3.4 Gyakorlat. Az előbb kapott összefüggés segítségével bizonyítsuk be, hogy a mátrix determinánsa nem változik, ha egy oszlopának számszorosát egy másik oszlopához hozzáadjuk. Használjuk fel a szorzatmátrix determinánsára tanultakat!

3.7. Példa

Igazoljuk, hogy az $\left|I+ab^T\right|$ determináns egyenlő $1+b^Ta$ -val!

Megoldás. Feltesszük, az a és b vektorok egyike sem zérus, mert különben a feladat triviális volna. Legyen az a vektor i -edik eleme $e_i^Ta=a_i\neq 0$, és tekintsük az $I-(a/a_i-e_i)e_i^T$ mátrixot. Ennek minden átlóeleme 1 és az i -edik oszlopában vannak még nemzérus elemek. De ezeket a nemzérus elemeket az i -edik sor valamely számszorosának hozzáadásával ki lehet nullázni, ebből következik, hogy a determinánsa 1. Most szorozzuk az $I+ab^T$ mátrixot balról $I-(a/a_i-e_i)e_i^T$ -vel. Ez az a vektort az a_ie_i vektorba viszi, így az eredmény: $I-(a/a_i-e_i)e_i^T+a_ie_ib^T$, amely már csak az i -edik sorában és oszlopában különbözik az egységmátrixtól. Most szorozzuk a kapott mátrix k -adik oszlopát a_k/a_i -vel és adjuk hozzá a i -edik $(i\neq k)$ oszlophoz (ld. 2.6 Példa):

$$\left(I-(\frac{a}{a_i}-e_i)e_i^T+a_ie_ib^T\right)\left(I+\frac{a_k}{a_i}e_ke_i^T\right)=I-\left(\frac{a-a_ke_k}{a_i}-e_i\right)e_i^T+a_ie_ib^T+a_kb_ke_ie_i^T.$$

Mint látjuk, az a vektor k -adik eleme kinullázódott, és az i -edik átlóelem $1+a_ib_i+a_kb_k$ lett. Ezt a műveletet minden $k\neq i$ -re végrehajtva az a/a_i vektor minden átlón kívüli eleme kinullázódik, az i -edik átlóelem $1+b^Ta$, a többi pedig 1-gyel egyenlő. A következő fázisban az e_k^T, $k\neq i$ sorvektorokkal az $a_ie_ib^T$ sorvektor nemdiagonális elemeit a determináns megváltozása nélkül kinullázhatjuk.

3.8. Diádösszegek, kifejtések

Az n -edrendű egységmátrix felírható diádösszegként: $I_n=\sum_{i=1}^n e_ie_i^T$. Ha ezt beírjuk két mátrix közé, akkor a szorzatmátrix diádösszeg-előállítását kapjuk:

$$AB=\sum_{i=1}^n Ae_ie_i^TB,$$

A oszlopai és B sorai képezik a diádokat, i -edik oszlop és i -edik sor.

3.5 Gyakorlat: Írjuk ki ADB diádösszeg előállítását, ahol $D=[d_i\delta_{ij}]$ diagonálmátrix, (csak a főátló elemei nemzérusok).

Tudjuk, az n -edrendű x vektor előállítása az egységvektorok segítségével $x=\sum_{i=1}^n e_i(e_i^Tx)$. Az előállítás hasonló a $\{q_i\}_{i=1}^n$ ortonormált vektorrendszerrel. Ugyanis vezessük be a $Q=[q_1q_2\ldots q_n]$ mátrixot. Ekkor $Q^TQ=I=QQ^T$ az ortonormáltság miatt, tehát írható $x=QQ^Tx=\sum_{i=1}^n q_i(q_i^Tx)$. Az ilyen tulajdonságú Q mátrixokat *ortogonális* (komplex megfelelője: *unitér*) mátrixoknak nevezzük.

3.9. Definíció

Az $\{a_i\}_{i=1}^n$ és $\{b_i\}_{i=1}^n$ rendszerek *biortogonális vektorrendszert* alkotnak, ha $a_i^Tb_j=\alpha_i\delta_{ij}$, $\alpha_i\neq 0$ teljesül bármely indexre. Ha n a vektorok dimenziója, akkor a rendszer *teljes.*

3.6 Gyakorlat. Az előbbi vektorokat gyűjtsük az $A=[a_1,a_2,\ldots,a_n]$ és $B=[b_1,b_2,\ldots,b_n]$ mátrixba. Ellenőrízzük, hogy A^TB diagonálmátrix! Ekkor az x vektor hogyan állítható elő az a_i vektorok lineáris kombinációjaként? És hogyan fejthető ki a b_i vektorok segítségével?

3.10. Tétel, mátrix egyszerű szorzatokra bontása

Minden nemszinguláris $A\in\mathbb{R}^{n\times n}$ mátrix felírható n egyszerű mátrix szorzataként, ahol egy tényező egy permutációból és egy $I+(a_i-e_i)e_i^T$ típusú tagból áll. A permutációra nincs mindig szükség.

Bizonyítás. Megadjuk az eljárást. Az első lépésben vizsgáljuk meg az A mátrix első oszlopát. Ha az első elem $a_{11}=e_1^TAe_1\neq 0$, akkor sorcserére nincs szükség. Ha az első elem zérus, akkor az oszlopban keresünk egy nemzérus elemet, majd ennek a sorát felcseréljük az első sorral. Ha az oszlop minden eleme zérus volna, akkor nem lenne invertálható a mátrix. Az első permutáció mátrixot jelüljük Π_1 -gyel és legyen $A_1=\Pi_1A$.

Most szorozzuk A_1 -et a $T_1=I-(A_1e_1-e_1)e_1^T/e_1^TA_1e_1$ mátrixszal. Tudjuk, ennek eredményeként az első oszlop e_1 -be megy át és $T_1^{-1}=I+(A_1e_1-e_1)e_1^T$. A második lépésben hasonlóan járunk el T_1A_1 második oszlopával: $A_2=\Pi_2T_1A_1$ olyan mátrix lesz, ahol a 22-es pozición nemzérus elem van. Így a $T_2=I-(A_2e_2-e_2)e_2^T/e_2^TA_2e_2$ mátrixszal szorozva a második oszlopot az e_2 vektorba visszük. Vegyük észre, T_2 az e_1 vektort helyben hagyja.

