1. Gépi szám, hibák

Áttekintjük a gépi aritmetika néhány jellegzetességét és szemügyre vesszük a számításokat terhelő hibafajtákat.

1.1. A gépi számok

A gépi számok leggyakrabban 2-es alapú (vagy bináris), előjeles normalizált számok, így elsősorban ezekkel fogunk foglalkozni. Alakjuk

$$\pm .101\ldots01 \cdot 2^{k} \;=\; \pm m \cdot 2^{k} \qquad (1.1)$$

előjel, t db bináris jegy ↖ kitevő

A nemzérus mantissza mindig 1-gyel kezdődik, emiatt $0.5 \le m < 1,\ \ m \neq 0$. Ha az alap nem 2, akkor a 10-es és a 16-os (hexadecimális) számok fordulnak még elő a gyakorlatban.

A bináris gépi számok halmazát jelölje $M(t,k^-,k^+)$, ahol t a mantisszahossz, k^- a legkisebb kitevő, k^+ pedig a legnagyobb kitevő. Az általunk használt PC-kben, - személyi számítógépekben a szimplapontos szám 4 *bájt* = 32 *bit* területet foglal el a memóriában és az egyes funkciók kiosztása a következő:

1	8	23

1 bit jut az előjelre, 8 bit a kitevőre és 23 a mantisszára. Ezen számok pontossága kb. 7 decimális jegynek felel meg ($23\log_{10}2 \approx 6.923$, azaz kb. 0.3-del szorzandó a bitek száma) és a nagyságrend 10^{-38}-tól 10^{38}-ig terjedhet. A duplapontos (kétszeres pontosságú) számok 64 biten helyezkednek el:

1	11	52

előjel: 1 bit, kitevő 11 bit és a mantisszahossz: 52 bit. Most a pontosság kb. 15 decimális jegy, és az ábrázolható számok nagyságrendje 10^{-307}-től 10^{307}-ig terjed. Egyes programnyelvek megengedik a négyszeres pontosságú számokat is.

A konkrét megvalósításban kihasználható, hogy a nemzérus mantissza első bitje mindig 1, emiatt elhagyható. Ezzel a fogással még plusz 1 bithez lehet jutni, aminek jelentősége az aritmetika tulajdonságainak javításában van. Ekkor viszont meg kell tudni különböztetni a zérust 0.5-től. Erre többféle lehetőség van, hiszen zérus mantissza mellett a kitevő bitjei extra információt hordozhatnak. Az igen nagy abszolút értékű, a gépi számokkal nem ábrázolható számok jelölésére is ki lehet alakítani egy bit-kombinációt. A már nem ábrázolható nagy számokra a ∞ jelet fogjuk használni. Szokás még az NaN jelölés: „not-a-number": *nem szám*, értsd: nem gépi szám. Egyes programnyelvekben ezt kapjuk eredményül, ha zérussal próbálunk osztani. Ha NaN-nel ezután bármilyen aritmetikai műveletet végzünk, az eredmény NaN, mégha zérussal szoroztunk, akkor is.

1.2. Nevezetes gépi számok

A legkisebb pozitív mantissza: ½. A legnagyobb mantissza: $.\overbrace{11\ldots1}^{t \text{ db } 1\text{-es}} = 1-2^{-t}$. $M(t,k^-,k^+)$-ban a legkisebb pozitív szám: $\varepsilon_0 = .10\ldots0 \cdot 2^{k^-} = 1/2 \cdot 2^{k^-}$.

A másik nevezetes szám ε_1, az a legkisebb pozitív szám, amelyet 1-hez hozzáadva 1-nél nagyobb gépi számot kapunk: $1+\varepsilon_1 = .10\ldots01 \cdot 2^{+1}$, innen $\varepsilon_1 = 2^{-t+1}$. A legnagyobb ábrázolható szám: $M_\infty = (.11\ldots1 \cdot 2^{k^+}) = (1-2^{-t})2^{k^+}$. A legkisebb szám ennek a negatívja.

Kezdés: $d'_1 = d_1, \quad b'_1 = b_1.$

$i = 2,3,\ldots,n$-re

$s := a_{i-1} / d'_{i-1}; \quad d'_i := d_i - c_{i-1} * s; \qquad b'_i := b_i - b'_{i-1} * s.$

$x_n := b'_n / d'_n;$

$i = n-1, n-2,\ldots,1$-re

$x_i := (b'_i - c_i * x_{i+1}) / d'_i.$

A másik módszer a megoldás második fázisában érvényes rekurziót veszi alapul:

$$x_i = f_i - g_i x_{i+1}.$$

Az egyenletrendszer első sorából $x_1 = (b_1 - c_1 x_2)/d_1$, ezzel $f_1 = b_1/d_1$ és $g_1 = c_1/d_1$. Ezután az i-edik sorba helyettesítve x_{i-1} kifejezését

$$a_{i-1}(f_{i-1} - g_{i-1}x_i) + d_i x_i + c_i x_{i+1} = b_i,$$

innen

$$x_i = \frac{b_i - a_{i-1}f_{i-1}}{d_i - a_{i-1}g_{i-1}} - \frac{c_i}{d_i - a_{i-1}g_{i-1}} x_{i+1} = f_i - g_i x_{i+1}.$$

ahonnan f_i és g_i előállítása kiolvasható. Ezzel az „üldözéses" vagy „passzázs" algoritmus:

Kezdés: $f_1 = b_1/d_1, \quad g_1 := c_1/d_1.$

$i = 2,3,\ldots,n$-re

$s := d_i - a_{i-1}g_{i-1}; \quad f_i := (b_i - a_{i-1}f_{i-1})/s; \qquad g_i := c_i/s.$

$x_n := f_n;$

$i = n-1, n-2,\ldots,1$-re

$x_i := f_i - g_i * x_{i+1}.$

5.3.3 Feladat

- Ha új jobboldal vektort kapunk, milyen részletszámításokat őrizzünk meg és mit számítsunk újra mindkét algoritmusban?

