„Táblán”

Numerikus módszerek 1. előadás (estis), 2020/2021 ősz

Lócsi Levente

Frissült: 2020. december 2.

Ebben az írásban a 2020/2021 őszi félév estis Numerikus módszerek 1. előadásának a diasorban nem szereplő, a „táblán” kifejezéssel jelölt és ennek megfelelően bemutatott bizonyításait foglaljuk össze.

Itt nem részletezzük az egyes tételekhez kapcsolódó definíciókat, fogalmakat. Látszólag csupán egymással nem összefüggő tételek kimondása és bizonyítása szerepel a következő oldalakon, viszont a tárgy honlapjáról elérhető diasorok kiegészítéseként forgatva hozzájárulhat a tárgy tananyagának alaposabb megértéséhez, az egyes tételek összefüggéseinek átlátásához.

Tartalomjegyzék

Tétel: az input hiba tétele

Minden $x \in \mathbb{R}$, $\varepsilon_0 \leq |x| \leq M_\infty$ *esetén*

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

illetve

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

Bizonyítás:

Ha $\varepsilon_0 \leq |x| \leq M_\infty$, akkor x kettes számrendszerbeli normalizált lebegőpontos alakban a k kitevőre $k^- \leq k \leq k^+$ igaz. Továbbá t kettedes-jegyre kerekítve kapjuk $fl(x)$ értékét. Tehát:

$$x = \pm\, 2^k \cdot 0.1 \_\ \_\ \ldots \_\ |\ \_ \ldots,$$
$$fl(x) = \pm\, 2^k \cdot 0.\underbrace{1\ \_\ \_\ \ldots\ \_}_{t \text{ db}}.$$

Ekkor x és $fl(x)$ eltérése legfeljebb $2^k \cdot 2^{-t-1}$. (Például a $0.10011 \cdot 2^3$ értéket kaphattuk 5 kettedes-jegyre kerekítve, ha a kerekített szám legalább $0.100101 \cdot 2^3$, de kisebb, mint $0.100111 \cdot 2^3$. Az eltérés valóban mindig legfeljebb $0.000001 \cdot 2^3 = 2^{-6} \cdot 2^3$.) Vagyis

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

Viszont x abszolút értékére, fenti alakját figyelembe véve $0.1 \cdot 2^k = 2^{k-1} \leq |x| < 2^k$ is teljesül, ezért a becslést így folytathatjuk:

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

Az állításban szereplő további alakokhoz idézzük fel, hogy $\varepsilon_1 = 2^{1-t} = 2 \cdot 2^{-t}$. $\square$

Tétel: függvényérték hibája

Ha $f \in D(k_{\Delta_a}(a))$*, akkor* $\Delta_{f(a)} = M_1 \cdot \Delta_a$*, ahol*

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

Bizonyítás:

Jelölje A a pontos értéket, a pedig a közelítő értéket, melyre $|A - a| < \Delta_a$. A függvényérték (pontos) hibájának felírásához a Lagrange-féle középértéket használjuk fel:

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

valamely $\xi \in k_{\Delta_a}(a)$ értékre.

Vizsgáljuk az abszolút hibát; erre jó felső becslést adva nyerjük az abszolút hibakorlátot:

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

hiszen M_1-nek a szóban forgó intervallumban tapasztalt deriváltértékek szuprémumát választottuk, Δ_a pedig definíciója szerint legalább akkora, mint $|\Delta a|$. Tehát az $M_1 \cdot \Delta_a$ érték választható az $f(a)$ függvényérték abszolút hibakorlátjának. $\square$

Tétel: a főelemkiválasztásról

Ha az $Ax = b$ egyenletrendszernek létezik egyértelmű megoldása, akkor előfordulhat, hogy főelemkiválasztásra van szükségünk a GE végrehajthatóságához, de a részleges főelemkiválasztás mindig elegendő.

Bizonyítás:

A bizonyítás a tétel két állításának megfelelően két részre tagolódik.

1. Ahhoz, hogy lássuk, hogy a Gauss-elimináció során szükség lehet (részleges) főelemkiválasztásra elég mutatnunk egy olyan lineáris egyenletrendszert, amelynek létezik egyértelmű megoldása, viszont a Gauss-elimináció algoritmusát nem tudjuk végrehajtani sorcserék nélkül. Ilyen például a következő:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \cdot x = \begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

Nyilván mindegy, hogy jobb oldalnak mit választunk...

2. Azt kell belátni, hogy a részleges főelemkiválasztás elegendő ahhoz, hogy a Gauss-elimináció algoritmusát végrehajthassuk egy invertálható mátrixú (egyértelmű megoldással rendelkező) lineáris egyenletrendszer esetén. Vizsgáljuk meg indirekt, mi történne, ha nem volna elegendő a részleges főelemkiválasztás a k-adik lépésben. Ekkor, $k-1$ lépés után az első $k-1$ oszlopot már kinulláztuk a főátló alatt:

$$A^{(k-1)} = \begin{pmatrix} U_{k-1} & B \\ 0 & C \end{pmatrix},$$

ahol U_{k-1}, B és C a Gauss-elimináció során kapott értékek mátrixai. A k-adik lépésben C első oszlopából kellene nem nulla főelemet választanunk. Ha ez nem sikerül, akkor C első oszlopa csupa nulla. Ekkor viszont – annak definíciója alapján – $A^{(k-1)}$ determinánsa (ami megegyezik A determinánsával) is 0. Tehát az egyenletrendszer mátrixa nem invertálható. Ezért ha az egyenletrendszer mátrixa invertálható, akkor a részleges főelemkiválasztás elegendő.