Hasonlóan folytatva, az i -edik lépésben $A_i=\Pi_iT_{i-1}A_{i-1}$ olyan mátrix, ahol az ii pozicióban nemzérus áll. (Ha az i -edik oszlop zérus volna, ismét ellentmondásba kerülnénk azzal a feltevéssel, hogy a mátrix nemszinguláris.) Ekkor a $T_i=I-(A_ie_i-e_i)e_i^T/e_i^TA_ie_i$ mátrixszal szorozva kapunk e_i -t az i -edik oszlopban és az eddig elkészült kisebb indexű egységvektorok sem romlottak el. A n -edik lépés után egységmátrixot kapunk, tehát végeredményben a mátrix inverzével szoroztunk. A szorzatokat összegyűjtve:

$$\Pi_1^TT_1^{-1}\Pi_2^TT_2^{-1}\ldots T_n^{-1}=A.$$

Figyeljük meg, T_i^{-1} megadásához elég, ha az i indexet és a benne szereplő $a_i=A_ie_i$ vektort ismerjük.

3.11. Háromszögmátrixok szorzatokra bontása

Az L mátrixot alsó háromszögmátrixnak nevezzük, ha a főátló feletti elemei mind zérust tartalmaznak. Hasonlóan az U mátrix felső háromszög mátrix, ha a főátló alatti elemek zérusok. A háromszögmátrixok szorzatokra bontása különösen egyszerű. Az előbbi tételt alkalmazva azonnal adódik az n -edrendű L alsó háromszögmátrix szorzat-előállítása:

$$L=\left(I+(L-I)e_1e_1^T\right)\left(I+(L-I)e_2e_2^T\right)\ldots\left(I+(L-I)e_ne_n^T\right),$$

ami tömören így is írható

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

ha megjegyezzük, hogy a tényezők növekvő indexek szerint mindig balról jobbra haladva írandók. A kifejezést közvetlenül is igazolhatjuk a j -edik oszlop meghatározásával. Jobbról az e_j vektorral szorozva az első e_j vektortól különböző eredményű szorzat $e_j+Le_j-e_j=Le_j$ az L mátrix j -edik

2.8. $A=\begin{bmatrix} 2 & -3 & 1 \\ -4 & -2 & 1 \end{bmatrix}$, $\|A\|_1=?$ $\|A\|_\infty=?$ $\|A\|_2=?$

2.9. $\|A\|_2 \le \sqrt{\|A\|_1 \|A\|_\infty}$.

2.10. Frobenius-norma: $\|A\|_F=\left(\sum_{i,j}|a_{ij}|^2\right)^{1/2}=\sqrt{\mathrm{tr}(A^TA)}=\sqrt{\mathrm{tr}(AA^T)}$. Igazoljuk, ez is mátrixnorma, de nem indukált norma, $\|I\|_F=?$ $\|Ax\|_2 \le \|A\|_F \|x\|_2$. (A 2-es normával illeszkedő mátrixnorma.)

2.11. $A=A^T$, akkor $\|A\|_2=\rho(A)$ = spektrál sugár, azaz szimmetrikus mátrixokra a spektrálnorma a minimális norma. ($\|\,.\,\|_2$ = spektrál norma).

2.12. $U^TU=I$ (ortogonális) $\rightarrow \|AU\|_F=\|A\|_F$.

3. A numerikus lineáris algebra egyszerű transzformációi

3.1. Jelölések

A mátrixokat latin nagybetűkkel: $A,B,C,\ldots$ a vektorokat latin kisbetűkkel: $a,b,c,\ldots$ jelöljük, kivéve az i,j,k,l,m,n betűket, amelyeket indexekben fogunk használni. A skalárokra görög kisbetűket alkalmazunk. Ha az A mátrixot az $a_1,a_2,\ldots$ oszlopvektorokból állítjuk össze, akkor ezt így jelöljük: $A=[a_1a_2\ldots a_n]$. A mátrix egy másik megadási formája $A=[a_{ij}]$, ekkor az ij -edik elemet adjuk meg általánosan. Az n -edrendű egységmátrix $I_n=[e_1\,e_2\ldots e_n]$, amely az $e_1,e_2,\ldots,e_n$ Descartes-egységvektorokat tartalmazza az oszlopaiban. A transzponált jelölése: a^T , komplex esetben a transzponált konjugált jelölése a^H .

3.2. A mátrixok szorzása

Az $A=[a_{ij}]\in\mathbb{R}^{m\times n}, B=[b_{jk}]\in\mathbb{R}^{n\times l}$ mátrixok összeszorzásának eredménye a $C=AB=[c_{ik}]=$ $=\left[\sum_{j=1}^{n}a_{ij}b_{jk}\right]\in\mathbb{R}^{m\times l}$ mátrix. A vektorok egy sorból vagy oszlopból álló speciális mátrixoknak tekinthetők, szorzásuk nem jelent újat. Az alkalmazásokban megkülönböztetjük a vektorok kétféle szorzási módját. Az egyik a *skaláris* szorzat, például a^Tb , amelynek eredménye egy skalár. A másik a *diadikus* szorzat, például ab^T , az eredmény egy 1-rangú mátrix. Vegyük észre, az első esetben szükséges, hogy a vektorok hossza azonos legyen, a második esetben nem.

3.1 Gyakorlat. Írjunk fel egy diádot. Indokoljuk meg, hogy a rangja tényleg 1. Hogyan egyszerűbb egy vektort diáddal szorozni? *a)* Képezzük $A=ab^T$ -t, majd Ax -et. *b)* Először kiszámítjuk b^Tx -et és ezzel a skalárral szorozzuk az a vektort.

A továbbiakban rátérünk speciális mátrixok ismertetésére.

$$(I-2P)(I-2P)=I-4P+4P=I\,,$$

és minden involutórius mátrix $(I-A)/2$ alakban meghatároz egy projektort:

$$(I-A)(I-A)/4=(2I-2A)/4=(I-A)/2.$$

Innen látható, az egységmátrixból végtelen sokféleképp lehet gyököt vonni.

3.10 Gyakorlat. Igazoljuk, hogy a $J=[e_ne_{n-1}\ldots e_1]$ mátrix, ahol az egységmátrix oszlopai fordított sorrendben vannak felsorolva, involutórius mátrix. Milyen projektort határoz meg ez a mátrix, ha $n=2,3$?

Az ab^T/b^Ta, $b^Ta\neq 0$ projektorral a következő involutórius mátrixot készíthetjük: $I-2ab^T/b^Ta$. Az 1.ábrából megállapíthatjuk, hogy ez a mátrix a b normálisú síkra való „ferde” tükrözést végzi, ami annyit jelent, hogy az a irány mentén eljutunk a síkig, majd azt keresztezve ugyanakkora távolságot teszünk meg a túloldalon. A tükrözés akkor merőleges a síkra, ha $a=b$.