A relatív hiba megállapításakor elég a mantissza hibáját tekinteni, mert a kitevő osztáskor kiesik. A kerekítéskor a mantisszában elkövetett hiba legfeljebb 2^{-t-1}. A relatív hibájának felső korlátját úgy kapjuk, hogy a lehetséges legkisebb pozitív mantissza-értékkel osztunk: ½-vel. Így kapjuk eredményül $\varepsilon_M = 2^{-t}$-t.

1.3 Gyakorlat. Hogyan módosulna a gépi epszilon, ha a kerekítés helyett csonkítást alkalmaznánk?

1.4. A gépi aritmetika

Vannak gépi számaink, a következő kérdés, hogy milyen tulajdonságú lesz a lebegőpontos számokkal megvalósított gépi aritmetika. A következő számpéldákban a tízes alapú számrendszert fogjuk használni, ahol van négy decimális jegyünk és a kitevő előjeles kétjegyű szám lehet. Ezen gépi számok halmazát egyszerűen M-mel fogjuk jelölni. Jelölés: $0.2543 \cdot 10^2 = 0.2543\ +02$

A gépi artimetikában nem lesz igaz minden, amit a valós számtestben megszoktunk. Az alábbiakban felsorolunk ilyen eltéréseket:

- Létezhet nemzérus $a,b \in M$, amelyre $a + b = a$. Ez a számok eltérő nagyságrendje miatt lehetséges. Például adjuk össze a következő számokat: 0.3460 +02 és 0.4524 –03:

$$\begin{array}{r} 0.3460\ +02 \\ \underline{0.000004524\ +02} \\ 0.3460\ +02 \end{array}$$

- Létezhet $a,b,c \in M$, amelyre $(a+b)+c \neq a+(b+c)$. Például

$$\begin{array}{r} 0.3460\ +02 \\ \underline{0.00004524\ +02} \\ 0.3460\ +02 \end{array} \qquad \begin{array}{r} 0.3460\ +02 \\ \underline{0.00003872\ +02} \\ 0.3460\ +02 \end{array}$$

de először a két kicsi számot összeadva

$$\begin{array}{r} 0.3872\ -02 \\ \underline{0.4524\ -02} \\ 0.8386\ -02 \end{array} \qquad \begin{array}{r} 0.3460\ +02 \\ \underline{0.00008386\ +02} \\ 0.3461\ +02 \end{array}$$

Ez arra int, hogyha sok számot összegzünk, akkor az abszolút érték szerinti kicsikkel érdemes kezdeni.

- Létezhet $a,b,c \in M$, amelyre $(ab)c \neq a(bc)$. Például

$$(0.1245\ +62 \cdot 0.4314\ -58) \cdot 0.4362\ -54 = .5371\ +03 \cdot 0.4362\ -54 = .2343\ -51,$$

míg a másik zárójelezés szerint a második és harmadik szám szorzata kisebb, mint a legkisebb ábrázolható gépi szám, így ez a szorzat zérus, ami a teljes szorzatra zérus eredményt ad. Így, ha sok számot kell összeszoroznunk, még nagyobb gondossággal kell eljárnunk, mert könnyen kerülhetünk abba a helyzetbe, hogy az eredmény, vagy valamely rész-szorzata kívül esik a számábrázolás tartományán. Ha az eredmény túl nagy, vagy túl kicsi, akkor egy lehetőség a gondok csökkentésére az eredmény logaritmusát számolni.

- Összevonás után az eredmény relatív hibája jelentősen megnőhet. Például

$$\begin{array}{r} 0.4693\ +02 \\ \underline{-0.4682\ +02} \\ 0.0011 + 02 \end{array}$$

ami egyenlő 0.1100 +00-val. Látjuk, itt már csak az első két jegy pontos. Ezt jelenséget *kivonási jegyveszteségnek* nevezzük. Néha adhatók fogások a kivonási jegyveszteség elkerülésére vagy

5.2. Főátló-dominancia

Sorok szerint főátló-domináns vagy diagonál-dominánsnak nevezzük a mátrixot, ha minden sorban a nemdiagonális sorelemek abszolút összege kisebb, mint a főátlóbeli elem abszolút értéke:

$$|a_{ii}| > \left\| e_i^T (A - \mathrm{diag}(A)) \right\|_\infty .$$

Lényegében főátló-domináns a mátrix, ha *nem minden sorban* a $\geq$ jel is megengedett és ezek a sorok nem zérus sorok. Az oszlopok szerint *főátló-domináns* mátrixok értelmezése hasonló. Itt $\mathrm{diag}(A) = D$, a mátrix főátlójából készített diagonálmátrixot jelöli.

5.2.1 Tétel, a főátló-dominancia megmaradása.

Amennyiben az A mátrix főátló-domináns, az LU-felbontás végrehatása során a még fel nem bontott jobb alsó részben a főátló-dominancia megmarad. Másképpen: a Schur-komplemens megőrzi a főátló-dominanciát.

Bizonyítás. Az *LU*-felbontás első lépése után a mátrix első oszlopa az $a_{11}e_1$ vektorba megy át és a k-adik sorvektor:

$$e_k^T (I - (a_1 / a_{11} - e_1) e_1^T) A = (e_k^T A - \frac{a_{k1}}{a_{11}} e_1^T A)(I - e_1 e_1^T), \quad k > 1 ,$$

ahol a hozzáírt $I - e_1 e_1^T$ vetítőmátrix az amúgy is zérus első sorelemet nullázza, így változást nem okoz. Az $e_k^T A(I - e_1 e_1^T)$ sorvektor rendelkezik a főátló-dominancia tulajdonsággal, mert csak az első a_{k1} elemet hagytuk el. A levont vektor sornormája pedig

$$\left\| a_{k1} e_1^T A(I - e_1 e_1^T) / a_{11} \right\|_\infty = |a_{k1}| \left\| e_1^T A(I - e_1 e_1^T) / a_{11} \right\|_\infty < |a_{k1}| ,$$

ha $a_{k1} \neq 0$. Itt az átlóelemmel osztott első sor normája kisebb 1-nél (főátló-dominancia!) és ez szorozza a_{k1}-et. Tehát a kivett a_{k1} helyébe egy kisebb abszolút értékű elem kerül az abszolút sorösszeg számításakor, így a *k*-adik sor főátló-dominanciája nem romolhat. A további lépésekben a helyzet hasonló. ■

A tétel következménye, hogy főátló-domináns mátrixoknál az átlóelem mindig alkalmas főelemnek.