□

Tétel: az LU-felbontás egyértelműségéről

Ha A-nak létezik LU-felbontása, továbbá $\det A \neq 0$*, akkor az LU-felbontás egyértelmű.*

Bizonyítás:

Indirekt tegyük fel, hogy az A invertálható mátrix LU-felbontása nem egyértelmű, azaz legalább kettő létezik:

$$A = L_1 \cdot U_1 = L_2 \cdot U_2.$$

Az egyenlőséget U_2^{-1}-zel jobbról, majd L_1^{-1}-zel balról szorozva kapjuk, hogy

$$U_1 \cdot U_2^{-1} = L_1^{-1} \cdot L_2.$$

A szóban forgó inverzek léteznek, hiszen L_1 determinánsa mindig 1 (nem nulla), valamint $\det U_2 = \det A \neq 0$. Viszont az egyenlőség bal oldalán egy felső háromszögmátrix, jobb oldalán pedig egy 1 főátlójú alsó háromszögmátrix áll. Ez csak úgy lehet, ha az egységmátrixról van szó. Tehát

$$\begin{aligned} U_1 \cdot U_2^{-1} = I \quad &\Longrightarrow \quad U_1 = U_2, \\ L_1^{-1} \cdot L_2 = I \quad &\Longrightarrow \quad L_1 = L_2. \end{aligned}$$

Vagyis az LU-felbontás egyértelmű. □

Tétel: az LU-felbontás „közvetlen” kiszámítása

Az L és U mátrixok elemei a következő képletekkel számolhatók:

$$i \leq j \text{ (felső)} \qquad u_{i,j} = a_{i,j} - \sum_{k=1}^{i-1} l_{i,k} \cdot u_{k,j},$$

$$i > j \text{ (alsó)} \qquad l_{i,j} = \frac{1}{u_{j,j}} \left(a_{i,j} - \sum_{k=1}^{j-1} l_{i,k} \cdot u_{k,j} \right).$$

Ha jó sorrendben számolunk, mindig ismert az egész jobb oldal.

Bizonyítás:

Írjuk fel az $A \in \mathbb{R}^{n \times n}$ mátrix mint mátrixszorzat i-edik sorának j-edik elemét feltéve, hogy $A = L \cdot U$. Használjuk ki, hogy háromszögmátrixokról van szó, majd válasszunk le egy tagot.

1. Ha $i \leq j$, azaz egy főátló feletti (vagy főátlóbeli) elemről van szó, akkor $k > i \Rightarrow l_{i,k} = 0$, valamint $l_{i,i} = 1$, és így

$$a_{i,j} = \sum_{k=1}^{n} l_{i,k} \cdot u_{k,j} = \sum_{k=1}^{i} l_{i,k} \cdot u_{k,j} = u_{i,j} + \sum_{k=1}^{i-1} l_{i,k} \cdot u_{k,j}.$$

Ebből $u_{i,j}$ kifejezhető, a tételben szereplő alakot nyerjük.

2. Ha $i > j$, azaz egy főátló alatti elemről van szó, akkor $k > j \Rightarrow u_{k,j} = 0$, és így

$$a_{i,j} = \sum_{k=1}^{n} l_{i,k} \cdot u_{k,j} = \sum_{k=1}^{j} l_{i,k} \cdot u_{k,j} = l_{i,j} \cdot u_{j,j} + \sum_{k=1}^{j-1} l_{i,k} \cdot u_{k,j}.$$

Ebből $l_{i,j}$ kifejezhető, a tételben szereplő alakot nyerjük.

Figyeljük meg, hogy ha valamely „jó sorrendben” (lásd az előadás diasorát) megyünk végig az (i, j) indexekkel A elemein, akkor az $l_{i,j}$ illetve $u_{i,j}$ értékét megadó egyenlőségek jobb oldalán minden mennyiség ismert. □

Tétel: Cholesky-felbontás egyértelmű létezéséről

Ha A szimmetrikus, pozitív definit mátrix, akkor egyértelműen létezik Cholesky-felbontása.

Bizonyítás:

Teljes indukcióval bizonyítunk n, a mátrix mérete szerint.

- Vizsgáljuk először az $n = 1$ esetet, azaz $A_1 = (a_{1,1}) \in \mathbb{R}^{1\times 1}$. Ekkor A pozitív definit volta csupán annyit jelent, hogy $a_{1,1} > 0$. Legyen $L_1 = (\sqrt{a_{1,1}})$, így nyilván $L_1 \cdot L_1^\top = A_1$.

- Tegyük fel, hogy az állítás igaz $n-1$ esetén, azaz tetszőleges $A_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$ szimmetrikus, pozitív definit mátrix esetén egyértelműen létezik olyan L_{n-1} alsó háromszögmátrix, amelyre $A_{n-1} = L_{n-1} \cdot L_{n-1}^\top$, valamint L_{n-1} főátlóbeli elemei pozitívak.

- Tekintsünk most n esetét, azaz legyen $A_n \in \mathbb{R}^{n\times n}$ szimmetrikus, pozitív definit mátrix. Ekkor az A_n mátrix $n-1$-edik főminora is szimmetrikus, pozitív definit, az indukciós feltevés alkalmazható rá, L_n-et, illetve a Cholesky-felbontást pedig a következő alakban keressük:

$$A_n = \begin{pmatrix} A_{n-1} & b \\ b^\top & a_{n,n} \end{pmatrix} = \begin{pmatrix} L_{n-1} & 0 \\ c^\top & \alpha \end{pmatrix} \cdot \begin{pmatrix} L_{n-1}^\top & c \\ 0 & \alpha \end{pmatrix} = L_n \cdot L_n^\top,$$

  ahol $b, c \in \mathbb{R}^{n-1}$, valamint $\alpha \in \mathbb{R}$. A mátrixszorzást elvégezve a következőket kapjuk:

  1. $A_{n-1} = L_{n-1} \cdot L_{n-1}^\top$, ez volt az indukciós feltételünk;
  2. $b = L_{n-1} \cdot c$, ebből – mivel L_{n-1} invertálható – c egyértelműen kifejezhető;
  3. $b^\top = c^\top \cdot L_{n-1}^\top$, ami ekvivalens az előző ponttal;
  4. $a_{n,n} = c^\top c + \alpha^2$, amiből $\alpha = \sqrt{a_{n,n} - c^\top c}$, azaz szintén létezik, egyértelmű és pozitív… de csak akkor, ha belátjuk, hogy a négyzetgyök alatti mennyiség – amit α^2-tel jelölünk – mindig pozitív lesz. Valóban:

$$0 < \det A_n = \det L_n \cdot L_n^\top = \alpha \cdot \det L_{n-1} \cdot \alpha \cdot \det L_{n-1}^\top = \alpha^2 \cdot (\det L_{n-1})^2.$$