3.11 Gyakorlat. Mutassuk meg, hogy az $I-2(x-y)(x-y)^T/(x-y)^T(x-y)$ mátrix az x és y vektorokat egymásba tükrözi, ha azok különbözőek és $x^Tx=y^Ty$.

3.12 Gyakorlat. Az előbbi tükröző mátrixszal lehetőségünk van arra, hogy az x vektort az $y=\pm\sigma e_1$ vektorba tükrözzük, ahol $\sigma^2=x^Tx$. Hogyan válasszuk meg σ előjelét ahhoz, hogy a kivonási jegyveszteséget biztosan elkerüljük?

3.14. Blokk mátrixok

A mátrixokat nemcsak skalár elemekből rakhatjuk össze, hanem kisebb méretű mátrixokból is. Az ilyen mátrix elemeit *blokkoknak* nevezzük, ha pedig egy mátrixot kisebb mátrixokra bontunk, akkor a mátrixot *blokkosítjuk*. A blokkosítás történhet a következőképp: Az egységmátrixot az oszlopok mentén felszeleteljük k részre: $I=[E_1,E_2,\ldots,E_k]$. Ha a sorokat ugyanilyen módon osztjuk fel blokkokra, akkor az ij -edik blokk $A_{ij}=E_i^TAE_j$ és a mátrix:

$$A=\begin{bmatrix} A_{11} & A_{12} & \ldots & A_{1k} \\ A_{21} & A_{22} & \ldots & A_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1} & A_{k2} & \ldots & A_{kk} \end{bmatrix}.$$

3.13 Gyakorlat. Legyen $F=I+UV^T$, U és V $n\times l$ -es mátrixok, azaz $l<n$ oszlopból állnak. Ha a kijelölt inverz létezik, ellenőrízzük: $F^{-1}=I-U(I_l+V^TU)^{-1}V^T$, ahol I_l $l\times l$ -es egységmátrix.

4. Mátrixok LU-felbontása, Gauss-Jordan algoritmus

Az *LU*-felbontás nem más, mint a Gauss-elimináció olyan átrendezése, ahol a részleteredményeket is elrakjuk. Ez úgy történik, hogy az A mátrixot felbontjuk egy L alsó és egy U felső háromszögmátrix szorzatára.

$$\rho(A) = \max_k |\lambda_k(A)|, \tag{2.12}$$

ahol $\lambda_k(A)$ az A mátrix sajátértéke. Az $\|A\|$ mátrixnorma és az $\|x\|$ vektornorma *illeszkedő,* ha eleget tesznek a (2.8) összefüggésnek. Ez utóbbi definició arra az esetre szól, amikor az alkalmazott mátrixnorma nem az $\|x\|$ vektornormából való indukálással készült. Igaz az összefüggés:

$$\rho(A) \le \|A\|, \tag{2.13}$$

ha $\|A\|$ indukált vagy illeszkedő norma, mert $Au_k = \lambda_k u_k$, $\|u_k\|=1$ mellett

$$\|A\| = \max_{\|x\|=1} \|Ax\| \ge \|Au_k\| = |\lambda_k|\|u_k\| = |\lambda_k|, \quad \forall k\text{-ra.}$$

2.9. A lineáris egyenletrendszer megoldásának perturbációi

Két esetet fogunk megvizsgálni. Az egyik, amikor az egyenletrendszer b jobboldalát perturbáljuk egy kis δb vektorral, a másik, amikor az együtthatómátrix perturbációját vizsgáljuk.

Az első esetben $A(x+\delta x) = b + \delta b$-ből következik $A\delta x = \delta b$ és illeszkedő normák esetén kapjuk a becslést:

$$\frac{1}{\|A\|\|A^{-1}\|}\frac{\|\delta b\|}{\|b\|} \le \frac{\|\delta x\|}{\|x\|} \le \|A\|\|A^{-1}\|\frac{\|\delta b\|}{\|b\|}. \tag{2.14}$$

Az eredeti és a perturbált értékekre vonatkozó egyenletekből

$$\begin{array}{ll} b = Ax, & \delta x = A^{-1}\delta b, \\ \downarrow & \downarrow \\ \|b\| \le \|A\|\|x\|, & \|\delta x\| \le \|A^{-1}\|\|\delta b\|. \end{array}$$

A kapott egyenlőtlenségek azonos oldalait összeszorozva kapjuk (2.14) jobboldali összefüggését. A baloldalit ugyanígy kapjuk, csak a mátrixot a másik oldalra rendezzük az induló egyenletekben:

$$\begin{array}{ll} x = A^{-1}b, & \delta b = A\delta x, \\ \downarrow & \downarrow \\ \|x\| \le \|A^{-1}\|\|b\|, & \|\delta b\| \le \|A\|\|\delta x\|. \end{array}$$

Lemma. *Ha $\|B\| < 1$, akkor $I + B$ invertálható és indukált normára érvényes*

$$\left\|(I+B)^{-1}\right\| \le \frac{1}{1-\|B\|}. \tag{2.15}$$

Az előző szakaszban látott norma és spektrál sugár összefüggése szerint most B spektrál sugara kisebb 1-nél, így minden sajátértéke is kisebb, azaz nem lehet $I + B$ egyik sajátértéke sem 0.

$$(I+B)^{-1} = (I+B-B)(I+B)^{-1} = I - B(I+B)^{-1},$$

ahonnan $\left\|(I+B)^{-1}\right\| \le 1 + \|B\|\left\|(I+B)^{-1}\right\|$, és innen átrendezéssel kapjuk az állítást. ■

Ha az együtthatómátrixot perturbáljuk δA-val: $(A+\delta A)(x+\delta x) = b \;\rightarrow (A+\delta A)\delta x = -\delta A x \;\rightarrow$
$\rightarrow \; \delta x = -(I + A^{-1}\delta A)^{-1}A^{-1}\delta A x$, innen kapjuk a becslést:

$$0 \le \frac{\|\delta x\|}{\|x\|} \le \left\|(I + A^{-1}\delta A)^{-1}\right\|\|A^{-1}\|\|A\|\frac{\|\delta A\|}{\|A\|} \le \|A\|\|A^{-1}\|\frac{\|\delta A\|}{\|A\|}\frac{1}{1-\|A^{-1}\delta A\|}, \tag{2.16}$$

$$e_i^T\left(I - \left(\frac{Ae_1}{a_{11}} - e_1\right)e_1^T\right)Ae_k = a_{ik} - \frac{a_{i1}a_{1k}}{a_{11}} = a_{ik} - \left(\frac{a_{i1}}{a_{11}}\right)a_{1k}.$$

