítása tapasztalatok alapján történik, szakkönyvek is ajánlanak értékeket. Jól kondicionált feladat esetén pedig az inverz hibából következtethetünk a direkt hibára.

A lineáris egyenletrendszerek általános alakja m egyenlet és n ismeretlen esetén:

$$\begin{aligned} a_{11}x_1 + \ldots + a_{1j}x_j + \ldots + a_{1n}x_n &= b_1 \\ &\vdots \\ a_{i1}x_1 + \ldots + a_{ij}x_j + \ldots + a_{in}x_n &= b_i \\ &\vdots \\ a_{m1}x_1 + \ldots + a_{mj}x_j + \ldots + a_{mn}x_n &= b_m \end{aligned}$$

Az egyenletrendszert megadhatjuk a tömörebb

$$Ax = b$$

formában is, ahol

$$A = [a_{ij}]_{i,j=1}^{m,n} \in \mathbb{R}^{m\times n},\ x \in \mathbb{R}^n,\ b \in \mathbb{R}^m.$$

Ha $m < n$, akkor **alulhatározott**,
ha $m > n$, akkor **túlhatározott** egyenletrendszerről beszélünk.
Az $m = n$ esetben pedig az egyenletrendszert **négyzetesnek** nevezzük. Az egyenletrendszerek geometriai tartalmát a következőképpen írhatjuk le:

Az $\mathbb{R}^n$ euklideszi tér $d \in \mathbb{R}^n$ normálvektorú és $x_0 \in R^n$ ponton átmenő hipersíkját az

$$(x - x_0)^T d = 0$$

egyenletet kielégítő $x \in R^n$ pontok határozzák meg.

A hipersík egyenlete más formában kifejezve:

$$x^T d = x_0^T d.$$

Az $Ax = b$ egyenletrendszert felírhatjuk az ekvivalens

$$\begin{aligned} a_1^T x &= b_1 \\ &\vdots \\ a_m^T x &= b_m \end{aligned}$$

alakban, ahol $a_i^T = [a_{i1}, \ldots, a_{in}]$,

Innen jól láthatjuk, hogy a lineáris egyenletrendszer megoldása m hipersík közös része. Ennek megfelelően három eset lehetséges:

- (i) az egyenletrendszernek nincs megoldása,
- (ii) az egyenletrendszernek pontosan egy megoldása van,
- (iii) az egyenletrendszernek végtelen sok megoldása van.

Definíció

Ha az $Ax = b$ lineáris egyenletrendszernek legalább egy megoldása van, akkor az egyenletrendszert **konzisztens**nek nevezzük. Ha az egyenletrendszernek nincs megoldása, akkor az egyenletrendszer **inkonzisztens**.

Az $Ax = b$ egyenletrendszert felírhatjuk az ekvivalens

$$\sum_{i=1}^{n} x_i a_i = x_1 a_1 + \ldots + x_n a_n = b$$

alakban is, ahol $a_i \in \mathbb{R}^n$ az A mátrix i-edik oszlopát jelöli (x_i pedig skalár: a megoldásvektor i-edik komponense). A $\sum_{i=1}^{n} x_i a_i$ összeg az $\{a_i\}_{i=1}^{n}$ vektorok lineáris kombinációja. Az egyenletrendszer akkor és csak akkor oldható meg, ha b kifejezhető az A oszlopvektorainak lineáris kombinációjaként.

A megoldhatóságot megállapíthatjuk a rang fogalmának segítségével is:
az $Ax = b$ egyenletrendszernek akkor és csak akkor van megoldása, ha $rank\,(A) = rank\,([A, b])$. Ha $rank\,(A) = rank\,([A, b]) = n$, akkor az $Ax = b$ egyenletrendszernek pontosan egy megoldása van.
A továbbiakban csak négyzetes egyenletrendszerekkel foglalkozunk. Feltesszük tehát, hogy $m = n$.

Tétel

Az $Ax = b$ ($A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$) egyenletrendszernek akkor és csak akkor van pontosan egy megoldása, ha létezik A^{-1}. Ekkor a megoldás $x = A^{-1}b$.