5.2.2 Feladatok. Mutassuk meg:

- A főátló-dominancia megmarad, ha a mátrixot balról nemszinguláris diagonálmátrixszal szorozzuk, vagy ha ugyanazt a két sort és oszlopot felcseréljük.
- Lényegében főátló-domináns mátrixok *LU*-felbontásakor a j-edik lépésben szigorú főátló-dominancia következik be a *k*-adik sorban, ha a j-edik sorban megvolt a szigorú főátló-dominancia és volt nemzérus $a_{jk}^{(j)}$, $j < k$ elem.
- Az oszlopok szerinti főátló-dominancia is öröklődik.

5.3. Két- és háromátlójú mátrixok

5.3.1 Speciális mátrixok

A kétátlójú vagy bidiagonális mátrixoknál csak a főátló és valamelyik mellette lévő átlóban vannak nemzérus elemek: $a_{ij} \neq 0$, $j - i \in \{0,1\}$, vagy $j - i \in \{0,-1\}$. Nevezetes képviselőjük a különbségképzés mátrixa:

A hibanalízis szempontjából fontosak az alapműveletek, $+,-,*,/$ hibái. Alább a baloldali összefüggések a hibákra, a jobboldaliak pedig a hibakorlátokra vonatkoznak:

$$\begin{aligned} &\Delta(x \pm y) = \Delta x \pm \Delta y, && \Delta_{x\pm y} = \Delta_x + \Delta_y , \\ &\Delta(xy) = x\Delta y + y\Delta x, && \Delta_{xy} = |x|\Delta_y + |y|\Delta_x , \\ &\Delta(x/y) = \frac{y\Delta x - x\Delta y}{y^2}, && \Delta_{x/y} = \frac{|y|\Delta_x + |x|\Delta_y}{|y^2|}. \end{aligned} \tag{1.5}$$

A hibaformulák hasonló módon származtathatók, mint az összeg-, szorzat-, és hányadosfüggvények differenciálási szabályai. Innen az is látható, hogy a formulák csak akkor tekinthetők jóknak, ha a hibák valóban kicsik, és a másodrendű hibatagok elhanyagolhatók. A jobboldali formulák a baloldaliakból következnek, akárcsak az alábbi, relatív hibákra vonatkozó kifejezések:

$$\begin{aligned} &\delta(x \pm y) = \frac{x\delta x \pm y\delta y}{x \pm y}, && \delta_{x\pm y} = \frac{|x|\delta_x + |y|\delta_y}{|x + y|}, \\ &\delta(xy) = \delta y + \delta x, && \delta_{xy} = \delta_y + \delta_x , \\ &\delta(x/y) = \delta x - \delta y, && \delta_{x/y} = \delta_x + \delta_y . \end{aligned} \tag{1.6}$$

A függvényértékek hibája. Legyen $f : \mathbb{R} \to \mathbb{R}$ legalább kétszer folytonosan differenciálható függvény. Ekkor a Lagrange középérték-tétel szerint létezik $\xi \in [x, \hat{x}]$, amelyre

$$f(\hat{x}) = f(x) - f'(x)\Delta x + f''(\xi)(\Delta x)^2 / 2 .$$

Innen a másodrendű kicsiny utolsó tag elhagyásával a *függvényérték hibája*:

$$f(\hat{x}) - f(x) = \Delta f \approx -f'(x)\Delta x .$$

Legyen $\max_{x \in [x-\Delta_x, x+\Delta_x]} |f'(x)| = M_1$, ezzel $\Delta_f = M_1 \Delta_x$, ahol vegyük tekintetbe, hogy a becslés x egy Δ_x sugarú környezetére vonatkozik. A relatív hibára kapjuk:

$$\delta f = \frac{\Delta f}{f(x)} \approx -\frac{xf'(x)}{f(x)} \frac{\Delta x}{x} = -\frac{xf'(x)}{f(x)} \delta x.$$

Az abszolút értékekre áttérve:

$$|\delta f| \approx c(f, x) |\delta x|, \tag{1.7}$$

ahol a $c(f, x) = |xf'(x) / f(x)|$ számot az f függvény x *pontbeli kondiciószámának* nevezzük. Ha ez a szám nagy, akkor a függvényt *instabilnak*, vagy *gyengén meghatározottnak* nevezzük, mert az argumentum kicsiny megváltozása nagy függvényérték-megváltozást eredményez. Túl nagy kondiciószám mellett a gépi számok kerekítési hibái is elviselhetetlenül nagy végső hibához vezetnek.

Fontos még az *inverz stabilitás* fogalma. Egy leképezés inverz stabil, ha az eredmény egy kissé perturbált kezdetiértékből pontos számítással megkapható.

5. Az LU-felbontás tulajdonságai, speciális inverzek

5.1. Szimetrikus pozitív definit mátrixok

Egy valós szimmetrikus A mátrixot *pozitív definitnek* nevezünk, ha $x^T Ax > 0$ teljesül minden $x \neq 0$ vektorra. *Pozitív szemidefinit* a mátrix, ha csak $x^T Ax \geq 0$ teljesül. A *negatív definit* és *negatív szemidefinit* tulajdonságot hasonlóképp értelmezzük, ha $x^T Ax < 0$ vagy $x^T Ax \leq 0$ valamelyike teljesül. *Indefinit* esetben a belső szorzat negatív és pozitív értékeket egyaránt felvehet.