Ez mutatja, hogy az $A - \frac{Ae_1}{e_1^T Ae_1}e_1^T A$ diádot kell számítanunk a jobb alsó $(n-1)$-edrendű blokkra. Az ebben szereplő oszlopvektor éppen L_1 első oszlopa, így célszerűen a következőképpen járhatunk el: kijelöljük a főelemet, vele leosztjuk az alatta lévő oszlopelemeket, a saját sorát pedig változatlanul átmásoljuk. A mátrix többi részében ebből a sorból és oszlopból készített diádot vonjuk le:

$$\begin{bmatrix} 2 & 0 & 3 & -1 \\ -4 & 5 & -2 & 3 \\ 6 & -5 & 4 & -3 \end{bmatrix} \rightarrow \begin{bmatrix} \boxed{2} & 0 & 3 & -1 \\ -2 & 5 & 4 & 1 \\ 3 & -5 & 5 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 0 & 3 & -1 \\ -2 & \boxed{5} & 4 & 1 \\ 3 & -1 & -1 & 1 \end{bmatrix}$$

$$L = \begin{bmatrix} 1 & & \\ -2 & 1 & \\ 3 & -1 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 2 & 0 & 3 \\ & 5 & 4 \\ & & -1 \end{bmatrix}.$$

A végén még megoldandó $Ux = [-1 \quad 1 \quad 1]^T$, ezt alulról felfelé megoldva $x = [1 \quad 1 \quad -1]^T$.

4.1 Gyakorlat. Oldjuk meg *LU*-felbontással a következő egyenletrendszert:

$$\begin{bmatrix} 2 & 2 & 3 \\ 4 & 3 & 7 \\ 6 & 7 & 5 \end{bmatrix} x = \begin{bmatrix} 1 \\ 5 \\ -3 \end{bmatrix}.$$

4.2. Az LU-felbontás műveletigénye.

Az első lépésben az oszlopvektor leosztása $n-1$ osztás, a diád levonása $(n-1)^2$ szorzást és összeadást igényel. Az aritmetikai műveletek mindegyike ugyanannyi idejűnek számít, emiatt az első lépés műveletigénye: $(n-1)(2n-1)$ flop (= *fl*oating point *op*eration, magyarul: lebegőpontos művelet). A következő lépés igénye $(n-2)(2n-3)$ flop, így a teljes műveletigény $\sum_{k=1}^{n-1}(k-1)(2k-1)$ flop. Ezt úgy közelítjük, hogy a legmagasabb fokú tagot integráljuk 0-tól n-ig: $2n^3/3$. A korrekciós tagok n kisebb hatványai, nem határozzuk meg őket, mert a legmagasabbfokú tag a domináns.

4.2 Gyakorlat. Mennyi Ax, LUx, $U^{-1}L^{-1}x$ műveletigénye? Az utolsó példánál alkalmazzuk a 2.11 szakaszban megismert faktorizációs összefüggéseket!

4.3. Blokk LU-felbontás

Néha célszerű a felbontást – vagy a mátrix invertálását – blokkosított formában végezni. Tipikusan ez a helyzet akkor, amikor az egyik elkülönített blokk egyszerűen invertálható, például azért mert egységmátrix, vagy alsó ill. felső háromszögmátrix. Mi most a blokk LU-felbontást a 2×2-es esetben fogjuk megnézni. A főelem ilyenkor blokk, amelyről fel kell tételeznünk, hogy létezik az inverze. Legyen az egységmátrix egy felosztása $I = [E_1, E_2]$, $A_{ij} = E_i^T A E_j$, ekkor az L mátrix a (4.1)-ben látható L_1 mátrix blokkos megfelelője (ld. még *3.13 Gyakorlat*)

$$L = I - \left(AE_1A_{11}^{-1} - E_1\right)E_1^T \tag{4.4}$$

és a mátrix blokkos felbontása a következő:

$$\|x\|_\infty = \max_j |x_j| \cdot \lim_{p\to\infty} \left\{ \sum_{i=1}^{n} \left| \frac{x_i}{\max_j |x_j|} \right|^p \right\}^{1/p} = \max_j |x_j|$$

a Csebisev-, ∞ -, vagy kocka-norma. Mint látjuk, (2.4) alapján itt a legnagyobb és legkisebb hatvány-normák szerepelnek, továbbá az ortogonális transzformációkkal szemben invariáns 2-es norma. Ezekre a normákra a definíciók alapján levezethetők a következő egyenlőtlenségek:

$$\begin{aligned} &\|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty, \\ &\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\,\|x\|_\infty, \\ &\frac{1}{\sqrt{n}}\|x\|_1 \le \|x\|_2 \le \|x\|_1. \end{aligned} \tag{2.5}$$

2.4. Konvergencia normában. A normák ekvivalenciája

A norma alkalmas arra, hogy segítségével egy vektorsorozat konvergenciáját értelmezzük. Ezek alapján $x^{(k)} \to x$ alatt azt értjük, hogy $\exists x \in \mathbb{R}^n$, $\lim_{k\to\infty} \|x^{(k)} - x\| = 0$.

Az $\|x\|_{(1)}$ és $\|x\|_{(2)}$ normákat *ekvivalensnek* nevezzük, ha $\exists c_1, c_2 > 0$ úgy, hogy

$$c_1\|x\|_{(1)} \le \|x\|_{(2)} \le c_2\|x\|_{(1)}.$$

6.5.1 Tétel (bizonyítás nélkül): Végesdimenziós vektortérben bármely két norma ekvivalens. Ez azt jelenti, hogy a normák akármennyire nem különbözhetnek egymástól. Így mindegy, milyen normában vizsgáljuk a konvergenciát.