Ezzel beláttuk, hogy tetszőleges méretű szimmetrikus, pozitív definit mátrixnak létezik egyértelmű Cholesky-felbontása. □

Tétel: QR-felbontás létezése és egyértelműsége

Ha $\det A \neq 0$*, akkor* A*-nak létezik* QR*-felbontása.*
Ha még feltesszük, hogy $r_{i,i} > 0$ $(\forall i)$*, akkor egyértelmű is.*

Bizonyítás:

A bizonyítást a Gram–Schmidt-féle ortogonalizációs eljárás adja: az A mátrix oszlopaiból – amelyek a feltétel értelmében lineárisan függetlenek – előállítjuk a Q oszlopait és R ismeretlen elemeit.

Tekintsük a $Q \cdot R = A$ mátrixszorzást:

$$\begin{pmatrix} q_1 & q_2 & \dots & q_n \end{pmatrix} \cdot \begin{pmatrix} r_{1,1} & r_{1,2} & \dots & r_{1,n} \\ 0 & r_{2,2} & \dots & r_{2,n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & r_{n,n} \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & \dots & a_n \end{pmatrix}.$$

- Tekintsük először a_1-et, A első oszlopát. A mátrixszorzásból $r_{1,1} \cdot q_1 = a_1$, amiből $q_1 = \frac{1}{r_{1,1}} \cdot a_1$. Mivel q_1-től azt várjuk el, hogy normált legyen, ezért $r_{1,1} := \|a_1\|_2$.

- Tegyük fel, hogy A első $k-1$ oszlopát már megvizsgáltuk, és így előállítottuk Q első $k-1$ oszlopát, melyek normáltak és egymásra ortogonálisak, valamint R első $k-1$ oszlopának elemeit is ismerjük.

- Tekintsük most a_k-t. A mátrixszorzásból felírhatjuk a_k-t, majd kifejezhetjük q_k-t:

$$a_k = \sum_{j=1}^{k} r_{j,k} \cdot q_j \quad \Longrightarrow \quad q_k = \frac{1}{r_{k,k}} \left(a_k - \sum_{j=1}^{k-1} r_{j,k} \cdot q_j \right)$$

  Az R-beli értékek meghatározásához szorozzuk be skalárisan mindkét oldalt q_i-vel rögzített i értékre $(i = 1, 2, \dots, k-1)$ és használjuk ki, hogy $\langle q_i, q_j \rangle = \delta_{i,j}$, valamint q_k-tól is azt várjuk, hogy merőleges legyen az összes eddigi q_i vektorra:

$$0 = \langle q_k, q_i \rangle = \frac{1}{r_{k,k}} \left(\langle a_k, q_i \rangle - \sum_{j=1}^{k-1} r_{j,k} \underbrace{\langle q_j, q_i \rangle}_{\delta_{i,j}} \right) = \frac{1}{r_{k,k}} \left(\langle a_k, q_i \rangle - r_{i,k} \right).$$

  Innen $r_{i,k} = \langle a_k, q_i \rangle$. Továbbá q_k-tól még azt várjuk el, hogy normált legyen, ezért $r_{k,k} = \left\| a_k - \sum_{j=1}^{k-1} r_{j,k} \cdot q_j \right\|_2$. Így megkaptuk az R mátrix k-adik oszlopának ismeretlen értékeit, az előállított q_k ortogonális az eddigi q_i-kre, valamint normált.

Így megkonstruáltuk A egyértelmű olyan QR-felbontását, amelyben R főátlóbeli elemei pozitívak.

További QR-felbontásokat nyerhetünk (az R főátlóbeli elemeire vonatkozó feltételt elhagyva) egy ± 1 értékeket tartalmazó diagonális mátrix által, vagyis:

$$D := \operatorname{diag}(\pm 1, \pm 1, \cdots \pm 1), \quad A = Q \cdot R = \overbrace{Q \cdot D} \cdot \overbrace{D \cdot R} = Q' \cdot R'.$$

□

Tétel: tetszőleges tükrözés Householder-mátrixszal

Legyen $a, b \in \mathbb{R}^n,\ a \neq b$ *és* $\|a\|_2 = \|b\|_2 \neq 0$. *Ekkor* a

$$v = \frac{a-b}{\|a-b\|_2} \text{ választással } H(v) \cdot a = b.$$

Bizonyítás:

Ismerve, hogy $H(v) = I - 2vv^\top$, számoljuk végig a $H(v) \cdot a$ szorzatot, várván, hogy b-t kapjuk. Közben használjuk ki, hogy $\|a\|_2 = \|b\|_2$, azaz $a^\top a = b^\top b$, valamint a skaláris szorzás kommutatív, azaz $a^\top b = b^\top a$.

$$\left(I - 2\frac{(a-b)(a-b)^\top}{\|a-b\|_2^2}\right) \cdot a = a - \frac{2(a-b)(a^\top a - b^\top a)}{(a-b)^\top (a-b)} =$$
$$= a - \frac{2(a-b)(a^\top a - b^\top a)}{a^\top a - a^\top b - b^\top a + b^\top b} = a - \frac{2(a-b)(a^\top a - b^\top a)}{2(a^\top a - b^\top a)} =$$
$$= a - (a-b) = b.$$

Tehát valóban, két különböző, de azonos hosszúságú vektor átvihető egymásba egy Householder-transzformáció által. (Egyébként v ilyen megválasztásával $H(v) \cdot b = a$ is teljesül.) $\square$

Tétel: az 1-es vektornormáról

A következő formula vektornormát definiál $\mathbb{R}^n$ felett:

$$\|x\|_1 := \sum_{i=1}^{n} |x_i| .$$

Bizonyítás:

Be kell látni, hogy az $\|.\|_1 : \mathbb{R}^n \to \mathbb{R}$ függvény teljesíti a vektornormák 4 axiómáját.