A pozitív definit tulajdonságnak adható még két másik ekvivalens definiciója. Az egyik szerint ekkor a mátrix minden sajátértéke pozitív, a másik szerint pedig a bal felső sarok aldeterminánsok (főminorok) mind pozitívak. Szemidefinit mátrixnak van zérus sajátértéke és zérus értékű sarok aldeterminánsa.

A nemszimmetrikus mátrixot pozitív definitnek mondjuk, ha a szimmetrikus része pozitív definit. A mátrix szimmetrikus része $A_+ = (A + A^T)/2$ és az antiszimmetrikus része $A_- = (A - A^T)/2$, $A = A_+ + A_-$. Vegyük észre, az antiszimmetrikus részhez tartozó belső szorzat $x^T A_- x$ mindig zérus.

Ha x-et e_i-nek választjuk, akkor a definicióból következik, hogy valós szimmetrikus pozitív definit mátrixra $a_{ii} > 0$ minden i-re, $x = e_i \pm e_j$ esetén pedig $a_{ii} + a_{jj} \pm 2a_{ij} > 0$-nak kell teljesülnie. Ezek az egyszerű feltételek néha hasznosak annak gyors eldöntésében, hogy a mátrix egyáltalán lehet-e pozitív definit. Például, ha a mátrix főátló-beli elemei mind 0-k és a főátlón kívüli elemek között vannak nemzérus elemek, akkor rögtön állítható, hogy a mátrix indefinit.

5.1.1 Tétel, elegendő feltétel pozitív definitségre.

Ha $A = V^T V$ alakban előállítható, ahol V oszlopai lineárisan függetlenek, akkor A pozitív definit.

Bizonyítás. A definició alapján minden nemzérus x-re $x^T Ax = x^T V^T Vx = \|Vx\|_2^2 > 0$ mert $Vx \neq 0$, ha V oszlopai lineárisan függetlenek. ■

5.1.2 Tétel, a pozitív definitség megőrződik az LU-felbontásban.

Pozitív definit A mátrix *LU*-felbontása megőrzi a pozitv definitséget, más szóval: minden lépés után a visszamaradó jobb alsó blokk pozitív definit marad. Az állítás blokk *LU*-felbontáskor is igaz.

Bizonyítás. Legyen A blokkos alakja

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad (A \mid A_{11}) = A_{22} - A_{21} A_{11}^{-1} A_{12},$$

ahol a blokk *LU*-felbontás egy lépése után visszamaradó blokk az $(A \mid A_{11})$ Schur-komplemens. Azt kell igazolni, hogy tetszőleges nemzérus x_2 vektorra $x_2^T (A \mid A_{11}) x_2 > 0$. Az állítást azzal bizonyítjuk, hogy megmutatjuk: létezik egy kiegészített $x^T = (x_1^T, x_2^T)$ vektor, amelyre $x^T Ax = x_2^T (A \mid A_{11}) x_2$. Ehhez válasszuk x_1-et úgy, hogy szorzáskor az első blokk-sor zérust adjon: $A_{11} x_1 + A_{12} x_2 = 0$. Ezzel $x_1 = -A_{11}^{-1} A_{12} x_2$ és $0 < x^T Ax = \begin{bmatrix} x_1^T & x_2^T \end{bmatrix} \begin{bmatrix} 0 \\ A_{21} x_1 + A_{22} x_2 \end{bmatrix} = x_2^T (A_{22} - A_{21} A_{11}^{-1} A_{12}) x_2$. ■

Megjegyzés. Ugyanígy látható be, hogy a felbontás során a pozitív szemidefinitség is megőrződik.

$$\alpha\beta \leq \int_0^\alpha x^{p-1} dx + \int_0^\beta y^{q-1} dy = \frac{\alpha^p}{p} + \frac{\beta^q}{q}$$

Ezután az

$$\alpha_i = \frac{|x_i|}{\|x\|_p}, \qquad \beta_i = \frac{|y_i|}{\|y\|_q}$$

helyettesítéssel és az i szerinti összegzés elvégézésével kapjuk (2.3) jobb oldali összefüggését.

(2.1) harmadik összefüggése, a háromszög-egyenlőtlenség úgy látható be, hogy $p/q = p - 1$ szem előtt tartása mellett a

$$\|x + y\|_p^p = \sum_{i=1}^n |x_i + y_i|^p \leq \sum_{i=1}^n \{|x_i| + |y_i|\} |x_i + y_i|^{p-1}$$

egyenlőtlenség jobb oldalának mindkét tagjára alkalmazzuk a Hölder-egyenlőtlenséget. Ekkor az első tagra a következő eredmény adódik:

$$\sum_{i=1}^n |x_i| |x_i + y_i|^{p-1} \leq \|x\|_p \left\{ \sum_{i=1}^n |x_i + y_i|^{(p-1)q} \right\}^{1/q} = \|x\|_p \|x + y\|_p^{p/q},$$

és a másik taggal is hasonló eredményre jutunk, a kettőt együtt rendezve kapjuk a kívánt egyenlőtlenséget, amit általánosan a p index mellett a *Minkowski-egyenlőtlenségnek* nevezünk.

2.3. A hatványnormák néhány tulajdonsága

A hatványnormákra teljesül:

$$\|x\|_{p+s} \leq \|x\|_p, \quad 1 \leq p, \quad 0 \leq s, \tag{2.4}$$

hiszen ez az összefüggés átírható a

$$\sum_{i=1}^n \left| \frac{x_i}{x_k} \right|^{p+s} \leq \left\{ \sum_{i=1}^n \left| \frac{x_i}{x_k} \right|^p \right\} \left\{ \sum_{i=1}^n \left| \frac{x_i}{x_k} \right|^p \right\}^{s/p}, \quad x_k \neq 0$$

alakba. Ha itt $|x_k| = \max_i |x_i|$ akkor a jobb oldal első tényezője tagról tagra nagyobb vagy egyenlő a bal oldalnál, a második tényező viszont biztosan nem kisebb 1-nél.