2.5. Mátrixnormák

A mátrix normája $\|A\|$: $\mathbb{R}^{m\times n} \to \mathbb{R}$ a következő tulajdonságokkal rendelkezik:

$$\begin{aligned} &i) && \|A\| = 0 \;\Leftrightarrow\; A = 0, \\ &ii) && \|\lambda A\| = |\lambda|\|A\|, \\ &iii) && \|A + B\| \le \|A\| + \|B\|, \\ &iv) && \|AB\| \le \|A\|\|B\|. \end{aligned} \tag{2.6}$$

Mivel a vektorok speciális mátrixoknak tekinthetők, minden mátrixnorma meghatároz egy vektor-normát, amelyet a mátrixnormával *kompatibilis* vektornormának nevezünk. Ez az út fordítva is bejárható, ugyanis minden vektornorma *indukál egy mátrixnormát* a következőképpen:

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|, \tag{2.7}$$

ahol $\|.\|$ vektornorma. Csak megjegyezzük, az általánosabb definícióban megengedhető, hogy más normák szerepeljenek a számlálóban és a nevezőben. A (2.7) definíció egyenes következménye

$$\|Ax\| \le \|A\|\|x\|. \tag{2.8}$$

2.6. Tétel

Az indukált mátrixnorma eleget tesz a (2.6) feltételeknek.

Bizonyítás. Ad 1. $A = 0 \to \|A\| = 0$. $\|A\| = 0 \to Ax = 0 \;\; \forall x - \text{re} \to A = 0$.

alkalmaznunk. Az harmadik tag azt mutatja, hogy az i -edik sort a főelemmel kell osztani, az első két tagból származó i -edik sor ugyanis zérus.

Az elmondottakat egy példán szemléltetjük. Invertálandó a $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 6 \end{bmatrix}$ mátrix. A kibővített mátrixban az első lépés egy sorcsere:

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 2 & 3 & 0 & 1 & 0 \\ 1 & 3 & 6 & 0 & 0 & 1 \end{bmatrix} \underset{\text{sorcsere}}{1\leftrightarrow 2} \begin{bmatrix} \boxed{1} & 2 & 3 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 3 & 6 & 0 & 0 & 1 \end{bmatrix} \overset{Tr1}{\to}$$

Az első transzformációs lépésben az első oszlop átmegy e_1 -be, az első sort végigosztjuk a főelemmel, a többi helyen pedig végrehajtjuk az első diád levonását:

$$\to \begin{bmatrix} 1 & 2 & 3 & 0 & 1 & 0 \\ 0 & \boxed{1} & 1 & 1 & 0 & 0 \\ 0 & 1 & 3 & 0 & -1 & 1 \end{bmatrix} \overset{Tr2}{\to} \begin{bmatrix} 1 & 0 & 1 & -2 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & \boxed{2} & -1 & -1 & 1 \end{bmatrix} \overset{Tr3}{\to} \begin{bmatrix} 1 & 0 & 0 & -3/2 & 3/2 & -1/2 \\ 0 & 1 & 0 & 3/2 & 1/2 & -1/2 \\ 0 & 0 & 1 & -1/2 & -1/2 & 1/2 \end{bmatrix}.$$

Az utolsó lépésben az induló egységmátrix helyén megjelent az inverz.

A „helyben” invertáláshoz azt kell észrevennünk, hogy minden lépésben összegyűjthető egy egységmátrix a kibővített mátrixból. Ezt szükségtelen tárolni. A jobboldali 3×3-as területen minden lépésben egy új vektor jelenik meg, a bal oldali 3×3-as területen pedig a távozó vektor helyére egy egységvektor lép be. A „tömör” algoritmusban a jobb oldalon belépő új vektort beírjuk a bal oldalon belépő egységvektor helyére. Az i -edik egységvektor helyén a jobb oldalról származó új vektor

$$\left(I - \frac{A_i e_i - e_i}{e_i^T A_i e_i} e_i^T \right) e_i = e_i - \frac{A_i e_i - e_i}{e_i^T A_i e_i} = \begin{cases} 1/e_i^T A_i e_i, & j = i, \\ -a_{ji}^{(i)} / a_{ii}^{(i)}, & j \neq i \end{cases}$$

Ez fog átkerülni a bal oldalon az i -edik oszlopba. Így a tömör algoritmusban a főelem helyére annak reciproka kerül, az oszlop többi eleme pedig negatív előjelet kap és osztódik a főelemmel. A levonandó diád kezelése ugyanaz, mint korábban. A bekeretezett elem jelöli ki azt a diádot (sor, oszlop), amelyből a levonandó diádot képezzük. Tehát a tömör algoritmus:

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 6 \end{bmatrix} \underset{\text{sorcsere}}{1\leftrightarrow 2} \begin{bmatrix} \boxed{1} & 2 & 3 \\ 0 & 1 & 1 \\ 1 & 3 & 6 \end{bmatrix} \overset{Tr1}{\to} \begin{bmatrix} 1 & 2 & 3 \\ 0 & \boxed{1} & 1 \\ -1 & 1 & 3 \end{bmatrix} \overset{Tr2}{\to} \begin{bmatrix} 1 & -2 & 1 \\ 0 & 1 & 1 \\ -1 & -1 & \boxed{2} \end{bmatrix} \overset{Tr3}{\to}$$

$$\to \begin{bmatrix} 3/2 & -3/2 & -1/2 \\ 1/2 & 3/2 & -1/2 \\ -1/2 & -1/2 & 1/2 \end{bmatrix} \underset{\text{oszlopcsere}}{1\leftrightarrow 2} \begin{bmatrix} -3/2 & 3/2 & -1/2 \\ 3/2 & 1/2 & -1/2 \\ -1/2 & -1/2 & 1/2 \end{bmatrix}.$$

A kezdeti sorcsere miatt nem az eredeti, hanem a ΠA mátrixot invertáltuk, ahol Π permutáció-mátrix. Ennek az inverze $A^{-1}\Pi^T$, mert $\Pi^{-1} = \Pi^T$. Így a kapott eredményt még szoroznunk kellett jobbról Π^T -vel, ami itt az $1\leftrightarrow 2$ oszlopcserét jelenti.

4.4 Gyakorlat. Mi a Gauss-Jordan invertáló módszer műveletigényében a domináns tag?

2. Normák, egyenlőtlenségek

Ebben a szakaszban vektorok és mátrixok között távolságfüggvényeket fogunk bevezetni.

1.1. Metrikus tér

Legyen $\mathcal{X}$ egy halmaz, amelynek elemei közt bevezetünk egy távolságfüggvényt $\delta:(\mathcal{X}\times\mathcal{X})\to\mathbb{R}$. Azt kívánjuk, $a,b\in\mathcal{X}$ -re rendelkezzen a következő tulajdonságokkal:

i) $\delta(a,b)=\delta(b,a)$, azaz a legyen olyan távolságra b-től, mint b a-tól (szimmetria).

ii) $\delta(a,b)=0 \Leftrightarrow a=b$, a távolság csak akkor legyen zérus, ha a két elem azonos.

iii) $\delta(a,c)\le\delta(a,b)+\delta(b,c)$, a háromszög-egyenlőtlenség. Azt fejezi ki, hogy két pont között legrövidebb út az egyenes.