1. Bármely $x \in \mathbb{R}^n$ vektorra $\|x\|_1 \geq 0$. Valóban, hiszen abszolút értékek összegéről van szó.

2. Ha $x = 0$, azaz nullvektor, akkor $\|x\|_1 = \sum_{i=1}^{n} 0 = 0$. Valamint megfordítva, ha $\|x\|_1 = 0$, ami nemnegatív számok (abszolút értékek) összege, akkor x minden koordinátája 0, azaz x nullvektor.

3. Vizsgáljuk meg a $\lambda \cdot x$ vektor normáját ($\lambda \in \mathbb{R}$):

$$\|\lambda x\|_1 = \sum_{i=1}^{n} |(\lambda x)_i| = \sum_{i=1}^{n} |\lambda \cdot x_i| = \sum_{i=1}^{n} |\lambda| \cdot |x_i| = |\lambda| \cdot \sum_{i=1}^{n} |x_i| = |\lambda| \cdot \|x\|_1 .$$

4. A háromszög-egyenlőtlenség belátásához vizsgáljuk $x + y$ normáját (két tetszőleges $\mathbb{R}^n$-beli vektor esetén):

$$\begin{aligned} \|x + y\|_1 = \sum_{i=1}^{n} |(x + y)_i| = \sum_{i=1}^{n} |x_i + y_i| \leq \\ \leq \sum_{i=1}^{n} (|x_i| + |y_i|) = \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i| = \|x\|_1 + \|y\|_1 . \end{aligned}$$

Ezzel megvizsgáltuk, az $\|.\|_1$ teljesíti a vektornormák axiómáit. □

Tétel: indukált mátrixnormákról

Az „indukált mátrixnormák" valóban mátrixnormák.

Bizonyítás:

Be kell látni, hogy tetszőleges $\|.\|_v : \mathbb{R}^n \to \mathbb{R}$ vektornorma esetén az

$$\|A\| := \sup_{x \neq 0} \frac{\|Ax\|_v}{\|x\|_v}$$

szabállyal definiált $\mathbb{R}^{n \times n} \to \mathbb{R}$ függvény teljesíti a mátrixnormák 5 axiómáját.

1. Az $\|A\|$ értéke nemnegatív, hiszen vektorok normájának (nemnegatív számok) hányadosainak szuprémuma.

2. Ha $A = 0$, azaz nullmátrix, akkor $\|Ax\|_v$ mindig 0, így a szuprémum értéke is. Valamint megfordítva, ha a szuprémum 0, akkor minden x-re Ax-nek nullvektornak kell lennie, ez csak úgy lehet, ha A nullmátrix.

3. $$\|\lambda A\| = \sup_{x \neq 0} \frac{\|\lambda Ax\|_v}{\|x\|_v} = \sup_{x \neq 0} \frac{|\lambda| \cdot \|Ax\|_v}{\|x\|_v} = |\lambda| \cdot \sup_{x \neq 0} \frac{\|Ax\|_v}{\|x\|_v} = |\lambda| \cdot \|A\| .$$

4. $$\|A + B\| = \sup_{x \neq 0} \frac{\|(A+B)x\|_v}{\|x\|_v} \leq \sup_{x \neq 0} \frac{\|Ax\|_v + \|Bx\|_v}{\|x\|_v} \leq \sup_{x \neq 0} \frac{\|Ax\|_v}{\|x\|_v} + \sup_{x \neq 0} \frac{\|Bx\|_v}{\|x\|_v}$$

5. Ha $B = 0$, akkor $\|B\| = 0$, valamint $A \cdot B = A \cdot 0 = 0$ és így $\|AB\| = 0$.

   Ha $B \neq 0$, akkor

   $$\|A \cdot B\| = \sup_{x \neq 0} \frac{\|ABx\|_v}{\|x\|_v} = \sup_{x \neq 0, Bx \neq 0} \frac{\|ABx\|_v}{\|Bx\|_v} \cdot \frac{\|Bx\|_v}{\|x\|_v} \leq$$
   $$\leq \sup_{Bx \neq 0} \frac{\|ABx\|_v}{\|Bx\|_v} \cdot \sup_{x \neq 0} \frac{\|Bx\|_v}{\|x\|_v} \leq \sup_{y \neq 0} \frac{\|Ay\|_v}{\|y\|_v} \cdot \sup_{x \neq 0} \frac{\|Bx\|_v}{\|x\|_v} = \|A\| \cdot \|B\| .$$

   Meggondolható, hogy a $Bx \neq 0$ feltétel nem változtatja meg a szuprémum értékét; közben bevezettük az $y := Bx$ jelölést. (Sőt, az $y \neq 0$ feltétel akár több vektort is megengedhet, mint a $Bx \neq 0$ feltétel, ha B nem invertálható.)

□

Tétel: a kettes vektornorma által indukált mátrixnorma

A $\|.\|_2$ vektornorma által indukált mátrixnorma:

$$\|A\| = \left(\max_{i=1}^{n} \lambda_i(A^\top A)\right)^{1/2},$$

ahol λ_i a mátrix i-edik sajátértékét jelöli.

Bizonyítás:

Először belátjuk, hogy $A^\top A$ szimmetrikus és sajátértékei nemnegatívak (azaz A pozitív szemidefinit).