A fontosabb hatványnormák a következők:

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

Ez az 1-es vagy oktaéder norma, mivel a 3-dimenziós térben az azonos normájú vektorok egy olyan oktaéderen helyezkednek el, amelynek csúcspontjai az $\|x\|_1 \{(\pm 1, 0, 0), (0, \pm 1, 0), (0, 0, \pm 1)\}$ vektorok.

$$\|x\|_2 = \left\{ \sum_{i=1}^n |x_i|^2 \right\}^{1/2}$$

az x vektor euklidészi, kettes vagy gömbnormája.

A $p \to \infty$ határesetben adódik

$$\begin{bmatrix} \boxed{A_{11}} & A_{12} \\ A_{21}A_{11}^{-1} & A_{22}-A_{21}A_{11}^{-1}A_{12} \end{bmatrix}, \text{ ahol } L=\begin{bmatrix} I_1 & 0 \\ A_{21}A_{11}^{-1} & I_2 \end{bmatrix},\ U=\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22}-A_{21}A_{11}^{-1}A_{12} \end{bmatrix}. \tag{4.5}$$

4.4. Schur-komplemens.

A felbontás jobb alsó sarkában megjelent mátrixot az A mátrix A_{11}-re vonatkozó Schur-komplemensének nevezzük és jelölése: $(A|A_{11}) = A_{22} - A_{21}A_{11}^{-1}A_{12}$. Természetesen létezik az A_{22}-re vonatkozó Schur-komplemens is. Ez az előbbiből úgy jön létre, hogy az $1 \leftrightarrow 2$ indexcserét elvégezzük.

4.5. Particionált inverz

A (4.5) felbontás alapján írhatjuk:

$$A=\begin{bmatrix} I_1 & 0 \\ A_{21}A_{11}^{-1} & I_2 \end{bmatrix}\begin{bmatrix} A_{11} & A_{12} \\ 0 & (A|A_{11}) \end{bmatrix}=\begin{bmatrix} I_1 & 0 \\ A_{21}A_{11}^{-1} & I_2 \end{bmatrix}\begin{bmatrix} A_{11} & \\ & (A|A_{11}) \end{bmatrix}\begin{bmatrix} I_1 & A_{11}^{-1}A_{12} \\ 0 & I_2 \end{bmatrix},$$

ahonnan

$$A^{-1}=\begin{bmatrix} I_1 & -A_{11}^{-1}A_{12} \\ 0 & I_2 \end{bmatrix}\begin{bmatrix} A_{11}^{-1} & \\ & (A|A_{11})^{-1} \end{bmatrix}\begin{bmatrix} I_1 & 0 \\ -A_{21}A_{11}^{-1} & I_2 \end{bmatrix}. \tag{4.6}$$

A diádösszeg kifejtés blokkos alakját felhasználva (ld. *3.5 Gyakorlat*) ez még a két blokk-oszlop és blokk-sor alapján kifejthető az

$$A^{-1}=\begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}+\begin{bmatrix} -A_{11}^{-1}A_{12} \\ I_2 \end{bmatrix}(A|A_{11})^{-1}\begin{bmatrix} -A_{21}A_{11}^{-1} & I_2 \end{bmatrix}. \tag{4.7}$$

alakban.

4.3 Gyakorlat. A *3.13 Gyakorlat* alapján mutassuk meg, hogy a (4.4) mátrix inverze úgy készíthető, hogy a 21-es blokk negatívját vesszük. A felső háromszögmátrixra vonatkozó eredmény innen transzponálással származtatható.

4.6. A Gauss-Jordan módszer az inverz mátrix kiszámítására

Láttuk, minden mátrix, amelynek van inverze, egyszerű mátrixok szorzatára bontható, ahol az n művelet mindegyike tartalmaz egy sorcserét – ha szükséges, és egy diáddal módosított egységmátrixszal való szorzást. Egy ilyen műveletsorozattal a mátrix az egységmátrixba transzformálható. Kézenfekvő az ötlet: a mátrixhoz hozzáírjuk az egységmátrixot: $A \to [A,I]$ és a kibővített mátrixra alkalmazzuk a transzformáció-sorozatot: $[TA,T]=[I,T]$. Világos, $T=A^{-1}$.

Ez a módszer alkalmas lineáris egyenletrendszer megoldására is, de a műveletszámlálás azt mutatja, hogy az LU-felbontás előnyösebb. Ha azonban a mátrix inverzét akarjuk előállítani, akkor a műveletigény ugyanakkora, sőt lehetőség van arra, hogy a mátrixot helyben invertáljuk.

Tegyük fel, az i-edik lépésben A_i már olyan, hogy a sorcserét végrehajtottuk, ha kellett. Az i-edik szorzás:

$$\left(I-\frac{A_ie_i-e_i}{e_i^TA_ie_i}e_i^T\right)A_i=A_i-\frac{A_ie_ie_i^TA_i}{e_i^TA_ie_i}+\frac{e_ie_i^TA_i}{e_i^TA_ie_i}.$$

Itt jobb oldalon az első két tag azt a diád-levonást jelenti, amit már megismertünk. Az LU-felbontáshoz képest azonban eltérés, hogy az i-edik sor és oszlop kivételével minden területre kell

Ad 2. $\|\lambda A\| = \sup_{\|x\|=1}\|\lambda Ax\| = |\lambda|\sup_{\|x\|=1}\|Ax\| = |\lambda|\|A\|$.