Ekkor a $(\delta,\mathcal{X})$ párt *metrikus térnek* nevezzük. A következőkben $\mathcal{X}$ gyanánt az $\mathbb{R}^n$ és $\mathbb{R}^{m\times n}$ halmazok kerülnek szóba, azaz vektorok és mátrixok között fogunk távolságfüggvényeket készíteni. Ez a δ nem lehet negatív, mert *iii)*-ból $0=\delta(a,a)\le\delta(a,b)+\delta(b,a)=2\delta(a,b)$ következmény.

2.1. A vektorok hatványnormája

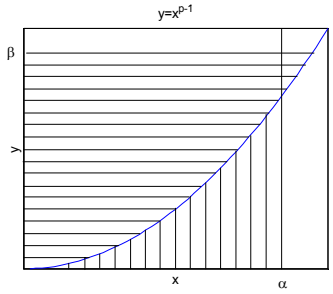
A vektor normája $\|x\|:\ \mathbb{R}^n\to\mathbb{R}$ a következő tulajdonságokkal rendelkezik:

$$\begin{aligned} &i) && \|x\|=0 \ \Leftrightarrow\ x=0,\\ &ii) && \|\lambda x\|=|\lambda|\|x\|,\\ &iii) && \|x+y\|\le\|x\|+\|y\|. \end{aligned} \tag{2.1}$$

Ekkor a $\delta(x,y)=\|x-y\|$ választás metrikát ad, mert a kívánt tulajdonságok teljesülnek. Az első két feltételt triviálisan kielégíti a hatványnorma:

$$\|x\|_p=\left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p},\qquad 1\le p\le\infty, \tag{2.2}$$

a harmadikat később fogjuk belátni.



2.2. A Hölder-egyenlőtlenség

A hatványnormákra fennáll a Hölder-egyenlőtlenség:

$$|y^Tx|\le\sum_{i=1}^{n}|x_i||y_i|\le\|x\|_p\|y\|_q,\qquad \frac{1}{p}+\frac{1}{q}=1, \tag{2.3}$$

ami $p=q=2$ -re a jólismert Cauchy-Bunyakovszkij egyenlőtlenségbe megy át. A p és q közötti összefüggés átrendezhető a $p-1=1/(q-1)$ alakba, amit szem előtt tartva könnyen belátható az alábbi egyenlőtlenség. Az alkalmazott függvény $y=x^{p-1}$, az első integrál a függőlegesen, a második a vízszintesen satírozott területet jelenti:

5.1.3 Tétel, pozitív szemidefinit mátrix felbonthatósága.

Ha A pozitív szemidefinit, akkor $A=LL^T$ alakban előállítható.

Bizonyítás. Láttuk, A főátlójában csak nemnegatív elemek lehetnek. Ha $a_{11}>0$, akkor készítsük el a következő

$$A_2=A-\frac{Ae_1e_1^TA}{e_1^TAe_1}, \tag{5.1}$$

mátrixot, amelyről tudjuk, hogy az első sora és oszlopa zérus. Válasszuk L első oszlopának $Le_1=Ae_1/\sqrt{a_{11}}$ -et, ezzel $A_2=A-Le_1e_1^TL$.

Ha az első diagonálelem zérus, akkor ugyanazon sor és oszlop cseréjével mozgassunk egy nemzérus diagonálelemet az 1,1 pozicióba és ugyanígy járjunk el.

Folytassuk az eljárárást a megmaradó 1-gyel kisebb méretű jobb alsó blokkal mindaddig, ameddig találunk pozitív diagonálelemet. Minden lépésben az L mátrix egy újabb oszlopát nyerjük. Ha olyan helyzethez értünk, ahol a megmaradt jobb alsó blokkban minden diagonálelem zérus, akkor a teljes blokknak zérusnak kell lennie, mert ha nem így volna, akkor a megmaradó blokk indefnit volna az 5.1.1 Tétel előtt tett megjegyzés szerint és ez ellentmondana annak, hogy a szemidefinitség megmarad.

Vegyük észre, az alkalmazott sor-oszlop cserék a felbontást csak annyiban befolyásolják, hogy $P^TAP=LL^T$ -et kellett volna írnunk, - P permutáció mátrix -, de ez átrendezhető az $A=\tilde{L}\tilde{L}^T$ alakba, ahol $\tilde{L}=PL$. ■

Szimmetrikus, pozitív definit mátrixra az $A=LL^T$ felbontást *Cholesky-felbontásnak* nevezzük. Itt most L főátlójában nem 1-esek állnak, mert például $Le_1=Ae_1/\sqrt{a_{11}}$ első eleme $\sqrt{a_{11}}$. A Cholesky-felbontás hasonlóképp készíthető, mint az LU-felbontás, csak most a főelemből gyököt kell vonni, és azzal végig kell osztani a saját sort és oszlopot. A számítógépes algoritmusban kihasználható, hogy a felső háromszög részt nem kell számítani, ezzel a műveletigény nagyjából megfeleződik.

5.1.4 Példa Choleski-felbontásra

$$\begin{bmatrix}4&-2&2\\-2&10&-7\\2&-7&21\end{bmatrix}\to\begin{bmatrix}\boxed{2}&-1&1\\-1&9&-6\\1&-6&20\end{bmatrix}\to\begin{bmatrix}2&&\\-1&\boxed{3}&-2\\1&-2&16\end{bmatrix}\to\begin{bmatrix}2&&\\-1&3&\\1&-2&\boxed{4}\end{bmatrix},$$

$$L=\begin{bmatrix}2&&\\-1&3&\\1&-2&4\end{bmatrix},\quad L^T=\begin{bmatrix}2&-1&1\\&3&-2\\&&4\end{bmatrix}.$$

Látható, a diádok levonása ugyanolyan módon történik, mint az LU-felbontásban.