- $(A^\top A)^\top = A^\top (A^\top)^\top = A^\top A$, azaz $A^\top A$ szimmetrikus, vagyis A sajátértékei valósak.
- Legyen $y \neq 0$ az $A^\top A$ mátrix λ-hoz tartozó sajátvektora, azaz $A^\top Ay = \lambda \cdot y$. Szorozzuk meg mindkét oldalt balról az $y^\top$ vektorral: $y^\top A^\top Ay = \lambda \cdot y^\top y$. Innét

$$\lambda = \frac{y^\top A^\top Ay}{y^\top y} = \frac{(Ay)^\top (Ay)}{y^\top y} = \frac{\|Ay\|_2^2}{\|y\|_2^2} \geq 0.$$

Ezután az indukált mátrixnormák definícióját követve Ax normáját fogjuk vizsgálni. Kihasználjuk, hogy $A^\top A$ szimmetrikus, és így (lásd lineáris algebra) létezik U unitér (ortogonális) mátrix, amire $A^\top A = U^\top DU$, ezért $UA^\top AU^\top = D$ valamely D diagonális mátrixra, azaz $A^\top A$ diagonalizálható mégpedig úgy, hogy a diagonálisban $A^\top A$ sajátértékei vannak (ezek nemnegatívak). Bevezetjük az $y = Ux$ jelölést.

$$\|Ax\|_2^2 = (Ax)^\top (Ax) = x^\top A^\top Ax = x^\top U^\top DUx = (Ux)^\top D(Ux) = y^\top Dy =$$
$$= \sum_{i=1}^{n} \underbrace{d_i}_{\geq 0} \cdot |y_i|^2 \leq \max_{i=1}^{n} d_i \cdot \sum_{i=1}^{n} |y_i|^2 = \max_{i=1}^{n} \lambda_i(A^\top A) \cdot \|y\|_2^2 .$$

Viszont $\|y\|_2^2 = y^\top y = (Ux)^\top (Ux) = x^\top U^\top Ux = x^\top x = \|x\|_2^2$, ezért a fenti számítást így folytathatjuk:

$$\|Ax\|_2^2 \leq \ldots \leq \max_{i=1}^{n} \lambda_i(A^\top A) \cdot \|x\|_2^2 .$$

Így viszont $x \neq 0$ esetén:

$$\frac{\|Ax\|_2}{\|x\|_2} \leq \left(\max_{i=1}^{n} \lambda_i(A^\top A)\right)^{1/2}$$

Még azt kell belátni, hogy van is olyan $x \neq 0$ vektor, amire a szuprémum felvétetik. Legyen $\lambda_m = \max \lambda_i(A^\top A)$ és $v_m \neq 0$, $\|v_m\|_2 = 1$ a hozzá tartozó sajátvektor. Erre:

$$\|Av_m\|_2^2 = (Av_m)^\top (Av_m) = v_m^\top \underbrace{A^\top Av_m}_{\lambda_m \cdot v_m} = \lambda_m \cdot \underbrace{v_m^\top v_m}_{=1} = \lambda_m.$$

□

Tétel: LER érzékenysége a bal oldal pontatlanságára

Ha A invertálható, $b \neq 0$ és $\|\Delta A\| \cdot \|A^{-1}\| < 1$, akkor

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A\| \cdot \|A^{-1}\|}{1 - \|\Delta A\| \cdot \|A^{-1}\|} \cdot \frac{\|\Delta A\|}{\|A\|}.$$

Lemma

Ha $\|M\| < 1$, akkor $(I + M)$ invertálható és

$$\left\|(I + M)^{-1}\right\| \leq \frac{1}{1 - \|M\|}.$$

Bizonyítás:

Előbb belátjuk a lemma állítását. (Indukált mátrixnormákkal dolgozunk.)

- Az $I + M$ mátrix tényleg invertálható, hiszen $\varrho(M) \leq \|M\| < 1$, azaz M sajátértékeire: $\lambda_i(M) < 1$. Meggondolható, hogy $I + M$ sajátvektorai ugyanazok, mint M sajátvektorai, a sajátértékekre pedig $\lambda_i(I + M) = 1 + \lambda_i(M)$ teljesül, így $I + M$ minden sajátértéke pozitív, következésképpen $I + M$ invertálható.

- Vizsgáljuk most $I + M$ inverzét, majd ennek normáját.

$$(I + M)^{-1} = I \cdot (I + M)^{-1} = (I + M - M)(I + M)^{-1} = I - M \cdot (I + M)^{-1},$$
$$\left\|(I + M)^{-1}\right\| \leq \|I\| + \|M\| \cdot \left\|(I + M)^{-1}\right\|,$$
$$(1 - \|M\|) \cdot \left\|(I + M)^{-1}\right\| \leq \|I\| = 1.$$

Ezután a tétel állítását bizonyítjuk.

Tudjuk, hogy $(A + \Delta A)(x + \Delta x) = b$, valamint $Ax = b$. A kettő különbségeként $(A + \Delta A) \cdot \Delta x + \Delta A \cdot x = 0$ adódik, avagy

$$(A + \Delta A) \cdot \Delta x = -\Delta A \cdot x,$$
$$A \cdot (I + A^{-1} \cdot \Delta A) \cdot \Delta x = -\Delta A \cdot x.$$

Mivel feltevésünk szerint $\|A^{-1} \cdot \Delta A\| \leq \|A^{-1}\| \cdot \|\Delta A\| < 1$, a lemma alapján mondhatjuk, hogy $(I + A^{-1} \cdot \Delta A)$ invertálható, így az inverz normájára adott becslésünket is felhasználva:

$$\Delta x = -(I + A^{-1} \cdot \Delta A)^{-1} \cdot A^{-1} \cdot \Delta A \cdot x$$
$$\|\Delta x\| \leq \left\|(I + A^{-1} \cdot \Delta A)^{-1}\right\| \cdot \left\|A^{-1}\right\| \cdot \|\Delta A\| \cdot \|x\|$$
$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{1}{1 - \|A^{-1} \cdot \Delta A\|} \cdot \left\|A^{-1}\right\| \cdot \|\Delta A\| \leq \frac{\|A\| \cdot \|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|\Delta A\|} \cdot \frac{\|\Delta A\|}{\|A\|}.$$

□

Tétel: Banach-féle fixponttétel $\mathbb{R}^n$-re

Ha a $\varphi\colon \mathbb{R}^n \to \mathbb{R}^n$ *függvény kontrakció* $\mathbb{R}^n$*-en* $0 \leq q < 1$ *kontrakciós együtthatóval, akkor*

1. $\exists x^* \in \mathbb{R}^n : x^* = \varphi(x^*)$*, azaz létezik fixpont,*
2. *a fixpont egyértelmű,*
3. $\forall x^{(0)} \in \mathbb{R}^n$ *esetén az* $x^{(k+1)} = \varphi(x^{(k)})$, $k \in \mathbb{N}_0$ *sorozat konvergens és* $\lim\limits_{k\to\infty} x^{(k)} = x^*$,
4. *és a következő hibabecslések teljesülnek:*
   - $\left\|x^{(k)} - x^*\right\| \leq q^k \cdot \left\|x^{(0)} - x^*\right\|$,
   - $\left\|x^{(k)} - x^*\right\| \leq \dfrac{q^k}{1-q} \cdot \left\|x^{(1)} - x^{(0)}\right\|$.