Tétel

Az $Ax = 0$ ($A \in \mathbb{R}^{n \times n}$) homogén lineáris egyenletrendszernek akkor és csak akkor van $x \neq 0$ nemtriviális megoldása, ha $\det(A) = 0$.

A Gauss-módszer két fázisból áll.
I. Azonos (a megoldást őrző) átalakításokkal az $Ax = b$ egyenletrendszert felső háromszög alak úra hozzuk:
II. A kapott felső háromszögmátrixú egyenletrendszert megoldjuk.

Az azonos átalakítást itt most úgy végezzük el, hogy egyik egyenletből kivonjuk a másik egyenlet alkalmasan megállapított konstansszorosát, így ott a szóban forgó ismeretlen együtthatója zérussá válik. Kiiktatjuk – idegen szóval eliminálјuk – az ismeretlent, ezért is nevezzük a módszert eliminációs eljárásnak. Az első lépésben az

$$\begin{array}{ccccccccc} a_{11}^{(2)}x_1 & + & a_{12}^{(2)}x_2 & + & \dots & + & a_{1n}^{(2)}x_n & = & b_1^{(2)} \\ & & a_{22}^{(2)}x_2 & + & \dots & + & a_{2n}^{(2)}x_n & = & b_2^{(2)} \\ & & \vdots & & & & \vdots & & \vdots \\ & & a_{n2}^{(2)}x_2 & + & \dots & + & a_{nn}^{(2)}x_n & = & b_n^{(2)} \end{array}$$

alakot kell tehát kapnunk,

amit ha $a_{11} \neq 0$,akkor elérhetünk úgy, hogy az i-edik sorból kivonjuk $(i = 2, \ldots, n)$ az első sor l_{i1}-szeresét:

$$(a_{i1} - l_{i1}a_{11})x_1 + (a_{i2} - l_{i1}a_{12})x_2 + \ldots + (a_{in} - l_{i1}a_{1n})x_n = b_i - l_{i1}b_1.$$

Az $a_{i1} - l_{i1}a_{11} = 0$ feltételből kapjuk, hogy $l_{i1} = \frac{a_{i1}}{a_{11}}$. Ha ezt a l_{i1} értéket behelyettesítjük az egyenletbe, akkor könnyen ellenőrizhető, hogy itt tényleg arról van szó, hogy az első egyenletből kifejezzük az x_1-et, és a kapott kifejezést behelyettesítjük a maradék többibe. A számítások során nem kell az x_i szimbólumokat magunkkal cipelni, elég, ha az együtthatómátrix elemein hajtjuk végre a megfelelő módosítást. Így könnyen programozható a kinullázás alábbi algoritmusa:

Legyen $a_{ij}^{(1)} = a_{ij}$. Ekkor minden $i = 2, \ldots, n$ és $j = 1, \ldots, n$ esetén

$$l_{i1} = a_{i1}^{(1)} / a_{11}^{(1)}$$

$$a_{ij}^{(2)} = a_{ij}^{(1)} - l_{i1} a_{1j}^{(1)}$$

$$b_i^{(2)} = b_i^{(1)} - l_{i1} b_1^{(1)}$$

A következő lépésben minden $i = 3, \ldots, n$ és $j = 2, \ldots, n$ esetén

$$l_{i2} = a_{i2}^{(2)} / a_{22}^{(2)}$$

$$a_{ij}^{(3)} = a_{ij}^{(2)} - l_{i2} a_{2j}^{(2)}$$

$$b_i^{(3)} = b_i^{(2)} - l_{i2} b_2^{(2)}$$

Általános (k-adik) lépés: minden $i = k+1, \ldots, n$ és $j = k, \ldots, n$ esetén

$$l_{ik} = a_{ik}^{(k)} / a_{kk}^{(k)}$$

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik} a_{kj}^{(k)}$$

$$b_{i}^{(k+1)} = b_{i}^{(k)} - l_{ik} b_{k}^{(k)}$$