Ad 3. $\|A+B\| = \sup_{\|x\|=1}\|(A+B)x\| \le \sup_{\|x\|=1}\{\|Ax\|+\|Bx\|\} \le \|A\|+\|B\|$.

Ad 4. $\exists x_0 \in \mathbb{R}^n$, $\|x_0\|=1$: $\|AB\| = \|ABx_0\| \le \|A\|\|Bx_0\| \le \|A\|\|B\|$. ■

2.7. Az indukált mátrixnormák meghatározása

$p=1$, *oszlopnorma:*

$$\|A\|_1 = \max_{(j)}\|Ae_j\|_1 = \max_{(j)}\sum_{i=1}^m|a_{ij}|. \tag{2.9}$$

Legyen $\|x\|_1=1$, ekkor $\|Ax\|_1 = \sum_{i=1}^m\left|\sum_{j=1}^n a_{ij}x_j\right| \le \sum_{i=1}^m\sum_{j=1}^n|a_{ij}||x_j| = \sum_{j=1}^n|x_j|\sum_{i=1}^m|a_{ij}| \le$ $\le\left(\sum_{j=1}^n|x_j|\right)\max_{(j)}\sum_{i=1}^m|a_{ij}| = \max_{(j)}\|Ae_j\|_1$. Ezt a felső korlátot valamely e_k-ra el is éri, így a maximumot találtuk meg.

$p=\infty$, *sornorma*:

$$\|A\|_\infty = \max_{(i)}\|e_i^TA\|_\infty = \max_{(i)}\|A^Te_i\|_1 = \max_{(i)}\sum_{j=1}^n|a_{ij}|. \tag{2.10}$$

Legyen $\|x\|_\infty=1$, ekkor $\|A\|_\infty = \max_{(i)}\left|\sum_{j=1}^n a_{ij}x_j\right| \le \max_{(i)}\sum_{j=1}^n|a_{ij}||x_j| \le \max_{(i)}\sum_{j=1}^n|a_{ij}|$. A kapott becslés az $x=[x_j]=[\overline{a}_{kj}/|a_{kj}|]$ vektorra valósul meg, ahol a felső vonás komplex konjugáltat jelent komplex számok esetére. Ekkor $\|x\|_\infty=1$ és $\|Ax\|_\infty$ éppen a megállapított felső korlát.

$p=2$, *spektrál norma:*

$$\|A\|_2 = \max_{(k)}\left(\lambda_k(A^TA)\right)^{1/2}. \tag{2.11}$$

Ekkor a következő maximumot keressük:

$$\|A\|_2^2 = \max\frac{\|Ax\|_2^2}{\|x\|_2^2} = \max\frac{x^TA^TAx}{x^Tx}.$$

Az itt látható hányados az A^TA mátrixra vonatkozó *Rayleigh-hányados*. Ha A^TA egy sajátvektora u_k λ_k sajátértékkel, akkor $x=u_k$ választással a Rayleigh-hányados értéke éppen λ_k lesz. Innen világos, a Rayleigh hányados legnagyobb értéke legalább $\lambda_{\max}=\max_k\lambda_k$. Megmutatjuk, nagyobb értéke nem lehet. Tudjuk, a szimmetrikus mátrix sajátvektorai teljes ortonormált rendszert alkotnak, így bármely x vektor kifejthető $x=\sum_{j=1}^n\alpha_ju_j$ alakban. Ezt helyettesítve a Rayleigh-hányadosba, a különbségre kapjuk:

$$\lambda_{\max}-\frac{x^TA^TAx}{x^Tx} = \lambda_{\max}-\frac{\sum_{j=1}^n\lambda_j\alpha_j^2}{\sum_{j=1}^n\alpha_j^2} = \frac{\sum_{j=1}^n(\lambda_{\max}-\lambda_j)\alpha_j^2}{\sum_{j=1}^n\alpha_j^2} \ge 0,$$

ami mutatja, hogy a maximumot találtuk meg.

2.8. A spektrálsugár és az indukált normák összefüggése

Egy mátrix *spektrál sugara* alatt a következőt értjük:

4.1. Tétel, LU-felbontás.

Ha $A \in \mathbb{R}^{n\times n}$ nemszinguláris mátrix, akkor a sorai mindig átrendezhetők egy P permutáció-mátrixszal PA-ba úgy, hogy az felbontható egy L alsó és U felső háromszögmátrix szorzatára. PA felbontása egyértelmű, ha L átlóelemeit 1-nek választjuk.

Bizonyítás. Tekintsük A első oszlopát. Ha a_{11} zérus, akkor keressünk az oszlopban egy nemzérus elemet és sorcserével mozgassuk az első sorba. A továbbiakban feltesszük, hogy $a_{11} \neq 0$. Ekkor szorozzuk A-t az $L_1^{-1} = I - (Ae_1 / a_{11} - e_1)e_1^T$ mátrixszal! A 3.7 példában láttuk, ennek a mátrixnak determinánsa és minden átlóeleme 1, következik, hogy az inverzét úgy kapjuk, ha a benne szereplő diádot pozitív előjellel vesszük. A szorzás eredményeként az Ae_1 oszlopvektor

$$\left(I - (Ae_1 / a_{11} - e_1)e_1^T\right) Ae_1 = Ae_1 - Ae_1 + a_{11}e_1 = a_{11}e_1 \tag{4.1}$$

-be megy át, tehát

$$L_1^{-1}A = \begin{bmatrix} a_{11} & * & \dots & * \\ 0 & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \dots & * \end{bmatrix}, \tag{4.2}$$

ahol a * egységesen nemzérus mátrixelemeket jelöl. Látjuk, a felső háromszögmátrix első oszlopa megjelent. $L_1 = I + (Ae_1 / a_{11} - e_1)e_1^T$ pedig a LU-felbontásban szereplő L mátrix első szozója, ahonnan kiolvashatjuk L első oszlopvektorát: Ae_1 / a_{11}-et.