Bizonyítás:

(a) A φ leképezés kontrakció voltából következik, hogy φ **folytonos** (sőt egyenletesen folytonos) is, ugyanis: $\|\varphi(x) - \varphi(y)\| \leq q\,\|x - y\|$ esetén $\varepsilon > 0$-hoz választható $\delta = \varepsilon/q$. Ekkor ha $\|x - y\| < \delta$, akkor $\|\varphi(x) - \varphi(y)\| < \varepsilon$.

(b) Belátjuk, hogy a tételben definiált $(x^{(k)})$ **Cauchy-sorozat,** így konvergens.

$$\left\|x^{(k+1)} - x^{(k)}\right\| = \left\|\varphi(x^{(k)}) - \varphi(x^{(k-1)})\right\| \leq q \cdot \left\|x^{(k)} - x^{(k-1)}\right\| \leq$$
$$\leq \ldots \leq q^k \cdot \left\|x^{(1)} - x^{(0)}\right\|.$$

Legyen $m \in \mathbb{N}$, $m \geq 1$, vizsgáljuk még két m távolságra lévő tag különbségét! A háromszög-egyenlőtlenséget és mértani sor összegképletét is felhasználva:

$$\left\|x^{(k+m)} - x^{(k)}\right\| \leq \left\|x^{(k+m)} - x^{(k+m-1)}\right\| + \ldots + \left\|x^{(k+1)} - x^{(k)}\right\| \leq$$
$$\leq (q^{m+k-1} + \ldots + q^k) \cdot \left\|x^{(1)} - x^{(0)}\right\| = q^k \cdot (q^{m-1} + \ldots + 1) \cdot \left\|x^{(1)} - x^{(0)}\right\| <$$
$$< \frac{q^k}{1-q} \cdot \left\|x^{(1)} - x^{(0)}\right\|.$$

Tehát $(x^{(k)})$ Cauchy-sorozat, mivel $(q^k) \to 0$, ha $k \to \infty$.

(c) Így $(x^{(k)})$ konvergens, $x^* := \lim(x^{(k)})$. Mivel φ folytonos, az átviteli elv értelmében $\varphi(x^*) = \lim \varphi(x^{(k)}) = \lim x^{(k+1)} = x^*$, azaz x^* **fixpontja** φ-nek.

(d) Az **egyértelműség** belátásához indirekt tegyük fel, hogy létezik két $x^* \neq x^{**}$ fixpont. Ekkor:

$$\|x^* - x^{**}\| = \|\varphi(x^*) - \varphi(x^{**})\| \leq q \cdot \|x^* - x^{**}\|$$

Ebből viszont $\|x^* - x^{**}\| = 0$, vagyis $x^* = x^{**}$ következik. Ellentmondás!

(e) A hibabecsléshez vizsgáljuk először a k-adik tag hibáját:

$$\left\|x^{(k)} - x^*\right\| = \left\|\varphi(x^{(k-1)}) - \varphi(x^*)\right\| \leq q \cdot \left\|x^{(k-1)} - x^*\right\| \leq \ldots \leq q^k \cdot \left\|x^{(0)} - x^*\right\|.$$

Valamint a korábbi képletben $m \to \infty$ esetén:

$$\left\|x^* - x^{(k)}\right\| \leq \frac{q^k}{1-q} \left\|x^{(1)} - x^{(0)}\right\|.$$

□

Tétel: a relaxált Jacobi-módszer konvergenciájáról

Ha egy mátrixra a $J(1)$ módszer konvergens, akkor $0 < \omega < 1$ esetén a $J(\omega)$ módszer is konvergens. (Az $\omega = 0$ esetben nem.)

Bizonyítás:

A Jacobi-módszer konvergenciája azzal ekvivalens, hogy $\varrho(B_{J(1)}) < 1$. Ez alapján kell belátnunk, hogy tetszőleges $0 < \omega < 1$ esetén $\varrho(B_{J(\omega)}) < 1$ is teljesül, vagyis a relaxált Jacobi-módszer is konvergál.

A Jacobi-iteráció mátrixa $B_{J(1)} = -D^{-1}(L+U)$, a módszer konvergenciája miatt ennek sajátértékeire tehát $\left|\lambda_i(B_{J(1)})\right| < 1$ teljesül. A relaxált Jacobi-módszer pedig így írható:

$$B_{J(\omega)} = (1-\omega)I - \omega D^{-1}(L+U) = (1-\omega)I + \omega B_{J(1)}.$$

Belátható, hogy sajátvektorai ugyanazok, mint $B_{J(1)}$-nek, a sajátértékekre pedig ez adódik: $\lambda_i(B_{J(\omega)}) = (1-\omega) + \omega \cdot \lambda_i(B_{J(1)})$. Ezeket vizsgálva, valamint kihasználva, hogy $0 < \omega < 1$, adódik a kívánt állítás:

$$\left|\lambda_i(B_{J(\omega)})\right| \leq |1-\omega| + \omega \cdot \underbrace{\left|\lambda_i(B_{J(1)})\right|}_{<1} < (1-\omega) + \omega = 1$$

□

Tétel: a relaxált Seidel-módszer konvergenciájáról

Ha egy mátrixra az $S(\omega)$ módszer konvergens, akkor $0 < \omega < 2$.