5.1.5 Feladatok.

- Legyen $A=LL^T$ egy Cholesky felbontás. Mennyi a műveletigénye x^TAx számításának, ha az eredeti mátrixot használjuk? Hogyan csökkenthető a műveletigény, ha az x^TLL^Tx alakot használjuk?
- A gyökvonást elkerülhetjük, ha az $A=LDL^T$ alakot használjuk, ahol L egységátlójú mátrix és D diagonálmátrix. Dolgozzuk ki ennek a felbontásnak a lépéseit! Ezt a módszert indefinit esetben is alkalmazhatjuk, ha nem adódik túlságosan kicsiny elem D-be.

csökkentésére, pl. ha $\sqrt{3472}-\sqrt{3471}$-et így számítjuk, kihasználva, hogy a gyök alatt egész számok vannak:

$$\frac{(\sqrt{3472}-\sqrt{3471})(\sqrt{3472}+\sqrt{3471})}{\sqrt{3472}+\sqrt{3471}}=\frac{1}{\sqrt{3472}+\sqrt{3471}}.$$

A *másodfokú egyenlet gyökeit* pedig az alábbi módon célszerű számítani:

$$x^2-2px+q=0 \text{ gyökei: } x_1=p+\operatorname{sign}(p)\sqrt{p^2-q},\quad x_2=q/x_1.$$

- Előfordulhat olyan eset, amikor a közbülső eredmény túlcsordul (nagyobb mint M_∞), emiatt rossz a program futása, pedig a végeredmény az ábrázolható számok közt van. Például legyen $a=0.3265+60$, $b=0.5671+02$ és számítandó $\sqrt{a^2+b^2}$. Az első szám kitevője négyzetre emeléskor 120, így túlcsordult számot kapunk. Ha viszont $s\sqrt{(a/s)^2+(b/s)^2}$-et számítjuk, ahol $s=\max(|a|,|b|)$, akkor ez nem következik be.
- Néha arra is számítani kell, hogy egy függvény nem adja olyan pontossággal vissza a helyettesítési értéket, mint amilyen pontossággal indultunk. Például tekintsük a sin függvényt. Ha az argumentum kicsi, akkor nincs semmi baj. Ha azonban x értéke nagy, például $x=2356$, akkor $\sin(2356)$ számításakor 2356 π-vel vett osztási maradékát kell vennünk. A maradékban már csak 1 jegy lesz pontos ha a fenti aritmetikát használjuk, így az eredménynél sem remélhetünk nagyobb pontosságot.

A mutatott példák alapján megállapíthatjuk, hogy a gépi aritmetika nemkívánatos jelenségei elsősorban akkor következnek be, ha a számok között túl nagy a nagyságrendi különbség, vagy egymáshoz nagyon közeli számokat vonunk ki egymásból.

1.5. Hibák

Az igényes számításoknál arra is kíváncsiak vagyunk, hogy az eredményt milyen pontosan tudtuk előállítani. Ehhez számba kell venni a lehetséges hibafajtákat. Az első a kiindulásul használt adatok *öröklött hibája*, nevezhetjük ezt *adathibának* is. Lehet, hogy a számítás során magunk is *tévedünk*, ezt gondos ellenőrzéssel magunknak kell felfedeznünk és kijavítanunk. A *képlethiba* az alkalmazott módszerhez tartozik. A *kerekítési hibák* részben bekövetkezhetnek a kézi számítás, adatelőkészítés során, de a gépi aritmetikának is mindig van ilyen hibája. A hibaelemzés során fel kell ismernünk, melyik az a hibafajta, ami az adott feladat szempontjából lényeges. Sok olyan számítás van, amikor az adathiba, vagy a képlethiba jelenti a fő hibaforrást. Az adathibát sokszor csak tudomásul vehetjük, de a képlethibát esetleg csökkenthetjük pontosabb módszer alkalmazásával.

A hibaszámítás alapmodellje szerint a közelítő értékekkel kapott pontos számítás eredményét közelítésnek tekintjük és azt vizsgáljuk, mekkora a hibája.

Jelölések. Az x mennyiség *pontos értéke* $\hat{x}$, hibája Δx: $\hat{x}=x+\Delta x$, ahol Δx előjeles szám. A relatív hiba $\delta x=\Delta x/x\approx\Delta x/\hat{x}$. Itt megjegyezzük, hogy egyes szerzők a relatív hibát a pontos értékkel definiálják, tehát az itt látható második formulát használják. A mi választásunk tudomásul veszi, hogy a pontos értéket nem ismerjük. A *hibakorlát* Δ_x egy nemnegatív szám, amellyel felülről becsüljük a hiba abszolút értékét: $|\Delta x|\le\Delta_x$. Hasonlóképp δ_x a *relatív hibakorlát*, amelyre $|\delta x|\le\delta_x$.

1.4 Gyakorlat. Mutassuk meg, hogy a relatív hiba kétféle megadása között a különbség másodrendű: $\delta x-\Delta x/\hat{x}=(\delta x)^2/(1+\delta x)$.

A valóságban a Δx hibát nem ismerjük, csak annak felső korlátját. Emiatt kiindulásul annyit tudunk, hogy $\hat{x}$ az x érték valamely Δ_x-sugarú környezetében van.

$$K=\begin{bmatrix}1&&&\\-1&1&&\\&\ddots&\ddots&\\&&-1&1\end{bmatrix},\quad K^{-1}=S=\begin{bmatrix}1&&&\\1&1&&\\\vdots&\vdots&\ddots&\\1&1&\cdots&1\end{bmatrix}.$$

Inverze éppen az összegzésmátrixot adja. E két mátrix segítségével egyszerűen megadható a gyakran előforduló

$$T=\begin{bmatrix}2&-1&&\\-1&2&\ddots&\\&\ddots&\ddots&-1\\&&-1&2\end{bmatrix} \tag{5.2}$$

mátrix inverze:

$$T^{-1}=\left(K+K^T\right)^{-1}=\left[K\left(S+S^T\right)K^T\right]^{-1}=K^{-T}\left(I+ee^T\right)^{-1}K^{-1}=K^{-T}\left(I-\frac{ee^T}{1+n}\right)K^{-1}, \tag{5.3}$$

ahol e a csupa 1-esből álló vektor. A $T^{-1}x$ vektor előállítása így $4n$ flop műveletet igényel.