A második lépésben ugyanezt ismételjük meg a kapott mátrix jobb alsó $(n-1)\times(n-1)$-es blokkjára, ahol az első lépés valamely nemzérus elemnek a 2,2 pozicióba mozgatása, ha szükséges:

$$A_2 = \begin{pmatrix} a_{11} & * & \dots & * \\ 0 & \boxed{*} & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \dots & * \end{pmatrix},$$

így L második oszlopában az első elem 0, a második elem 1. Az eljárást hasonlóan folytatva végül

$$L = L_1 L_2 \dots L_{n-1}, \quad U = L_{n-1}^{-1} L_{n-2}^{-1} \dots L_1^{-1} PA = \begin{pmatrix} * & * & \dots & * \\ & * & \dots & * \\ & & \ddots & \vdots \\ & & & * \end{pmatrix}. \tag{4.3}$$

■

Ha az $Ax = b$ egyenletrendszert oldjuk meg, akkor a b vektort célszerű az A mátrix mellé venni: $[A,b]$, mert b-re is ugyanazok a transzformációk hatnak. Például legyen az egyenletrendszer:

$$\begin{bmatrix} 2 & 0 & 3 \\ -4 & 5 & -2 \\ 6 & -5 & 4 \end{bmatrix} x = \begin{bmatrix} -1 \\ 3 \\ -3 \end{bmatrix}.$$

Vegyük észre, az L_1^{-1}-gyel való szorzás a mátrix első sorát nem változtatja meg: $e_1^T L_1^{-1} = e_1^T$. A jobb alsó $(n-1)$-edrendű blokkban pedig a következőket kell számítani, $k,i > 1$:

az utolsó lépésben felhasználtuk az előbbi lemmát.

2.10. A mátrix kondiciószáma

Az előbbi becslések azt mutatják, hogy a megoldás relatív megváltozása arányos a $\mathrm{cond}(A) = \|A\|\|A^{-1}\|$ számmal, ezért ezt a számot a mátrix kondiciószámának nevezzük. Szokás még a $\kappa(A)$ jelölés használata is. Ha az egyenletrendszer együtthatómátrixának kondiciószáma nagy, akkor az egyenletrendszert *gyengén meghatározottnak* nevezzük.

2.11. A relatív maradék

A $\|\delta x\|/\|x\|$ szám nem jellemzi a megoldó módszer stabilitását, mert a megoldó módszertől függetlenül nagy lehet, ha $\mathrm{cond}(A)$ nagy. Erre a célra alkalmasabb a maradékvektor. Tegyük fel, az $\tilde{x}$ közelítő megoldást kaptuk, ekkor a maradékvektor: $r = b - A\tilde{x}$, amit szokás még reziduum vektornak is nevezni. A relatív maradékot a következő formulával készítjük:

$$\eta = \frac{\|r\|}{\|A\|\|\tilde{x}\|}. \tag{2.17}$$

A stabilitás inverz megfogalmazása szerint a megoldó módszer stabil, ha a kapott eredmény egy kissé perturbált kiinduló eredményhez tartozik: $(A + \delta A)\tilde{x} = b$, ahol $\|\delta A\|/\|A\|$ kicsi.

Meg lehet mutatni: ha η nagy, $\|\delta A\|/\|A\|$ is nagy. Ugyanis $0 = b - (A + \delta A)\tilde{x} = r - \delta A\tilde{x}$, ahonnan $\|r\| \le \|\delta A\|\|\tilde{x}\|$ és innen következik

$$\eta = \frac{\|r\|}{\|A\|\|\tilde{x}\|} \le \frac{\|\delta A\|}{\|A\|}.$$

Másrészt, ha η kicsi, akkor 2-es normában a mátrix relatív megváltozása is kicsi. Ugyanis δA-ra megoldás:

$$\delta A = \frac{r\tilde{x}^T}{x^T x}, \quad \text{mert} \quad b - \left(A + \frac{r\tilde{x}^T}{x^T x}\right)\tilde{x} = b - A\tilde{x} - r = 0. \tag{2.18}$$

Ekkor 2-es normában $\|r\tilde{x}^T\|_2 = \|r\|_2 \|\tilde{x}^T\|_2$ (l. 6.5 gyakorlat), s ezzel $\dfrac{\|\delta A\|_2}{\|A\|_2} = \dfrac{\|r\|_2}{\|A\|_2\|\tilde{x}\|_2}$.

2.12. Gyakorlatok

2.1. Mutassuk meg: indukált normára $\|I\| = 1$.

2.2. Ha A invertálható, akkor $\|x\|_A = \|Ax\|$ is vektornorma.

2.3. A mátrix kondiciószáma indukált normánál nem lehet kisebb 1-nél.

2.4. 2-es normánál az ortogonális vagy unitér mátrixok kondiciószáma 1.

2.5. $\|ab^T\|_2 = \|a\|_2 \|b\|_2$.

2.6. $U^T U = I$ (ortogonális) $\rightarrow \|AU\|_2 = \|A\|_2$.

2.7. $\|\|A\| - \|B\|\| \le \|A \pm B\|$.

oszlopa. A többi szorzatban lévő e_k, $k<j$ vektorral ennek a skaláris szorzata zérus, emiatt a végeredmény Le_j. Felírható a sorvektorokkal is a szorzatokra bontás:

$$L=\prod_{i=1}^{n}\left(I+e_ie_i^T(L-I)\right).$$

Ellenőrízzük, hogy ennek a j-edik sora visszaadja L j-edik sorát!

A U felső háromszögmátrixra vonatkozó hasonló összefüggések:

$$U=\prod_{i=n\ (-1)}^{1}\left(I+(U-I)e_ie_i^T\right)=\prod_{i=n\ (-1)}^{1}\left(I+e_ie_i^T(U-I)\right),$$

ahol a tényezők balról jobbra az indexek szerint csökkenő sorrendben írandók.