Lemma

$$\det B = \prod_{i=1}^{n} \lambda_i(B)$$

Bizonyítás:

Előbb belátjuk a lemma állítását. Írjuk fel a B mátrix karakterisztikus polinomját, amelyről tudjuk, hogy gyökei a mátrix sajátértékei; majd rendezzük λ hatványai szerint:

$$p(\lambda) = \det(B - \lambda I) = \prod_{i=1}^{n}(\lambda_i - \lambda) = (-1)^n \cdot \lambda^n + \ldots + \prod_{i=1}^{n} \lambda_i.$$

A $\lambda = 0$ értéket behelyettesítve a konstans tagot kapjuk, amire:

$$p(0) = \det(B) = \prod_{i=1}^{n} \lambda_i.$$

Ezután a tétel állítását bizonyítjuk. A konvergencia tényéből, azaz a $\varrho(B_{S(\omega)}) < 1$ állításból kell ω kívánt becslését előállítanunk. Egyrészt

$$\varrho(B_{S(\omega)}) < 1 \implies \left|\lambda_i(B_{S(\omega)})\right| < 1 \implies \left|\prod_{i=1}^{n} \lambda_i(B_{S(\omega)})\right| < 1 \implies \left|\det(B_{S(\omega)})\right| < 1.$$

Másrészt, mivel az iteráció mátrixa $B_{S(\omega)} = (D + \omega L)^{-1}[(1-\omega)D - \omega U]$, valamint kihasználva, hogy háromszögmátrixok determinánsa a főátlóbeli elemek szorzata (tehát nem függ a diagonálison kívüli elemektől),

$$\left|\det(B_{S(\omega)})\right| = \underbrace{\left|\det((D + \omega L)^{-1})\right|}_{1/|\det(D)|} \cdot \underbrace{|\det((1-\omega)D - \omega U)|}_{|1-\omega|^n \cdot |\det(D)|} = |1-\omega|^n < 1$$

Ebből pedig $|1 - \omega| < 1$ következik, ami ekvivalens a $0 < \omega < 2$ becsléssel. □

Tétel: p-edrendben konvergens iterációk

Legyen $\varphi\colon \mathbb{R} \to \mathbb{R}$, $\varphi \in C^p[a,b]$ *és az* $x_{k+1} = \varphi(x_k)$ *sorozat konvergens. Ha* $\varphi'(x^*) = \cdots = \varphi^{(p-1)}(x^*) = 0$, *de* $\varphi^{(p)}(x^*) \neq 0$, *akkor a konvergencia* p*-edrendű és hibabecslése:*

$$|x_{k+1} - x^*| \leq \frac{M_p}{p!} |x_k - x^*|^p ,$$

ahol $M_p = \max\limits_{\xi \in [a,b]} \left|\varphi^{(p)}(\xi)\right|$.

Bizonyítás:

Írjuk fel a φ függvény x^* körüli Taylor-polinomját a maradéktaggal.
$\exists \xi \in (x, x^*)$ (vagy (x^*, x)):

$$\varphi(x) = \varphi(x^*) + \varphi'(x^*)(x - x^*) + \cdots + \frac{\varphi^{(p-1)}(x^*)}{(p-1)!}(x - x^*)^{p-1} + \frac{\varphi^{(p)}(\xi)}{p!}(x - x^*)^p$$

Vizsgáljuk ezt az $x = x_k$ helyen, kihasználva a deriváltak zérus voltát is. ($\exists \xi_k$):

$$x_{k+1} = \varphi(x_k) = \underbrace{\varphi(x^*)}_{x^*} + \frac{\varphi^{(p)}(\xi_k)}{p!}(x_k - x^*)^p$$

Ezt átrendezve, vegyük szemügyre a $k+1$-edik és a k-adik tag hibáját.

$$|x_{k+1} - x^*| = \frac{\left|\varphi^{(p)}(\xi_k)\right|}{p!} \cdot |x_k - x^*|^p \leq \frac{M_p}{p!} |x_k - x^*|^p ,$$

ahol $M_p = \max\limits_{\xi \in [a,b]} \left|\varphi^{(p)}(\xi)\right|$. Tehát (x_k) egy p-adrendben konvergens sorozat. □

Tétel: a Newton-módszer lokális konvergencia tétele

Ha $f \in C^2[a,b]$ *és*

1. $\exists x^* \in [a,b] : f(x^*) = 0$, *azaz van gyök,*
2. f' *állandó előjelű,*
3. $x_0 \in [a,b] : |x_0 - x^*| < r := \min\left\{ \frac{1}{M}, |x^* - a|, |x^* - b| \right\}$,

akkor az x_0 *pontból indított Newton-módszer másodrendben konvergál a gyökhöz, és az*

$$|x_{k+1} - x^*| \leq M \cdot |x_k - x^*|^2$$

hibabecslés érvényes, ahol

$$M = \frac{M_2}{2 \cdot m_1}, \quad m_1 = \inf_{x \in [a,b]} |f'(x)|, \quad M_2 = \sup_{x \in [a,b]} |f''(x)|.$$

Bizonyítás:

Alkalmazzuk az f függvényre a Taylor-formulát, x_k középponttal az x^* helyen, másodfokú maradéktaggal. $\exists \xi_k \in (x_k, x^*)$ (vagy (x^*, x_k)):

$$0 = f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{f''(\xi_k)}{2}(x^* - x_k)^2.$$

Mindkét oldalt $f'(x_k)$-val osztva, majd átrendezve és a Newton-módszer képletét felismerve kapjuk, hogy

$$0 = \frac{f(x_k)}{f'(x_k)} + x^* - x_k + \frac{f''(\xi_k)}{2 \cdot f'(x_k)}(x^* - x_k)^2,$$
$$\left(x_k - \frac{f(x_k)}{f'(x_k)} \right) - x^* = x_{k+1} - x^* = \frac{f''(\xi_k)}{2 \cdot f'(x_k)}(x^* - x_k)^2,$$
$$|x_{k+1} - x^*| \leq \frac{M_2}{2 \cdot m_1} \cdot |x_k - x^*|^2 = M \cdot |x_k - x^*|^2,$$

ahol M, m_1, M_2 a tételben definiált mennyiségek. Bevezetve az $\varepsilon_k := x_k - x^*$ jelölést, így is írhatjuk: $|\varepsilon_{k+1}| \leq M \cdot |\varepsilon_k|^2$. Ezzel beláttuk, hogy ha (x_k) konvergál, akkor másodrendben teszi azt és határértéke x^*.