5.3.2 Főátló-domináns háromátlós mátrix

Láttuk, ebben az esetben nem kell a főelemválasztással foglalkozni az *LU*-felbontás során. Ha a felbontást a mutatott módon hajtjuk végre, akkor a lineáris egyenletrendszer megoldásának műveletigénye lényegében $9n$ flop. Háromátlós esetben van azonban két módszer is, amellyel $8n$ flop művelettel célba érünk. A következőkben ezeket ismertetjük. Az első módszert hívhatjuk gyors *LU*-felbontásnak. Vegyük fel a háromátlós mátrixú egyenletrendszert a következő alakban:

$$Hx=\begin{bmatrix}d_1&c_1&&\\a_1&d_2&\ddots&\\&\ddots&\ddots&c_{n-1}\\&&a_{n-1}&d_n\end{bmatrix}x=\begin{bmatrix}b_1\\b_2\\\vdots\\b_n\end{bmatrix}. \tag{5.4}$$

Az LU-felbontás első lépése csak a második sort változtatja meg:

$$\begin{bmatrix}a_1/d_1 & d_2-a_1c_1/d_1 & c_2 & \dots & 0\end{bmatrix}x=b_2-b_1a_1/d_1.$$

Eredményül kaptunk egy 1-gyel kisebb méretű háromátlójú mátrixot, amire az eljárást megismételhetjük. Tovább folytatva végül a főelemek és jobboldalak a következők lesznek:

$$\begin{aligned}&d_1'=d_1;\ \ d_i'=d_i-a_{i-1}c_{i-1}/d_{i-1}',\quad i=2,3,\dots,n,\\&b_1'=b_1;\ \ b_i'=b_i-a_{i-1}b_{i-1}'/d_{i-1}',\quad i=2,3,\dots,n.\end{aligned} \tag{5.5}$$

Most a felbontás U mátrixa felső bidiagonális - kétátlójú mátrix - és a megoldandó egyenletrendszer:

$$\begin{bmatrix}d_1&c_1&&\\0&d_2'&\ddots&\\&\ddots&\ddots&c_{n-1}\\&&0&d_n'\end{bmatrix}x=\begin{bmatrix}b_1\\b_2'\\\vdots\\b_n'\end{bmatrix},\quad x_n=b_n'/d_n';\ x_i=b_i'-c_ix_{i+1},\ \ i=n-1,n-2,\dots,1.$$

Látjuk: az *L* mátrix nem is kell a megoldáshoz, másrészt (5.5) mindkét sorában szerepel a_{i-1}/d_{i-1}', amit elegendő egyszer előállítani. Ezzel a megoldási algoritmus:

Például legyen a gépi számok halmaza $M(5,-4,3)$. Ekkor a legnagyobb mantissza: $.11111 = 1-2^{-5}$, a legkisebb mantissza ½. Az első pozitív gépi szám: $\varepsilon_0 = 1/2 \cdot 2^{-4} = 2^{-5}$. Az 1 után következő első gépi szám távolsága 1-től: $\varepsilon_1 = 2^{-t+1} = 2^{-4}$. A legnagyobb ábrázolható szám: $M_\infty = (1-2^{-t}) \cdot 2^{k^+} =$ $= (1-2^{-5})2^3 = 8 - 1/4$.

1.3. Valós számok konverziója gépi számmá

A következő kérdés: a valós számokat hogyan alakítsuk át gépi számokká. Az ezt megvalósító input függvényt fl-lel jelöljük (a *floating point number* kifejezés kezdőbetűi), $\mathrm{fl}: \mathbb{R} \to M$. Megadása a következő:

$$\mathrm{fl}(x) = \begin{cases} \infty, & \text{ha} \quad |x| > M_\infty \\ 0, & \text{ha} \quad |x| < \varepsilon_0 \\ x\text{-hez legközelebbi gépi szám, ha } \varepsilon_0 \le |x| \le M_\infty \end{cases}, \tag{1.2}$$

ahol az x-hez legközelebbi gépi szám a kerekítés szabályai szerint értendő.

Például alakítsuk át 10.87-et 8-jegyű bináris számmá. Ezt célszerűen úgy tesszük, hogy az egész részt 2-vel osztjuk, és jegyezzük a maradékokat. A sorrendet megfordítva kapjuk a bináris jegyeket. A tört részt 2-vel szorozzuk. A kijövő egész részt nem szorozzuk tovább, hanem bináris jegyként megőrízzük. Az utolsó jegyet már abból meg tudjuk állapítani, hogy a tört rész kisebb-e 0.5-nél. Ha kisebb, az adódó jegy 0, egyébként 1.

$$\begin{array}{r|l} 10 & 0 \\ 5 & 1 \\ 2 & 0 \\ 1 & 1 \end{array} \to 10_2 = 1010 \qquad \begin{array}{r|l} . & 87 \\ 1 & 74 \\ 1 & 48 \\ 0 & 96 \end{array} \to 0.87_2 = .1101\ldots$$

Kaptuk: $10.87_2 = 1010.1101\ldots$. Ez nem kerekítéssel, hanem csonkítással kapott eredmény. A kerekített szám megállapításához még egy jegyet meg kell határozni. Ha a következő jegy 1, akkor az utolsó bináris jegyhez 1-et adunk, egyébként változatlanul hagyjuk. Esetünkben a következő (kilencedik) jegy 1, így a kerekített érték: 1010.1110. Ha 10.87-et az előbbi példában szereplő $M(5,-4,3)$ halmazra kívánjuk leképezni az fl függvénnyel, akkor $\mathrm{fl}(11.87) = \infty$, mert $M_\infty < 10.87$.

1.1 Gyakorlat. Legyen a gépi számok halmaza $M(5,-4,4)$. Határozzuk meg a nevezetes számait! Mi lesz a következő számok leképezése a halmazba: 1/50, 0.37, 3.67, 7.2, 21.78?

1.2 Gyakorlat. Hogyan konvertálnánk 10.87-et 3-as alapú számrendszerbe?

Feltesszük, hogy x-et pontosan ismerjük. Ekkor $\mathrm{fl}(x)$ hibája a következőképp becsülhető:

$$|x - \mathrm{fl}(x)| \le \begin{cases} \infty, & \text{ha} \quad |x| > M_\infty \\ \varepsilon_0, & \text{ha} \quad |x| < \varepsilon_0 \\ \varepsilon_M |x|, & \text{ha} \quad \varepsilon_0 \le |x| \le M_\infty \end{cases}, \tag{1.3}$$

ahol $\varepsilon_M = \varepsilon_1 / 2 = 2^{-t}$ a *gépi epszilon*, ez adja az ε_0 és M_∞ közé eső szám ábrázolásának relatív hibáját. Itt az első sornak csak jelzés értéke van. A második sor önmagáért beszél, egyedül a harmadik sor kíván némi magyarázatot. Azt fejezi ki, hogy az ábrázolt szám hibája nem nagyobb, mint a t-edik bináris jegyben elkövetett hiba. A harmadik sor átrendezése a relatív hiba korlátját adja:

$$\frac{|x - \mathrm{fl}(x)|}{|x|} \le \varepsilon_M . \tag{1.4}$$