3.12. Vetítőmátrixok

Tekintsük a

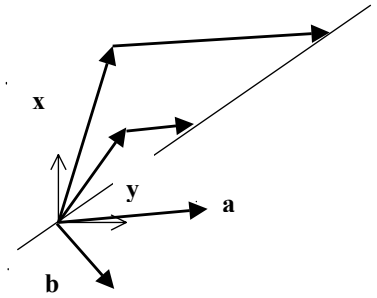
$$P=I-ab^T \tag{3.3}$$

mátrixot, ahol $b^Ta=1$. Ennek determinánsa 0, így az inverze nem létezik. Van azonban egy érdekes tulajdonsága: önmagával szorozva visszaadja saját magát:

$$\left(I-ab^T\right)\left(I-ab^T\right)=I-2ab^T+ab^Tab^T=I-ab^T.$$

Az $P^2=P$ feltételt kielégítő mátrixokat *vetítő-mátrixoknak* vagy *projektoroknak* nevezzük.

Ha $a=b$, akkor a mátrix szimmetrikus. A szimmetrikus vetítőmátrixok *ortogonális* vetítők, mert egy altérre merőleges vetítést valósítanak meg. Ha a és b nem egyirányú, akkor *ferde* vetítésről beszélünk. Szokás még a projektorokat *idempotens* mátrixoknak nevezni arra a tulajdonságukra utalva, hogy a mátrix minden hatványa önmaga. Vegyük észre, (3.3)-ból: $Pa=0$ és $b^TP=0$.



1. Ábra

Az 1. ábra azt szemlélteti, a (3.3) projektor hogy vetíti az x és y vektort az a irány mentén a b normálisú síkba, amely áthalad az origón. Ha a iránya megegyezne b irányával, akkor a síkba vetítés merőlegesen történne.

3.7 Gyakorlat. Ellenőrízzük: ha P projektor, akkor $I-P$ is az.

3.8 Gyakorlat. Egy sík normálvektora s, egyenlete $s^Tx=\sigma$. Legyen a vetítőmátrix $P=I-ss^T/s^Ts$. Mutassuk meg, a tér bármely y vektorára a $Py+\sigma s/s^Ts$ művelet egy síkbeli vektort állít elő.

3.9 Gyakorlat. Mutassuk meg, az előbbi P mátrixszal $Py\perp s$. Adjuk meg a $Py+\sigma s/s^Ts$ vektort és az y vektort összekötő vektort!

3.13. Involutórius mátrixok

Az A mátrixot *involutóriusnak* nevezzük, ha eleget tesz az $A^2=I$ összefüggésnek. Minden projektor $A=I-2P$ alakban meghatároz egy involutórius mátrixot:

3.3. Permutáció-mátrix

Úgy kapjuk, ha az egységmátrix sorait vagy oszlopait permutáljuk, emiatt minden sor és oszlopban csak egy 1-es fordulhat elő, a többi elem 0. Az ábrázolásukhoz nem szükséges a mátrixot kitölteni, elég egy egész (számokból álló) vektor.

Tegyük fel, egy mátrix sorait cserélgetjük és ezt szeretnénk egy vektorban feljegyezni, ami a permutációmátrixot reprezentálja. Kezdetben a vektor k-adik eleme legyen egyenlő k-val. A cserék során ennek a vektornak az elemeit cserélgessük ugyanúgy, mint a mátrix sorait (mintha oszlopvektorként a mátrixhoz csatoltuk volna). Így a végén mindegyik sorról meg tudjuk állapítani, hogy hova került. Ha például az első elem 5-ös, akkor ez azt jelenti, hogy az ötödik sor az elsőbe került.

3.2 Gyakorlat. Tekintsük a $\Pi=[e_2,e_4,e_3,e_1]$ permutáció-mátrixot és ellenőrízzük, hogy az inverze a transzponáltja! Ezt a tényt általánosan bizonyítsuk be! Hogyan tároljuk a fenti mátrixot egy 4-elemű vektorban?

3.4. Diáddal módosított egységmátrix

A numerikus lineáris algebrában különösen fontos szerepet játszanak az olyan egyszerű mátrixok, amelyek az egységmátrixtól csak egy diádban különböznek:

$$F=I+ab^T \tag{3.1}$$

Segítségükkel a különféle lineáris algebrai transzformációk egyszerűen végezhetők, a bennük szereplő a és b vektorok megválasztása mindig az elérendő céltól függ.

Ennek a mátrixnak az inverze könnyen meghatározható. Feltételezve, hogy $F^{-1}=I+\alpha ab^T$, az $FF^{-1}=I$ összefüggésből adódik: $\alpha=-1/(1+b^Ta)$, így

$$F^{-1}=I-\frac{ab^T}{1+b^Ta}. \tag{3.2}$$

Az inverz nem létezik, ha $1+b^Ta=0$, ebből már sejthetjük, hogy a nevező nem más, mint F determinánsa.

3.5. Példa

Ha az egységmátrixból kivesszük az i-edik oszlopot és a helyére betesszük az a vektort:

$$F=I+(a-e_i)e_i^T.$$

Az inverze:

$$F^{-1}=I-\frac{(a-e_i)e_i^T}{1+e_i^T(a-e_i)}=I-\frac{(a-e_i)e_i^T}{e_i^Ta}.$$

Az ilyen típusú mátrixok fontosak a lineáris egyenletrendszer-megoldó algoritmusoknál.

3.3 Gyakorlat. Ellenőrízzük: $F^{-1}a=e_i$.

3.6. Példa

A következő műveletet végezzük: az A mátrix i-edik oszlopát α-val szorozzuk és hozzáadjuk a k-adik oszlopához. Írjuk fel azt a mátrixot, amellyel szorozva A-t, pont ez történik!

Megoldás. $A+\alpha Ae_ie_k^T=A(I+\alpha e_ie_k^T)$.