A konvergencia bizonyításához belátjuk, hogy az $|\varepsilon_k|$ hibakorlátok sorozata 0-hoz tart. Bevezetjük a $d_k := M \cdot |\varepsilon_k|$ jelölést.

$$|\varepsilon_{k+1}| \leq M \cdot |\varepsilon_k|^2 \quad \Longrightarrow \quad M \cdot |\varepsilon_{k+1}| \leq (M \cdot |\varepsilon_k|)^2,$$
$$d_{k+1} \leq d_k^2 \quad \Longrightarrow \quad d_k \leq d_{k-1}^2 \leq d_{k-2}^{2 \cdot 2} \leq \ldots \leq d_0^{2^k},$$
$$M \cdot |\varepsilon_k| \leq (M \cdot |\varepsilon_0|)^{2^k} \quad \Longrightarrow \quad |\varepsilon_k| \leq \frac{1}{M} \cdot (M \cdot |\varepsilon_0|)^{2^k}.$$

De mivel $|\varepsilon_0| = |x_0 - x^*| < \frac{1}{M}$, így $M \cdot |\varepsilon_0| < 1$, ezért a fenti becslésre tekintettel $|\varepsilon_k| \to 0$. Tehát az (x_k) sorozat konvergens, másodrendben tart x^*-hoz. $\square$

Tétel: becslés polinom gyökeire

A $P(x) = a_n \cdot x^n + a_{n-1} \cdot x^{n-1} + \cdots + a_1 \cdot x + a_0$ *polinom esetén, ha* $a_0 \neq 0$ *és* $a_n \neq 0$*, akkor* P *bármely* x_k *gyökére:*

1. $|x_k| < \underbrace{1 + \frac{\max\limits_{i=0}^{n-1} |a_i|}{|a_n|}}_{R}$,

2. $|x_k| > \underbrace{\left(1 + \frac{\max\limits_{i=1}^{n} |a_i|}{|a_0|}\right)^{-1}}_{r}$.

Bizonyítás:

Komplex gyökök is szóba jöhetnek, így akár komplex együtthatós polinomokat is megengedhetünk, a bizonyítás menetén nem változhat, gondolhatunk valós együtthatós polinomokra is.

1. A felső becslés belátásához induljunk ki a háromszög-egyenlőtlenségből. Megmutatjuk, hogy ha $|x| \geq R$, akkor $|P(x)| > 0$. Feltehetjük, hogy $|x| > 1$.

$$a_n \cdot x^n = P(x) - \sum_{k=0}^{n-1} a_k \cdot x^k \quad \Longrightarrow \quad |a_n \cdot x^n| \leq |P(x)| + \left|\sum_{k=0}^{n-1} a_k \cdot x^k\right|.$$

Innen $|P(x)| \geq |a_n| \cdot |x|^n - |a_{n-1}x^{n-1} + \ldots + a_0|$. Hogy a becslést folytathassuk, növeljük a kivonandó összeget.

$$\left|a_{n-1}x^{n-1} + \ldots + a_0\right| \leq |a_{n-1}| \cdot |x|^{n-1} + \ldots + |a_0| \leq$$
$$\leq \left(\max_{i=0}^{n-1} |a_i|\right) \cdot (|x|^{n-1} + \ldots + 1) = \left(\max_{i=0}^{n-1} |a_i|\right) \cdot \frac{|x|^n - 1}{|x| - 1} <$$
$$< \left(\max_{i=0}^{n-1} |a_i|\right) \cdot \frac{|x|^n}{|x| - 1}.$$

Folytassuk $|P(x)|$ becslését, majd vizsgáljuk meg, mikor pozitív.

$$P(x) > |a_n| \cdot |x|^n - \left(\max_{i=0}^{n-1} |a_i|\right) \cdot \frac{|x|^n}{|x| - 1} \geq 0$$

Ezt az egyenlőtlenséget vizsgálva kapjuk, hogy

$$|a_n| \cdot |x|^n \geq \left(\max_{i=0}^{n-1} |a_i|\right) \cdot \frac{|x|^n}{|x| - 1} \quad \Longleftrightarrow \quad |x| \geq 1 + \frac{\max\limits_{i=0}^{n-1} |a_i|}{|a_n|} =: R.$$

Tehát ha $|x| \geq R$, akkor $|P(x)| > 0$; ennélfogva P minden gyökére $|x| < R$ teljesül.

2. Az alsó becslést pedig azáltal nyerjük, hogy az imént belátott becslést alkalmaz-

zuk $P(x)$ reciprok-polinomjára. ($x \neq 0$)

$$y := \tfrac{1}{x}, \qquad P(x) = P(\tfrac{1}{y}) = a_n(\tfrac{1}{y})^n + a_{n-1}(\tfrac{1}{y})^{n-1} + \ldots + a_1(\tfrac{1}{y}) + a_0 =$$
$$= (\tfrac{1}{y})^n \cdot \underbrace{(a_n + a_{n-1}y + \ldots + a_1 y^{n-1} + a_0 y^n)}_{Q(y)} = x^n \cdot Q(\tfrac{1}{x}).$$

A Q polinomot a P *reciprok-polinomjának* nevezzük. Ekkor $P(x_k) = 0 \iff Q(\frac{1}{x_k}) = 0$, vagyis Q gyökei P gyökeinek reciprokai. Alkalmazzuk a már belátott becslésünket Q-ra:

$$\frac{1}{|x_k|} < 1 + \frac{\max\limits_{i=1}^{n} |a_i|}{|a_0|} = \frac{1}{r} \quad \Longrightarrow \quad |x_k| > r.$$

□